This book covers algorithms and discretization procedures for the solution of nonlinear programming, semi-infinite optimization, and optimal control problems. Among the important features included are a theory of algorithms represented as point-to-set maps; the treatment of finite- and infinite-dimensional min-max problems with and without constraints; a theory of consistent approximations dealing with the convergence of approximating problems and master algorithms that call standard nonlinear programming algorithms as subroutines, which provides a framework for the solution of semi-infinite optimization, optimal control, and shape optimization problems with very general constraints; and the completeness with which algorithms are analyzed. Chapter 5 contains mathematical results needed in optimization from a large assortment of sources.

Readers will find of particular interest the exhaustive modern treatment of optimality conditions and algorithms for min-max problems, as well as the newly developed theory of consistent approximations and the treatment of semi-infinite optimization and optimal control problems in this framework.

This book presents the first rigorous treatment of implementable optimization algorithms for optimal control problems with state-trajectory and control constraints, and fully accounts for all the approximations that one must make in their solution. It is also the first to make use of the concepts of epi-convergence and optimality functions in the construction of consistent approximations to infinite-dimensional problems.

Graduate students, university teachers, and optimization practitioners in applied mathematics, engineering, and economics will find this book useful.

ISBN 0-387-94971-2

ISBN 0-387-94971-2

EAN 9 780387 949710

124

Polak

Optimization

Algorithms and Consistent Approximations

Applied Mathematical Sciences 124

Elijah Polak

Optimization

Algorithms and Consistent Approximations

Springer

Applied Mathematical Sciences
Volume 124

Springer
New York
Berlin
Heidelberg
Barcelona
Budapest
Hong Kong
London
Milan
Paris
Santa Clara
Singapore
Tokyo

Applied Mathematical Sciences

1. *John:* Partial Differential Equations, 4th ed.
2. *Sirovich:* Techniques of Asymptotic Analysis.
3. *Hale:* Theory of Functional Differential Equations, 2nd ed.
4. *Percus:* Combinatorial Methods.
5. *von Mises/Friedrichs:* Fluid Dynamics.
6. *Freiberger/Grenander:* A Short Course in Computational Probability and Statistics.
7. *Pipkin:* Lectures on Viscoelasticity Theory.
8. *Giacoglia:* Perturbation Methods in Non-linear Systems.
9. *Friedrichs:* Spectral Theory of Operators in Hilbert Space.
10. *Stroud:* Numerical Quadrature and Solution of Ordinary Differential Equations.
11. *Wolovich:* Linear Multivariable Systems.
12. *Berkovitz:* Optimal Control Theory.
13. *Bluman/Cole:* Similarity Methods for Differential Equations.
14. *Yoshizawa:* Stability Theory and the Existence of Periodic Solution and Almost Periodic Solutions.
15. *Braun:* Differential Equations and Their Applications, 3rd ed.
16. *Lefschetz:* Applications of Algebraic Topology.
17. *Collatz/Wetterling:* Optimization Problems.
18. *Grenander:* Pattern Synthesis: Lectures in Pattern Theory, Vol. I.
19. *Marsden/McCracken:* Hopf Bifurcation and Its Applications.
20. *Driver:* Ordinary and Delay Differential Equations.
21. *Courant/Friedrichs:* Supersonic Flow and Shock Waves.
22. *Rouche/Habets/Laloy:* Stability Theory by Liapunov's Direct Method.
23. *Lamperti:* Stochastic Processes: A Survey of the Mathematical Theory.
24. *Grenander:* Pattern Analysis: Lectures in Pattern Theory, Vol. II.
25. *Davies:* Integral Transforms and Their Applications, 2nd ed.
26. *Kushner/Clark:* Stochastic Approximation Methods for Constrained and Unconstrained Systems.
27. *de Boor:* A Practical Guide to Splines.
28. *Keilson:* Markov Chain Models—Rarity and Exponentiality.
29. *de Veubeke:* A Course in Elasticity.
30. *Shiatycki:* Geometric Quantization and Quantum Mechanics.
31. *Reid:* Sturmian Theory for Ordinary Differential Equations.
32. *Meis/Markowitz:* Numerical Solution of Partial Differential Equations.
33. *Grenander:* Regular Structures: Lectures in Pattern Theory, Vol. III.
34. *Kevorkian/Cole:* Perturbation Methods in Applied Mathematics.
35. *Carr:* Applications of Centre Manifold Theory.
36. *Bengtsson/Ghil/Källén:* Dynamic Meteorology: Data Assimilation Methods.
37. *Saperstone:* Semidynamical Systems in Infinite Dimensional Spaces.
38. *Lichtenberg/Lieberman:* Regular and Chaotic Dynamics, 2nd ed.
39. *Piccini/Stampacchia/Vidossich:* Ordinary Differential Equations in $\mathbf{R}^n$.
40. *Naylor/Sell:* Linear Operator Theory in Engineering and Science.
41. *Sparrow:* The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors.
42. *Guckenheimer/Holmes:* Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields.
43. *Ockendon/Taylor:* Inviscid Fluid Flows.
44. *Pazy:* Semigroups of Linear Operators and Applications to Partial Differential Equations.
45. *Glashoff/Gustafson:* Linear Operations and Approximation: An Introduction to the Theoretical Analysis and Numerical Treatment of Semi-Infinite Programs.
46. *Wilcox:* Scattering Theory for Diffraction Gratings.
47. *Hale et al:* An Introduction to Infinite Dimensional Dynamical Systems—Geometric Theory.
48. *Murray:* Asymptotic Analysis.
49. *Ladyzhenskaya:* The Boundary-Value Problems of Mathematical Physics.
50. *Wilcox:* Sound Propagation in Stratified Fluids.
51. *Golubitsky/Schaeffer:* Bifurcation and Groups in Bifurcation Theory, Vol. I.
52. *Chipot:* Variational Inequalities and Flow in Porous Media.
53. *Majda:* Compressible Fluid Flow and System of Conservation Laws in Several Space Variables.
54. *Wasow:* Linear Turning Point Theory.
55. *Yosida:* Operational Calculus: A Theory of Hyperfunctions.
56. *Chang/Howes:* Nonlinear Singular Perturbation Phenomena: Theory and Applications.
57. *Reinhardt:* Analysis of Approximation Methods for Differential and Integral Equations.
58. *Dwoyer/Hussaini/Voigt (eds):* Theoretical Approaches to Turbulence.
59. *Sanders/Verhulst:* Averaging Methods in Nonlinear Dynamical Systems.
60. *Ghil/Childress:* Topics in Geophysical Dynamics: Atmospheric Dynamics, Dynamo Theory and Climate Dynamics.

(continued following index)

Elijah Polak

Optimization

Algorithms and Consistent Approximations

With 32 Illustrations

Springer

Elijah Polak
Department of Electrical Engineering
and Computer Science
University of California
Berkeley, CA 94720-1770
USA

Editors

J.E. Marsden
Control and Dynamical Systems, 104-44
California Institute of Technology
Pasadena, CA 91125
USA

L. Sirovich
Division of Applied Mathematics
Brown University
Providence, RI 02912
USA

Mathematics Subject Classification (1991): 90C30, 49Lxx, 9002, 65L05

Library of Congress Cataloging-in-Publication Data
Polak, E. (Elijah), 1931–
Optimization : algorithms and consistent approximations / Elijah Polak.
p. cm. — (Applied mathematical sciences ; 124)
Includes bibliographical references and index.
ISBN 0-387-94971-2 (alk. paper)
1. Mathematical optimization. 2. Algorithms. I. Title.
II. Series: Applied mathematical sciences (Springer-Verlag New York Inc.) ; v. 124.
QA1.A647 vol. 124
[QA402.5]
510 s—dc21
[519.3] 97-2158

Printed on acid-free paper.

Production managed by Steven Pisano; manufacturing supervised by Johanna Tschebull.
Photocomposed pages prepared from the author's troff files.
Printed and bound by Maple-Vail Book Manufacturing Group, York, PA.
Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-94971-2 Springer-Verlag New York Berlin Heidelberg SPIN 10567981

To my grandchildren:
Alexander, Arielle, and Rachel

Preface

This book deals with optimality conditions, algorithms, and discretization techniques for nonlinear programming, semi-infinite optimization, and optimal control problems. The unifying thread in the presentation consists of an abstract theory, within which optimality conditions are expressed in the form of zeros of *optimality functions*, algorithms are characterized by point-to-set iteration maps, and all the numerical approximations required in the solution of semi-infinite optimization and optimal control problems are treated within the context of consistent approximations and algorithm implementation techniques.

Traditionally, necessary optimality conditions for optimization problems are presented in Lagrange, F. John, or Karush-Kuhn-Tucker multiplier forms, with gradients used for smooth problems and subgradients for nonsmooth problems. We present these classical optimality conditions and show that they are satisfied at a point if and only if this point is a zero of an upper semicontinuous *optimality function*. The use of optimality functions has several advantages. First, optimality functions can be used in an abstract study of optimization algorithms. Second, many optimization algorithms can be shown to use search directions that are obtained in evaluating optimality functions, thus establishing a clear relationship between optimality conditions and algorithms. Third, establishing optimality conditions for highly complex problems, such as optimal control problems with control and trajectory constraints, is much easier in terms of optimality functions than in the classical manner. In addition, the relationship between optimality conditions for finite-dimensional problems and semi-infinite optimization and optimal control problems becomes transparent. Finally, optimality functions are an important part of our theory of consistent approximations that we use to construct efficient, implementable algorithms for semi-infinite optimization and optimal control.

Before attempting to analyze the great variety of optimization algorithms in current use, we present an abstract theory of optimization algorithms. The first part of this theory deals with a fairly small number of algorithm models, expressed in terms of point-to-set iteration maps, for which it provides convergence theorems. These theorems highlight the basic properties that an optimization algorithm must have to be convergent. As the reader will see, our algorithm models and convergence theorems fit most existing optimization

algorithms and hence provide a powerful unifying framework for the study of their convergence properties.

In addition, our algorithm models and convergence theorems are very useful in the construction of new optimization algorithms, since they provide guidance for the transformation of heuristically conceived algorithms into algorithms with proven convergence properties.

The second part of our algorithm theory deals with algorithm implementation. When a nonlinear programming algorithm is formally extended to the solution of semi-infinite programming, optimal control, or shape optimization problems, the result is a *conceptual* algorithm since it involves many operations, such as the solution of a differential equation, that cannot be carried out exactly and must be approximated. Our algorithm implementation theory provides "feedback rules" for adjusting the precision of these approximations, so that the resulting *implementable* algorithm is efficient and convergent.

The third part of our algorithm theory deals with consistent approximations, and is aimed at semi-infinite optimization, optimal control, and shape optimization problems. In contrast to the approach used in conceptual algorithm implementation, the consistent approximations approach does not approximate a conceptual algorithm but, instead, requires that an intractable, infinite-dimensional problem be replaced by an infinite sequence of finite-dimensional problems, obtained by numerical approximation. The theory provides consistency conditions ensuring that the global minimizers, local minimizers, and stationary points of the approximating problems converge to global minimizers, local minimizers, and stationary points, respectively, of the original problem. Since finite-dimensional problems are solvable by standard nonlinear programming software, the consistent approximations approach offers the significant practical advantage of eliminating the need for writing new semi-infinite optimization or optimal control computer programs.

In addition, the theory of consistent approximations provides master algorithm models and corresponding convergence theorems. The master algorithms construct a consistent approximation, initialize a nonlinear programming algorithm on this approximation, and terminate its operation when a test is satisfied. At this point a finer consistent approximation is constructed, and the whole process is repeated, using the last iterate, obtained from the coarser approximation, as a "warm start". Some of the master algorithms lead to specific algorithms that retain the rate of convergence of the nonlinear programming algorithms used as subroutines.

Algorithm models and convergence theorems for unconstrained optimization are presented in Section 1.2, while those for constrained optimization can be found in Section 2.3. Finally, the theory of consistent approximations is described in Section 3.3.

Optimality conditions and algorithms for nonlinear programming problems are treated in the first two chapters, with an emphasis on "classical" algorithms and completeness of analysis. As a result, due to space considerations, some of the emerging and rather promising algorithms, such as interior-point methods and modified barrier function methods are not included, while other algorithms, such as trust region methods, and algorithms with nonmonotone descent properties, are only treated abstractly. The selection of a particular algorithm in a class, such as variable metric, penalty function, or sequential quadratic programming methods, was made on the basis of the ease with which it can be presented efficiently within the framework adopted in this book.

Optimality conditions and algorithms for semi-infinite optimization and optimal control problems are presented in the third and fourth chapter. A great deal of this material is original, partly because of the use of optimality functions to express optimality conditions, and partly because the algorithms are developed within the framework of consistent approximations. As a result, most of the semi-infinite optimization and optimal control algorithms presented are in the form of a master algorithm that controls precision of approximation and calls nonlinear programming software as a subroutine. In our experience, the consistent approximations approach has considerable advantages over conceptual algorithm implementation, both because the resulting algorithms compute faster and because the tedious task of writing computer code for implementations of conceptual algorithms is eliminated.

Finally, the fifth chapter presents an essential collection of mathematical results basic to the understanding of the material in this book. These include results from functional analysis, convex analysis, theory of set-valued maps, nonsmooth analysis, duality, and minimax theory.

The bibliography dealing with the subjects in this book is enormous, and so, not surprisingly, the list of references at the end of the book is representative, but far from complete. Fortunately, search tools, such as the *University of California's MELVYL® library system*, and the *MELVYL® INSPEC® data base* simplify the examination of bibliographic data bases and are recommended to the reader interested in carrying out an exhaustive search.

This book grew out of two sets of graduate level class notes. The first set contained material from the first two chapters of this book and some from the fifth chapter and was used for many years in a first year, one semester, graduate course on finite-dimensional optimization. The second set of notes contained material on semi-infinite programming and optimal control, as well as some of the more advanced mathematical background from the fifth chapter of this book and was used in a second year, one semester, graduate course. Hence this book is usable as a text for two one-semester graduate courses on optimization. In addition, because of the depth of coverage, it should prove to be a very useful reference text.

The author is grateful to Jim Burke, John Dennis, Asen Dontchev, Joseph Dunn, Olvi Mangasarian, Boris Mordukhovich, Kurt Overley, Andrew Philpott, Stephen Robinson, Terry Rockafellar, Andre Tits, Richard Vinter, Yorai Wardi, Roger Wets, and Jian Zhou, who have offered comments on the first draft of the book or advice on particular topics.

Particular thanks are due to Bradley Bell and Stephen Wright for their help in cleaning up the theory of consistent approximations, to Daniel Ralph for his help in developing a presentation of SQP methods in the Josephy framework, to Daniel Sorensen for his assistance with the trust region model, and to Carlos Kirjner, David Mayne, Adam Schwartz, and Thomas Yang, who invested many hours of their time in removing errors and ambiguities from the text.

The contents of this book reflect the author's collaboration with David Mayne and Clovis Gonzaga as well as with his former students Ted Baker, Limin He, Joe Higgins, Carlos Kirjner, Robert Klessig, Hiro Mukai, Olivier Pironneau, Adam Schwartz, Andre Tits, Yorai Wardi, and Thomas Yang.

Finally, I am grateful to my wife, Ginette, for her support and encouragement in this project, as well as for her help in proofreading the manuscript.

The preparation of this volume involved a great amount of research which would have been impossible without the support received from the National Science Foundation, the Air Force Office of Scientific Research, and the University of California at Berkeley. This support is gratefully acknowledged.

Elijah Polak

Contents

Conventions and Symbols

1 Numbering and Cross-Referencing System

The following system of numbering and cross-referencing is used in this book.

Definitions, lemmas, propositions, theorems, corollaries, and remarks are numbered in order of occurrence, using a three number system *(a.b.c)*, where a is the chapter number, b is the section number, and c is the item number. This numbering system does not distinguish between definitions, theorems, lemmas, etc.

Within each section, equations are numbered consecutively, using a single number system, and are referred to by a single number. Equations from other sections are referred to using a three number system $(a.b.c)$, where a is the chapter number, b is the section number, and c is the item number. When we refer to Theorem x.y.z*(a)*, we mean part *(a)* of Theorem x.y.z.

2 Conventions

1. $\mathbb{R}^n$ denotes the Euclidean space of n–tuplets of real numbers. Elements of $\mathbb{R}^n$ are denoted by lower case letters. The vectors $x \in \mathbb{R}^n$ are always treated as column vectors, and their components are denoted by superscripts; since no ambiguity can arise, we will write $x = (x^1, x^2, \ldots, x^n)$ to denote the column vector x, without a transpose sign. The inner product in $\mathbb{R}^n$ (Euclidean n–space) is denoted by $\langle \cdot, \cdot \rangle$ and is defined by $\langle x, y \rangle \triangleq \sum_{i=1}^n x_i y_i$. The norm in $\mathbb{R}^n$ is denoted by $\|\cdot\|$ and is defined by $\|x\| \triangleq \langle x, x \rangle^{1/2}$.

2. $L_{\infty,2}^k[0,1]$ denotes the pre-Hilbert space consisting of equivalence classes of Lebesgue measurable, essentially bounded functions from $[0,1]$ to $\mathbb{R}^k$, with with inner product $\langle u_1, u_2 \rangle_2 \triangleq \int_0^\infty \langle u_1(t), u_2(t) \rangle \, dt$ and norm $\|\cdot\|_2$, defined by $\|u\|_2 = \langle u, u \rangle_2^{1/2}$.

3. $L_2^k[0,1]$ denotes the Hilbert space consisting of equivalence classes of Lebesgue square-integrable functions from $[0,1]$ to $\mathbb{R}^k$, with inner product $\langle u_1, u_2 \rangle_2 \triangleq \int_0^\infty \langle u_1(t), u_2(t) \rangle \, dt$ and norm $\|\cdot\|_2$, defined by $\|u\|_2 = \langle u, u \rangle_2^{1/2}$.

4. $L_\infty^k[0,1]$ denotes the Banach space consisting of equivalence classes of Lebesgue measurable, essentially bounded functions from $[0,1]$ to $\mathbb{R}^k$, with norm $\|\cdot\|_\infty$, defined by $\|u\|_\infty = \operatorname{ess\,sup}_{t\in[0,1]} \|u(t)\|$.

5. $f(\cdot)$ denotes a function, with the dot standing for the undesignated variable; $f(x)$ denotes the value of $f(\cdot)$ at the point x. We write $f: A \to B$ to indicate that the domain of $f(\cdot)$ is in the space A and that its range is in the space B. When $B = \mathbb{R}^k$, $f(x) = (f^1(x), f^2(x), \ldots, f^k(x))$, is a column vector.

6. Given a function $g: \mathbb{R}^n \to \mathbb{R}^m$, we denote its Jacobian matrix by $g_x(x)$ or by $\partial g(x)/\partial x$. This is a matrix whose ijth element is $\partial g^i(z)/\partial x^j$. Similarly, when $g: \mathbb{R}^n \to \mathbb{R}^1$, we denote its Hessian by $g_{xx}(x)$. This is a matrix whose ijth element is $\partial^2 g(z)/\partial x^i x^j$.

7. Superscript -1 denotes the inverse of a matrix or an operator, e.g., A^{-1}.

8. Superscript T denotes the transpose of a matrix or the adjoint of an operator, e.g., A^T.

9. When we say that a function $f(\cdot)$ is (Lipschitz) continuously differentiable on bounded sets, we mean that it is continuously differentiable and that its derivative is (Lipschitz) continuous on bounded sets.

3 Symbols

Spaces

$\mathbb{R}^n$	Euclidean n-space
$\mathbb{R}^n_+$	$\{x \in \mathbb{R}^n \mid x^i \geq 0,\ \forall i\}$
$L^m_{\infty,2}[0,1]$	$(L^m_\infty[0,1], \langle\cdot,\cdot\rangle_{L^m_2[0,1]}, \|\cdot\|_{L^m_2[0,1]})$
H_2	$\mathbb{R}^n \times L^m_2[0,1]$
$H_{\infty,2}$	$\mathbb{R}^n \times L^m_{\infty,2}[0,1]$
L_N	finite-dimensional subspace of $L^m_{\infty,2}[0,1]$
$\bar{L}_N$	$\mathbb{R}^m \times \cdots \times \mathbb{R}^m$
H_N	$\mathbb{R}^n \times L_N \subset H_{\infty,2}$
$\bar{H}_N$	$\mathbb{R}^n \times \bar{L}_N$

Elements

u	$u \in L^m_{\infty,2}[0,1]$
u_N	$u_N \in L_N$
$\bar{u}_N$	$\bar{u}_N = (\bar{u}_0, \ldots, \bar{u}_{N-1}) \in \bar{L}_N$
η	$\eta = (\xi, u) \in H_{\infty,2}$
η_N	$\eta_N = (\xi, u) \in H_N$
$\bar{\eta}_N$	$\bar{\eta}_N = (\xi, \bar{u}) \in \bar{H}_N$

Sets

$B(x,\rho)$	$\{x' \in \mathcal{B} \mid \|x' - x\|_{\mathcal{B}} \leq \rho\}$
$2^{\mathcal{B}}$	set of all subsets of $\mathcal{B}$
$\operatorname{aff} A$	affine hull of A
$\operatorname{co} A$	convex hull of A
$\operatorname{int} A$	interior of A
$L_\alpha(f)$	level set of $f(\cdot)$
$\partial L_\alpha(f)$	boundary of level set of $f(\cdot)$
QS	set of quasi-stationary points
S	set of stationary points
$\mathbf{q}$	$\{1,2,\ldots,q\}$
$\bar{\mathbf{q}}$	$\{0,1,2,\ldots,q\}$
$\hat{\mathbf{q}}(x)$	$\{j \in \mathbf{q} \mid f^j(x) = \psi(x)\}$
$\mathbf{q}_A(x)$	$\{j \in \mathbf{q} \mid f^j(x) \geq 0\}$
$\mathbf{U}$	$\{u \in L^m_{\infty,2}[0,1] \mid \|u\|_\infty \leq \rho_{\max}\}$
$\mathbf{U}^o$	$\{u \in L^m_{\infty,2}[0,1] \mid \|u\|_\infty < \gamma\rho_{\max}\}$
U	$U \subset \mathbb{R}^m$ pointwise control constraint set
$\mathbf{U}_c$	$\{u \in \mathbf{U}^o \mid u(t) \in U \subset B(0,\gamma\rho_{\max}),\ \forall t \in [0,1]\}$
$\mathbf{H}$	$\mathbb{R}^n \times \mathbf{U}$
$\mathbf{H}^o$	$\mathbb{R}^n \times \mathbf{U}^o \subset \mathbf{H}$
$\mathbf{H}^o_{cl}$	closure of $\mathbf{H}^o$
$\mathbf{H}_c$	$\mathbb{R}^n \times \mathbf{U}_c$
$\mathbf{H}_N$	$\mathbf{H} \cap H_N$
$\mathbf{H}^o_N$	$\mathbf{H}^o \cap H_N$
$\mathbf{H}^o_{cl,N}$	closure of $\mathbf{H}^o_N$
$\bar{\mathbf{H}}_N$	$\mathbb{R}^n \times \mathbb{R}^{mN}$
$\mathbf{H}_{c,N}$	$\mathbf{H}_c \cap H_N$
$\mathbb{N}$	$\{0,1,2,\ldots\}$
$\mathcal{N}$	$\{2^N\}_{N=1}^\infty$
$\mathbb{Z}$	$\{\ldots,-2,-1,0,1,2,\ldots\}$
Σ_q	$\{\mu \mid \mu \in \mathbb{R}^q_+,\ \sum_{j=1}^q \mu^j = 1\}$
Σ^0_q	$\{(\mu^0,\mu) \mid \mu^0 \in \mathbb{R}_+,\ \mu \in \mathbb{R}^q_+,\ \sum_{j=0}^q \mu^j = 1\}$

Functions

$\lfloor i \rfloor$	$\max\{k \in \mathbb{N} \mid k \leq i\}$
$\langle\cdot,\cdot\rangle_{\mathcal{H}}$	scalar product in the space $\mathcal{H}$
$\|\cdot\|_{\mathcal{B}}$	norm in space $\mathcal{B}$
$\det A$	determinant of matrix A
x_+	$x \in \mathbb{R}$, $\max\{0,x\}$
y_+	$y \in \mathbb{R}^n$, $(y^1_+,\ldots,y^m_+)$
$\nabla g(x)$	$g_x(x)^T$

$\partial\psi(x)$	subgradient of $\psi(\cdot)$ at x
$\psi(\cdot)$	a max function
$Nr\,[\partial\psi(x)]$	$\arg\min\{\|\xi\| \mid \xi \in \partial\psi(x)\}$
$d\psi(x;h)$	directional derivative
$d_2F(x,y;h)$	directional derivative with respect to second argument
$\theta(\cdot)$	optimality function for unconstrained or constrained problem
$\theta_u(\cdot)$	optimality function for unconstrained problem
$\theta_c(\cdot)$	optimality function for constrained problem

Sequences

$\underline{x}_i$	$\{x_j\}_{j=0}^{i}$
$\{x_i\}_{i=0}^{\infty}$	a sequence
$x_i \to^K x$	$\{x_i\}_{i\in K} \subset \{x_i\}_{i=0}^{\infty}$ converges to x.
Σ_q	$\{\mu \mid \mu \in \mathbb{R}^{q_+}, \sum_{j=1}^{q}\mu^j = 1\}$
Σ_q^0	$\{(\mu^0,\mu) \mid \mu^0 \in \mathbb{R}_+, \mu \in \mathbb{R}^{q_+}, \sum_{j=0}^{q}\mu^j = 1\}$
$\overline{\lim}\, x_i$	limit superior of a scalar sequence
$\underline{\lim}\, x_i$	limit inferior of a scalar sequence

Differential and Difference Equations

$x^\eta(t)$	solution of differential equation
$\bar{x}_{N,k}^{\eta}$	solution of Euler difference equation
$x_N^\eta(t)$	linear interpolation of $\bar{x}_{N,k}^{\eta}$
$f_t(\eta)$	$F(\xi, x^\eta(t))$
$f^j(\eta)$	$F^j(\xi, x^\eta(1))$
$\phi^j(\eta,t)$	$\tilde{\phi}^j(\xi, x^\eta(t), t)$
$f_{N,k}(\eta)$	$F(\xi, x_N^\eta(k/N))$
$\bar{f}_{N,k}(\bar{\eta})$	$F(\xi, \bar{x}_{N,k}^{\bar{\eta}})$
$f_N(\eta)$	$F(x_N^\eta(1))$
$\bar{f}_N(\eta)$	$F(\bar{x}_{N,N}^{\bar{\eta}})$

Abbreviations

i.s.c.	inner semicontinuous
o.s.c.	outer semicontinuous
l.s.c.	lower semicontinuous
u.s.c.	upper semicontinuous
l.L.c.	locally Lipschitz continuous

Miscellaneous

$\triangleq$	equal by definition
<=>	if and only if
=>	if part of proof
<=	only if part of proof
$\square$	end of proof, end of example, end of remark, etc

Chapter 1

Unconstrained Optimization

We devote this chapter primarily to optimality conditions and algorithms for solving unconstrained optimization problems of the form $\min_{x\in\mathbb{R}^n} f(x)$, where $f(\cdot)$ is a continuously differentiable cost function defined on $\mathbb{R}^n$. When convenient, we will extend our results to problems of the form $\min_{x\in X} f(x)$, where $f(\cdot)$ is a continuously differentiable cost function, defined on $\mathbb{R}^n$, and $X \subset \mathbb{R}^n$ is an "unstructured", convex, constraint set. We expect the reader to be familiar with the mathematical background contained in the first four sections of Chapter 5.

1.1 Optimality Conditions

It is impossible to tell whether a piece of yellow metal is gold without submitting it to physical and chemical tests. It is equally impossible to tell whether a vector is a solution of an optimization problem without checking if it satisfies optimality conditions. As we will shortly see, it is often easier to identify gold than a solution to an optimization problem. This fact has profound consequences on the extent to which one can hope to "solve" an optimization problem.

We will present two kinds of optimality conditions. *Necessary conditions* are those that must be satisfied by any local minimizer. *Sufficient conditions* are those implying that a point is a local minimizer. Throughout this book, we will present necessary optimality conditions in three forms. The first is the most basic, and expresses the fact that, to first or second order, the cost must increase in the vicinity of a local minimizer. The second form is a consequence of the first and consists of an equation involving gradients or a quadratic inequality involving second derivatives, and possibly also gradients. When one tries to define a method for verifying whether necessary conditions in either of the above two forms are satisfied at a point, one is led to the third form which

characterizes local minimizers as zeros of an easily evaluated *optimality function*. As we will see later, our favorite optimality functions are based on a strictly convex, first order local model for an optimization problem and have two important advantages: *(i)* their evaluation yields a continuous cost descent direction, and *(ii)* their value can be used to compute upper and lower bounds on the minimum value being sought.

As we have mentioned above, we will consider two problems in this chapter:

$$\min_{x \in \mathbb{R}^n} f(x)\,, \tag{1a}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, and

$$\min_{x \in X} f(x)\,, \tag{1b}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, and X is a convex subset of $\mathbb{R}^n$.

Definition 1.1.1.

(a) We will say that $\hat{x}$ is a global *minimizer for problem (1a) if*

$$f(\hat{x}) \le f(x), \quad \forall\, x \in \mathbb{R}^n\,. \tag{2a}$$

(b) We will say that $\hat{x}$ is a local minimizer *for problem (1a) if there exists a $\hat{\rho} > 0$ such that*

$$f(\hat{x}) \le f(x), \quad \forall\, x \in B(\hat{x}\,, \hat{\rho})\,, \tag{2b}$$

where

$$B(\hat{x}\,, \hat{\rho}) \triangleq \{x \in \mathbb{R}^n \mid \|x - \hat{x}\| \le \hat{\rho}\}. \tag{2c}$$

(c) We will say that $\hat{x}$ is a strict local minimizer *for (1a) if there exists a $\hat{\rho} > 0$ such that $f(\hat{x}) < f(x)$ for all $x \in B(\hat{x}\,, \hat{\rho}) \cap X$, with $x \neq \hat{x}$.*

(d) We will say that $\hat{x}$ is a global *minimizer for problem (1b) if*

$$f(\hat{x}) \le f(x), \quad \forall\, x \in X\,. \tag{2d}$$

(e) We will say that $\hat{x}$ is a local minimizer *for problem (1b) if there exists a $\hat{\rho} > 0$ such that*

$$f(\hat{x}) \le f(x), \quad \forall\, x \in B(\hat{x}\,, \hat{\rho}) \cap X\,. \tag{2e}$$

(f) We will say that $\hat{x}$ is a strict local minimizer *for problem (1b) if there exists a $\hat{\rho} > 0$ such that $f(\hat{x}) < f(x)$ for all $x \in B(\hat{x}\,, \hat{\rho}) \cap X$, with $x \neq \hat{x}$.* □

Note that the above definitions make sense even when the cost function $f(\cdot)$ is not continuous, e.g., it can be lower semicontinuous.

1.1.1 First- and Second-Order Necessary Conditions

We begin with *necessary conditions of optimality*, i.e., those that must be satisfied by any solution to (1a) or (1b), under increasing assumptions of differentiability. As a minimum, we will assume that the cost function $f(\cdot)$ has one-sided directional derivatives in all directions. Note that this implies that $f(\cdot)$ is at least continuous. The most basic optimality condition for problems (1a) and (1b) stems from the observation that at a local minimizer $\hat{x}$ all the permissible directions must lead "up hill", at least to first order. For problem (1a) all directions are permissible, while for problem (1b) only directions that lead from $\hat{x}$ into the set X are permissible. When formalized, these observations lead to the following results.

Theorem 1.1.2. *Consider problem (1a), and suppose that, for all $x\,,\ h \in \mathbb{R}^n$, the one-sided directional derivative $df(x\,; h)$ exists. If $\hat{x}$ is a local minimizer for (1a), with associated radius $\hat{\rho} > 0$, as in Definition 1.1.1, then*

$$df(\hat{x}\,; h) \ge 0\,, \quad \forall\, h \in \mathbb{R}^n\,. \tag{3}$$

Proof. To obtain a contradiction, suppose that there exists an $h \in \mathbb{R}^n$ such that $df(\hat{x}\,; h) < 0$; clearly, $h \neq 0$. Since, by definition of the one-sided directional derivative,

$$\lim_{\lambda \downarrow 0} \left[\frac{f(\hat{x} + \lambda h) - f(\hat{x})}{\lambda} - df(\hat{x}\,; h) \right] = 0\,,$$

there exists a $\lambda^* \in (0\,, \hat{\rho} / \|h\|]$ such that $\hat{x} + \lambda^* h \in B(\hat{x}\,, \hat{\rho})$ and

$$\left[\frac{f(\hat{x} + \lambda^* h) - f(\hat{x})}{\lambda^*} - df(\hat{x}\,; h) \right] \le -\tfrac{1}{2} df(\hat{x}\,; h)\,.$$

Consequently,

$$f(\hat{x} + \lambda^* h) - f(\hat{x}) \le \tfrac{1}{2} \lambda^* df(\hat{x}\,; h) < 0\,,$$

which contradicts the optimality of $\hat{x}$. □

Corollary 1.1.3. *Suppose that $f(\cdot)$ in (1a) is continuously differentiable and that $\hat{x}$ is a local minimizer for (1a). Then*

$$\nabla f(\hat{x}) = 0\,. \tag{4}$$

Proof. By Theorem 1.1.2, $df(\hat{x}\,; h) = \langle \nabla f(\hat{x})\,, h \rangle \ge 0$, for all $h \in \mathbb{R}^n$. Since $v = 0$ is the only vector in $\mathbb{R}^n$ with the property that $\langle v\,, h \rangle \ge 0$ for all $h \in \mathbb{R}^n$, the desired result follows. □

Exercise 1.1.4. Consider the problem (1b) and suppose that $X \subseteq \mathbb{R}^n$ is a convex set and that, for all $x \in X$, $h \in \mathbb{R}^n$, the directional derivative $df(x;h)$ exists. Show that if $\hat{x} \in X$ is a local minimizer for (1b), then

$$df(\hat{x};x-\hat{x}) \geq 0, \quad \forall x \in X. \tag{5}$$

□

Theorem 1.1.5. *Suppose that $f(\cdot)$ in (1a,b) is twice continuously differentiable, and let $f_{xx}(x) \triangleq \partial^2 f(x)/\partial x^2$.*

(a) If $\hat{x}$ is a local minimizer for problem (1a), then

$$\langle h, f_{xx}(\hat{x})h \rangle \geq 0, \quad \forall h \in \mathbb{R}^n. \tag{6a}$$

(b) If $\hat{x}$ is a local minimizer for problem (1b), then

$$\langle x-\hat{x}, f_{xx}(\hat{x})(x-\hat{x}) \rangle \geq 0, \tag{6b}$$

for all $x \in X$ such that $\langle \nabla f(\hat{x}), x-\hat{x} \rangle = 0$.

Proof. *(a)* Let $B(\hat{x},\hat{\rho})$ be the ball associated with $\hat{x}$. Since, by Corollary 1.1.3, $\nabla f(\hat{x}) = 0$, it follows from the optimality of $\hat{x}$ and (5.1.17d) that, for any $h \in \mathbb{R}^n$ and $\lambda > 0$ such that $\lambda \|h\| \leq \hat{\rho}$, there exists an $s_\lambda \in [0,1]$ such that

$$f(\hat{x}+\lambda h) - f(\hat{x}) = \tfrac{1}{2}\lambda^2 \langle h, f_{xx}(\hat{x}+s_\lambda \lambda h)h \rangle \geq 0, \tag{7a}$$

which implies that $\langle h, f_{xx}(\hat{x}+s_\lambda\lambda h)h \rangle \geq 0$. Letting $\lambda \to 0$, we conclude that

$$\langle h, f_{xx}(\hat{x})h \rangle \geq 0. \tag{7b}$$

Since (7b) holds for all $h \in \mathbb{R}^n$, the desired result follows.

(b) Let $B(\hat{x},\hat{\rho})$ be the ball associated with $\hat{x}$, and let $x \in X$ be such that $\langle \nabla f(\hat{x}), x-\hat{x} \rangle = 0$. Let $h = x-\hat{x}$. Then, for all $\lambda \in [0,1]$, $\hat{x}+\lambda h \in X$, and for any $\lambda \in [0,1]$ such that $\lambda\|h\| \leq \hat{\rho}$, (7a) must hold, for some $s_\lambda \in [0,1]$. Letting $\lambda \to 0$, we obtain the desired result. □

It is easy to express (3) and (5) in alternative ways. We begin with (3). Thus, suppose that, for the function $f(\cdot)$, in the problems (1a,b), the one-sided directional derivative $df(x;h)$ exists for all $x, h \in \mathbb{R}^n$ and that $df(x;\cdot)$ is continuous and convex.† Let $\theta : \mathbb{R}^n \to \mathbb{R}$ be defined by

† When $f(\cdot)$ is locally Lipschitz continuous and its generalized directional derivative equals the "ordinary" directional derivative, as in all the problems considered in this book, $df(x;\cdot)$ is guaranteed to be Lipschitz continuous and convex (see Theorem 5.1.32).

$$\theta(x) \triangleq \min_{h \in B(0,1)} df(x;h). \tag{8a}$$

Then, because the ball $B(0,1)$ is compact, it follows from Corollary 5.1.25 that $\theta(\cdot)$ is well defined. Furthermore, it follows by inspection that (3) holds if and only if $\theta(\hat{x}) = 0$. Thus, the evaluation of $\theta(\hat{x})$ provides a constructive way of determining whether (3) is satisfied or not. When (3) is not satisfied at $\hat{x}$, then $\theta(\hat{x}) < 0$, and the vector

$$h(\hat{x}) \triangleq \arg \min_{h \in B(0,1)} df(\hat{x};h) \tag{8b}$$

defines a "steepest descent" direction for the cost function $f(\cdot)$ at $\hat{x}$. Note that when $f(\cdot)$ is continuously differentiable, $h(\hat{x}) = -\nabla f(\hat{x})/\|\nabla f(\hat{x})\|$.

Next, consider (5). Let $\theta_X : \mathbb{R}^n \to \mathbb{R}$ be defined by

$$\theta_X(x) \triangleq \min_{h \in B(0,1)\cap\{X-x\}} df(x;h). \tag{8c}$$

Then it follows from Corollary 5.1.25 that $\theta_X(\cdot)$ is well defined, and it follows by inspection that (5) holds if and only if $\theta_X(\hat{x}) = 0$. Thus, the evaluation of $\theta_X(\hat{x})$ provides a constructive way of determining whether (5) is satisfied or not. When (5) is not satisfied at $\hat{x}$, $\theta_X(\hat{x}) < 0$, and the vector

$$h_X(\hat{x}) \triangleq \arg \min_{x' \in B(0,1)\cap\{X-\hat{x}\}} df(\hat{x};h) \tag{8d}$$

defines a "steepest feasible descent" direction for the cost function $f(\cdot)$ at $\hat{x}$.

We will call the above, nonpositive-valued functions $\theta(\cdot)$ and $\theta_X(\cdot)$ *optimality functions*. When the function $f(\cdot)$ is continuously differentiable, it follows from Corollary 5.4.2 that these optimality functions are continuous. The above examples illustrate the advantages of expressing optimality conditions in terms of optimality functions: *(a)* optimality functions provide a convenient method for verifying whether a point x satisfies the "basic" optimality condition of nondecrease, to first or second order, in cost function value in a ball about x, and *(b)* when a point x does not satisfy the "basic" optimality condition, their evaluation produces a (feasible) *descent direction*† $h_X(x)$, i.e., there is a $\lambda_x \in (0,1]$ such that $f(x+\lambda h_X(x)) < f(x)$, (and $x+\lambda h_X(x) \in X$) *for all* $\lambda \in (0,\lambda_x]$. Generically, we will denote optimality functions by $\theta(\cdot)$ and the associated descent directions by $h(x)$. When there is a possibility of confusion, we will use subscripts to distinguish optimality and descent direction function

† It is common to refer to descent direction functions as *search direction* functions, even though, in general, search directions need not be descent directions.

from one problem to another. We will require that optimality functions be nonpositive valued and upper semicontinuous (u.s.c). The reason for this is that if we construct a sequence $\{x_i\}_{i=0}^{\infty}$ such that $x_i \to \hat{x}$, as $i \to \infty$, and $\theta(x_i) \to 0$, as $i \to \infty$, then the fact that $\theta(x_i) \leq 0$ for all i together with the upper semicontinuity of $\theta(\cdot)$ ensure that $\theta(\hat{x}) = 0$.

It is possible to associate more than one optimality function with a particular optimality condition. A particularly fruitful method for constructing optimality functions consists of replacing the original optimization problem by a strictly convex, first-order, quadratic local approximation, or model, and defining the optimality function as the minimum value of the model. For example, assuming that the function $f(\cdot)$ in (1a) and (1b) is continuously differentiable, we can define a local quadratic model[†] of the cost function at x, $\tilde{f}(x,\cdot): \mathbb{R}^n \to \mathbb{R}$, by

$$\tilde{f}(x,x') \triangleq f(x) + \langle \nabla f(x), x' - x \rangle + \tfrac{1}{2}\|x' - x\|^2 . \tag{8e}$$

The quadratic term $\frac{1}{2}\|x' - x\|^2$ in (8e) removes the need to use a ball in the definition of optimality functions and makes them easier to evaluate. Hence, for problem (1a), we define the optimality function $\theta: \mathbb{R}^n \to \mathbb{R}$ by

$$\begin{aligned}\theta(x) &\triangleq \min_{x' \in \mathbb{R}^n} \langle \nabla f(x), x' - x \rangle + \tfrac{1}{2}\|x' - x\|^2 \\ &= \min_{h \in \mathbb{R}^n} \langle \nabla f(x), h \rangle + \tfrac{1}{2}\|h\|^2 = -\tfrac{1}{2}\|\nabla f(x)\|^2 ,\end{aligned} \tag{8f}$$

and we define its associated descent direction function $h: \mathbb{R}^n \to \mathbb{R}^n$ by[‡]

$$h(x) \triangleq \arg\min_{h \in \mathbb{R}^n} \langle \nabla f(x), h \rangle + \tfrac{1}{2}\|h\|^2 = -\nabla f(x) . \tag{8g}$$

For problem (1b), we define the optimality function $\theta_X: \mathbb{R}^n \to \mathbb{R}$ by

$$\begin{aligned}\theta_X(x) &\triangleq \min_{x' \in X} \langle \nabla f(x), x' - x \rangle + \tfrac{1}{2}\|x' - x\|^2 , \\ &= \min_{h \in X - x} \langle \nabla f(x), h \rangle + \tfrac{1}{2}\|h\|^2 ,\end{aligned} \tag{8h}$$

and its associated descent direction function $h: \mathbb{R}^n \to \mathbb{R}^n$ by

$$h_X(x) \triangleq \arg\min_{h \in X - x} \langle \nabla f(x), h \rangle + \tfrac{1}{2}\|h\|^2 , \tag{8i}$$

where $X - x \triangleq \{h \in \mathbb{R}^n \mid x + h \in X\}$.

[†] The quadratic function $\tilde{f}(x,\cdot)$ is only a first-order approximation to $f(\cdot)$ at x, see (2.1.9k).

[‡] Note that in (8f) and in the right-hand side of (8g), h is an arbitrary vector in $\mathbb{R}^n$, but in the left-hand side of (8g), $h(x)$ is the solution of the minimization problem defining $\theta(x)$ and hence a function of x.

Since the the $X - x$ is convex and the function $\langle \nabla f(x), h \rangle + \frac{1}{2}\|h\|^2$ is strictly convex in h and tends to infinity as $\|h\| \to \infty$, it follows that both $\theta_X(x)$ and $h_X(x)$ are well defined, i.e., (8h) has a unique solution $h_X(x)$.

Proposition 1.1.6. *Consider the problems (1a) and (1b). Suppose that $f: \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, that X is a closed convex subset of $\mathbb{R}^n$ and that $\theta: \mathbb{R}^n \to \mathbb{R}$, $\theta_X: X \to \mathbb{R}$, $h: \mathbb{R}^n \to \mathbb{R}^n$, and $h_X: \mathbb{R}^n \to \mathbb{R}^n$ are defined by (8f), (8h), (8g), and (8i), respectively. Then,*

(a) $\theta(x) \leq 0$ for all $x \in \mathbb{R}^n$, and $\theta_X(x) \leq 0$ for all $x \in X$;

(b) the functions $\theta(\cdot)$, $\theta_X(\cdot)$, $h(\cdot)$, and $h_X(\cdot)$ are all continuous;

(c) for any $x \in \mathbb{R}^n$, $df(x; h(x)) \leq \theta(x)$, and for any $x \in X$, $df(x; h_X(x)) \leq \theta_X(x)$;

(d) for any $x \in \mathbb{R}^n$, $\theta(x) = 0$ if and only if $\nabla f(x) = 0$, and for any $x \in X$, $\theta_X(x) = 0$ if and only if (5) holds;

(e) if the function $f(\cdot)$ is twice continuously differentiable and there exist $0 < m \leq M < \infty$ such that

$$m\|h\|^2 \leq \langle h, f_{xx}(x)h \rangle \leq M\|h\|^2 , \tag{9a}$$

for all x, $h \in \mathbb{R}^n$, then for any $x \in \mathbb{R}^n$,

$$\theta(x)/m \leq f(\hat{x}) - f(x) \leq \theta(x)/M , \tag{9b}$$

where $\hat{x}$ is the global minimizer[†] for (1a);

(f) if the function $f(\cdot)$ is twice continuously differentiable and there exist $0 < m \leq 1 \leq M < \infty$[‡] such that for any $x \in X$ and $h \in \mathbb{R}^n$ (9a) holds, then for any $x \in X$,

$$\theta_X(x)/m \leq f(\hat{x}_X) - f(x) \leq \theta_X(x)/M , \tag{9c}$$

where $\hat{x}_X$ is the global minimizer for (1b).

Proof. (a) The fact that $\theta(x) \leq 0$ for all $x \in \mathbb{R}^n$ and $\theta_X(x) \leq 0$ for all $x \in X$ follows from the fact that $h = 0$ is an admissible selection in (8f) and that $x' = x$ is an admissible selection in (8h).

[†] Referring to Theorem 5.2.14, we see that it follows from (9a) that the function $f(\cdot)$ is strictly convex. Hence (1a) and (1b) (with X convex) must have unique global minimizers; see Theorem 1.1.11.

[‡] It is possible to remove the restriction $0 < m \leq 1 \leq M < \infty$, which can adversely affect the estimate (9c), by redefining $\theta_X(\cdot)$ by $\theta_X(x) \triangleq \min_{x' \in X} \langle \nabla f(x), h \rangle + \frac{1}{2}\delta\|x' - x\|^2$, with $\delta \in [m, M]$.

(b) First we will deal with $\theta(\cdot)$ and $h(\cdot)$. It follows directly from Corollary 1.1.3 applied to the minimization problem (8f) that $h(x) = -\nabla f(x)$ and hence that $\theta(x) = -\frac{1}{2}\|\nabla f(x)\|^2$. Since by assumption, $\nabla f(\cdot)$ is continuous, it follows that both $\theta(\cdot)$ and $h(\cdot)$ are continuous.

Next, we turn to $\theta_X(\cdot)$ and $h_X(\cdot)$. Let $x^* \in X$ and $\rho > 0$ be arbitrary. Then there exists a constant $b < \infty$ such that, for all $x \in X \cap B(x^*, \rho)$, $\|\nabla f(x)\| \leq b$. Hence, for all $x \in X \cap B(x^*, \rho)$ and $x' \in X$,

$$\langle \nabla f(x), x' - x \rangle + \tfrac{1}{2}\|x' - x\|^2 > 0, \tag{10a}$$

if $\|x' - x\| > 2b$. Let $X^* \triangleq X \cap B(x^*, \rho+2b)$. Then we see that X^* is a convex, compact set, and that for all $x \in B(x^*, \rho)$, $B(x, 2b) \subset B(x^*, \rho+2b)$. Therefore, for all $x \in B(x^*, \rho)$,

$$\theta_X(x) = \min_{x' \in X^*} \langle \nabla f(x), x' - x \rangle + \tfrac{1}{2}\|x' - x\|^2$$

$$= \min_{h \in X^* - x} \langle \nabla f(x), h \rangle + \tfrac{1}{2}\|h\|^2. \tag{10b}$$

It follows from Corollary 5.4.2 and Theorem 5.4.3 that $\theta_X(\cdot)$ and $h_X(\cdot)$ are both continuous.

(c) We note that $df(x; h(x)) = \langle \nabla f(x), h(x) \rangle \leq \theta(x)$, and $df(x; h_X(x)) = \langle \nabla f(x), h_X(x) \rangle \leq \theta_X(x)$ always holds.

(d) Clearly, since $\theta(x) = -\frac{1}{2}\|\nabla f(x)\|^2$, $\theta(x) = 0$ if and only if $\nabla f(x) = 0$. Next, suppose that $x \in X$ is such that $\theta_X(x) < 0$. Then, by *(c)*, $df(x; h_X(x)) < 0$. Consequently, by contraposition, (5) implies that $\theta_X(x) = 0$. Now suppose that $x \in X$ is such that, for some $x' \in X$, $df(x; x' - x) < 0$. Since $x(\lambda) = \hat{x} + \lambda(x' - x) \in X$ for all $\lambda \in [0, 1]$ and since $\hat{x} + \lambda(x' - x) - x = \lambda(x' - x)$, it follows that there exists a $\lambda^* \in (0, 1]$ such that

$$\langle \nabla f(x), x(\lambda^*) - x \rangle + \tfrac{1}{2}\|x(\lambda^*) - x\|^2 = \lambda^* \langle \nabla f(x), x' - x \rangle + \tfrac{1}{2}\lambda^{*2}\|x' - x\|^2 < 0,$$

which shows that $\theta_X(\hat{x}) < 0$. Therefore we conclude, by contraposition, that $\theta_X(\hat{x}) = 0$ implies that (5) holds, which completes our proof.

(e) By Theorem 5.1.28, for any $x', x \in \mathbb{R}^n$, there exists an $s \in [0, 1]$ such that

$$f(x') - f(x) = \langle \nabla f(x), x' - x \rangle + \tfrac{1}{2}\langle x' - x, f_{xx}(x + s(x' - x))(x' - x) \rangle$$

It follows from (9a) that

$$f(x') - f(x) \leq \langle \nabla f(x), x' - x \rangle + \tfrac{1}{2}M\|x' - x\|^2$$

$$= \frac{1}{M}\{\langle \nabla f(x), M(x' - x) \rangle + \tfrac{1}{2}\|M(x' - x)\|^2\}, \tag{10c}$$

and also that

$$f(x') - f(x) \geq \langle \nabla f(x), x' - x \rangle + \tfrac{1}{2}m\|x' - x\|^2$$

$$= \frac{1}{m}\{\langle \nabla f(x), m(x' - x) \rangle + \tfrac{1}{2}\|m(x' - x)\|^2\}. \tag{10d}$$

Minimizing first the left-hand side and then the right-hand side of (10c), with respect to x', we find that $f(\hat{x}) - f(x) \leq \theta(x)/M$. Next, minimizing first the right-hand side and then the left-hand side of (10c), with respect to x', we find that $f(\hat{x}) - f(x) \geq \theta(x)/m$. Thus we see that (9a) holds.

(f) Since, by assumption, $M \geq 1$, and since $0 \in X - x$, we conclude that $X - x \subset M(X - x)$. Hence, first, setting $h = M(x' - x)$ and minimizing the left-hand side of (10c), over $h \in M(X - x)$, and then minimizing the right-hand side of (10c), over $x' \in X$, we conclude that

$$f(\hat{x}) - f(x) \leq \frac{1}{M}\min_{h \in M(X-x)}\{\langle \nabla f(x), h \rangle + \tfrac{1}{2}\|h\|^2\} \leq \frac{1}{M}\theta_X(x). \tag{10e}$$

Next, since by assumption $m \leq 1$, we have that $X - x \supset m(X - x)$. Hence, first setting $h = m(x' - x)$ and minimizing the right-hand side of (10d), over $h \in m(X - x)$, and then minimizing the left-hand side of (10d) over $x' \in X$, we find that

$$\frac{1}{m}\theta_X(x) \leq \frac{1}{m}\min_{h \in m(X-x)}\{\langle \nabla f(x), h \rangle + \tfrac{1}{2}\|h\|^2\} \leq f(\hat{x}) - f(x), \tag{10f}$$

which shows that (9c) is true. □

It is possible to give a geometric interpretation of the function $\theta(\cdot)$ in (8f) as follows. Suppose that the point x is given. First, referring to Fig. 1.1.1, note

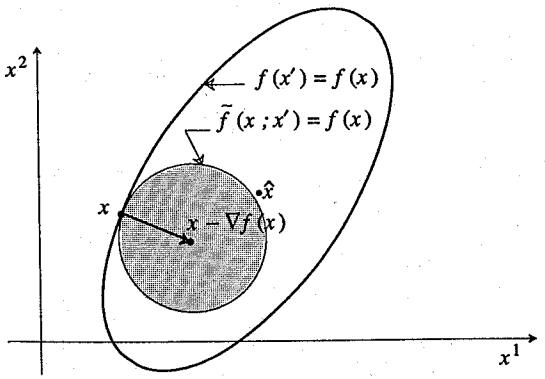


Fig. 1.1.1. A geometric interpretation of $\theta(x)$.

that the equal cost contours $\{x' \in \mathbb{R}^n \mid \tilde{f}(x, x') = \alpha\}$ of the quadratic function $\tilde{f}(x, \cdot)$, defined in (8e), are spheres with center at $x - \nabla f(x)$ and radius r given by $r^2 = 2(\alpha - f(x)) + \|\nabla f(x)\|^2$. Next, for $\alpha = f(x)$, we see that the corresponding sphere is tangent to the equal cost contour $\{x' \in \mathbb{R}^n \mid f(x') = f(x)\}$ at $x' = x$. Thus we can think of $\tilde{f}(x, x')$ as a first-order *convex* approximation to $f(x')$ and of its center $x - \nabla f(x)$ as an estimate of a minimizer, $\hat{x}$, of the function $f(\cdot)$. Referring to (8f), we see that $\theta(x) = \min_{x' \in \mathbb{R}^n} \tilde{f}(x, x') - f(x)$, i.e., it is an estimate of the decrease in cost potentially achievable from the current value $f(x)$.

The optimality functions $\theta(\cdot)$ and $\theta_X(\cdot)$, defined in (8f) and (8h) respectively, are first-order optimality functions because they correspond to first-order optimality conditions. Equation (11a) below defines a second-order optimality function, which corresponds to the second-order necessary condition in (6a).

Proposition 1.1.7. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, and that $\theta : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ are defined by*†

$$\theta(x) \triangleq \min_{h \in B(0,1)} \langle \nabla f(x), h \rangle + \tfrac{1}{2}\langle h, f_{xx}(x)h \rangle \tag{11a}$$

and

$$h(x) \triangleq \arg \min_{h \in B(0,1)} \langle \nabla f(x), h \rangle + \tfrac{1}{2}\langle h, f_{xx}(x)h \rangle . \tag{11b}$$

Then,

(a) $\theta(\cdot)$ is a continuous, negative valued function ($\theta(x) \leq 0$, $\forall x \in \mathbb{R}^n$);

(b) $\theta(x) = 0$ if and only if $\{\nabla f(x) = 0$ and $\langle h, f_{xx}(x)h \rangle \geq 0$ for all $h \in \mathbb{R}^n\}$; and

(c) the set-valued function $h(\cdot)$ is outer semicontinuous, and $df(x; h) \leq 0$ for all $h \in h(x)$.

Proof. *(a)* The fact that $\theta(\cdot)$ is continuous follows directly from Corollary 5.4.2. Since $0 \in B(0, 1)$, it is clear that $\theta(x) \leq 0$ for all $x \in \mathbb{R}^n$.

(b) $\Longleftarrow$ Suppose that $\nabla f(x) = 0$ and $\langle h, f_{xx}(x)h \rangle \geq 0$ for all $h \in \mathbb{R}^n$. Then it follows by inspection that $\theta(x) = 0$.

$\Longrightarrow$ Suppose that $\theta(x) = 0$. For the sake of contradiction, suppose that $\nabla f(x) \neq 0$. Then we see that there must exist a $\lambda \in (0, 1/\|\nabla f(x)\|]$ such that $-\lambda \nabla f(x) \in B(0, 1)$ and $-\lambda \|\nabla f(x)\|^2 + \tfrac{1}{2}\lambda^2 \langle \nabla f(x), f_{xx}(x) \nabla f(x) \rangle < 0$, which contradicts the fact that $\theta(x) = 0$. Hence, $\theta(x) = 0$ implies that $\nabla f(x) = 0$. It

† Note that when $f_{xx}(x)$ is not positive-definite, the solution to the minimization problem (11a) may not be unique.

now follows directly from the definition of $\theta(x)$ that, if $\theta(x) = 0$ and $\nabla f(x) = 0$, then $f_{xx}(x)$ is a positive-semidefinite matrix.

(c) The fact that $h(\cdot)$ is outer semicontinuous follows from Theorem 5.4.3. Suppose that $h \in h(x)$ and that $df(x; h) = \langle \nabla f(x), h \rangle > 0$. Then

$$\langle \nabla f(x), (-h) \rangle + \tfrac{1}{2}\langle (-h), f_{xx}(x)(-h) \rangle < \langle \nabla f(x), h \rangle + \tfrac{1}{2}\langle h, f_{xx}(x)h \rangle$$

must hold, contradicting our assumption that $h \in h(x)$. □

1.1.2 Sufficient Conditions

Next, we present *sufficient* conditions of optimality. Any point $\hat{x} \in \mathbb{R}^n$ which satisfies part *(a)* of these conditions must be a local minimizer for (1a), while any point $\hat{x} \in \mathbb{R}^n$ which satisfies either part *(b)* or part *(c)* of these condition must be a local minimizer for (1b).

Theorem 1.1.8. *Suppose that $f(\cdot)$ is twice continuously differentiable.*

(a) If $\hat{x} \in \mathbb{R}^n$ is such that $\nabla f(\hat{x}) = 0$, and there exists an $m > 0$ such that

$$\langle h, f_{xx}(\hat{x})h \rangle \geq m\|h\|^2, \quad \forall h \in \mathbb{R}^n , \tag{12a}$$

then $\hat{x}$ is a strict local minimizer for problem (1a).

(b) If $\hat{x} \in X$ is such that

$$\langle \nabla f(\hat{x}), x - \hat{x} \rangle \geq m\|x - \hat{x}\|, \quad \forall x \in X , \tag{12b}$$

for some $m > 0$, then $\hat{x}$ is a strict local minimizer for problem (1b).

(c) If $\hat{x} \in X$ is such that (5) holds and there exists an $m > 0$ such that

$$\langle h, f_{xx}(\hat{x})h \rangle \geq m\|h\|^2 , \tag{12c}$$

for all $h \in \mathbb{R}^n$ such that $\langle \nabla f(\hat{x}), h \rangle = 0$, then $\hat{x}$ is a strict local minimizer for problem (1b).

Proof. *(a)* Let $S \triangleq \{h \in \mathbb{R}^n \mid \|h\| = 1\}$. Since, by assumption, $f_{xx}(\cdot)$ is continuous, there exists a $\hat{\rho} > 0$ such that, for all $x \in B(\hat{x}, \hat{\rho})$ and $h \in S$,

$$|\langle h, f_{xx}(x)h \rangle - \langle h, f_{xx}(\hat{x})h \rangle| \leq \|f_{xx}(x) - f_{xx}(\hat{x})\| \leq \frac{m}{2} , \tag{13a}$$

which, because of (12a), leads to the conclusion that, for all $x \in B(\hat{x}, \hat{\rho})$ and $h \in S$,

$$\langle h, f_{xx}(x)h \rangle \geq \frac{m}{2} . \tag{13b}$$

Now, by the Mean-Value Theorem 5.1.28, part *(b)*, for any $x \in B(\hat{x}, \hat{\rho})$, there exists an $s \in [0,1]$ such that

$$f(x) - f(\hat{x}) = \langle \nabla f(\hat{x}), x - \hat{x} \rangle + \tfrac{1}{2} \langle (x - \hat{x}), f_{xx}(\hat{x} + s(x - \hat{x}))(x - \hat{x}) \rangle$$

$$= \tfrac{1}{2} \langle (x - \hat{x}), f_{xx}(\hat{x} + s(x - \hat{x}))(x - \hat{x}) \rangle \geq \frac{m}{4} \|x - \hat{x}\|^2, \qquad (13c)$$

where the last inequality follows from (13b) because $\hat{x} + s(x - \hat{x}) \in B(\hat{x}, \hat{\rho})$. Hence, we see that $\hat{x}$ is a strict local minimizer with associated radius $\hat{\rho}$.

(b) Suppose that (12b) holds and that $\hat{x}$ is not a strict local minimizer for (1b). Then there exists a sequence $\{x_i\}_{i=0}^{\infty}$ in X such that $x_i \to \hat{x}$, as $i \to \infty$, and

$$f(x_i) \leq f(\hat{x}), \quad \forall i \in \mathbb{N}. \qquad (13d)$$

It follows from (12b), (13d), the Mean-Value Theorem 5.1.28 (see (5.1.17b)), and the continuity of $\nabla f(\cdot)$ that there exists an i_0, and $s_i \in [0,1]$ such that, for all $i \geq i_0$,

$$0 \geq \langle \nabla f(\hat{x} + s_i(x_i - \hat{x})), x_i - \hat{x} \rangle$$

$$= \langle \nabla f(\hat{x}), x_i - \hat{x} \rangle + \langle \nabla f(\hat{x} + s_i(x_i - \hat{x})) - \nabla f(\hat{x}), x_i - \hat{x} \rangle$$

$$\geq \|x_i - \hat{x}\| [m - \|\nabla f(\hat{x} + s_i(x_i - \hat{x}) - \nabla f(\hat{x})\|] \geq \tfrac{1}{2} m, \qquad (13e)$$

which is clearly impossible. Hence, we have a contradiction, and we see that $\hat{x}$ is a strict local minimizer for (1b).

(c) Recall that by (5), $\langle \nabla f(\hat{x}), x - \hat{x} \rangle \geq 0$ for all $x \in X$. If $\nabla f(\hat{x}) = 0$, then the proof of this part follows directly from part *(b)*. Hence, suppose that $\nabla(\hat{x}) \neq 0$ and let $h \in S$ be arbitrary, where $S \triangleq \{h \in \mathbb{R}^n \mid \|h\| = 1\}$. We can write $h = h' + h''$, with $\langle \nabla f(\hat{x}), h' \rangle = 0$, and $\langle h', h'' \rangle = 0$. Hence it follows that $h'' = \delta \nabla f(\hat{x})$, for some $\delta \in \mathbb{R}$. Now suppose that $|\langle \nabla f(\hat{x}), h \rangle| \leq \varepsilon$ for some $\varepsilon > 0$. Then, because

$$|\langle \nabla f(\hat{x}), h \rangle| = |\langle \nabla f(\hat{x}), h'' \rangle| = \|\nabla f(\hat{x})\| \|h''\|, \qquad (13f)$$

$\|h''\| \leq \varepsilon'' \triangleq \varepsilon / \|\nabla f(\hat{x})\|$ must hold. Hence, in view of (12c), if follows that

$$\langle f_{xx}(\hat{x})h, h \rangle = \langle f_{xx}(\hat{x})(h' + h''), (h' + h'') \rangle$$

$$= \langle f_{xx}(\hat{x})h', h' \rangle + \langle f_{xx}(\hat{x})h'', h' + h \rangle$$

$$\geq m\|h'\|^2 - 2\varepsilon'' \|f_{xx}(\hat{x})\|, \qquad (13g)$$

where we have used the fact that $\|h'\| \leq \|h\| = 1$. Since $h \in S$, $\|h'\| \to \|h\|$, as

$\varepsilon \to 0$ (and hence $\varepsilon'' \to 0$). Therefore there exists an $\varepsilon > 0$ such that, for all $h \in S$ satisfying

$$|\langle \nabla f(\hat{x}), h \rangle| \leq \varepsilon, \qquad (13h)$$

$$\langle h, f_{xx}(\hat{x})h \rangle \geq \tfrac{1}{2} m \qquad (13i)$$

must hold. Let $\varepsilon > 0$ be as above, and let $\rho \in (0,1]$ be such that, for all $x \in B(\hat{x}, \rho)$, $\|\nabla f(\hat{x}) - \nabla f(x)\| \leq \varepsilon/2$, and also $\|f_{xx}(\hat{x}) - f_{xx}(x)\| \leq m/4$. Let the point $x \in X \cap B(\hat{x}, \rho)$, $x \neq \hat{x}$, be arbitrary, and let $h = x - \hat{x}$. Then $\hat{x} + \lambda h \in X \cap B(\hat{x}, \rho)$ for all $\lambda \in [0,1]$. If $\langle \nabla f(\hat{x}), h \rangle \geq \varepsilon$, then, by the Mean-Value Theorem 5.1.28, part *(a)*, for every $\lambda \in [0,1]$, there exists an s_λ such that

$$f(\hat{x} + \lambda h) - f(\hat{x}) = \lambda \{\langle \nabla f(\hat{x}), h \rangle + [\langle \nabla f(\hat{x} + s_\lambda h), h \rangle - \langle \nabla f(\hat{x}), h \rangle]\}$$

$$\geq \lambda(\varepsilon - \varepsilon/2) = \lambda\varepsilon/2. \qquad (13j)$$

By assumption, $\langle \nabla f(\hat{x}), h \rangle \geq 0$. Hence, if (13h) holds, then, for any $\lambda \in [0,1]$, there exists an s_λ such that

$$f(\hat{x} + \lambda h) - f(\hat{x}) \geq \tfrac{1}{2}\lambda^2 \{\langle h, f_{xx}(\hat{x})h \rangle$$

$$+ [\langle h, f_{xx}(\hat{x} + s_\lambda h)h \rangle - \langle h, f_{xx}(\hat{x})h \rangle]\}$$

$$\geq \lambda(m/2 - m/4)\|h\|^2 = \lambda m \|h\|^2/4. \qquad (13k)$$

In view of (13j) and (13k), we conclude that $\hat{x}$ is a strict local minimizer with associated radius ρ. □

1.1.3 The Convex Case

It is clear from the above that in the general case we have tests only for identifying *local* minimizers. We have no reasonable tests for determining whether a point is a *global* minimizer. In the case of convex functions, the situation is different.

Theorem 1.1.9. *Consider the problems (1a) and suppose that $f(\cdot)$ is convex and that its directional derivative $df(x; h)$ exists for all x, $h \in \mathbb{R}^n$.*

(a) If $\hat{x}$ is such that (3) holds, then $\hat{x}$ is a global minimizer for (1a).

(b) If $\hat{x}$ is such that (5) holds, then $\hat{x}$ is a global minimizer for (1b).

Proof. *(a)* Let $x \in \mathbb{R}^n$ be arbitrary. Then, because $f(\cdot)$ is convex, for any $\lambda \in (0,1]$, by (5.2.6b),

$$\frac{f(\hat{x}+\lambda(x-\hat{x}))-f(\hat{x})}{\lambda} \leq f(x)-f(\hat{x}), \quad \forall \lambda \in [0,1]. \tag{14}$$

Taking the limit in (14), as $\lambda \to 0$, we conclude that $f(x)-f(\hat{x}) \geq df(\hat{x};x-\hat{x}) \geq 0$ and hence that $\hat{x}$ is a global minimizer for (1a).

(b) Let $x \in X$ be arbitrary. Then, for any $\lambda \in (0,1]$, because X is convex, $\hat{x}+\lambda(x-\hat{x}) \in X$, and, because $f(\cdot)$ is convex, (14) holds. Taking the limit in (14), as $\lambda \to 0$, we conclude that $f(x)-f(\hat{x}) \geq df(\hat{x};x-\hat{x}) \geq 0$ and hence that $\hat{x}$ is a global minimizer for (1b). □

The following result is just a special case of *(a)* in the above theorem.

Corollary 1.1.10. *Suppose that $f:\mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and convex. If $\hat{x} \in \mathbb{R}^n$ is such that $\nabla f(\hat{x})=0$, then $\hat{x}$ is a global minimizer of $f(\cdot)$.* □

Theorem 1.1.11. *Consider the problems (1a) and (1b). If $f:\mathbb{R}^n \to \mathbb{R}$ is strictly convex, then the problems (1a) and (1b) can have, at most, one global minimizer.*

Proof. First, consider problem (1b). Suppose that x^*, $x^{**} \in X$ are two global minimizers for problem (1b). Then $f(x^*)=f(x^{**})$ must hold. Since, by assumption, X is convex, for any $\lambda \in [0,1]$, $\lambda x^{**}+(1-\lambda)x^* \in X$ and since $f(\cdot)$ is strictly convex, for any $\lambda \in (0,1)$,

$$f(\lambda x^{**}+(1-\lambda)x^*) < \lambda f(x^{**})+(1-\lambda)f(x^*) = f(x^*), \tag{15}$$

which contradicts the fact that x^* is a global minimizer.

Now consider problem (1a). Letting $X=\mathbb{R}^n$ in (1b), we obtain the desired result from the above. □

Corollary 1.1.12. *Consider the problems (1a) and (1b), with X in (1b) a closed, convex set. Suppose that $f:\mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and that there exists an $m>0$ such that, for all $x \in \mathbb{R}^n$, $h \in \mathbb{R}^n$, $\langle h, f_{xx}(x)h\rangle \geq m\|h\|^2$. Then the problems (1a) and (1b) have unique global minimizers.*

Proof. First, consider problem (1b). It follows from Theorem 5.2.14 that $f(\cdot)$ is strictly convex. Hence, if problem (1b) has a global minimizer, that minimizer is unique. Thus, we need to prove only that a global minimizer exists. Let $x_0 \in X$ be arbitrary. Then, by Proposition 5.2.15, the level set $L \triangleq \{x \mid f(x) \leq f(x_0)\}$ is compact. Because L is compact, $L \cap X$ is compact, and the existence of a global minimizer now follows from Corollary 5.1.25 and the fact that

$$\inf_{x \in X} f(x) = \inf_{x \in L \cap X} f(x) = \min_{x \in L \cap X} f(x). \tag{16}$$

Letting $X=\mathbb{R}^n$ in problem (1b), we obtain the desired result for (1a). □

1.2 Algorithm Models and Convergence Conditions I

Since, in the absence of convexity, we do not have tests for identifying global minimizers of a problem of the form

$$\min_{x \in \mathbb{R}^n} f(x), \tag{1a}$$

even when $f:\mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, the most we can expect is to find a point that satisfies a local optimality condition. In general, even such a point cannot be obtained explicitly. Needless to say, the same statement remains true for problems of the form

$$\min_{x \in X} f(x), \tag{1b}$$

where $f:\mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and X is a convex, closed subset of $\mathbb{R}^n$. Hence, problems of the form (1a) and (1b) are usually "solved" by iterative methods, which construct infinite sequences, $\{x_i\}_{i=0}^{\infty}$, of progressively better approximations to a "solution", i.e., a point satisfying an optimality condition. We will refer to such iterative methods as *optimization algorithms*.

This section is devoted to an abstract theory of optimization algorithms and to elementary results on the rate of convergence of sequences and algorithm efficiency. The abstract theory formulates a model problem to be solved and a set of algorithm models for its solution. The model problem reflects the fact that, as a practical matter, quite often points satisfying an optimality condition must be accepted as "solutions" to an optimization problem. The theory provides conditions which ensure that every accumulation point of a sequence constructed by an algorithm model is a solution to the model problem. These convergence conditions are satisfied by many optimization algorithms in the literature and are very helpful both in analyzing existing algorithms and in improvising new ones.

In the remainder of this chapter and in parts of the other chapters, we will establish the convergence properties of a particular algorithm by showing that it satisfies the convergence conditions for a corresponding algorithm model in this section. Hence the reader has two choices. The first is to study this section, in its entirety, before attempting to read the rest of Chapter 1. The second, which may be preferable for the novice, is to read only the first two subsections, which present a rationale for the abstract theory and a few basic models (omitting Algorithm Model 1.2.17), and the last subsection which deals with rate of

convergence. The remaining material in this section can then be read when required by the analysis of a particular algorithm. Thus, the Wolf theorem will be needed in the study of conjugate gradient methods. The Polak-Sargent-Sebastian theorem will be needed in the study of gradient, global Newton and quasi-Newton methods. Trust regions are used to construct global versions of Newton and quasi-Newton methods, and implementation theory will be needed in the study of optimization algorithms that use discretization or approximation techniques to evaluate, approximately, functions and their gradients.

1.2.1 Geometry of Descent Methods

An algorithm for solving problem (1a) (or (1b)) will be called a *descent method* if it constructs sequences $\{x_i\}_{i=0}^{\infty}$ such that $f(x_{i+1}) < f(x_i)$ for all $i \in \mathbb{N}$ (and $x_i \in X$ for all $i \in \mathbb{N}$). To display the geometry of a descent method, we need to introduce the concept of a level set.

Definition 1.2.1. *Given a function $f : \mathbb{R}^n \to \mathbb{R}$ and an $\alpha \in \mathbb{R}$, such that $\alpha \geq \inf_{x \in \mathbb{R}^n} f(x)$, we will say that the set $L_\alpha(f) \subset \mathbb{R}^n$, defined by*

$$L_\alpha(f) \triangleq \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}, \tag{2}$$

is a level set *of $f(\cdot)$, parametrized by α.* □

The boundary $\partial L_\alpha(f)$ of a level set $L_\alpha(f)$ (see Fig. 1.2.1a) can be visualized as a constant altitude line on a topographical map. Points on the boundary of $L_\alpha(f)$ satisfy the equation $f(x) = \alpha$. Again, referring to Fig. 1.2.1a, we see that, if $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by a descent method in solving (1b),

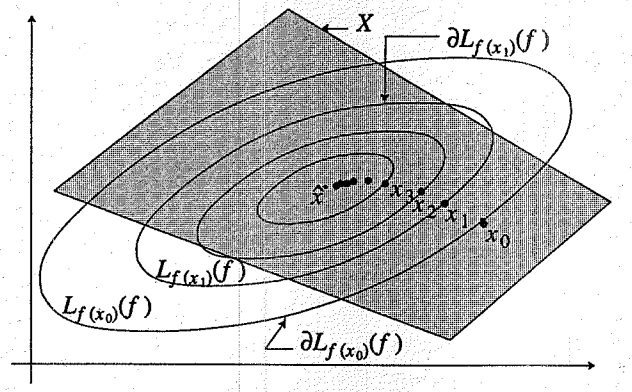


Fig. 1.2.1a. The geometry of a descent method for problem (1b).

then for every i, the "tail", $\{x_j\}_{j=i+1}^{\infty}$, of this sequence must satisfy $x_j \in L_{f(x_i)}(f) \cap X$, for all $j \geq i+1$.

Descent methods are usually based on *descent directions* which are elements of a descent cone (see Fig. 1.2.1b), defined as follows.

Definition 1.2.2. *Consider the function $f : \mathbb{R}^n \to \mathbb{R}$ and suppose that it is directionally differentiable. Given a point $x \in \mathbb{R}^n$, we define the* descent cone *for $f(\cdot)$ at x by*

$$DC_f(x) \triangleq \{h \in \mathbb{R}^n \mid f(x + \lambda h) < f(x),\ \forall \lambda \in (0, \lambda_h],\ \lambda_h > 0\}, \tag{3a}$$

and we define the first-order descent cone *for $f(\cdot)$ at x by*

$$DC'_f(x) \triangleq \{h \in \mathbb{R}^n \mid df(x; h) < 0\}. \tag{3b}$$

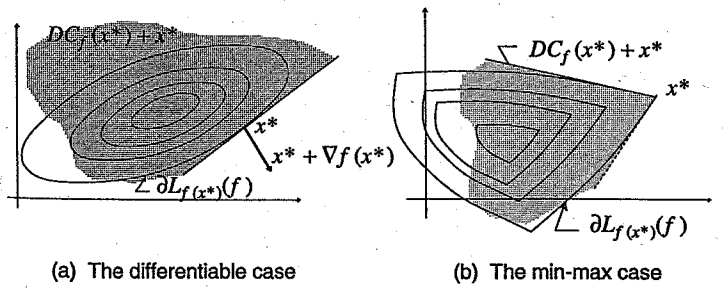
□

From the definition of $DC_f(x)$, it follows that if $h \in DC_f(x)$, then there exists a $\lambda_h > 0$ such that $x + \lambda h \in L_{f(x)}(f)$ for all $\lambda \in [0, \lambda_h]$. Referring to Fig. 1.2.1b, we see that, when the function $f(\cdot)$ is differentiable, the set $DC_f(x^*)$ is a half space with normal $\nabla f(x^*)$, while when the function $f(\cdot)$ is the maximum of a finite number of convex functions, i.e., when $f(x) = \max_{j \in q} f^j(x)$, $DC_f(x^*)$ is a convex cone. The following is obvious:

Proposition 1.2.3. *The level sets of a continuous function $f : \mathbb{R}^n \to \mathbb{R}$ have the following properties:*

(a) If $\alpha_1 > \alpha_2$, then $L_{\alpha_2}(f) \subset L_{\alpha_1}(f)$, i.e., the level sets are nested.

(b) If $\hat{\alpha} = \min_{x \in \mathbb{R}^n} f(x)$, then $L_{\hat{\alpha}}(f)$ is the set of global minimizers for the problem (1a). Similarly, if $\hat{\alpha} = \min_{x \in X} f(x)$, then $L_{\hat{\alpha}}(f)$ is the set of global minimizers for the problem (1b). □



(a) The differentiable case (b) The min-max case

Fig. 1.2.1b. The descent cones.

The following proposition establishes a relation between the actual and first-order descent cones.

Proposition 1.2.4. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ has a directional derivative $df(x;h)$, at any $x \in \mathbb{R}^n$, in any direction $h \in \mathbb{R}^n$. Then $DC'_f(x) \subset DC_f(x)$.*

Proof. Suppose that $h \in DC'_f(x)$. Then, because

$$0 > df(x;h) = \lim_{\lambda \downarrow 0} \frac{f(x+\lambda h) - f(x)}{\lambda},$$

it follows that there exists a $\lambda^* > 0$ such that, for all $\lambda \in (0, \lambda^*]$,

$$f(x+\lambda h) - f(x) \leq \tfrac{1}{2}\lambda\, df(x;h) < 0,$$

which completes our proof. □

Exercise 1.2.5. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function that has a directional derivative $df(x;h)$ at every $x \in \mathbb{R}^n$ in any direction $h \in \mathbb{R}^n$. Show that, in this case, $DC'_f(x) = DC_f(x)$. □

1.2.2 Basic Algorithm Models

The fact that an algorithm is a descent method is not sufficient to guarantee that it has useful properties. For example, suppose that $f(x) \triangleq \|x\|^2$, and consider an algorithm which, starting from an initial point x_0 such that $\|x_0\| = 1$, constructs a sequence, $\{x_i\}_{i=0}^{\infty}$, according to the recursion rule

$$x_{i+1} = x_i + \lambda_i h_i, \quad i = 1, 2, \ldots, \tag{4a}$$

where

$$\lambda_i = \arg\min_{\lambda \geq 0} f(x_i + \lambda h_i) \tag{4b}$$

and h_i is such that

$$\langle \nabla f(x_i), h_i \rangle = -\frac{2}{3(2i+1)} \|\nabla f(x_i)\| \|h_i\|. \tag{4c}$$

Note that this recursion rule guarantees that $\{f(x_i)\}_{i=0}^{\infty}$ is strictly monotonically decreasing.

Since, for this example, $\nabla f(x) = 2x$ and $f_{xx}(x) \equiv 2I$, it follows from (4c) that

$$f(x_i + \lambda h_i) - f(x_i) = \lambda \langle \nabla f(x_i), h_i \rangle + \lambda^2 \|h_i\|^2$$

$$= -\lambda \frac{2}{3(2i+1)} \|\nabla f(x_i)\| \|h_i\| + \lambda^2 \|h_i\|^2.$$

$$= -\lambda \frac{4}{3(2i+1)} \|x_i\| \|h_i\| + \lambda^2 \|h_i\|^2. \tag{5a}$$

Hence, after some rearranging of terms, it follows from (4b), that

$$f(x_{i+1}) = \|x_{i+1}\|^2 = \left[1 - \frac{4}{9(2i+1)^2}\right] \|x_i\|^2, \quad i = 1, 2, 3, \ldots, \tag{5b}$$

and, therefore, since $\|x_0\| = 1$, we conclude that $\|x_i\|^2 \to \Pi_{i=0}^{\infty} [1 - 4/9(2i+1)^2]$, as $i \to \infty$. Since, for any θ, $\cos\theta = \Pi_{i=0}^{\infty} [1 - 4\theta^2/(2i+1)^2\pi^2]$, it follows that $\cos \pi/3 = \frac{1}{2} = \Pi_{i=1}^{\infty} [1 - 4/9(2i+1)^2]$, and hence we conclude that $\|x_i\|^2 \to \frac{1}{2}$, which is nowhere near the minimum cost of 0. Furthermore, since every accumulation point $\hat{x}$ of the above constructed sequence has norm $(\frac{1}{2})^{\frac{1}{2}}$, we see that not only does it fail to be a minimizer for our problem, but it also fails to satisfy the first-order necessary condition $\nabla f(\hat{x}) = 0$.

Since we do not have tests for identifying global minimizers, their computation, in the absence of convexity can be dealt with only in a probabilistic setting. Hence we will replace problem (1b) with the problem of computing a stationary point, i.e., a point satisfying an optimality condition. Because both local minimizers and local maximizers can sometimes satisfy the same optimality condition, it is desirable to use descent methods for this task, since they cannot converge to a local maximizer, although they can (but are not likely to) converge to a saddle point.

In the remainder of this section, we will establish conditions that ensure satisfactory convergence properties for descent methods used to find stationary points for problems of the form (1b). Because problem (1b) reduces to problem (1a) when $X = \mathbb{R}^n$, these conditions will apply equally to descent methods for solving problem (1a).

Our first step is to define formally the problem that we wish to solve. We will assume that we are given an upper semicontinuous, nonpositive-valued *optimality function* $\theta : \mathbb{R}^n \to \mathbb{R}$, which vanishes at all the local minimizers of the problem (1b). We have seen several examples of optimality functions in Section 1. In particular, when assumptions permit, $\theta(\cdot)$ can be chosen to be $\theta_X(\cdot)$, defined as in (1.1.8h). We will call the zeros of $\theta(\cdot)$ *quasi-stationary points*. We will denote the set of all the *quasi-stationary points* by $\mathcal{QS}$, i.e.,

$$\mathcal{QS} \triangleq \{x \in \mathbb{R}^n \mid \theta(x) = 0\}. \tag{6a}$$

It is entirely possible that $\theta(x) = 0$ at a point $x \notin X$ or at points in $x \in X$ that are not local minimizers of problem (1b).

Since it is important that the points that we compute also satisfy constraints, we must introduce the set of *feasible* quasi-stationary points, which we will call *stationary points*. We will denote by $\mathcal{S}$ the set of all stationary points, i.e.,

$$S \triangleq \{x \in X \mid \theta(x) = 0\}. \tag{6b}$$

We are now ready to state formally the problem that we wish to solve:

Model Problem 1.2.6.

Given:

(a) In (1b), $f : \mathbb{R}^n \to \mathbb{R}$ is continuous.

(b) In (1b) and in (6b), X is either a closed, convex subset of $\mathbb{R}^n$ or $X = \mathbb{R}^n$.

(c) In (6b),

(i) $\theta : \mathbb{R}^n \to \mathbb{R}$ is upper semicontinuous;

(ii) For all $x \in \mathbb{R}^n$, $\theta(x) \leq 0$;

(iii) If $\hat{x}$ is a local minimizer of (1b), then $\theta(\hat{x}) = 0$.

Construct:

a sequence $\{x_i\}_{i=0}^{k} \subset X$ such that

(i) $f(x_{i+1}) < f(x_i)$ for all i, and

(ii) if $k < \infty$, then $x_k \in S$, and, if $k = \infty$, then every accumulation point of $\{x_i\}_{i=0}^{\infty}$ is in S.

The simplest descent methods for solving the Model Problem 1.2.6 utilize an *algorithm function* $a : X \to X$ and are of the following form:

Algorithm Model 1.2.7.

Data. $x_0 \in X$.

Step 0. Set $i = 0$.

Step 1. Compute $x_{i+1} = a(x_i)$.

Step 2. If $\theta(x_{i+1}) = 0$, stop. Else, replace i by $i + 1$, and go to Step 1.

Note that Algorithm Model 1.2.7 must be initialized with a point $x_0 \in X$ and that it constructs sequences $\{x_i\}_{i=0}^{\infty}$ contained in X. Hence, any accumulation point of such a sequence that is quasi-stationary is automatically stationary. This will also be the case with all the other algorithm models in this section. Such algorithm models correspond to various unconstrained optimization algorithms, as well as projected gradient, projected conjugate gradient, projected quasi-Newton, and projected Newton type algorithms. In addition, various "phase II" type algorithms, can be analyzed using these models.

Algorithm Model 1.2.7 can be considered a discrete-time, dynamical system, i.e., one whose behavior is governed by the nonlinear difference equation $x_{i+1} = a(x_i)$, $i \in \mathbb{N}$, for which the stable equilibrium points ought to be the quasi-stationary points† of problem (1b). There is a considerable body of literature dealing with the stability of such systems. The assumptions in the theorem below and in convergence theorems to follow were inspired by Lyapunov stability theory (see [LaL.61]). Thus, in the theorems below, the cost function plays the role of a Lyapunov function. To ensure convergence of Algorithm Model 1.2.7, we must postulate an appropriate descent property for the algorithm function $a(\cdot)$, with respect to the cost function $f(\cdot)$ in (1b). Extrapolating from the assumptions appearing in stability theorems, we will assume that our algorithm has a *monotone uniform descent* (MUD) property, i.e., every every non-quasi-stationary point x has a neighborhood from which the algorithm decreases the cost by a guaranteed amount, as illustrated in Fig. 1.2.2a. When this condition is stated technically, we obtain the following result.

Theorem 1.2.8. *Consider Algorithm Model 1.2.7 together with Model Problem 1.2.6. Suppose that, for every $x \in X$ that is not stationary ($\theta(x) < 0$), there exist a $\rho_x > 0$ and a $\delta_x > 0$ such that*

$$f(a(x')) - f(x') \leq -\delta_x < 0, \quad \forall x' \in B(x, \rho_x) \cap X. \tag{7}$$

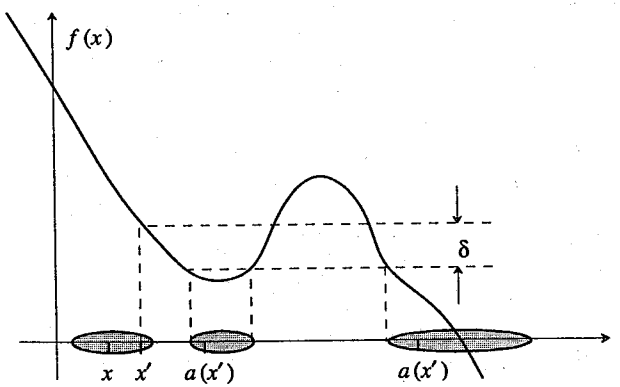


Fig. 1.2.2a. Possible locations of $a(x')$.

† Since, by assumption, the function $a(\cdot)$ maps X into X and $x_0 \in X$, these quasi-stationary points are automatically stationary.

Then, either the sequence $\{x_i\}$ constructed by Algorithm Model 1.2.7 is finite and its last element is stationary, or else it is infinite and every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ is stationary (i.e., $\hat{x} \in X$, and $\theta(\hat{x}) = 0$).

Proof. By assumption, $a(\cdot)$ maps X into X, and $x_0 \in X$. Hence the sequence $\{x_i\}$ constructed by Algorithm Model 1.2.7 is contained in X, and so are its accumulation points. Clearly, by the test in Step 2 of Algorithm Model 1.2.7, the sequence $\{x_i\}$ is finite if and only if its last element is quasi-stationary, and since it is in X, this last element is stationary.

Next we note that (7) implies that $f(a(x)) < f(x)$ for all $x \in X$ that are not quasi-stationary. Hence, suppose that the sequence $\{x_i\}_{i=0}^{\infty}$ is infinite and that it has a subsequence that converges to the point $\hat{x} \in X$, i.e., $x_i \to^K \hat{x}$ as $i \to \infty$, for some infinite subset $K \subset \mathbb{N}$. If $\hat{x}$ is not quasi-stationary, then, by assumption, there exist a $\hat{\rho} > 0$, a $\hat{\delta} > 0$, and an $i_o \in K$ such that, for all $i \geq i_o$, $i \in K$, $\|x_i - \hat{x}\| \leq \hat{\rho}$, and hence for all $i \geq i_o, i \in K$,

$$f(x_{i+1}) - f(x_i) \leq -\hat{\delta}. \tag{8}$$

Now, because $f(\cdot)$ is continuous, $f(x_i) \to^K f(\hat{x})$, as $i \to \infty$, and, by construction, $f(x_{i+1}) < f(x_i)$ for all i. Therefore it follows from the Monotone Sequences Proposition 5.1.16 that $f(x_i) \to f(\hat{x})$ as $i \to \infty$. Since this is contradicted by (8), our proof is complete. □

It is informative to note the following consequence of the MUD assumption in Theorem 1.2.8. Suppose that $x \in X$ is not quasi-stationary (i.e., $\theta(x) < 0$) and that $\delta > 0$ and $\rho > 0$ are associated constants satisfying (7). Since $f(\cdot)$ is continuous, there exists a $\rho^* \in (0, \rho]$ such that, for all $x' \in B(x, \rho^*)$, $|f(x') - f(x)| \leq \delta/4$. Consequently, it follows from (7) that, for all $x' \in B(x, \rho^*)$, $f(a(x')) \leq f(x) - 3\delta/4$, as illustrated in Fig.1.2.2b. It should be clear from Fig.1.2.2b that, once Algorithm Model 1.2.7 has constructed a point $x_{i_0} \in B(x, \rho^*)$, then, for all $i > i_0$, $x_i \notin B(x, \rho^*)$ must hold. Hence, x cannot be an accumulation point of the sequence generated by the Algorithm Model 1.2.7.

In many applications, it is more natural to use the following consequence of Theorem 1.2.8.

Corollary 1.2.9. *Consider Algorithm Model 1.2.7 together with Model Problem 1.2.6. Suppose that, for every $x \in X$, there exist a $\rho_x > 0$ and a $\gamma_x > 0$ such that*

$$f(a(x')) - f(x') \leq \gamma_x \theta(x'), \quad \forall x' \in B(x, \rho_x) \cap X. \tag{9}$$

Then, either the sequence $\{x_i\}$ constructed by Algorithm Model 1.2.7 is finite

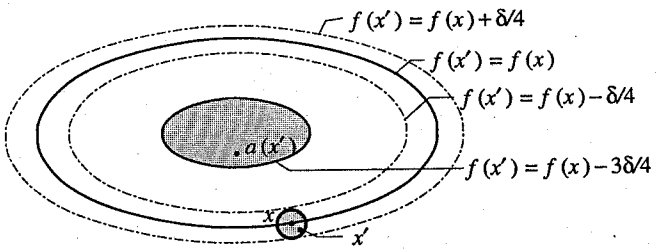


Fig. 1.2.2b. Illustration of the MUD property.

and its last element is stationary, or else it is infinite and every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ is stationary (i.e., $\hat{x} \in X$, and $\theta(\hat{x}) = 0$). □

It should be obvious that the existence of $\rho_x > 0$ and a $\gamma_x > 0$ such that (9) holds implies that, for some $\rho'_x \in (0, \rho_x]$, (7) holds with $\delta_x \triangleq -\gamma_x \theta(x)/2$.

Exercise 1.2.10. Consider Algorithm Model 1.2.7 together with Model Problem 1.2.6. Suppose that the function $a(\cdot)$ is continuous and that $f(a(x)) \geq f(x)$ if and only if $x \in X$ is stationary ($\theta(x) = 0$). Show that in this case the MUD assumption (7) in Theorem 1.2.8 is satisfied. □

Because many algorithms contain step-size or search direction scaling parameters that may be varied from iteration to iteration and because some algorithms use subprocedures, such as the simplex algorithm for linear programming, which may return different answers depending on the code being used, it is necessary to consider algorithm models that use a *set-valued* algorithm function A, mapping X into the set of all nonempty subsets of X (which we write as $A : X \to 2^X$). Then we get the following obvious modification of Algorithm Model 1.2.7.

Algorithm Model 1.2.11.

Data. $x_0 \in X$.

Step 0. Set $i = 0$.

Step 1. Compute an $x_{i+1} \in A(x_i)$.

Step 2. If $\theta(x_{i+1}) = 0$, stop. Else, replace i by $i + 1$, and go to Step 1.

Theorem 1.2.12. *Consider Algorithm Model 1.2.11 together with Model Problem 1.2.6. Suppose that, for every $x \in X$ which is not quasi-stationary (i.e. $\theta(x) < 0$), there exist a $\rho_x > 0$ and a $\delta_x > 0$ such that*

$$f(x'') - f(x') \leq -\delta_x < 0\,, \tag{10}$$

for all $x' \in B(x, \rho_x) \cap X$, *for all* $x'' \in A(x')$.

(a) If Algorithm Model 1.2.11 constructs a finite sequence $\{x_i\}_{i=0}^k$, *then its last element is stationary, i.e.,* $x_k \in S$.

(b) If Algorithm Model 1.2.11 constructs an infinite sequence $\{x_i\}_{i=0}^\infty$, *then every accumulation point of* $\hat{x}$ *of* $\{x_i\}_{i=0}^\infty$ *is stationary, i.e.,* $\hat{x} \in S$. □

Corollary 1.2.13. *Consider Algorithm Model 1.2.11 together with Model Problem 1.2.6. Suppose that for every* $x \in X$ *there exist a* $\rho_x > 0$ *and a* $\gamma_x > 0$ *such that*

$$f(x'') - f(x') \leq \gamma_x \theta(x')\,, \tag{11}$$

for all $x' \in B(x, \rho_x) \cap X$, *for all* $x'' \in A(x')$.

Then, either the sequence $\{x_i\}$ *constructed by Algorithm Model 1.2.11 is finite and its last element is stationary, or else it is infinite and every accumulation point* $\hat{x}$ *of* $\{x_i\}_{i=0}^\infty$ *is stationary (i.e.,* $\hat{x} \in X$ *and* $\theta(\hat{x}) = 0$*).* □

Exercise 1.2.14. Prove Theorem 1.2.12 and Corollary 1.2.13. □

Exercise 1.2.15. Consider Algorithm Model 1.2.11 together with Model Problem 1.2.6. Suppose that *(a)* the set-valued function $A(\cdot)$ is outer semicontinuous (see Definition 5.3.1) and compact-valued, and *(b)* $f(x') \geq f(x)$ for some $x' \in A(x)$ if and only if $x \in X$ is quasi-stationary, i.e, $\theta(x) = 0$.

Show that, if Algorithm Model 1.2.11 constructs a bounded infinite sequence $\{x_i\}_{i=0}^\infty$ such that $\theta(x_i) < 0$ for all $i \in \mathbb{N}$ and $\hat{x}$ is an accumulation point of $\{x_i\}_{i=0}^\infty$, then $\hat{x} \in X$ and $\theta(\hat{x}) = 0$, i.e., $\hat{x}$ is stationary, i.e. $\hat{x} \in S$. □

Remark 1.2.16. The reader should be careful not to read more into the statements of the convergence theorems above than they actually say. Note that these theorems state only that, *if* a convergent subsequence exists, then its limit point will be stationary. To ensure that accumulation points exist, it is necessary to make some additional assumptions. For example, one may assume that the set X is compact or that the level set $\{x \in X \mid f(x) \leq f(x_0)\}$ is compact, where x_0 is the starting point for the algorithm. The reason for not including such assumptions in the statement of theorems such as 1.2.8 and 1.2.12 is that it is usually better to use information associated with a specific problem to determine whether an algorithm will produce bounded sequences or not. □

The two algorithm models above apply to "one-point" algorithms, i.e., algorithms that construct x_{i+1} using only the information contained in x_i. As we will see, these include gradient methods, Newton's method, and minimax algorithms. There are also many "multi-point" algorithms, including secant

methods, quasi-Newton methods and conjugate gradient methods, which use either some or all of the preceding points in the construction of x_{i+1}. We proceed to develop an algorithm model for dealing with such methods.

First, we need to simplify our notation for finite sequences in X. Given a finite sequence $\{x_i\}_{i=0}^k \subset X$, we let

$$\underline{x}_k \triangleq \{x_i\}_{i=0}^k, \tag{12a}$$

and we define $\underline{X}$ as the set of all finite sequences in X, i.e.,

$$\underline{X} \triangleq \{\underline{x}_k \mid \underline{x}_k \subset X\,,\ k \in \mathbb{N}\}. \tag{12b}$$

Next, we introduce a multi-point algorithm model that uses an algorithm function $\underline{A} : \underline{X} \to 2^X$:

Algorithm Model 1.2.17.

Data. $x_0 \in X$.

Step 0. Set $i = 0$.

Step 1. Compute an $x_{i+1} \in \underline{A}(\underline{x}_i)$.

Step 2. If $\theta(x_{i+1}) = 0$, stop. Else, replace i by $i + 1$, and go to Step 1.

It should be clear that Algorithm Model 1.2.17 includes Algorithm Model 1.2.11 as a special case, which is obtained by defining $\underline{A}(\underline{x}_i) \triangleq A(x_i)$.

Theorem 1.2.18. *Consider Algorithm Model 1.2.17 together with Model Problem 1.2.6. Suppose that, for every* $x \in X$ *that is not stationary, at least one of the following two statements holds:*

(a) There exist a $\rho_x > 0$ *and a* $\delta_x > 0$ *such that, if* $\underline{x}_i = \{x_j\}_{j=0}^i$ *is constructed by Algorithm Model 1.2.17 and* $x_i \in B(x, \rho_x)$, *then*

$$f(x') - f(x_i) \leq -\delta_x < 0\,,\ \forall\, x' \in \underline{A}(\underline{x}_i)\,. \tag{13a}$$

(b) There exist a $\rho_x > 0$ *and a* $\gamma_x > 0$ *such that, if* $\underline{x}_i = \{x_j\}_{j=0}^i$ *is constructed by Algorithm Model 1.2.17 and* $x_i \in B(x, \rho_x)$, *then*

$$f(x') - f(x_i) \leq -\gamma_x \theta(x_i) < 0\,,\ \forall\, x' \in \underline{A}(\underline{x}_i)\,. \tag{13b}$$

Then, either the sequence $\{x_i\}$ *constructed by Algorithm Model 1.2.17 is finite and its last element is stationary, or else it is infinite and every accumulation point* $\hat{x}$ *of* $\{x_i\}_{i=0}^\infty$ *is stationary (i.e.,* $\hat{x} \in X$, *and* $\theta(\hat{x}) = 0$*).*

Proof. Since the existence of a $\rho_x > 0$ and a $\gamma_x > 0$ such that (13b) holds implies the existence of a $\rho_x > 0$ and a $\delta_x > 0$ such that (13a) holds, we need to establish the theorem only under hypothesis (13a). First, by assumption on $\underline{A}(\cdot)$,

the sequence $\{x_i\}$ constructed by Algorithm Model 1.2.17 is contained in X, and hence so are its accumulation points.

Clearly, by the test in Step 2 of Algorithm Model 1.2.17, the sequence $\{x_i\}$ is finite, if and only if its last element is quasi-stationary and hence, since it is in X, stationary.

Next, suppose that the sequence $\{x_i\}_{i=0}^{\infty}$ is infinite and that it has a subsequence which converges to the point $\hat{x} \in X$, i.e., $x_i \to^K \hat{x}$, as $i \to \infty$, for some infinite subset $K \subset \mathbb{N}$. If $\hat{x}$ is not quasi-stationary, then, by assumption, there exist a $\hat{\rho} > 0$, a $\hat{\delta} > 0$, and a $i_o \in K$ such that, for all $i \geq i_o$, $i \in K$, $\|x_i - \hat{x}\| \leq \hat{\rho}$, and hence for all $i \geq i_o$, $i \in K$,

$$f(x_{i+1}) - f(x_i) \leq -\hat{\delta}. \tag{14}$$

Now, because $f(\cdot)$ is continuous, $f(x_i) \to^K f(\hat{x})$, as $i \to \infty$, and, by (13a), $f(x_{i+1}) < f(x_i)$ for all $i \in \mathbb{N}$. Hence, it follows from the Monotone Sequences Proposition 5.1.16 that $f(x_i) \to^K f(\hat{x})$, as $i \to \infty$. Since this is contradicted by (14), our proof is complete. □

In the case of a multi-point algorithm, the cost sequence that it constructs need not be monotone decreasing for the algorithm to be convergent, as the following theorem shows.

Theorem 1.2.19. *Consider Algorithm Model 1.2.17 together with Model Problem 1.2.6. Let $q \geq 0$ be a given integer, for any $i \in \mathbb{N}$, let $q(i) \triangleq \min\{i, q\}$, and, for any $\underline{x}_i = \{x_j\}_{j=0}^{i}$, let*

$$\psi_q(\underline{x}_i) \triangleq \max_{0 \leq j \leq q(i)} f(x_{i-j}). \tag{15a}$$

Suppose that, for every bounded subset $S \subset X$, there exist an $\alpha_S \in (0, \infty)$ such that, if $\underline{x}_i = \{x_j\}_{j=0}^{i}$ is constructed by Algorithm Model 1.2.17 and $x_i \in S$, then

$$f(x') - \psi_q(\underline{x}_i) \leq \alpha_S \theta(x_i), \quad \forall x' \in \underline{A}(\underline{x}_i). \tag{15b}$$

If Algorithm Model 1.2.17 constructs a bounded infinite sequence $\{x_i\}_{i=0}^{\infty}$, then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ is stationary (i.e., $\hat{x} \in X$, and $\theta(\hat{x}) = 0$). Furthermore, if there exist $\beta, \gamma \in (0, \infty)$ such that, for all $i \in \mathbb{N}$,

$$\|x_{i+1} - x_i\| \leq \beta \theta(x_i)^{\gamma}, \tag{15c}$$

then the cost sequence $\{f(x_i)\}_{i=0}^{\infty}$ converges, and no accumulation point of $\{x_i\}_{i=0}^{\infty}$ is a local maximizer of $f(\cdot)$.

Proof. Suppose that Algorithm Model 1.2.17 has constructed a bounded sequence $\{x_i\}_{i=0}^{\infty}$. Then, by assumption, there exists an $\alpha \in (0, \infty)$ such that,

for all i, by (15b),

$$f(x_{i+1}) \leq \psi_q(\underline{x}_i) + \alpha\theta(x_i). \tag{16a}$$

Hence, for all $i \geq q$,

$$\psi_q(\underline{x}_{i+1}) = \max\{f(x_{i+1}), f(x_i), \ldots, f(x_{i+1-q})\}$$

$$\leq \max\{f(x_{i+1}), \psi_q(\underline{x}_i)\} \leq \psi_q(\underline{x}_i). \tag{16b}$$

Hence, since $\{f(x_i)\}_{i=0}^{\infty}$ is bounded and, consequently, the sequence $\{\psi_q(\underline{x}_i)\}_{i=0}^{\infty}$ is bounded, because $\{x_i\}_{i=0}^{\infty}$ is bounded, by assumption, and $f(\cdot)$ is continuous, it follows from the Monotone Sequences Proposition 5.1.16 that there exists a $\psi^* > -\infty$ such that $\psi_q(\underline{x}_i) \to \psi^*$, as $i \to \infty$.

Next, it follows from (16a) that, for $i \geq q$,

$$\psi_q(\underline{x}_{i+1+q}) \leq \psi_q(\underline{x}_i) + \alpha \min_{0 \leq j \leq q} \theta(x_{i+q-j}). \tag{16c}$$

Since $\psi_q(\underline{x}_{i+1+q}) - \psi_q(\underline{x}_i) \to 0$, as $i \to \infty$, and since $\theta(x_i) < 0$ for all i, it follows from (16c) that

$$\lim_{i \to \infty} \min_{0 \leq j \leq q} \theta(x_{i+q-j}) = \lim_{i \to \infty} \theta(x_i) = 0. \tag{16d}$$

Since $\overline{\lim_{i \to \infty}} \theta_i) \leq 0$, it follows that $\theta(x_i) \to 0$, as $i \to \infty$. Therefore, since $\theta(\cdot)$ is upper semicontinuous by assumption, if $\hat{x}$ is an accumulation point of $\{x_i\}_{i=0}^{\infty}$, then $\theta(\hat{x}) = 0$, which shows that $\hat{x}$ is *quasi–stationary*. Since, by assumption on $\underline{A}(\cdot)$, the sequence $\{x_i\}_{i=0}^{\infty}$ is contained in X, it follows that $\hat{x} \in X$ and hence it is stationary.

Now suppose that (15c) holds. Then, for all $i \geq q$, let $p(i) \in \{i, i-1, \ldots, i-q\}$ be such that

$$\psi_q(\underline{x}_i) = f(x_{p(i)}). \tag{16e}$$

Then $f(x_{p(i)}) \to \psi^*$, as $i \to \infty$. Now, for any $i \geq q$,

$$\|x_{p(i+2+q)} - x_i\| \leq \sum_{j=0}^{p(i+2+q)-(i+1)} \|x_{p(i+2+q)-j} - x_{p(i+2+q)-(j+1)}\|, \tag{16f}$$

and $1 \leq p(i+2+q) - (i+1) \leq q+1$. Therefore it follows from (15c), (16d), and (16f) that $\|x_{p(i+2+q)} - x_i\| \to 0$, as $i \to \infty$. Now, for all $i \geq q$,

$$|f(x_i) - \psi^*| \leq |f(x_i) - f(x_{p(i+2+q)})| + |f(x_{p(i+2+q)}) - \psi^*|. \tag{16g}$$

Now, $|f(x_{p(i+2+q)}) - \psi^*| \to 0$, as $i \to \infty$, because $f(x_{p(i)}) \to \psi^*$, as $i \to \infty$, and $|f(x_i) - f(x_{p(i+2+q)})| \to 0$, as $i \to \infty$ because of (16f) and because $f(\cdot)$ is uniformly continuous on bounded sets. Hence, $f(x_i) \to \psi^*$, as $i \to \infty$. Clearly, $f(\hat{x}) = \psi^*$ must hold.

Now, suppose that $\hat{x}$ is an accumulation point of $\{x_i\}_{i=0}^{\infty}$, i.e., that for some infinite subset $K \subset \mathbb{N}$, $x_i \to^K \hat{x}$, as $i \to \infty$. It follows from (16f) that $x_{p(i+2+q)} \to^K \hat{x}$, as $i \to \infty$, and (16c) implies that $f(x_{p(i+2+q)}) > f(\hat{x})$ for all i. Hence, $\hat{x}$ cannot be a local maximizer of $f(\cdot)$. □

1.2.3 The Wolfe and Polak-Sargent-Sebastian Theorems

We are now ready to return to problem (1a), under the additional assumption that the function $f(\cdot)$ is continuously differentiable. The use of Algorithm Model 1.2.17 and Theorem 1.2.18, in analyzing specific algorithms that solve problem (1a), is made easier when the algorithm function $\underline{A}(\cdot)$ is endowed with more structure than is found in Algorithm Model 1.2.17. In particular, it is useful to decompose the evaluation of the algorithm function $\underline{A}(\cdot)$ into two operations: in the first, a search direction is computed, in the second the step-size is computed, corresponding to the assumption that $\underline{A}(x_i) = x_i + \lambda(x_i)h(x_i)$, with $h(x_i)$ the search direction and $\lambda(x_i)$ the step-size. We begin with an algorithm model that uses a step-size determined by an exact line search.

Algorithm Model with Exact Line Search 1.2.20.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. If $\nabla f(x_i) = 0$, stop. Else, compute a *search direction* $h_i \in \mathbb{R}^n$.

Step 2. Compute a *step-size*

$$\lambda_i \in \arg\min_{\lambda \geq 0} f(x_i + \lambda h_i). \tag{17}$$

Step 3. Set $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i+1$, and go to step 1.

The following theorem assumes that the angle between the search direction h_i and the cost gradient $\nabla f(x_i)$ is obtuse and bounded away from 90°. In the form stated, it is a special case of a more general result established in [Wol.69, Wol.71].

Wolfe Theorem 1.2.21. *Suppose that $f: \mathbb{R}^n \to \mathbb{R}$ in (1a) is Lipschitz continuously differentiable on bounded sets and that there exists a continuous function $\kappa: \mathbb{R}^n \to [0,1]$ such that $\kappa(x) > 0$ for all $x \in \mathbb{R}^n$ satisfying $\nabla f(x) \neq 0$. Furthermore, suppose that, for all $i \in \mathbb{N}$, the search direction vectors h_i constructed by Algorithm Model 1.2.20 satisfy $h_i \neq 0$ if $\nabla f(x_i) \neq 0$ and*

$$\langle \nabla f(x_i), h_i \rangle \leq -\kappa(x_i)\|\nabla f(x_i)\|\|h_i\|. \tag{18}$$

If $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm Model 1.2.20, then any accumulation point $\hat{x}$ of this sequence satisfies $\nabla f(\hat{x}) = 0$.

Proof. Suppose that $\hat{x} \in \mathbb{R}^n$ is an accumulation point of a sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm Model 1.2.20 and that $\nabla f(\hat{x}) \neq 0$. Then $\kappa(\hat{x}) > 0$, and hence there exists a $\rho_0 > 0$ such that $\kappa(x) \geq \kappa(\hat{x})/2$ for all $x \in B(\hat{x}, \rho_0)$. Let $L < \infty$ be a Lipschitz constant for $\nabla f(\cdot)$ that is valid on the ball $B(\hat{x}, 2\rho_0)$, and let $\rho_1 \in (0, \rho_0]$ be such that $\|\nabla f(x)\| \geq \frac{1}{2}\|\nabla f(\hat{x})\|$ for all $x \in B(\hat{x}, \rho_1)$. Then, for any $x_i \in B(\hat{x}, \rho_1)$ and $\lambda > 0$ such that $\lambda\|h_i\| \leq \rho_0$, $x_i + \lambda h_i \in B(\hat{x}, 2\rho_0)$ and hence, using (18) and the Mean-Value Theorem 5.1.28(*b*), we find that, for some $s \in [0,1]$,

$$f(x_i + \lambda h_i) - f(x_i) = \lambda\langle \nabla f(x_i), h_i \rangle + \lambda[\langle \nabla f(x_i + s\lambda h_i) - \nabla f(x_i), h_i \rangle]$$

$$\leq -\tfrac{1}{4}\kappa(\hat{x})(\lambda\|h_i\|)\|\nabla f(\hat{x})\| + L(\lambda\|h_i\|)^2. \tag{19a}$$

Now, the minimum value of the right-hand side of (19a) is attained at

$$\lambda\|h_i\| = s^* \triangleq \kappa(\hat{x})\|\nabla f(\hat{x})\|/8L. \tag{19b}$$

Let $\hat{s} = \min\{s^*, \rho_0\}$, and let $-\hat{\delta} \triangleq -\kappa(\hat{x})\hat{s}\|\nabla f(\hat{x})\| + L\hat{s}^2$. Then $-\hat{\delta} < 0$, and hence, since the value of the right-hand side in (19a) for $\lambda\|h_i\| = \hat{s}$ must be equal to or larger than the minimum value of the left-hand side of (19a), with respect to λ, we find that, if $x_i \in B(\hat{x}, \hat{\rho})$, then

$$f(x_{i+1}) - f(x_i) \leq -\hat{\delta}. \tag{19c}$$

Since $\{f(x_i)\}_{i=0}^{\infty}$ is monotone decreasing by construction and since, by continuity of $f(\cdot)$, it has an accumulation point $f(\hat{x})$, if follows from the Monotone Sequences Proposition 5.1.16 that $f(x_i) \to f(\hat{x})$, as $i \to \infty$. Since (19c) implies that $f(x_i) \to -\infty$, as $i \to \infty$, we have a contradiction. □

Exercise 1.2.22. Prove Theorem 1.2.21 under the weaker assumption that $f(\cdot)$ is only continuously differentiable. □

The step-size rule shown in (17) is problematic because the set $\arg\min_{\lambda \geq 0} f(x_i + \lambda h_i)$ may be empty or impossible to compute exactly, and hence many algorithms use a more efficient step-size rule, first proposed by Armijo [Arm.66]. This rule uses two parameters α, $\beta \in (0,1)$. The geometry of this rule is illustrated in Fig. 1.2.3. It can be implemented in two different versions which we select by assigning the values of 0 or 1 to the step-size rule selector s in the algorithm model below, where $\mathbb{Z}$ denotes the set of all integers,

both positive and negative. Referring to Fig. 1.2.3, when $s = 0$, the selected step-size will be λ_i, while when $s = 1$, the selected step-size will be either λ_i or λ'_i. The advantage of using the parameter $s = 1$ is that it results in a very inexpensive step-size calculation, but, as shown in Fig. 1.2.3, it can result in a considerably smaller cost decrease than the use of the parameter $s = 0$. The only disadvantage to the use of the parameter $s = 0$ is that it can lead to expensive step-size calculations.

Algorithm Model with Armijo Line Search 1.2.23.

Parameters. $\alpha, \beta \in (0,1), k^* \in \mathbb{Z}, s \in \{0,1\}$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. If $\nabla f(x_i) = 0$, stop. Else, compute a *descent direction* h_i.

Step 2. If $s = 0$, set $K^* \triangleq \{k \in \mathbb{Z} \mid k \geq k^*\}$, and compute the *step-size*

$$\lambda_i = \beta^{k_i} \triangleq \arg\max_{k \in K^*} \{\beta^k \mid f(x_i + \beta^k h_i) - f(x_i) \leq \beta^k \alpha \langle \nabla f(x_i), h_i \rangle\}. \quad (20a)$$

If $s = 1$, compute the *step-size* $\lambda_i = \beta^{k_i}$, where $k_i \in \mathbb{Z}$ is any integer such that

$$f(x_i + \beta^{k_i} h_i) - f(x_i) \leq \beta^{k_i} \alpha \langle \nabla f(x_i), h_i \rangle \quad (20b)$$

and

$$f(x_i + \beta^{k_i - 1} h_i) - f(x_i) > \beta^{k_i - 1} \alpha \langle \nabla f(x_i), h_i \rangle. \quad (20c)$$

Step 3. Set $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i + 1$, and go to step 1.

The selection $s = 0$ is normally used with Newton-like algorithms, with $k^* = 0$ to ensure superlinear convergence. In such algorithms, after a finite number of iterations, the choice $k_i = 0$ is usually accepted for all subsequent

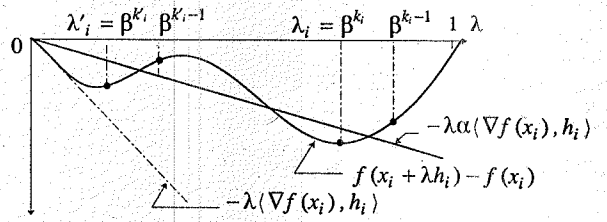


Fig. 1.2.3. The Armijo step-size rule.

iterations. The selection $s = 0$ is not very good for first-order algorithms because, on the average, it requires considerably more function evaluations then the selection $s = 1$. Furthermore, as can be deduced from the rate of convergence analysis of the Armijo Gradient Algorithm, in the next section, a selection of the parameter k^* that limits the step-size too severely, relative to some hard to estimate problem parameters, may increase the rate of convergence constant, i.e., it can cause slower convergence. On the other hand, decreasing k^* makes the search for the step-size more costly. In view of these considerations, it appears that $s = 1$ should be used in first-order algorithms.

Note that, provided that the function $f(\cdot)$ is bounded from below, it is very easy to find a k_i satisfying (20b) and (20c), using the following subprocedure, which uses the last used step length $\lambda_{i-1} = \beta^{k_{i-1}}$, as the starting point for the computation of the next one.

Step-Size Subprocedure 1.2.23a.

Step 1. If $i = 0$, set $k' = k^*$. Else set $k' = k_{i-1}$.

Step 2. If $k_i = k'$ satisfies (20b) and (20c), stop.

Step 3. If $k_i = k'$ satisfies (20b) but not (20c), replace k' by $k' - 1$, and go to Step 2.
If $k_i = k'$ satisfies (20c) but not (20b), replace k' by $k' + 1$, and go to Step 2.

From Theorems 1.2.24a and 1.2.24b, we will see that Algorithm 1.2.23 is well defined, i.e., whenever $\nabla f(x_i) \neq 0$, the search for a step-size λ_i is a finite process, whether $s = 0$ or $s = 1$. In practice, only a very small number of iterations of Subprocedure 1.2.23a are required to compute the Armijo step-size. However, occasionally, once a very small step-size has occurred, the use of Subprocedure 1.2.23a, can trap an algorithm into using a very small step-size for all subsequent iterations, a rather undersirable effect. Hence, when a very small step-size occurs for several iterations, it may be wise to revert to setting $s = 0$ for one or two iterations.

Although it is possible to deal with all the cases of interest of Algorithm 1.2.23 within a single framework, it may be instructive to consider a simple case first.

Theorem 1.2.24a. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ in (1a) is continuously differentiable and bounded from below and that Algorithm Model 1.2.23 sets $h_i = h(x_i)$, where $h : \mathbb{R}^n \to \mathbb{R}^n$ is a continuous function such that, for all $x \in \mathbb{R}^n$ satisfying $\nabla f(x) \neq 0$,*

$$\langle \nabla f(x), h(x) \rangle < 0 . \tag{21a}$$

(a) If x_i is such that $\nabla f(x_i) \neq 0$, then λ_i is computed by Algorithm Model 1.2.23 using a finite number of function evaluations.

(b) If $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm Model 1.2.23, then any accumulation point $\hat{x}$ of this sequence satisfies $\nabla f(\hat{x}) = 0$.

Proof. Under the assumptions stated, Algorithm Model 1.2.23 is a special case of Algorithm Model 1.2.7. We will show that the assumptions of Corollary 1.2.9 (and hence also of Theorem 1.2.8) are satisfied. Thus, let $\theta(x) \triangleq \langle \nabla f(x), h(x) \rangle$. Then we see that $\theta(\cdot)$ is continuous, that $\theta(x) \leq 0$ for all $x \in \mathbb{R}^n$, and that $\theta(x) = 0$ if and only if $\nabla f(x) = 0$, i.e., that $\theta(\cdot)$ is an optimality function for the problem (1a).

Next, suppose that $\hat{x}$ is such that $\nabla f(\hat{x}) \neq 0$. Then it follows from the definition of the directional derivative $df(\hat{x}; h(\hat{x})) = \langle \nabla f(\hat{x}), h(\hat{x}) \rangle$ that, for any $\varepsilon > 0$ such that $\alpha + \varepsilon < 1$, there exists a $k_\varepsilon \in \mathbb{N}$ such that, for all $\lambda \in (0, \beta^{k_\varepsilon}]$,

$$f(\hat{x} + \lambda h(\hat{x})) - f(\hat{x}) \leq \lambda(\alpha + \varepsilon) df(\hat{x}; h(\hat{x})) < 0 . \tag{21b}$$

Rearranging terms in (21b) and substituting for $df(\hat{x}; h(\hat{x}))$, we find that, for all $\lambda \in (0, \beta^{k_\varepsilon}]$,

$$f(\hat{x} + \lambda h(\hat{x})) - f(\hat{x}) - \lambda\alpha \langle \nabla f(\hat{x}), h(\hat{x}) \rangle \leq \lambda\varepsilon \langle \nabla f(\hat{x}), h(\hat{x}) \rangle < 0 . \tag{21c}$$

First, consider the case when $s = 0$. In this case, it follows from (21c) that, at $\hat{x}$, Algorithm Model 1.2.23 would compute a step-size $\hat{\lambda} = \beta^{\hat{k}} \geq \beta^{k_\varepsilon}$ in a finite number of operations. Next consider the case when $s = 1$. In this case, because $f(\cdot)$ is bounded from below, it follows from (21c) that there exists a finite value of k_i such that (20b) is satisfied. Because $f(\cdot)$ is bounded from below, it is clear that Step-Size Subprocedure 1.2.23a cannot decrease the value of k' an infinite number of times without satisfying both (20b) and (20c). Hence, it follows, also, in the case of $s = 1$, that at $\hat{x}$ Algorithm Model 1.2.23 would compute a step size $\hat{\lambda} = \beta^{\hat{k}} \geq \beta^{k_\varepsilon}$ in a finite number of operations. Next consider the case when $s = 1$. Hence *(a)* is proved.

Next, from the continuity of $f(\cdot)$, $\nabla f(\cdot)$, and $h(\cdot)$, it follows that there exists a $\hat{\rho} > 0$ such that, for all $x_i \in B(\hat{x}, \hat{\rho})$,

$$\langle \nabla f(x_i), h(x_i) \rangle \leq \tfrac{1}{2} \langle \nabla f(\hat{x}), h(\hat{x}) \rangle , \tag{21d}$$

and

$$f(x_i + \beta^{k_\varepsilon} h(x_i)) - f(x_i) - \beta^{k_\varepsilon} \alpha \langle \nabla f(x_i), h(x_i) \rangle < 0 , \tag{21e}$$

which shows that, for all such x_i, $\lambda_i \geq \beta^{k_\varepsilon}$ and hence that

$$f(x_{i+1}) - f(x_i) \leq \lambda_i \alpha \langle \nabla f(x_i), h(x_i) \rangle \leq \tfrac{1}{2} \beta^{k_\varepsilon} \alpha \langle \nabla f(\hat{x}), h(\hat{x}) \rangle < 0 . \tag{21f}$$

Since (21f) shows that the assumptions of Corollary 1.2.9 (as well as of Theorem 1.2.8) are satisfied, *(b)* follows immediately. □

The following much more general theorem subsumes Theorem 1.2.24a, because it does not require that the search directions h_i be generated using a continuous function $h(\cdot)$. Instead, it requires only that the search direction h_i be bounded from above, and it imposes a restriction on the angle between h_i and $\nabla f(x_i)$. It is a slight generalization of the result established in [PSS.74].

Polak-Sargent-Sebastian Theorem 1.2.24b. *Suppose that*

(i) the function $f : \mathbb{R}^n \to \mathbb{R}$ in (1a) is Lipschitz continuously differentiable on bounded sets;

(ii) the sequences $\{x_i\}_{i=0}^{\infty}$ and $\{h_i\}_{i=0}^{\infty}$ were constructed by Algorithm Model 1.2.23 in solving the problem (1a);

(iii) there exist two continuous functions $N_1 : \mathbb{R}^n \to \mathbb{R}$ and $N_2 : \mathbb{R}^n \to \mathbb{R}$ such that

(a) for all x satisfying $\nabla f(x) \neq 0$, $N_1(x) > 0$, $N_2(x) > 0$, and $N_1(x) = 0$ if and only if $\nabla f(x) = 0$, and,

(b) for all $i \in \mathbb{N}$, the x_i and h_i satisfy the inequalities

$$\langle \nabla f(x_i), h_i \rangle \leq -N_1(x_i) \tag{22a}$$

and

$$\|h_i\| \leq N_2(x_i) . \tag{22b}$$

Under these assumptions,

(a) if x_i is such that $\nabla f(x_i) \neq 0$, then λ_i is computed by Algorithm Model 1.2.23 using a finite number of function evaluations, and

(b) any accumulation point $\hat{x}$ of the sequence $\{x_i\}_{i=0}^{\infty}$ satisfies $\nabla f(\hat{x}) = 0$.

Proof. First, Algorithm Model 1.2.23 can be viewed as an expanded version of Algorithm Model 1.2.17. Hence, to prove our theorem, we need to show only that the assumptions of Theorem 1.2.18 are satisfied. To this end, we define $\theta : \mathbb{R}^n \to \mathbb{R}$ by $\theta(x) \triangleq -N_1(x)$, so that $\theta(\cdot)$ is continuous, $\theta(x) \leq 0$ for all $x \in \mathbb{R}^n$, and $\theta(x) = 0$ if and only if $\nabla f(x) = 0$. Clearly, $\theta(\cdot)$ is an optimality function for the problem (1a).

(a) The proof of this part is identical to that of part *(a)* in Theorem 1.2.24a and hence is omitted.

(b) Suppose that $\hat{x} \in \mathbb{R}^n$ is such that $\theta(\hat{x}) < 0$. Let $\rho \in (0,1]$ be arbitrary. Then, because $N_i(\cdot)$ and $N_2(\cdot)$ are continuous and $f(\cdot)$ is bounded from below, there exists a $k' \in \mathbb{N}$ such that for all $x_i \in B(\hat{x}, \rho)$ and $k \leq k'$,

$$f(x_i + \beta^k h_i) - f(x_i) \geq -\beta^k \alpha N_1(x_i) \geq \beta^k \alpha \langle \nabla f(x_i), h_i \rangle, \tag{22c}$$

which shows that $k_i \geq k'$ for all $i \in \mathbb{N}$ such that $x_i \in B(\hat{x}, \rho)$, regardless whether $s = 0$ or $s = 1$. Let L be a local Lipschitz constant for $\nabla f(\cdot)$, which is valid on the ball $B(\hat{x}, \beta^{k'} 3N_2(\hat{x}))$, and let $\rho_0 \in (0, \beta^{k'} N_2(\hat{x})]$ be such that, for all $x \in B(\hat{x}, \rho_0)$, $N_1(x) \geq \frac{1}{2} N_1(\hat{x})$ and $N_2(x) \leq \frac{3}{2} N_2(\hat{x})$. Let $\hat{\rho} = \min\{\rho, \rho_0\}$. If $x_i \in B(\hat{x}, \hat{\rho})$, then for any $\lambda \in (0, \beta^{k'}]$, $x_i + \lambda h_i \in B(\hat{x}, \beta^{k'} 3N_2(\hat{x}))$ must hold, and hence (examine the calculation in (20a)), for any $\lambda \in (0, \beta^{k'}]$,

$$\begin{aligned} & f(x_i + \lambda h_i) - f(x_i) - \lambda\alpha \langle \nabla f(x_i), h_i \rangle \\ & = \lambda(1-\alpha) \langle \nabla f(x_i), h_i \rangle + \lambda \int_0^1 \langle \nabla f(x_i + s\lambda h_i) - \nabla f(x_i), h_i \rangle \, ds \\ & \leq -\lambda(1-\alpha) N_1(x_i) + \tfrac{1}{2}\lambda^2 L N_2(x_i)^2 \\ & \leq -\tfrac{1}{2}\lambda(1-\alpha) N_1(\hat{x}) + \tfrac{9}{8}\lambda^2 L N_2(\hat{x})^2 . \end{aligned} \tag{22d}$$

Let

$$\gamma \triangleq 4(1-\alpha) N_1(\hat{x}) / 9 L N_2(\hat{x})^2 . \tag{22e}$$

Clearly, the last expression in (22d) is negative for all $\lambda \in (0, \gamma]$. Hence for any k such that $\beta^k \leq \omega \triangleq \min\{\beta^{\hat{k}}, \gamma\}$, β^k satisfies the inequality (20b). Now let $\hat{k} \in \mathbb{N}$ be such that $\beta^{\hat{k}} \leq \omega < \beta^{\hat{k}-1}$, then $\beta^{\hat{k}} > \beta\omega$. Hence it follows that, for all $x_i \in B(\hat{x}, \hat{\rho})$, the step-size λ_i must satisfy $\lambda_i \geq \beta^{\hat{k}} \geq \beta\omega$. In view of this, it follows from (20a,b) that

$$f(x_{i+1}) - f(x_i) \leq \lambda_i \alpha \langle \nabla f(x_i), h(x_i) \rangle \leq -\beta\omega\alpha N_1(x_i), \tag{22f}$$

which is of the form of (14b). Hence, part *(b)* now follows from Theorem 1.2.18. □

Exercise 1.2.25. *(a)* Use Corollary 5.1.23 to show that the Polak-Sargent-Sebastian Theorem 1.2.24b remains valid under the weaker assumption that $f(\cdot)$ is only continuously differentiable.

(b) Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ in (1a) is Lipschitz continuously differentiable on bounded sets, that the sequences $\{x_i\}_{i=0}^{in}$ and $\{h_i\}_{i=0}^{\infty}$ were constructed by Algorithm Model 1.2.23 in solving problem (1a), and that there exist three continuous functions $\kappa : \mathbb{R}^n \to [0,1]$, $N_1 : \mathbb{R}^n \to \mathbb{R}$, $N_2 : \mathbb{R}^n \to \mathbb{R}$ such that, for all x satisfying $\nabla f(x) \neq 0$, $\kappa(x) > 0$, $N_1(x) > 0$, $N_2(x) > 0$, and, for all $i \in \mathbb{N}$,

$$\langle \nabla f(x_i), h_i \rangle \leq -\kappa(x_i) \|\nabla f(x_i)\| \|h_i\| ,$$

$$N_1(x_i) \leq \|h_i\| \leq N_2(x_i) .$$

Show that

(i) if x_i is such that $\nabla f(x_i) \neq 0$, then λ_i is computed by Algorithm Model 1.2.23 using a finite number of function evaluations, and

(ii) if $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm Model 1.2.23, then any accumulation point $\hat{x}$ of this sequence satisfies $\nabla f(\hat{x}) = 0$. □

1.2.4 A Trust Region Model

There are a number of trust region algorithm models in the literature that combine the computation of the search direction with the step-size calculation into a single operation; see, e.g. [Mor.83, Pow.75, Pow.84, Sor.82, BSS.87] for a small sample. These algorithm models have been developed to assist in constructing globally converging versions of Newton's method, as well as various approximations to Newton's method, and, in particular, quasi-Newton methods. One of the earliest applications was in the stabilization of the Levenberg-Marquardt algorithm (see [Lev.44, Mar.63]), which replaces a near singular or non-positive-definite Hessian $f_{xx}(x)$ by a positive-definite matrix of the form $f_{xx}(x) + \lambda I$, with I an identity matrix. The algorithm models in the literature are generally multi-point, i.e., they have memory. In keeping with the approach taken in this book, we will extract, from the existing results, a one-step, i.e. memoryless, algorithm model.

Thus, consider problem (1a) under the assumption that $f : \mathbb{R}^n \to \mathbb{R}$ is twice Lipschitz continuously differentiable on bounded sets. The trust region algorithm models in the literature have this much in common: Given a point $x_i \in \mathbb{R}^n$, they construct a symmetric, $n \times n$ matrix H_i, which is an approximation to the Hessian, $f_{xx}(x_i)$, and a quadratic approximation, or local model, $\Delta_i(h)$, to $f(x_i + h) - f(x_i)$, given by

$$\Delta_i(h) \triangleq \langle g_i, h \rangle + \tfrac{1}{2}\langle h, H_i h \rangle , \tag{23a}$$

where

$$g_i \triangleq \nabla f(x_i) . \tag{23b}$$

To compute the next iterate, they assume that the local model is "trustworthy" for all h in the "trust region"

$$T_i \triangleq \{h \in \mathbb{R}^n \mid \|D_i h\| \leq \sigma_i\} , \tag{23c}$$

where $\sigma_i > 0$ and D_i is a symmetric, positive-definite scaling matrix, and compute an approximate solution h_i to the problem

$$\min \{ \Delta_i(h) \mid h \in T_i \}. \tag{23d}$$

If $f(x_i + h_i) - f(x_i) \leq \alpha \Delta_i(h_i)$, with $\alpha \in (0,1)$, they either set $x_{i+1} = x_i + h_i$ or try to make a larger step by increasing σ_i. If $f(x_i + h_i) - f(x_i) > \alpha \Delta_i(h_i)$, they decrease σ_i and try again.

The algorithm model below also uses these elements.

Algorithm Model 1.2.26.

Parameters. $\alpha, \beta \in (0,1)$, $\sigma_{\max} > 0$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute an $n \times n$ symmetric matrix H_i, and construct the function $\Delta_i(\cdot)$, defined in (23a), (23b).

Step 2. Set $\sigma_i = \sigma_{\max}$.

Step 3. Compute a nonsingular, $n \times n$, scaling matrix D_i and define $\tilde{\Delta}_i : \mathbb{R}^n \to \mathbb{R}^n$ by

$$\tilde{\Delta}_i(s) \triangleq \Delta_i(D_i^{-1}s) = \langle \tilde{g}_i, s \rangle + \tfrac{1}{2}\langle s, \tilde{H}_i s \rangle, \tag{24a}$$

where

$$\tilde{g}_i \triangleq (D_i^{-1})^T g_i, \quad \tilde{H}_i \triangleq (D_i^{-1})^T H_i D_i^{-1}. \tag{24b}$$

Step 4. Compute an $h_i \in \mathbb{R}^n$ such that $\|D_i h_i\| \leq \sigma_i$ and

$$\Delta_i(h_i) \leq \min_{\tau} \{ \tilde{\Delta}_i(-\tau \tilde{g}_i) \mid \tau \|\tilde{g}_i\| \leq \sigma_i \}. \tag{24c}$$

Step 5. If

$$f(x_i + h_i) - f(x_i) \leq \alpha \Delta_i(h_i), \tag{24d}$$

set

$$x_{i+1} = x_i + h_i, \tag{24d}$$

replace i by $i+1$, and go to Step 1.

Else, replace σ_i by $\beta\sigma_i$, and go to Step 4.

Note that Algorithm Model 1.2.26 specifies that an approximate solution h_i to problem (23d) is good enough if $\Delta_i(h_i)$ is smaller or equal to the constrained minimum value of $\tilde{\Delta}_i(\cdot)$ along its negative gradient $-\tilde{g}_i$.

Assumption 1.2.27 *We will assume that*

(i) the function $f : \mathbb{R}^n \to \mathbb{R}$ is twice Lipschitz continuously differentiable on bounded sets, and

(ii) there exist constants $c_k \in (0, \infty)$, $k = 1, 2$, such that for all $i \in \mathbb{N}$, and x_i, H_i, and D_i constructed by Algorithm Model 1.2.26,

$$\|H_i\| \leq c_1, \quad \max \{ \|D_i\|, \|D_i^{-1}\| \} \leq c_2. \tag{25}$$

□

The following result can be found in [Mor.83].

Lemma 1.2.28 *Suppose that $g_i \neq 0$. If $h_i \in \mathbb{R}^n$ satisfies (24c), then*

$$\Delta_i(h_i) \leq -\tfrac{1}{2}\|\tilde{g}_i\| \min \{ \sigma_i, \|\tilde{g}_i\|/\|\tilde{H}_i\| \}. \tag{26}$$

Proof. Let $\psi : \mathbb{R} \to \mathbb{R}$ be defined by $\psi(\tau) \triangleq \Delta_i(-\tau \tilde{g}_i)$, so that

$$\psi(\tau) \triangleq -\tau \|\tilde{g}_i\|^2 + \tfrac{1}{2}\tau^2 \langle \tilde{g}_i, \tilde{H}_i \tilde{g}_i \rangle, \tag{27a}$$

and let τ_i be the minimizer of $\psi(\cdot)$ over $[0, \sigma_i/\|\tilde{g}_i\|]$.

First suppose that $\langle \tilde{g}_i, \tilde{H}_i \tilde{g}_i \rangle > 0$, and let $\hat{\tau} = \|\tilde{g}_i\|^2/\langle \tilde{g}_i, \tilde{H}_i \tilde{g}_i \rangle$ be the unconstrained minimizer of $\psi(\tau)$. If $\hat{\tau} < \sigma_i/\|\tilde{g}_i\|$, then $\tau_i = \hat{\tau}$ and hence

$$\psi(\tau_i) = -\tfrac{1}{2}\|\tilde{g}_i\|^4/\langle \tilde{g}_i, \tilde{H}_i \tilde{g}_i \rangle \leq -\tfrac{1}{2}\|\tilde{g}_i\|^2/\|\tilde{H}_i\|. \tag{27b}$$

If $\hat{\tau} \geq \sigma_i/\|\tilde{g}_i\|$, then $\tau_i = \sigma_i/\|\tilde{g}_i\|$. In this case, we must have that

$$\hat{\tau} = \|\tilde{g}_i\|^2/\langle \tilde{g}_i, \tilde{H}_i \tilde{g}_i \rangle \geq \sigma_i/\|\tilde{g}_i\|. \tag{27c}$$

and hence that

$$\sigma_i \langle \tilde{g}_i, \tilde{H}_i \tilde{g}_i \rangle/\|\tilde{g}_i\|^2 \leq \|g_i\|. \tag{27d}$$

Therefore, in this case,

$$\psi(\tau_i) = \psi(\sigma_i/\|\tilde{g}_i\|) \leq -\tfrac{1}{2}\sigma_i \|\tilde{g}_i\|. \tag{27e}$$

Finally, suppose that $\langle \tilde{g}_i, \tilde{H}_i \tilde{g}_i \rangle \leq 0$. In this case, $\tau_i = \sigma_i/\|\tilde{g}_i\|$ must hold, and hence (27d) follows, by inspection, from (27a). □

Theorem 1.2.29 *Suppose that Assumption 1.2.27 is satisfied and that Algorithm Model 1.2.26 has constructed a sequence $\{x_i\}_{i=0}^{\infty}$, together with the matrices $\{D_i\}_{i=0}^{\infty}$ and $\{H_i\}_{i=0}^{\infty}$ and vectors $\{g_i\}_{i=0}^{\infty}$. If $\hat{x}$ is an accumulation point of $\{x_i\}_{i=0}^{\infty}$, then $\nabla f(\hat{x}) = 0$.*

Proof. Suppose that $\hat{x}$ is an accumulation point of $\{x_i\}_{i=0}^{\infty}$, i.e., there exists an infinite subset $K \subset \mathbb{N}$ such that $x_i \to^K \hat{x}$, as $i \to \infty$. By construction, the sequence $\{f(x_i)\}_{i=0}^{\infty}$ is monotone decreasing. Since $f(\cdot)$ is continuous, $f(x_i) \to^K f(\hat{x})$, as $i \to \infty$. Therefore, it follows from Proposition 5.1.16 that

$f(x_i) \to f(\hat{x})$, as $i \to \infty$. Now, for the sake of contradiction, suppose that $\nabla f(\hat{x}) \neq 0$. Then there exists a $\rho > 0$ such that, for all $x \in B(\hat{x}, \rho)$, $\|\nabla f(x)\| \geq \frac{1}{2}\|\nabla f(\hat{x})\| \triangleq \omega_0 > 0$. Hence, for all $x_i \in B(\hat{x}, \rho)$,

$$\|\tilde{g}_i\| \geq \|g_i\| / \|D_i^T\| \geq \|g_i\| / c_2 \geq \omega_0 / c_2 . \tag{28a}$$

Let $\omega_1 \triangleq \omega_0 / c_2$. It now follows from (26) that for all $x_i \in B(\hat{x}, \rho)$,

$$\Delta_i(h_i) \leq -\tfrac{1}{2}\omega_1 \min\{\sigma_i, \omega_1 / c_1\}. \tag{28b}$$

Now, in view of (25), for any $h \in T_i$, $\|h\| = \|D_i^{-1} D_i h\| \leq c_2 \|D_i h\| \leq c_2 \sigma_i \leq c_2 \sigma_{\max}$. Hence there exists a $K < \infty$ such that for all $x_i \in B(\hat{x}, \rho)$ and $h \in T_i$,

$$\begin{aligned} f(x_i + h) - f(x_i) &= \langle g_i, h \rangle + \tfrac{1}{2}\langle h, H_i h \rangle \\ &\quad + \int_0^1 (1 - s)\langle h, [f_{xx}(x_i + sh) - H_i]h \rangle \, ds \\ &\leq \Delta_i(h) + K\|h\|^2 . \end{aligned} \tag{28c}$$

Hence, in view of (28b) and the fact that $\|h_i\| \leq c_2 \sigma_i$, we conclude that

$$\begin{aligned} f(x_i + h_i) - f(x_i) - \alpha\Delta_i(h_i) &\leq (1 - \alpha)\Delta_i(h_i) + K\|h_i\|^2 \\ &\leq -(1 - \alpha)\omega_1 \min\{\sigma_i, \omega_1 / c_1\} + K c_2^2 \sigma_i^2 . \end{aligned} \tag{28d}$$

It follows now directly from (28d) that there exists a $\hat{\sigma} = \beta^k \sigma_{\max}$, with $k \geq 0$, such that, for all $\sigma_i \leq \hat{\sigma}$ and $h_i \in T_i$,

$$f(x_i + h_i) - f(x_i) - \alpha\Delta_i(h_i) \leq 0 . \tag{28e}$$

Hence, for all $x \in B(\hat{x}, \rho)$, (24d) holds with $\sigma_i \geq \hat{\sigma}$. Therefore, because of (28b), for all $x_i \in B(\hat{x}, \rho)$,

$$f(x_{i+1} - f(x_i) \leq -\tfrac{1}{2}\alpha\omega_1 \min\{\tilde{\sigma}, \omega_1 / c_1\} \triangleq -\delta < 0 . \tag{28f}$$

Since $\{f(x_i)\}_{i=0}^{\infty}$ is monotonically decreasing and since, by continuity of $f(\cdot)$, it has an accumulation point $f(\hat{x})$, it follows from Proposition 5.1.16 that $f(x_i) \to f(\hat{x})$ as $i \to \infty$. However, (28f) implies that $f(x_i) \to -\infty$, as $i \to \infty$, and hence we have a contradiction, which completes our proof. □

To complete our discussion of trust region methods, we will now sketch out an approach to the computation of a "good" h_i in Step 4 of Algorithm Model 1.2.26. Suppose that $\tilde{g}_i \neq 0$, and let $\hat{s}$ be a solution to

$$\min\{\tilde{\Delta}_i(s) \mid \|s\|^2 \leq \sigma_i^2\}. \tag{29a}$$

Then it is clear that $\hat{h} \triangleq D_i^{-1}\hat{s}$ is a solution to (23d), i.e., to $\min\{\Delta_i(h) \mid \|D_i h\| \leq \sigma_i\}$, and hence we need to deal only with the solution of (29a). Since $\tilde{g}_i \neq 0$, it follows that $\hat{s} \neq 0$, and hence it follows from Corollary

2.2.5 that there exists a scalar $\hat{\mu} \geq 0$ such that

$$\tilde{g}_i + (\tilde{H}_i + 2\hat{\mu}I)\hat{s} = 0 , \tag{29b}$$

where I is the $n \times n$ identity matrix and

$$\hat{\mu}(\|\hat{s}\|^2 - \sigma_i^2) = 0 . \tag{29c}$$

It is shown in [Sor.92] that (29b,c) is both a necessary and sufficient condition, i.e., that $\hat{s}$ is a solution of (29a) if and only if (29b,c) hold, with $\tilde{H}_i + 2\hat{\mu}I$ positive-semidefinite.

Since H_i is symmetric, it has a set of orthonormal eigenvectors $\{u_j\}_{j=1}^n$, corresponding to real eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, and hence has a spectral decomposition of the form:

$$\tilde{H}_i = \sum_{j=1}^{n} \lambda_j u_j u_j^T . \tag{29d}$$

First suppose that $\|\hat{s}\| < \sigma_i$. Then $\hat{\mu} = 0$ and $\hat{s}$ is an unconstrained minimizer of $\tilde{\Delta}_i(\cdot)$. In that case, by Theorem 1.1.5, $\tilde{H}_i \geq 0$, so that $\lambda_j \geq 0$, for $j = 1, \ldots, n$, and, in addition, we must have that $\langle \tilde{g}_i, u_j \rangle = 0$, for all $j \in \{1, \ldots, n\}$ such that $\lambda_j = 0$, because otherwise $\inf_{s \in \mathbb{R}^n} \tilde{\Delta}_i(s) = -\infty$. Let $\lambda_k > 0$, $k \leq n$, be the smallest nonzero eigenvalue of $\tilde{H}_i$. Then it follows from (29b) that

$$\hat{s} = -U\Lambda^{-1}U^T\tilde{g}_i , \tag{29e}$$

where $U = [u_1, \ldots, u_k]$ and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_k)$.

Thus, assuming that $\tilde{H}_i \neq 0$, we must first check if all $\lambda_j \geq 0$ and if $\|U\Lambda^{-1}U^T\tilde{g}_i\| \leq \sigma_i$, with U and Λ defined as in (29e) (since $\tilde{H}_i \neq 0$, there must be a $k \leq n$ such that $\lambda_k > 0$). If this turns out to be the case, then $\hat{s}$ is determined by (29e).

Next, suppose that $\|\hat{s}\| = \sigma_i$ and that $\hat{s}$ must be determined by (29b,c) with $\hat{\mu} \geq 0$. Since the case considered above, where $\hat{\mu} = 0$ corresponds to $\hat{s}$ being an unconstrained minimizer of $\tilde{\Delta}_i(\cdot)$, we need to deal only with the case where $\hat{\mu} > 0$. Note that the matrix

$$H_i + \nu I = \sum_{j=1}^{n} (\lambda_j + \nu) u_j u_j^T , \tag{29f}$$

is positive-definite, and hence nonsingular for all $\nu \in (-\lambda_n, \infty)$. Hence, since the eigenvectors of $\tilde{H}_i$ are orthonormal, we can define the function $\phi : (-\lambda_n, \infty) \to \mathbb{R}$ by

$$\phi(\nu) \triangleq \|(\tilde{H}_i + \nu I)^{-1}\tilde{g}_i\|^2 = \sum_{j=1}^{n} \frac{1}{(\lambda_j+\nu)^2}\langle u_j, \tilde{g}_j\rangle^2. \tag{29g}$$

Now $(d\phi/d\nu)(\nu) < 0$ for all $\nu \in (-\lambda_n, \infty)$, $\phi(\nu) \to 0$, as $\nu \to \infty$, and $\phi(\nu) \to \infty$, as $\nu \to -\lambda_n$. Hence there is a unique $\hat{\nu} \in (-\lambda_n, \infty)$ such that $\phi(\hat{\nu}) = \sigma_i^2$, and therefore $\hat{s} = -(\tilde{H}_i + \hat{\mu}I)^{-1}\tilde{g}_i$, with $\hat{\mu} = \hat{\nu}/2$, is a solution of (29a).

Efficient numerical methods for solving (29a), based on the above observations, are described in [Gan.81, MoS.83, GoM.91, Sor.94]. The simplest method uses a bisection technique, as follows. First, let $a_0 = \max\{0, -\lambda_n\}$, let $\delta > 0$, let $k \in \mathbb{N}$ be the smallest integer such that $\phi(k\delta) \leq \sigma_i^2$, and let $b_0 = k\delta$. Then it is clear that $\hat{\nu} \in [a_0, b_0]$. We now proceed recursively. Let $[a_j, b_j] \subset [a_0, b_0]$ be an interval containing $\hat{\nu}$. If $\phi((a_j + b_j)/2) \geq \sigma_i^2$, then $\hat{\nu} \in [a_{j+1}, b_{j+1}]$, with $a_{j+1} = (a_j + b_j)/2$ and $b_{j+1} = b_j$, else, $\hat{\nu} \in [a_{j+1}, b_{j+1}]$, with $b_{j+1} = (a_j + b_j)/2$ and $a_{j+1} = a_j$. We note that $b_{j+1} - a_{j+1} = (b_j - a_j)/2$, i.e., each iteration of this process halves the length of the interval containing $\hat{\nu}$. When $(b_j - a_j)/b_j$ is small, b_j is a good approximation to $\hat{\nu}$, and we can set $h_i = -D_i^{-1}[\tilde{H}_i + (b_j/2)I]^{-1}\tilde{g}_i$, provided that (24c) is satisfied.

1.2.5 Algorithm Implementation Theory

Throughout this book we will make a distinction between *conceptual* and *implementable* algorithms based on the following criterion, which assumes that all computations are carried out on an infinite precision computer. Thus, each iteration of a *conceptual algorithm* may be made up of an *arbitrary* number of infinite precision arithmetical operations and function evaluations, while each iteration of an *implementable algorithm* must be made up of a *finite* number of arithmetical operations and function evaluations.

It should be obvious that our distinction between conceptual and implementable algorithms is to some degree subjective, for the reader may decide to call the finding of a root of a transcendental equation a function evaluation, while we insist that it should be regarded as an infinite subprocedure and hence inadmissible as an operation to be performed in the course of each iteration of an implementable algorithm. As a rule of thumb, we assume that the computation of an acceptable approximation, to the values of such functions as e^x and $\sin x$, is performable in finite time. However, we assume that the computation of an "acceptable" approximation to the smallest positive root of the equation $e^{-x} + \sin x = 0$ requires infinite time because any practical procedure for computing this root, such as Newton's method, will require a very large number of values of e^{-x} and of $\sin x$.

To be consistent with the above, we must consider algorithms of the form of Algorithm Model 1.2.20 to be conceptual, because the computation of λ_i, according to (17) usually requires the use of a subprocedure which calls for the evaluation of $f(x_i + \lambda h_i)$ an infinite number of times. The use of infinite subprocedures is not the only factor which can make an algorithm conceptual. The need to compute gradients $\nabla f(x)$ by finite differences has a similar effect and can also render algorithms of the form of Algorithm Model 1.2.23 conceptual.

In view of the discussion above, it is clear that algorithms corresponding to the algorithm models discussed so far may, at least on some problems, be only conceptual. It is also plausible to assume (and has been observed in practice) that the indiscriminate use of approximations may destroy the convergence properties promised by the theorems that we have seen. Hence, we adopt a systematic approach to the implementation of conceptual algorithms. For this purpose, we return now to the general setting of the Model Problem 1.2.6. We will consider two situations. In the first, the value of the cost function $f(x_i)$ can be computed exactly, but the value of the optimality function $\theta(x_i)$ and points in the set $\underline{A}(x_i)$, specified by a conceptual algorithm, can only be approximated. In the second situation, everything must be approximated: $f(x_i)$, $\theta(x_i)$, and points in the set $\underline{A}(x_i)$. The second situation is commonly encountered in the solution of semi-infinite minimax problems, such as $\min_{x \in \mathbb{R}^n} \max_{y \in Y} \phi(x, y)$, optimal control, shape optimization, and structural optimization. For example, suppose that it is necessary to compute the radius at which to place six symmetrically spaced point-supports for a horizontal antenna dish, so as to minimize the least squares distortion of the dish caused by its weight. In this case the cost function is computed by solving a partial differential equation by the finite element method, a process based on approximation of the surface by adjacent triangular elements. This fact also applies to the evaluation of an optimality function, and points in a set, defined by an algorithm function, $\underline{A}(x_i)$.

We begin by addressing the first situation, viz., we assume that the cost function can be evaluated exactly without any difficulty. The following multipoint algorithm implementation model uses an algorithm map $\underline{A} : \mathbb{R}_+ \times \underline{X} \to 2^X$, where $\underline{X}$ was defined in (12b). We think of the set $\underline{A}(\varepsilon, x_i)$ as an approximation to the set $\underline{A}(x_i)$ in Algorithm Model 1.2.17, with ε acting as a precision parameter. We assume that precision is increased as $\varepsilon \downarrow 0$, so that $\underline{A}(0, x_i) = \underline{A}(x_i)$. To be consistent with the assumption that the algorithm function must be approximated, we also assume that the optimality function must be approximated (i.e., we can compute only an approximation $\theta_\varepsilon(x)$ to $\theta(x)$), as is the case when gradients must be computed by finite differences. Since $\theta_\varepsilon(c) = 0$ is not a meaningful stopping criterion, we do not include a stopping criterion in the algorithm model below. Please note that the α and β below are not related to α and β in Algorithm Model 1.2.23.

Algorithm Model 1.2.30.

Parameters. $\varepsilon_0 > 0, \alpha > 0, \beta \in (0,1)$.

Data. $x_0 \in X$.

Step 0. Set $i = 0$.

Step 1. Set $\varepsilon = \varepsilon_0$.

Step 2. Compute a $y \in \underline{A}(\varepsilon, x_i)$.

Step 3. If

$$f(y) - f(x_i) \leq -\alpha\varepsilon\,, \tag{30}$$

set $x_{i+1} = y$, replace i by $i+1$, and go to Step 1.

Else, replace ε by $\beta\varepsilon$, and go to Step 2.

Theorem 1.2.31. *Consider Algorithm Model 1.2.30 together with Model Problem 1.2.6. Suppose that, for every $x \in X$ that is not stationary, at least one of the following two statements holds:*

(i) There exist a $\rho_x > 0$, a $\delta_x > 0$, and an $\varepsilon_x > 0$, such that if $\{x_j\}_{j=0}^{i}$ is a finite sequence generated by Algorithm Model 1.2.30 and $x_i \in B(x, \rho_x)$, then

$$f(y) - f(x_i) \leq -\delta_x < 0, \forall\, y \in \underline{A}(\varepsilon, x_i),\ \forall\, \varepsilon \in (0, \varepsilon_x]\,. \tag{31a}$$

(ii) There exist a $\rho_x > 0$, a $\gamma_x > 0$, and an $\varepsilon_x > 0$, such that if $\{x_j\}_{j=0}^{i}$ is a finite sequence generated by Algorithm Model 1.2.30 and $x_i \in B(x, \rho_x)$, then

$$f(x') - f(x_i) \leq -\gamma_x \theta(x_i) < 0, \forall\, x' \in \underline{A}(\varepsilon, x_i),\ \forall\, \varepsilon \in (0, \varepsilon_x]\,. \tag{31b}$$

Then, either the sequence $\{x_i\}$, constructed by Algorithm Model 1.2.30, is finite (i.e., Algorithm Model 1.2.30 jams up at a point x_k, cycling in the loop defined by Steps 2 and 3) and its last element x_k is stationary, or else it is infinite and every accumulation point of $\{x_i\}_{i=0}^{\infty}$ is stationary.

Proof. Since the existence of a $\rho_x > 0$, a $\gamma_x > 0$, and an $\varepsilon_x > 0$ such that (31b) holds implies the existence of a $\rho_x > 0$, a $\delta_x > 0$, and an $\varepsilon_x > 0$ such that (31a) holds, we need only to establish the theorem under hypothesis (31a). First we must show that Algorithm Model 1.2.30 cannot jam up at a non-quasi-stationary point $x_k \in X$, cycling in the loop defined by Step 2 and Step 3. Thus, suppose that the algorithm jams up at x_k, a nonstationary point. Then the algorithm must be producing an infinite sequence of vectors $y_j \in \underline{A}(\varepsilon_0\beta^j, x_k)$, $j = 0, 1, 2, \ldots$, in Step 2, such that

$$f(y_j) - f(x_k) > -\alpha\,\varepsilon_0\beta^j\,, \tag{32a}$$

which prevents the construction of x_{k+1}. However since x_k is not stationary, by assumption, there exist an $\rho_k > 0$, a $\delta_k > 0$, and an $\varepsilon_k > 0$ for which (31a) holds, with $x = x_k$, $\delta_x = \delta_k$, and $\varepsilon_x = \varepsilon_k$. Also, there exists an integer $j_0 > 0$ such that, for all $j \in \mathbb{N}$, $j \geq j_0$, $\varepsilon_0\beta^j \leq \min\{\delta_k/\alpha\,,\ \varepsilon_k\}$, and hence, from (31a), we must have that

$$f(y_j) - f(x_k) \leq \delta_k \leq -\alpha\varepsilon_0\beta^j\,, \tag{32b}$$

for all $j \geq j_0$, which contradicts our hypothesis that the algorithm jams up at x_k. Consequently, Algorithm Model 1.2.30 cannot jam up at a non-quasi-stationary point x_k.

Hence, let us suppose that the sequence $\{x_i\}$ is infinite, that $x_i \to^K \hat{x}$, as $i \to \infty$, and that the accumulation point $\hat{x}$ is not stationary. Since, by construction, $\{x_i\}_{i=0}^{\infty}$ is contained in X and X is closed, $\hat{x} \in X$. Then, by assumption, there exist a $\hat{\rho} > 0$, a $\hat{\delta} > 0$, and an $\hat{\varepsilon} > 0$ for which (31a) holds. Since $x_i \to^K \hat{x}$, as $i \to \infty$, there must exist an $i_0 \in K$ such that $x_i \in B(\hat{x}, \hat{\rho})$ for all $i \in K$, $i \geq i_0$, and there is also an integer k such that $\varepsilon_0\beta^k \leq \min\{\hat{\delta}/\alpha\,, \hat{\varepsilon}\}$. Hence, for any $i \in K$ such that $i \geq i_0$, we must have that $\varepsilon_i \geq \varepsilon_0\beta^k$ and hence that

$$f(x_{i+1}) - f(x_i) \leq -\delta_{\hat{x}} \leq -\alpha\varepsilon_0\beta^k \leq -\alpha\varepsilon_i\,. \tag{32c}$$

Now, by construction, the sequence $\{f(x_i)\}_{i=0}^{\infty}$ is monotone decreasing, and, by continuity of $f(\cdot)$, it has an accumulation point $f(\hat{x})$. Therefore, it follows from the Monotone Sequences Proposition 5.1.16 that this sequence converges to $f(\hat{x})$. However, this is contradicted by (32c) and hence our proof is complete. □

Algorithm Model 1.2.30 is *time-invariant* because it always resets ε to ε_0 in Step 1 in each iteration. The advantage to this approach is that one never uses more precision than is absolutely necessary. However, when the computation of $y \in \underline{A}(\varepsilon, x_i)$ is costly, repeating many cycles in the loop defined by Step 2 and Step 3 in Algorithm Model 1.2.30 may be unacceptable. In such a case, it is preferable not to reset ε to ε_0, but to use the substitute Step 3, outlined below.

Exercise 1.2.32. Consider the time-varying version of Algorithm Model 1.2.30, in which Step 3 is replaced by Step 3' below, i.e., ε is not reset to ε_0 at the beginning of each iteration. Show that under the assumptions stated in Theorem 1.2.31, its conclusions are also valid for this time-varying version:

Step 3'. If

$$f(y) - f(x_i) \leq -\alpha\varepsilon\,, \tag{30}$$

set $x_{i+1} = y$, replace i by $i+1$, and go to Step 2.

Else, set $\varepsilon = \beta\varepsilon$, and go to Step 2. □

Algorithm Model 1.2.30 uses a feedback mechanism for precision adjustment, which demands less precision when x_i is far from a stationary point than when it is near a stationary point. This fact results in considerable computational savings over schemes that use very high precision throughout the entire computation.

Let ε_i denote the value of ε accepted in iteration i of Algorithm Model 1.2.30. Then it is obvious from (30) that $f(x_{i+1}) - f(x_i) \leq -\alpha\varepsilon_i$ for all i and hence, if Algorithm Model 1.2.30 (or its time-varying version) constructs a bounded infinite sequence $\{x_i\}_{i=0}^{\infty}$, then $\varepsilon_i \to 0$, as $i \to \infty$.

The reader may have wondered whether it is absolutely necessary to have a *feedback* mechanism, such as (30) in Algorithm Model 1.2.30, for precision adjustment. We will show now that one can also use an *open-loop* approach which adjusts the precision of approximation at a preselected rate. In practice, algorithms of the form given below are very efficient in situations where one solves essentially the same problem over and over again. In such a case, one would spend a certain amount of time carefully selecting an open-loop precision improvement strategy which could be made more efficient than a feedback strategy, particularly when the parameter α in (30) is poorly selected. A near optimal scheme for precision adjustment in a class of algorithms can be found in [HeP.90].

We will say that a function $t : \mathbb{N} \to \mathbb{N}$ is a *truncation function* if it is monotone increasing and $t(k) \to \infty$, as $k \to \infty$. The algorithm model below uses an algorithm function $A : \mathbb{N} \times X \to 2^X$ and two truncation functions, $t_1(\cdot)$, $t_2(\cdot)$. We assume that the larger the value of the integer j, the the greater the accuracy with which $A(j, \cdot)$ approximates a conceptual algorithm map $A(\cdot)$. The algorithm model below uses two counters, one for the iteration index i and one for the precision control parameter j.

Algorithm Model 1.2.33.

Data. $x_0 \in X$.

Step 0. Set $i = 0$.

Step 1. Set $x = x_i$ and set $j = t_1(i)$.

Step 2. Compute a $y \in A(j, x)$.

Step 3. If $f(y) < f(x)$, set $x_{i+1} = y$, set $i = i + 1$, and go to step 1.
Else, set $x_{i+1} = x_i$, set $j = t_2(i)$, replace i by $i + 1$, and go to step 2.

Theorem 1.2.34. *Consider Algorithm Model 1.2.33 together with Model Problem 1.2.6. Suppose that for every $x \in X$ which is not quasi-stationary, there exist an $\varepsilon_x > 0$, a $\delta_x > 0$ and a $k_x \in N$ such that*

$$f(x'') - f(x') \leq -\delta_x < 0, \tag{33}$$

for all $x' \in B(x, \varepsilon_x)$, and for all $x'' \in A(j, x')$ and $j \geq k_x$. Then every accumulation point of an infinite sequence $\{x_i\}_{i=0}^{\infty}$, constructed by Algorithm Model 1.2.33, is stationary (i.e., $\hat{x} \in X$ and $\theta(\hat{x}) = 0$). □

Exercise 1.2.35. Prove Theorem 1.2.34. □

Next, we turn to the situation where everything must be approximated: $f(x_i)$, $\theta(x_i)$, and points in the set $\underline{A}(x_i)$. Thus, we assume that, in addition to an approximating algorithm function $\underline{A}(\varepsilon, x_i)$, we are also given an approximating cost function $f^*(\varepsilon, x_i)$. In this case, Algorithm Model 1.2.30 becomes extended to the following form:

Master Algorithm Model 1.2.36.

Parameters: $\varepsilon_0 > 0$ $\alpha, \beta, \gamma \in (0, 1)$.

Data. $x_0 \in X$.

Step 0. Set $i = 0$.

Step 1. Set $\varepsilon = \varepsilon_0$.

Step 2. Compute a $y \in \underline{A}(\varepsilon, x_i)$

Step 3. If

$$f^*(\varepsilon, y) - f^*(\varepsilon, x_i) \leq -\alpha\varepsilon^{\gamma}, \tag{34}$$

set $x_{i+1} = y$, replace i by $i + 1$, and go to Step 2.
Else, replace ε by $\beta\varepsilon$, and go to Step 2.

Note that for technical reasons, the test in (34) had to be made less stringent than the test in (30).

Theorem 1.2.37. *Consider Algorithm Model 1.2.36 together with Model Problem 1.2.6, and suppose that*

(i) for any bounded subset $S \subset \mathbb{R}^n$, there exists constants $\varepsilon_S > 0$ and $K_S \in (0, \infty)$ such that for all $x \in S$ and $\varepsilon \in [0, \varepsilon_S]$,

$$|f^*(\varepsilon, x) - f(x)| \leq K_S\varepsilon; \tag{35a}$$

(ii) for every $x \in X$ that is not stationary, there exist a $\rho_x > 0$, a $\delta_x > 0$, and an $\varepsilon_x > 0$ such that if $\{x_j\}_{j=0}^{i}$ is a sequence constructed by Algorithm Model

1.2.36, and $x_i \in B(x, \rho_x)$, *then*

$$f^*(\varepsilon, y) - f^*(\varepsilon, x_i) \leq -\delta_x , \tag{35b}$$

for all $y \in \underline{A}(\varepsilon, x_i)$ *and* $\varepsilon \in (0, \varepsilon_x]$.

Then, either the sequence $\{x_i\}$ *constructed by Algorithm Model 1.2.36 is finite (i.e., Algorithm Model 1.2.36 jams up at a point* x_k, *cycling in the loop defined by Steps 2 and 3) and its last element* x_k *is stationary, or else it is infinite and every accumulation point of* $\{x_i\}_{i=0}^{\infty}$ *is stationary.* □

We omit a proof of Theorem 1.2.37 since its proof can be deduced from the proof of Theorem 3.3.19.

Our final result establishes a sufficient condition for a sequence $\{x_i\}_{i=0}^{\infty}$ constructed by any one of our algorithm models to converge to a stationary point.

Proposition 1.2.38. *Suppose that* $\{x_i\}_{i=0}^{\infty}$ *is a bounded sequence constructed by any one of our Algorithm Models, that* $\|x_{i+1} - x_i\| \to 0$, *as* $i \to \infty$, *and that it has an isolated accumulation point* $\hat{x} \in S$ *(where* S *is the set of stationary points), i.e., if* $\inf\{\|x - \hat{x}\| \mid x \in S,\ x \neq \hat{x}\} > 0$, *then* $x_i \to \hat{x}$, *as* $i \to \infty$.

Proof. Suppose that $\hat{x}$ is as stipulated and that $\rho \triangleq \inf\{\|x - \hat{x}\| \mid x \in S\}$. By assumption, $\rho > 0$. Since, by the property of the Algorithm Models, all the accumulation points of $\{x_i\}_{i=0}^{\infty}$ are in S, there must exist an i_0 such that for all $i \geq i_0$, $\inf\{\|x - x_i\| \mid x \in QS\} \leq \rho/3$. Let $i_1 \geq i_0$ be such that $\|x_{i+1} - x_i\| \leq \rho/4$ for all $i \geq i_1$, and let $i_2 \geq i_1$ be such that $x_{i_2} \in B(\hat{x}, \rho/3)$. Then

$$\|x_{i_2+1} - \hat{x}\| = \|(x_{i_2+1} - x_{i_2}) + (x_{i_2} - \hat{x})\| \leq \rho/4 + \rho/3 = 7\rho/12 . \tag{36a}$$

Hence, for all $x \in S, x \neq \hat{x}$,

$$\begin{aligned} \|x - x_{i_2+1}\| &= \|(x - \hat{x}) - (x_{i_2+1} - \hat{x})\| \\ &\geq \|x - \hat{x}\| - \|x_{i_2+1} - \hat{x}\| \\ &> \rho - 7\rho/12 > \rho/3 . \end{aligned} \tag{36b}$$

Since we must have that $\inf\{\|x - x_{i_2+1}\| \mid x \in S\} \leq \rho/3$, it follows that $x_{i_2+1} \in B(\hat{x}, \rho/3)$. It now follows by induction that $x_i \in B(\hat{x}, \rho/3)$ for all $i \geq i_2$. Since $\hat{x}$ is the only possible accumulation point in $B(\hat{x}, \rho/3)$, the desired result follows. □

1.2.6 Rate of Convergence of Sequences

The most reliable way of evaluating the relative merits of two algorithms is to apply both to a set of problems of interest and to compare the cpu times needed to solve these problems. Such an exhaustive comparison is not always possible. Hence it is useful to have some mathematical measures of algorithm performance that can be used to make qualitative distinctions between algorithms. We will present an elementary exposition of two of these performance measures. For an in-depth treatment, please see [OrR.70] and [NeY.83].

Definition 1.2.39.

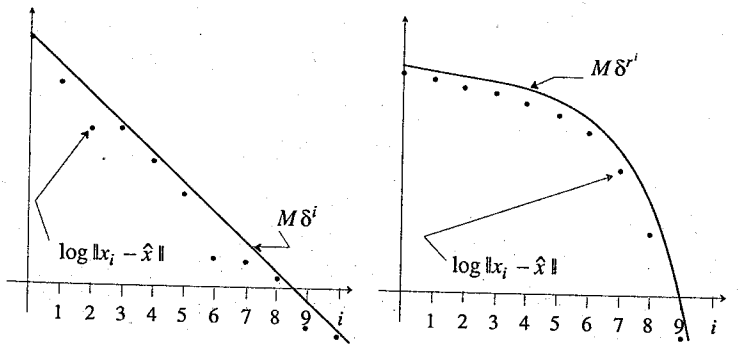
(a) We say that a sequence $\{x_i\}_{i=0}^{\infty}$ *in* $\mathbb{R}^n$ *converges to a point* $\hat{x}$ *at least* R-linearly *(i.e., with* root rate *(*R*-rate) 1) if there exist* $\kappa \in (0, \infty)$, $c \in (0, 1)$, *and* $i_0 \in \mathbb{N}$ *such that for all* $i \geq i_0$,

$$\|x_i - \hat{x}\| \leq \kappa c^i . \tag{37a}$$

We say that a sequence $\{x_i\}_{i=0}^{\infty}$ *in* $\mathbb{R}^n$ *converges to a point* $\hat{x}$ R-superlinearly *if there exist a* $\kappa \in (0, \infty)$ *and a sequence* $\{c_i\}_{i=0}^{\infty}$ *such that* $c_i > 0$ *for all* $i \in \mathbb{N}$, $c_i \to 0$, *as* $i \to \infty$, *and*

$$\|x_i - \hat{x}\| \leq \kappa \prod_{k=0}^{i} c_k . \tag{37b}$$

We say that a sequence $\{x_i\}_{i=0}^{\infty}$ *in* $\mathbb{R}^n$ *converges to a point* $\hat{x}$ *at least with* root rate *(*R-rate*)* $r > 1$ *if there exist a* $\kappa \in [0, \infty)$, *a* $c \in [0, 1]$, *and an* $i_0 \in \mathbb{N}$ *such that for all* $i \geq i_0$,



(a) R-linear convergence (b) R-superlinear convergence

Fig. 1.2.4. Illustration of R-linear and R-superlinear convergence.

$$\|x_i - \hat{x}\| \le \kappa c^{r^i} . \tag{37c}$$

(b) We say that a sequence $\{x_i\}_{i=0}^{\infty}$ *in* $\mathbb{R}^n$ *converges to a point* $\hat{x}$ *at least* Q-linearly *(i.e., with* quotient rate *(*Q*-rate) 1 if there exist a* $c \in (0,1)$ *and an* $i_0 \in \mathbb{N}$ *such that for all* $i \ge i_0$,

$$\frac{\|x_{i+1} - \hat{x}\|}{\|x_i - \hat{x}\|} \le c . \tag{37d}$$

We say that a sequence $\{x_i\}_{i=0}^{\infty}$ *in* $\mathbb{R}^n$ *converges to a point* $\hat{x}$ *at least* Q-superlinearly *if there exists a sequence* $\{c_i\}_{i=0}^{\infty}$ *such that* $c_i > 0$ *for all* $i \in \mathbb{N}$, $c_i \to 0$, *as* $i \to \infty$, *and*

$$\frac{\|x_{i+1} - \hat{x}\|}{\|x_i - \hat{x}\|} \le c_i \ \forall\, i \in \mathbb{N} . \tag{37e}$$

We say that a sequence $\{x_i\}_{i=0}^{\infty}$ *in* $\mathbb{R}^n$ *converges to a point* $\hat{x}$ *at least with* quotient rate *(*Q-rate*)* $r > 1$ if there exist a $\kappa \in [0,\infty)$ and an $i_0 \in \mathbb{N}$ such that for all $i \ge i_0$,

$$\frac{\|x_{i+1} - \hat{x}\|}{\|x_i - \hat{x}\|^r} \le \kappa . \tag{37f}$$

□

The naming of the quotient rate is obvious from the relations (37d,e,f). The naming of the root rate stems from the fact that, if $x_i \to \hat{x}$, as $i \to \infty$, at least R–linearly, then there exists a $c \in [0,1)$ such that $\overline{\lim} \|x_i - \hat{x}\|^{1/i} \le c$. When $x_i \to \hat{x}$, as $i \to \infty$, R-superlinearly, $\overline{\lim} \|x_i - \hat{x}\|^{1/i} = 0$, and when $x_i \to \hat{x}$, as $i \to \infty$, R-superlinearly with R–rate at least $r > 1$, then there exists a $c \in [0,1)$ such that $\overline{\lim} \|x_i - \hat{x}\|^{1/r^i} \le c$.

When plotted on a semilog scale, a sequence converging at least R–linearly produces a graph as shown in Fig. 1.2.4, while a R–superlinearly converging sequence produces a graph as shown in Fig. 1.2.4b.

Remark 1.2.40. *(a)* It is possible for a sequence $\{x_i\}_{i=0}^{\infty}$ to converge to a point $\hat{x}$ *slower* then R–linearly, e.g., when $\|x_i - \hat{x}\| = k/i$. In this case $\overline{\lim} \|x_i - \hat{x}\|^{1/i} = 1$.

(b) Consider the sequence of real numbers $\{e_i\}_{i=0}^{\infty}$, with $e_i = (\frac{1}{2})^i$ for i even and $e_i = (\frac{1}{4})^i$ for i odd. Clearly, $e_i \to 0$, as $i \to \infty$, R-linearly, since $e_i \le (\frac{1}{2})^i$ for all i. However, this sequence does not converge Q-linearly.

(c) Although, towards the end, any superlinearly converging sequence approaches its limit much faster than any linearly converging sequence, the initial progress may be slower for the superlinearly converging sequence, as illustrated in Fig. 1.2.4b. □

The following results are basic to the study of the rate of convergence of "one-point" algorithms such as gradient methods, secant methods, and various forms of Newton's method.

Theorem 1.2.41. *Let* $\{x_i\}_{i=0}^{\infty}$ *be a sequence in* $\mathbb{R}^n$ *such that* $x_i \to \hat{x}$, *as* $i \to \infty$, *for some* $\hat{x} \in \mathbb{R}^n$.

(a) If $x_i \to \hat{x}$, *as* $i \to \infty$, *at least* Q*-linearly, i.e., there exists a* $c \in (0,1)$ *and an* $i_0 \in \mathbb{N}$ *such that*

$$\|x_{i+1} - \hat{x}\| \le c \|x_i - \hat{x}\|, \quad \forall\, i \ge i_0, \tag{38a}$$

then $x_i \to \hat{x}$, *as* $i \to \infty$, *at least* R*–linearly.*

(b) Suppose that $x_i \to \hat{x}$, *as* $i \to \infty$, Q*-superlinearly, with* Q*-rate* $r > 1$, *i.e., there exist a* $\kappa \in (0,\infty)$ *and an* $r > 1$ *such that*

$$\|x_{i+1} - \hat{x}\| \le \kappa \|x_i - \hat{x}\|^r, \quad \forall\, i \ge i_0, \tag{38b}$$

then $x_i \to \hat{x}$, *as* $i \to \infty$, R*–superlinearly with* R*-rate at least* r.

Proof. (a) For all $i \in \mathbb{N}$, let $e_i \triangleq \|x_i - \hat{x}\|$. Then, by induction, for all $i \ge i_0$,

$$e_i \le e_{i_0} c^{i-i_0} = (c^{-i_0} e_{i_0})\, c^i , \tag{39a}$$

which completes the proof of *(a)*.

(b) Again let $e_i \triangleq \|x_i - \hat{x}\|$. Since $e_i \to 0$, as $i \to \infty$, by assumption, there exists an $i_1 \ge i_0$ such that

$$\kappa^{[1/(r-1)]} e_{i_1} < 1 . \tag{39b}$$

Next, multiplying both sides of (38b) by $\kappa^{[1/(r-1)]}$, we find that

$$\kappa^{[1/(r-1)]} e_{i+1} \le \{\kappa^{[1/(r-1)]} e_i\}^r, \quad \forall\, i \ge i_1 . \tag{39c}$$

Let $\varepsilon_i \triangleq \kappa^{[1/(r-1)]} e_i$. Then, in view of (39b), $\varepsilon_{i_1} < 1$, and (39c) becomes

$$\varepsilon_{i+1} \le \varepsilon_i^r, \quad \forall\, i \ge i_0 . \tag{39d}$$

Hence, for all $i \ge i_1 + 1$,

$$\varepsilon_i \le \varepsilon_{i-1}^r \le \varepsilon_{i-2}^{r^2} \le \varepsilon_{i-3}^{r^3} \le \cdots \le \varepsilon_{i_1}^{r^{(i-i_1)}} , \tag{39e}$$

and therefore for all $i \ge i_1$,

$$e_i \le \frac{1}{\kappa^{[1/(r-1)]}}(\kappa^{[1/(r-1)]}e_{i_1})^{r^{(i-i_1)}} = \frac{1}{\kappa^{[1/(r-1)]}}\left((\kappa^{[1/(r-1)]}e_{i_1})^{(1/r)^{i_1}}\right)^{r^i} . \quad (39f)$$

Since, by assumption, $\kappa^{[1/(r-1)]}e_{i_1} < 1$, the desired result now follows, with $M = (1/\kappa^{[1/(r-1)]})$ and $c = (\kappa^{[1/(r-1)]}e_{i_1})^{(1/r)^{i_1}}$. □

Theorem 1.2.42. *Let $\{x_i\}_{i=0}^{\infty}$ be a sequence in $\mathbb{R}^n$.*

(a) If there exists a $c \in (0,1)$ and an $i_0 \in \mathbb{N}$ such that

$$\|x_{i+1} - x_i\| \le c\,\|x_i - x_{i-1}\| , \quad \forall\ i \ge i_0+1 , \quad (40a)$$

then there exists an $\hat{x} \in \mathbb{R}^n$ such that $x_i \to \hat{x}$, as $i \to \infty$, at least R-linearly.

(b) If there exists a $\kappa \in (0,\infty)$, an $r > 1$, and an $i_0 \in \mathbb{N}$ such that

$$\kappa^{[1/(r-1)]}\,\|x_{i_0+1} - x_{i_0}\| < 1 , \quad (40b)$$

and

$$\|x_{i+1} - x_i\| \le \kappa\|x_i - x_{i-1}\|^r , \quad \forall\ i \ge i_0+1 , \quad (40c)$$

then there exists an $\hat{x} \in \mathbb{R}^n$ such that $x_i \to \hat{x}$, as $i \to \infty$, R-superlinearly, with R-rate at least r.

Proof. *(a)* For $i = 0,1,2,\ldots$, let $e_i \triangleq \|x_{i+1} - x_i\|$. Then, by induction, it follows from (40a) that for all $i \ge i_0$, $e_i \le e_{i_0}c^{i-i_0}$, and hence, since $c \in (0,1)$, that $e_i \to 0$, as $i \to \infty$, R-linearly. Therefore, for any $j > k \ge i_0$, we find that

$$\|x_j - x_k\| = \|(x_j - x_{j-1}) + (x_{j-1} - x_{j-2}) + \ldots + (x_{k+1} - x_k)\|$$

$$\le \sum_{i=k}^{j-1} e_i \le \sum_{i=k}^{\infty} e_i \le \sum_{j=k-i_0}^{\infty} e_{i_0}c^j = c^{k-i_0}\left(\frac{e_{i_0}}{1-c}\right) , \quad (41a)$$

i.e., $\|x_j - x_k\| \to 0$ for $j > k$, as $k \to \infty$, uniformly in j. Hence the sequence $\{x_i\}_{i=0}^{\infty}$ is Cauchy, and therefore, it must converge to a point $\hat{x}$.

Next, letting $j \to \infty$, we can replace x_j by $\hat{x}$ in (41a) to obtain

$$\|\hat{x} - x_k\| \le \left[\frac{e_{i_0}c^{-i_0}}{1-c}\right] c^k , \quad (41b)$$

which proves that $x_k \to \hat{x}$, as $k \to \infty$ R-linearly.

(b) Next, again with $e_i = \|x_{i+1} - x_i\|$, proceeding as in part *(b)* of the proof of Theorem 1.2.41, we deduce from (40c) that with $M \triangleq (1/\kappa)^{[1/(r-1)]}$ and $c \triangleq (\kappa^{[1/(r-1)]}e_{i_0})^{(1/r)^{i_0}}$,

$$e_i \le Mc^{r^i} , \quad \forall\ i \ge i_0 . \quad (41c)$$

Next we note that because of (40b), $c \in (0,1)$. Therefore, since $r > 1$, there

exists an $i_r \ge i_0$ such that $c^{(r^i - r^k)} \le c^{(i-k)}$ for all $i \ge k \ge i_r$, and hence, by arguments analogous to the ones used in obtaining (41a), we conclude that, for all $j > k \ge i_0$,

$$\|x_j - x_k\| \le \sum_{i=k}^{\infty} e_i \le M\sum_{i=k}^{\infty} c^{r^i} \le Mc^{r^k}\sum_{i=k}^{\infty} c^{(r^i - r^k)} = Mc^{r^k}\sum_{l=0}^{\infty} c^{r^k(r^l-1)} . \quad (41d)$$

Now $(d/dy)r^r(r^y - 1) = r^{k+y}\ln r$. Hence, if $i_1 \ge i_0$ is such that $r^{i_1}\ln r \ge 1$, then for all $k \ge i_1$ and $l \in \mathbb{N}$, $l \ge 1$, $r^k(r^l - 1) \ge l$ and, therefore, for all $j \ge k$,

$$\|x_j - x_k\| \le \frac{M}{1-c}c^{r^k} , \quad (41e)$$

which proves that the sequence $\{x_k\}_{k\in\mathbb{N}}$ is Cauchy, so that it must have a limit point $\hat{x}$. Letting $j \to \infty$ in (41e) and setting $M' \triangleq M/(1-c)$, we obtain

$$\|\hat{x} - x_k\| \le M'c^{r^k} , \quad \forall\ k \ge i_1 , \quad (41f)$$

which shows that $x_k \to \hat{x}$, as $i \to \infty$, with R-rate at least r. □

There is a broad class of multi-point methods related to Newton's method. These include the iterated Newton method, which updates the Hessian only periodically, as well as a variety of secant methods. The study of their rate of convergence is considerably simplified by the following rather general result.

Theorem 1.2.43. *Let $\{x_i\}_{i=0}^{\infty}$ be a sequence in $\mathbb{R}^n$. Suppose that there exist an $\hat{x} \in \mathbb{R}^n$, a $p \in \mathbb{N}$, and $\gamma_j \ge 0$, $j \in \mathbf{p}$, not all zero, such that for all $i \in \mathbb{N}$, $i \ge p$,*

$$\|x_{i+1} - \hat{x}\| \le \|x_i - \hat{x}\|\sum_{j=0}^{p}\gamma_j\|x_{i-j} - \hat{x}\| . \quad (42a)$$

If there exists an $\eta \in (0,1)$ such that for $i = 0,1,\ldots,p$,

$$\|x_i - \hat{x}\| \le \eta/\sum_{j=0}^{p}\gamma_j , \quad (42b)$$

then $x_i \to \hat{x}$, as $i \to \infty$, with R-rate τ_p, where τ_p is the unique positive root of the equation

$$t^{p+1} - t^p - 1 = 0 . \quad (42c)$$

Proof. *(a) (convergence)* Let $\gamma \triangleq \sum_{j=0}^{p}\gamma_j$, let $\delta_j \triangleq \gamma_j/\gamma$, for $j = 0,1,\ldots,p$, and let $\varepsilon_i \triangleq \gamma\|x_i - \hat{x}\|$. Then (42a) transforms into

$$\varepsilon_{i+1} \le \varepsilon_i\sum_{j=0}^{p}\delta_j\varepsilon_{i-j} . \quad (43a)$$

Now, by assumption, $\|x_i - \hat{x}\| \le \eta/\gamma$, for $i = 0, 1, 2, \ldots, p$. Hence it follows that $\varepsilon_i \le \eta < 1$ for $i = 0, 1, 2, \ldots, p$. Next, since $\sum_{j=0}^{p} \delta_j = 1$,

$$\varepsilon_{p+1} \le \varepsilon_p \sum_{j=0}^{p} \delta_j \varepsilon_{p-j} \le \eta \sum_{j=0}^{p} \delta_j \eta = \eta^2 ,$$

$$\varepsilon_{p+2} \le \varepsilon_{p+1} \sum_{j=0}^{p} \delta_j \varepsilon_{p+1-j} \le \eta^2 \sum_{j=0}^{p} \delta_j \eta = \eta^3 ,$$

$$\varepsilon_{p+3} \le \varepsilon_{p+2} \sum_{j=0}^{p} \delta_j \, \varepsilon_{p+2-j} \le \eta^3 \eta = \eta^4 .$$

Continuing in this manner, we conclude that

$$\varepsilon_{p+k} \le \eta^{k+1}, \quad k = 1, 2, \ldots \tag{43b}$$

Hence $\varepsilon_i \to 0$, as $i \to \infty$, which shows that $x_i \to \hat{x}$, as $i \to \infty$.

(b) (rate of convergence) It is clear from (43b) that $\varepsilon_i \le \eta^{\mu_i}$, for some integer $\mu_i \ge i - p + 1$. We will show that for $i \ge 0$, this relation is satisfied with $\mu_i \ge \tau_p^i/\tau_p^p$. Because $1 < \tau_p < 2$ holds, $\tau_p^i/\tau_p^p \le 1$ for $i = 0, 1, \ldots, p$, and hence $\eta \le \eta^{(\tau_p^i/\tau_p^p)}$ for $i = 0, 1, \ldots, p$. Consequently, we may set

$$\mu_0 = \mu_i = \cdots = \mu_p = 1 . \tag{43c}$$

Next we proceed by induction. Suppose that $\mu_i \ge \tau_p^i/\tau_p^p$ for $i = 0, 1, 2, \ldots, k$, with $k \ge p$. Since, by (43a),

$$\varepsilon_{k+1} \le \varepsilon_k \sum_{j=0}^{p} \delta^j \varepsilon_{k-j} \le \eta^{\mu_k} \sum_{j=0}^{p} \delta^j \, \eta^{\mu_{k-j}} \le \eta^{\mu_k + \mu_{k-p}} , \tag{43d}$$

it follows that

$$\mu_{k+1} \ge \mu_k + \mu_{k-p} . \tag{43e}$$

Since, by our induction hypothesis, $\mu_i \ge \tau_p^i/\tau_p^p$ for $i = 0, 1, 2, \ldots, k$, it follows that

$$\mu_{k+1} \ge \frac{1}{\tau_p^p}[\tau_p^k + \tau_p^{k-p}] = \frac{\tau_p^{k+1}}{\tau_p^p}[\tau_p^{-1} + \tau_p^{-(p+1)}] . \tag{43f}$$

Now, from (42c),

$$\tau_p^{p+1}[1 - \tau_p^{-1} - \tau_p^{-(p+1)}] = 0 , \tag{43g}$$

which implies that

$$\tau_p^{-1} + \tau_p^{-(p+1)} = 1 . \tag{43h}$$

Substituting from (43h) into (43g), we find that

$$\mu_{k+1} \ge \tau_p^{k+1}/\tau_p^p , \tag{43i}$$

which completes the induction step. Thus we have shown that, for all $\in \mathbb{N}$,

$$\|x_i - \hat{x}\| \le \frac{1}{\gamma}\eta^{\tau_p^i/\tau_p^p} = \frac{1}{\gamma}(\eta^{1/\tau_p^p})^{\tau_p^i} . \tag{43j}$$

It follows by inspection, from (43j), that $x_i \to \hat{x}$, as $i \to \infty$ with R-rate at least τ_p. □

The following table gives a few values of τ_p:

p	1	2	3	4	5	10	50
τ_p	1.618	1.466	1.380	1.325	1.285	1.184	1.0584

Table 1.2.1. Values of τ_p

1.2.7 Algorithm Efficiency

Next, we turn to the task of estimating the work needed to solve an optimization problem to prescribed precision.

Thus, suppose that in solving a particular problem, an optimization algorithm produces a sequence $\{x_i\}_{i=0}^{\infty}$ which converges to a solution $\hat{x}$ at least Q-linearly, with $i_0 = 0$, i.e., there exist a $c \in [0, 1)$ such that for all $i \in \mathbb{N}$,

$$\|x_{i+1} - \hat{x}\| \le c \|x_i - \hat{x}\| . \tag{44a}$$

Hence for any $i \in \mathbb{N}$,

$$\|x_i - \hat{x}\| \le c^i \|x_0 - \hat{x}\| . \tag{44b}$$

To reduce the error from an initial value of $\|x_0 - \hat{x}\|$ to $\alpha\|x_0 - \hat{x}\|$, for a given $\alpha \in (0, 1)$, requires a certain number of iterations. An estimate of this number is given by the smallest solution $i^* \in \mathbb{N}$ of the inequality

$$c^i \le \alpha . \tag{44c}$$

Taking logarithms of both sides, (44c) yields (since both c and $\alpha \in (0, 1)$) that i^* is the smallest integer such that

$$i^* \ge \frac{\ln \alpha}{\ln c} . \tag{44d}$$

Assuming that it takes w units of work (say cpu seconds) to construct x_i and that $\ln\alpha/\ln c$ is much greater than 1, so that $i^* \approx \ln\alpha/\ln c$, an estimate for the total work performed in reducing the error by the factor α is given by

$$W \approx \frac{w}{\ln c} \ln \alpha . \tag{44e}$$

The factor

$$\eta \triangleq -\frac{\ln c}{w} > 0 \tag{44f}$$

is called the *efficiency* of the process that constructed the sequence $\{x_i\}_{i=0}^{\infty}$.

Now consider the case where $x_i \to \hat{x}$, as $i \to \infty$, Q-superlinearly, with Q-rate at least $r > 1$, with $i_0 = 0$, i.e., there exist a $\kappa < \infty$ such that for all $i \geq 0$,

$$\|x_{i+1} - \hat{x}\| \leq \kappa \|x_i - \hat{x}\|^r . \tag{44g}$$

For all $i \in \mathbb{N}$, let the normalized error be defined by $\varepsilon_i \triangleq \kappa^{[1/(r-1)]}\|x_i - \hat{x}\|$. Then it follows from the proof of Theorem 1.2.41 that, after k iterations of the process that constructed the sequence $\{x_i\}_{i=0}^{\infty}$, the normalized error ε_k satisfies the inequality

$$\varepsilon_k \leq \varepsilon_0^{r^k} . \tag{44h}$$

Now suppose that the initial normalized error $\varepsilon_0 < 1$, and suppose that we want to reduce the normalized error to the value ε^m, where $m > 1$ is an integer. Then it is clear from (44h) that if $k \geq \ln m / \ln r$, then $\varepsilon_k \leq \varepsilon_0^m$. Hence, given that the work involved in constructing each x_i is w, the total work expended in producing the required reduction in normalized error is given by

$$W \approx \frac{w}{\ln r} \ln m . \tag{44i}$$

Hence it is natural to define the efficiency of the process constructing the sequence $\{x_i\}_{i=0}^{\infty}$ by

$$\eta \triangleq \frac{\ln r}{w} . \tag{44j}$$

1.2.8 Notes

The idea of analyzing optimization algorithms in the abstract setting of point-to-set algorithm maps seems to have been first proposed by Zangwill in [Zan.69, Zan.69a] and depended on the outer semicontinuity of the point-to-set algorithm map. Extensions of Zangwill's work appear in [Hua.75, Hua.79]. Other examples of general convergence theorems for optimization algorithms can be found in [Pol.69, Pol.71, Mey.76, Mey.77]. In [TPK.79] we find a comparison of various convergence theorems in the literature.

Viewing optimization algorithms as discrete-time dynamical systems, Polak developed a series of convergence theorems that were inspired by Lyapunov stability theory (see [Pol.69, Pol.71]) and are considerably more general than Zangwill's result in their applicability because they did not require that the point-to-set algorithm map be

outer semicontinuous. In [Pol.71, Pol.71a], Polak also seems to be the first to make a distinction between "conceptual" and implementable algorithms, presenting implementation schemes. A number of the results in [Pol.71] resulted from the collaboration with Klessig [KlP.70, KlP.72, KlP.73].

The Armijo step-size rule (20a) is not the only one currently being used. Two other rules are encountered in the literature. The first, used only occasionally, is a "two-line" rule proposed by Goldstein [Gol.67]. This rule uses two parameters $0 < \alpha < \beta < 1$ and replaces (20a) by

$$\lambda_i \in \{\lambda \mid \lambda\beta\langle \nabla f(x_i), h_i \rangle \leq f(x_i + \lambda h_i) - f(x_i) \leq \lambda\alpha\langle \nabla f(x_i), h_i \rangle\} . \tag{45a}$$

Much more common, particularly in quasi-Newton type methods, is the Wolfe rule [Wol.69, Wol.71], which also uses two parameters $0 < \alpha < \beta < 1$ and replaces (20a) by

$$\lambda_i \in \{\lambda \mid f(x_i + \lambda h_i) - f(x_i) \leq \lambda\alpha\langle \nabla f(x_i), h_i \rangle ,$$

$$\langle \nabla f(x_i + \lambda h_i), h_i \rangle \geq \beta\langle \nabla f(x_i), h_i \rangle\} . \tag{45b}$$

It is possible to construct convergence theorems for trust region methods without assuming that the approximations to the Hessian H_i are bounded. Thus Powell [Pow.77] shows that, if the function $f(\cdot)$ is bounded from below, its gradient $\nabla f(\cdot)$ is uniformly continuous, and the H_i satisfy

$$\|H_i\| \leq d_1 + d_2 i , \ i \in \mathbb{N} , \tag{45c}$$

with $d_i < \infty$, $i = 1, 2$, then for any sequence $\{x_i\}_{i=0}^{\infty}$, constructed by Algorithm Model 1.2.26, $\underline{\lim}\, \|\nabla f(x_i)\| = 0$. Note that this statement does not imply that every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies $\nabla f(\hat{x}) = 0$. Rather (assuming that $\{x_i\}_{i=0}^{\infty}$ has accumulation points), it implies that at least one accumulation point $\hat{x}$ satisfies $\nabla f(\hat{x}) = 0$.

With the exception of Theorem 1.2.19, all the convergence theorems that we have presented are distinguished by the fact that they require that algorithms construct sequences on which the cost decreases monotonically. There is a small, but important class of algorithms, which globalize some version of Newton's method, which do not have this property and use nonmonotone line searches to speed up Newton's method or to avoid the Maratos effect. For examples, see [CLP.82, GLL.86, PoY.86, ZhT.93]. Theorem 1.2.19 is an abstraction of the following generalization, by Grippo, Lampariello, and Lucidi [GLL.86], of the Polak-Sargent-Sebastian Theorem 1.2.24b.

Theorem 1.2.44. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ in (1a) is Lipschitz continuously differentiable on bounded sets. Let $\alpha, \beta \in (0, 1)$ and $k^* \in \mathbb{Z}$, $M \in \mathbb{N}$, be given, and suppose that $\{x_i\}_{i=0}^{\infty}$ is a sequence defined by*

$$x_{i+1} = x_i + \lambda_i h_i , \ i \in \mathbb{N} , \ h_i \neq 0 , \tag{46a}$$

with

$$\lambda_i = \beta^{k_i} = \max_{\substack{k \in \mathbb{Z} \\ k \geq k^*}} \{\beta^k \mid \max_{0 \leq j \leq m(i)} f(x_{i-j}) + \alpha\beta^k \langle \nabla f(x_i), h_i \rangle\} , \tag{46b}$$

where $m(0) = 0$ and $0 \leq m(k) \leq \min\{m(k-1) + 1, M\}$, $k \geq 1$.

If the level set $L_{f(x_0)}(f)$ is compact and there exist $c_1, c_2 \in (0, \infty)$ such that for all $i \in \mathbb{N}$,

$$\langle \nabla f(x_i), h_i \rangle \leq -c_2 \|\nabla f(x_i)\|^2 , \tag{46c}$$

$$\|d_i\| \leq c_2 \|\nabla f(x_i)\| , \tag{46d}$$

then (a) the sequence $\{x_i\}_{i=0}^{\infty}$ *is contained in* $L_{f(x_0)}(f)$*; (b) every accumulation point* $\hat{x}$ *of this sequence satisfies* $\nabla f(\hat{x}) = 0$*; (c) no accumulation point of* $\{x_i\}_{i=0}^{\infty}$ *is a local maximizer; and (d) if the number of stationary points in* $L_{f(x_0)}(f)$ *is finite, then* $\{x_i\}_{i=0}^{\infty}$ *converges.* □

In recent years a considerable literature has grown dealing with the notion of problem complexity and the efficiency of optimization algorithms. We recommend the book by Nemirovsky and Yudin [NeY.83] for in-depth study of this topic.

1.3 Gradient Methods

We devote this section to three gradient methods: the steepest descent algorithm, the Armijo gradient algorithm for unconstrained optimization, and an extension of the Armijo gradient algorithm to a projected gradient type algorithm for a simple class of constrained problems. We will deal with their convergence, implementation, rate of convergence, and efficiency.

Gradient methods solve problems of the form

$$\min_{x \in \mathbb{R}^n} f(x) , \tag{1}$$

with $f : \mathbb{R}^n \to \mathbb{R}$ continuously differentiable.

1.3.1 Method of Steepest Descent

The earliest gradient method known is the steepest descent algorithm, invented by Cauchy more than 100 years ago [Cau.847]. It corresponds to Algorithm Model 1.2.20 with the search direction h_i defined by $h_i = -\nabla f(x_i)$.

Steepest Descent Algorithm 1.3.1.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute the *search direction*

$$h_i = -\nabla f(x_i) . \tag{2a}$$

Stop if $\nabla f(x_i) = 0$.

Step 2. Compute the *step-size*

$$\lambda_i \in \lambda(x_i) \triangleq \arg\min_{\lambda \geq 0} f(x_i + \lambda h_i) . \tag{2b}$$

Step 3. Set

$$x_{i+1} = x_i + \lambda_i h_i , \tag{2c}$$

replace i by $i + 1$, and go to Step 1.

In (2b) there is the implicit assumption that $\lambda(x_i)$ exists. This rules out cases, such as $f(x_i + \lambda h_i) = e^{-\lambda}$ or $f(x_i + \lambda h_i) = -\lambda$.

Since in the case of the above algorithm we may set $\kappa(x_i) \equiv 1$ in (1.2.18), the following result follows directly from the Wolfe Theorem 1.2.21:

Theorem 1.3.2. *If* $\{x_i\}_{i=0}^{\infty}$ *is an infinite sequence constructed by Algorithm 1.3.1 in solving (1), then every accumulation point* $\hat{x}$ *of* $\{x_i\}_{i=0}^{\infty}$ *satisfies* $\nabla f(\hat{x}) = 0$. □

The step-size rule (2b) implies that

$$(d/d\lambda) f(x_i + \lambda_i h_i) = \langle \nabla f(x_i + \lambda_i h_i), h_i \rangle = 0 ,$$

i.e., that $\nabla f(x_{i+1})$ is perpendicular to $\nabla f(x_i)$. Hence the method of steepest descent constructs trajectories as shown in Fig. 1.3.1.

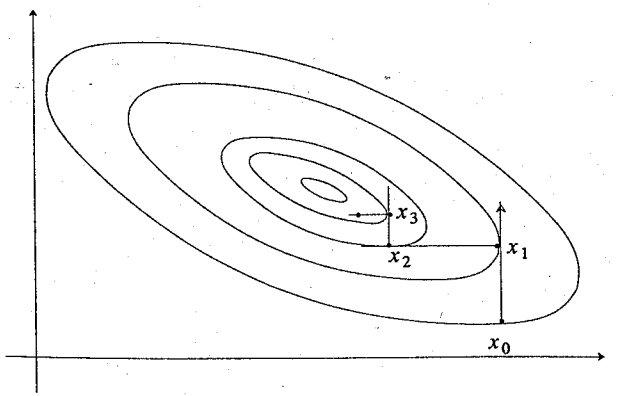


Fig. 1.3.1. Steepest descent trajectories.

1.3.2 Armijo Gradient Method

Clearly, the method of steepest descent contains a nonimplementable step-size rule. Although it is possible to construct implementations of this step-size rule, the implied requirement of gradually better and better precision in its approximation makes implementable steepest descent algorithms noncompetitive with the more recent Armijo gradient method [Arm.66]. The Armijo gradient method is patterned on Algorithm Model 1.2.23.

Armijo Gradient Algorithm 1.3.3.

Parameters: $\alpha, \beta \in (0,1), k^* \in \mathbb{Z}$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute *the search direction*

$$h_i = -\nabla f(x_i). \tag{3a}$$

Stop if $\nabla f(x_i) = 0$.

Step 2. Use Subprocedure 1.2.23a (which requires k^*) to compute the *step-size* $\lambda_i = \beta^{k_i}$, where $k_i \in \mathbb{Z}$ is such that

$$f(x_i + \beta^{k_i} h_i) - f(x_i) \leq -\beta^{k_i} \alpha \|\nabla f(x_i)\|^2 \tag{3b}$$

and

$$f(x_i + \beta^{k_i - 1} h_i) - f(x_i) > -\beta^{k_i - 1} \alpha \|\nabla f(x_i)\|^2. \tag{3c}$$

Step 3. Set $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i+1$, and go to Step 1.

Theorem 1.3.4. *Suppose that the function $f(\cdot)$ in (1) is continuously differentiable.*

(a) If $x_i \in \mathbb{R}^n$ is such that $\nabla f(x_i) \neq 0$, then the step-size λ_i, defined by (3b) and (3c), is constructed using a finite number of function evaluations.

(b) If $\{x_i\}_{i=0}^{\infty}$ is an infinite sequence constructed by Algorithm 1.3.3 in solving (1), then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies $\nabla f(\hat{x}) = 0$.

(c) If $x^* \neq x^{**}$ are two accumulation points of a sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 1.3.3, then $f(x^*) = f(x^{**})$.

(d) If the set $\{x \in \mathbb{R}^n \mid \nabla f(x) = 0\}$ contains only a finite number *of points, then any bounded sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 1.3.3* must *converge to a point $\hat{x}$ such that $\nabla f(\hat{x}) = 0$.*

(e) If $\{x_i\}_{i=0}^{\infty}$ is an infinite sequence constructed by Algorithm 1.3.3 in solving (1) and it has an accumulation point $\hat{x}$ such that $\nabla f(\hat{x}) = 0$ and, for some $m > 0$, $\langle h, f_{xx}(\hat{x})h \rangle \geq m\|h\|^2$ for all $h \in \mathbb{R}^n$, then $x_i \to \hat{x}$, as $i \to \infty$.

Proof. By inspection, *(a)* and *(b)* follow directly from Theorem 1.2.24a.

Part *(c)* follows from the fact that, since $f(\cdot)$ is continuous, both $f(x^*)$ and $f(x^{**})$ must be accumulation points of the sequence $\{f(x_i)\}_{i=0}^{\infty}$, and hence, since the $\{f(x_i)\}_{i=0}^{\infty}$ is monotone decreasing, the fact that $f(x^*) = f(x^{**})$ follows from the Monotone Sequences Proposition 5.1.16.

Part *(d)* follows from Proposition 1.2.38, and part *(e)* follows from Proposition 1.2.38 and the fact that because $f_{xx}(\hat{x})$ is strictly positive-definite, there must exist a $\rho > 0$ such that for all $x \in B(\hat{x}, \rho), x \neq \hat{x}, \nabla f(\hat{x}) \neq 0$. □

Next we will establish a bound on the rate of convergence of Algorithm 1.3.3. under the following assumption which guarantees strict convexity of $f(\cdot)$ in (1).

Assumption 1.3.5. *The function $f(\cdot)$ in (1) is twice continuously differentiable and there exist constants $0 < m \leq M < \infty$ such that for all $x, y \in \mathbb{R}^n$,*

$$m\|y\|^2 \leq \langle y, f_{xx}(x)y \rangle \leq M\|y\|^2. \tag{4}$$

□

Assumption 1.3.5 is stronger than absolutely needed. It is enough to assume that, for every bounded set S, there exist constants $0 < m \leq M < \infty$ such that for all $x \in S$, $y \in \mathbb{R}^n$, (4) is satisfied. However, all rate of convergence proofs become more complex because of the need to keep track of a bounded set containing points of interest.

When Assumption 1.3.5 is satisfied, it follows from Theorem 5.2.14 that $f(\cdot)$ is strictly convex and from Corollary 1.1.12 that it has a unique global minimizer $\hat{x}$.

Lemma 1.3.6. *Consider the function $f(\cdot)$ in (1) and suppose that Assumption 1.3.5 holds. If $\hat{x}$ is the unique minimizer of $f(\cdot)$, then for any x satisfying Assumption 1.3.5,*

(a)
$$f(\hat{x}) - f(x) \geq -\frac{1}{2m}\|\nabla f(x)\|^2, \tag{5a}$$

and

(b)
$$\frac{m}{2}\|x - \hat{x}\|^2 \leq f(x) - f(\hat{x}) \leq \frac{M}{2}\|x - \hat{x}\|^2. \tag{5b}$$

Proof. *(a)* Referring to Proposition 1.1.6, with $\theta(\cdot)$ defined by (1.1.8f), we see that (5a) holds.

(b) Next, using the fact that $\nabla f(\hat{x}) = 0$, (4), and the fact that by Theorem 5.1.28*(b)*,

$$f(x) - f(\hat{x}) = \langle \nabla f(\hat{x}), (x - \hat{x}) \rangle + \tfrac{1}{2}\langle x - \hat{x}, f_{xx}(x + s(x - \hat{x}))(x - \hat{x}) \rangle, \quad (6)$$

for some $s \in [0, 1]$, we obtain (5b). □

As we will show, it is easy to obtain a loose bound on the rate of convergence of the Armijo Gradient Method.

Theorem 1.3.7. *Suppose that Assumption 1.3.5 holds. If $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 1.3.3 in solving (1), then*

(a) $x_i \to \hat{x}$, *as* $i \to \infty$, *with* $\hat{x}$ *the unique minimizer of* $f(\cdot)$, *and*

(b) *for all* $i \in \mathbb{N}$,

$$f(x_{i+1}) - f(\hat{x}) \le c[f(x_i) - f(\hat{x})], \quad (7a)$$

$$\|x_i - \hat{x}\| \le \left\{ \frac{2}{m}[f(x_0) - f(\hat{x})] \right\}^{1/2} (c^{1/2})^i, \quad (7b)$$

where the linear rate of convergence constant c is given by

$$c = 1 - \frac{4m\beta\alpha(1-\alpha)}{M}. \quad (7c)$$

Proof. *(a)* By Proposition 5.2.15, the level sets of $f(\cdot)$ are compact and by Corollary 1.1.12, problem (1) has a unique global minimizer $\hat{x}$. Hence every sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 1.3.3 must have accumulation points, and since by Theorem 1.3.4, every accumulation point x^* of this sequence must satisfy $\nabla f(x^*) = 0$, it follows that $\hat{x}$ is the only accumulation point of this sequence, i.e., $x_i \to \hat{x}$, as $i \to \infty$.

(b) Next we turn to rate of convergence. Expanding the formula used in the Armijo step-size calculation (3b) and (3c), to second order, we find that for some $s \in (0, 1)$ (see Fig. 1.3.2),

$$f(x_i - \lambda \nabla f(x_i)) - f(x_i) + \lambda\alpha\|\nabla f(x_i)\|^2$$

$$= -\lambda\left[(1-\alpha)\|\nabla f(x_i)\|^2 - \tfrac{1}{2}\lambda \langle \nabla f(x_i), f_{xx}(x_i - s\nabla f(x_i))\nabla f(x_i) \rangle \right]$$

$$\le -\lambda\|\nabla f(x_i)\|^2 [(1-\alpha) - \tfrac{1}{2}\lambda M]. \quad (8a)$$

The right-hand side of (8a) is negative for all $\lambda \in [0, 2(1-\alpha)/M]$. Hence it follows that

$$\beta^{k_i} \ge \frac{2\beta}{M}(1-\alpha), \quad (8b)$$

because we must have that $k_i \le \hat{k}$, where $\hat{k}$ is such that $\beta^{\hat{k}} \le 2(1-\alpha)/M$ and $\beta^{\hat{k}-1} > 2(1-\alpha)/M$. Consequently we deduce from (3b) that

$$f(x_{i+1}) - f(x_i) \le -\frac{2\beta\alpha(1-\alpha)}{M}\|\nabla f(x_i)\|^2, \quad (8c)$$

for all $i \in \mathbb{N}$. Substituting for $\|\nabla f(x_i)\|^2$ in (5a) (where we replace x by x_i) from (8c), we get that

$$f(x_{i+1}) - f(x_i) \le 4\beta\alpha(1-\alpha)\frac{m}{M}[f(\hat{x}) - f(x_i)]. \quad (8d)$$

Subtracting $f(\hat{x})$ from both sides of (8d) and rearranging terms, we conclude that (7a) holds for all $i \in \mathbb{N}$, with $c = 1 - 4m\beta\alpha(1-\alpha)/M$.

Next, it follows by recursion on (7a) that, for all $i \ge 0$,

$$0 \le f(x_i) - f(\hat{x}) \le [f(x_0) - f(\hat{x})]c^i. \quad (8e)$$

Hence, using (5b) we obtain (7b), which completes our proof. □

Exercise 1.3.8. Show that (7c) suggests that $\alpha = \frac{1}{2}$ is the best choice for the parameter α in the Armijo Gradient Method 1.3.3. Setting $\alpha = \frac{1}{2}$ and $\beta = 1$, deduce from (7c) that, under Assumption 1.3.5, Algorithm 1.3.1 satisfies (7a,b) with $c = 1 - m/M$. (Note that, as $\beta \to 1$, the number of tries needed to compute the Armijo step-size is likely to go to infinity). □

Exercise 1.3.9. Referring to Fig. 1.3.2, we see that when Assumption 1.3.5 is satisfied and $\alpha = \frac{1}{2}$, then the Armijo step length satisfies the inequality $\beta/M \le \beta^{k_i} \le 1/m$ and hence that the work per iteration is bounded. Obtain an estimate of the efficiency of the Armijo Gradient Method 1.3.3, using the number of function evaluations per iteration as a measure of work. □

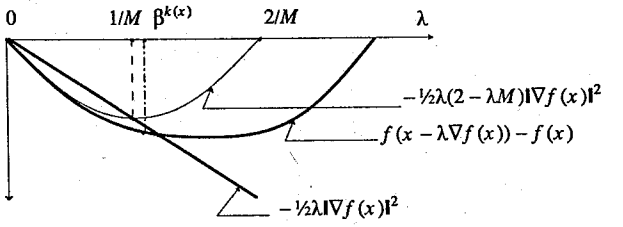


Fig. 1.3.2. Bounds on the Armijo step-size $\beta^{k(x)}$ for $\alpha = \frac{1}{2}$.

By keeping more terms in the expansions used to obtain the inequalities (5a), (8c), and (8d), it is possible to obtain a sharper bound on the rate of convergence of both the method of steepest descent and the Armijo gradient method. In addition, one needs to use the following Kantorovich inequality [KaA.59].

Kantorovich Inequality 1.3.10. *Let H be any positive-definite, symmetric, $n \times n$ matrix, and let $0 < m \leq M$ be such that*

$$m\|y\|^2 \leq \langle y, Hy \rangle \leq M\|y\|^2, \quad \forall\, y \in \mathbb{R}^n . \tag{9a}$$

Then

$$\frac{\|y\|^4}{\langle y, Hy \rangle \langle y, H^{-1}y \rangle} \geq \frac{4mM}{(m+M)^2}, \quad \forall\, y \in \mathbb{R}^n . \tag{9b}$$

□

Theorem 1.3.11. *Suppose that Assumption 1.3.5 holds and, in addition, that $H(\cdot) \triangleq f_{xx}(\cdot)$ is Lipschitz continuous with global Lipschitz constant L.*

(a) If $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 1.3.1 in solving (1), then

(i) $x_i \to \hat{x}$, *as* $i \to \infty$, *with* $\hat{x}$ *the unique minimizer of* $f(\cdot)$, *and*

(ii) for all $i \in \mathbb{N}$,

$$f(x_{i+1}) - f(\hat{x}) \leq c\,[f(x_i) - f(\hat{x})], \tag{10a}$$

where the linear rate of convergence constant c is given by

$$c = 1 - \frac{4mM}{(m+M)^2} = \left[\frac{M-m}{M+m}\right]^2 . \tag{10b}$$

(b) If $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 1.3.3 in solving (1), then

(i) $x_i \to \hat{x}$, *as* $i \to \infty$, *with* $\hat{x}$ *the unique minimizer of* $f(\cdot)$, *and*

(ii) for every $\varepsilon \in (0,1)$, *there exists an* i_ε *such that for all* $i \geq i_\varepsilon$,

$$f(x_{i+1}) - f(\hat{x}) \leq c_\varepsilon [f(x_i) - f(\hat{x})], \tag{10c}$$

where the linear rate of convergence constant c is given by

$$c_\varepsilon = 1 - 4\varepsilon\beta\alpha(1-\alpha)\frac{4mM}{(m+M)^2} , \tag{10d}$$

i.e., the asymptotic *linear rate of convergence constant for the Armijo method (obtained by letting $\varepsilon \to 1$) is $c = 1 - 16\beta\alpha(1-\alpha)mM/(m+M)^2$. (Note that when $\alpha = \frac{1}{2}$ and $\beta = \varepsilon = 1$, the expressions (10b) and (10d) become identical).*

Proof. Part *(a)* is only of academic interest. Therefore we will prove only the somewhat more difficult part *(b)*. Since *(i)* was already proved in Theorem 1.3.7, we need to prove only *(ii)*. Note that the form of the Kantorovich inequality forces us to develop expressions involving both $H(x)$ and $H(x)^{-1}$. To simplify our expressions, for any $x \in \mathbb{R}^n$, we define $g(x) \triangleq \nabla f(x)$,

$$H_1(x) \triangleq \int_0^1 H(\hat{x} + s(x - \hat{x}))ds , \tag{11a}$$

$$H_2(x) \triangleq 2\int_0^1 (1-s)H(\hat{x} + s(x - \hat{x}))\,ds , \tag{11b}$$

and

$$\Delta(x) \triangleq H_2(x)H_1(x)^{-1} - I . \tag{11c}$$

Then we see that both $H_1(x) \to H(\hat{x})$, $H_2(x) \to H(\hat{x})$, as $x \to \hat{x}$, and $\Delta(x) \to 0$, as $x \to \hat{x}$. Furthermore since for any $x \in \mathbb{R}^n$, $g(x) = H_1(x)(x - \hat{x})$, because $g(\hat{x}) = 0$, it follows that $x - \hat{x} = H_1(x)^{-1}g(x)$.

First we refine the bound given by (5a). Clearly, by the second order expansion formula (5.1.18b), for any $x \in \mathbb{R}^n$,

$$\begin{aligned} f(\hat{x}) - f(x) &= -\langle g(x), x - \hat{x} \rangle + \tfrac{1}{2}\langle (x - \hat{x}), H_2(x)(x - \hat{x}) \rangle \\ &= -\langle g(x), H_1(x)^{-1}g(x) \rangle \\ &\quad + \tfrac{1}{2}\langle H_1(x)^{-1}g(x), H_2(x)H_1(x)^{-1}g(x) \rangle \\ &= -\tfrac{1}{2}\langle g(x), H_1(x)^{-1}g(x) \rangle [1 - D(x)] , \end{aligned} \tag{11d}$$

where

$$D(x) \triangleq \frac{\langle H_1(x)^{-1}g(x), \Delta(x)g(x) \rangle}{\langle g(x), H_1(x)^{-1}g(x) \rangle} . \tag{11e}$$

Hence $D(x) \to 0$, as $x \to \hat{x}$.

Next we will refine (8c). Thus, again using (5.1.18b) and the fact that $H(\cdot)$ is Lipschitz continuous, we find (c.f. (8a)) that for any $i \in \mathbb{N}$,

$$\begin{aligned} & f(x_i - \lambda g(x_i)) - f(x_i) + \lambda\alpha\|g(x_i)\|^2 \\ &= -\lambda\left[(1-\alpha)\|g(x_i)\|^2 - \tfrac{1}{2}\lambda\langle g(x_i), H(x_i - sg(x_i))g(x_i) \rangle\right] \\ &\leq -\lambda(1-\alpha)\|g(x_i)\|^2 + \frac{\lambda^2}{2}\langle g(x_i), H(x_i)g(x_i) \rangle + \frac{\lambda^2 L}{6}\|g(x_i)\|^3 \end{aligned}$$

$$\triangleq r_i(\lambda). \tag{11f}$$

Now, the sum of the first two terms in $r_i(\lambda)$ is negative for all $\lambda \in [0, \bar{\lambda}_i]$, where $\bar{\lambda}_i \triangleq 2(1-\alpha)\|g(x_i)\|^2 / \langle g(x_i), H(x_i)g(x_i)\rangle$. Hence, for any $\varepsilon \in (0,1)$,

$$r_i(\varepsilon\bar{\lambda}_i) = -\varepsilon\bar{\lambda}_i \{(1-\varepsilon)(1-\alpha) - \tfrac{1}{2}L(\varepsilon\bar{\lambda}_i)^2\|g(x_i)\|\}. \tag{11g}$$

Hence, since $\|g(x_i)\| \to 0$,, as $i \to \infty$, and $\bar{\lambda}_i$ is bounded, for any $\varepsilon \in (0,1)$, there exists an i_ε such that $r_i(\varepsilon\bar{\lambda}_i) \leq 0$ for all $i \geq i_\varepsilon$. It follows that, for all $i \geq i_\varepsilon$, $\lambda_i \geq \beta\varepsilon\bar{\lambda}_i$, and hence that for all $i \geq i_\varepsilon$,

$$f(x_{i+1}) - f(x_i) \leq -2\varepsilon\beta\alpha(1-\alpha)\frac{\|g(x_i)\|^4}{\langle g(x_i), H(x_i)g(x_i)\rangle}. \tag{11h}$$

Without loss of generality, we can assume that $1 > 1 - D(x_i) > 0$ for all $i \geq i_\varepsilon$. Hence, combining (11h) with (11d) we deduce that, for all $i \geq i_\varepsilon$,

$$\frac{f(x_{i+1}) - f(x_i)}{f(\hat{x}) - f(x_i)} \geq 4\varepsilon\beta\alpha(1-\alpha)\frac{\|g(x_i)\|^4}{\langle g(x_i), H^{-1}(x_i)g(x_i)\rangle \langle g(x_i), H(x_i)g(x_i)\rangle}$$

$$\geq 4\varepsilon\beta\alpha(1-\alpha)\frac{4mM}{(m+M)^2} = 1 - c_\varepsilon, \tag{11i}$$

which implies that, for all $i \geq i_\varepsilon$,

$$f(x_{i+1}) - f(x_i) \leq -(1-c_\varepsilon)[f(x_i) - f(\hat{x})]. \tag{11j}$$

Adding $f(x_i) - f(\hat{x})$ to both sides of (11j), we obtain the desired result. □

Exercise 1.3.12. Prove Theorem 1.3.11 without the additional assumption that $H(\cdot)$ is Lipschitz continuous. □

In structural optimization, as well as in other applications where one uses complex numerical integration code for function evaluations, gradients must often be computed using finite differences. Therefore we present a version of the Armijo gradient method based on Algorithm Model 1.2.30, especially tailored for such applications. In the algorithm below, the search direction h_i is a finite difference approximation to $-\nabla f(x_i)$, with the parameter ε controlling the precision of this approximation.

Discrete Armijo Gradient Algorithm 1.3.13.

Parameters: $\alpha, \beta \in (0,1), \gamma \in (0,\infty), k^*, k_0 \in \mathbb{Z}$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0, \varepsilon = \beta^{k_0}$.

Step 1. Compute *the search direction* h_i with components determined by

$$h_i^j = -\frac{1}{\varepsilon}[f(x_i + \varepsilon e_j) - f(x_i)],\ j = 1,2,\ldots,n, \tag{12a}$$

where e_j is the jth column of the $n \times n$ unit matrix.

Step 2. Compute

$$\Delta(x_i; h_i) = \frac{1}{\varepsilon}[f(x_i + \varepsilon h_i) - f(x_i)]. \tag{12b}$$

Step 3. If $\Delta(x_i; h_i) > 0$, replace ε by $\beta\varepsilon$, and go to Step 1.

Else, use Subprocedure 1.2.23a (which requires k^*) to compute the *step-size* $\lambda_i = \beta^{k_i}$, where $k_i \in \mathbb{Z}$ is such that

$$f(x_i + \beta^{k_i}h_i) - f(x_i) \leq \beta^{k_i}\alpha\Delta(x_i; h_i) \tag{12c}$$

and

$$f(x_i + \beta^{k_i-1}h_i) - f(x_i) > \beta^{k_i-1}\alpha\Delta(x_i; h_i). \tag{12d}$$

Step 4. If $f(x_i + \lambda_i h_i) - f(x_i) > -\gamma\varepsilon$, replace ε by $\beta\varepsilon$, and go to Step 1.

Step 5. Set $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i+1$, and go to step 1.

Exercise 1.3.14. Prove that the following assertions are true:

(a) If Algorithm 1.3.13 jams at x_i, cycling indefinitely in the loop defined by Steps 1 - 3 or in the loop defined by Steps 1 - 4, then $\nabla f(x_i) = 0$.

(b) If $\{x_i\}_{i=0}^{\infty}$ is an infinite sequence constructed by Algorithm 1.3.13 in solving (1), then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies $\nabla f(\hat{x}) = 0$.

Hint: Make use of Theorem 1.2.31. □

Exercise 1.3.15. Let $\omega \in (1,2)$. Suppose that Step 3 in Algorithm 1.3.13 is replaced by

Step 3. If $\|h_i\|^\omega < \varepsilon$ or $\Delta(x_i; h_i) > 0$, replace ε by $\beta\varepsilon$, and go to Step 1.

Else, use Subprocedure 1.2.23a (which requires k^*) to compute the *step-size* $\lambda_i = \beta^{k_i}$, where $k_i \in \mathbb{N}$ is such that

$$f(x_i + \beta^{k_i} h_i) - f(x_i) \le \beta^{k_i} \alpha \Delta(x_i ; h_i) \tag{13a}$$

and

$$f(x_i + \beta^{k_i - 1} h_i) - f(x_i) > \beta^{k_i - 1} \alpha \Delta(x_i ; h_i) . \tag{13b}$$

Show that if Assumption 1.3.5 is satisfied and $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by the modified Algorithm 1.3.13, with $\alpha = \frac{1}{2}$, then for every $\rho > 0$, there exists an $i_\rho \in \mathbb{N}$, and $b < \infty$ such that, for all $i \ge i_\rho$,

$$f(x_{i+1}) - f(\hat{x}) \le (c + \rho)[f(x_i) - f(\hat{x})] , \tag{13c}$$

$$\|x_i - \hat{x}\| \le b\, [(c + \rho)^{1/2}]^i , \tag{13d}$$

where $c = 1 - \beta m / M$. □

1.3.3 Projected Gradient Method

It is very easy to extend the Armijo gradient method to a class of simple constrained optimization problems of the form

$$\min_{x \in X} f(x) , \tag{14a}$$

with $f : \mathbb{R}^n \to \mathbb{R}$ continuously differentiable and $X \subset \mathbb{R}^n$ convex and compact. The search direction subprocedure in the method which we will describe is implementable only when the set X is defined by affine inequalities such as

$$X \triangleq \{ x \in \mathbb{R}^n \mid Ax \le b \} , \tag{14b}$$

where A is an $m \times n$ matrix and $b \in \mathbb{R}^m$ and, perhaps, in a few other special cases.

We will suppose that $\delta > 0$ is given. Let $\theta_X : X \to \mathbb{R}$ and $\eta_X : X \to X$ be defined by

$$\theta_X(x) \triangleq \min_{x' \in X} \langle \nabla f(x), x' - x \rangle + \tfrac{1}{2}\delta \|x' - x\|^2 , \tag{15a}$$

and

$$\eta_X(x) \triangleq \arg\min_{x' \in X} \langle \nabla f(x), x' - x \rangle + \tfrac{1}{2}\delta \|x' - x\|^2 . \tag{15b}$$

From a trivial extension of Proposition 1.1.6, it follows that both $\theta_X(\cdot)$ and $\eta_X(\cdot)$ are continuous, that $\theta_X(\hat{x}) = 0$ if and only if (1.1.5) holds and hence, if $\hat{x}$ is a local minimizer for (14a), then $\theta_X(\hat{x}) = 0$.

Now we can state an obvious extension of the Armijo Gradient Algorithm 1.3.3. Note that an alternative formula for $\theta_X(x)$ is given by

$$\theta_X(x) = \min_{x' \in X} \frac{\delta}{2} \|x' - [x - \frac{1}{\delta}\nabla f(x)]\|^2 - \frac{1}{2\delta}\|\nabla f(x)\|^2 . \tag{15c}$$

Thus, the computation of $\theta_X(x)$ yields a vector x' that is the projection of the vector $x - \frac{1}{\delta}\nabla f(x)$ onto the set X. When $\delta = 1$, the algorithm, below, is a standard projected gradient algorithm. When $\delta \ne 1$, it becomes a scaled projected gradient algorithm. In the algorithm below, the step-size is kept in the interval $[0, 1]$, so as never to violate the problem constraints.

Projected Gradient Algorithm 1.3.16.

Parameters : $\alpha \in (0, 1]$, $\beta \in (0, 1)$, $\delta > 0$ (for $\theta_X(\cdot)$).

Data. $x_0 \in X$.

Step 0. Set $i = 0$.

Step 1. Compute the *optimality function* $\theta_X(x_i)$ and the corresponding minimizer $\eta_X(x_i)$, using (15a) and (15b), and *the search direction* h_i given by

$$h_i = \eta_X(x_i) - x_i . \tag{16a}$$

Stop if $\theta_X(x_i) = 0$.

Step 2. Compute the *step-size* $\lambda_i = \beta^{k_i}$,

$$\lambda_i = \beta^{k_i} \triangleq \arg\max_{k \in \mathbb{N}} \{ \beta^k \mid f(x_i + \beta^{k_i} h_i) - f(x_i) \le \beta^{k_i} \alpha \theta_X(x_i) \} . \tag{16b}$$

Step 3. Set $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i + 1$, and go to step 1.

Exercise 1.3.17. Suppose that the function $f(\cdot)$ in (14a) is continuously differentiable. Use Algorithm Model 1.2.7 and Theorem 1.2.8 to show that

(a) if $f(\cdot)$ is as in (14a) and $\theta_X(x_i) \ne 0$, then the step-size λ_i, defined by (16b), is constructed using a finite number of function evaluations;

(b) if $\{x_i\}_{i=0}^{\infty}$ is an infinite sequence constructed by Algorithm 1.3.16 in solving (14a), then $x_i \in X$ for all $i \in \mathbb{N}$, and every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies $\theta_X(\hat{x}) = 0$. (To simplify the proof, you may choose to assume that $\nabla f(\cdot)$ is Lipschitz continuous on bounded sets.)

(c) if $x^* \ne x^{**}$ are two accumulation points of a sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 1.3.16, then $f(x^*) = f(x^{**})$;

(d) if the set $\{x \in X \mid \theta_X(x) = 0\}$ contains only a *finite number* of points, then any sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 1.3.16 *must converge* to a point $\hat{x}$ such that $\theta_X(\hat{x}) = 0$. □

Theorem 1.3.18. *Suppose that Assumption 1.3.5 holds and that in (15a,b) δ satisfies $m \le \delta \le M$. If $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 1.3.16 in solving (14a), then*

(a) $x_i \to \hat{x}$, *as* $i \to \infty$, *with* $\hat{x}$ *the unique solution of (14a), and*

(b) for all $i \in \mathbb{N}$,

$$f(x_{i+1}) - f(\hat{x}) \le c\,[f(x_i) - f(\hat{x})]\,, \tag{17a}$$

where the rate of convergence constant

$$c = 1 - \frac{m\beta\alpha}{M}\,. \tag{17b}$$

Proof. *(a)* Since X is convex and compact, it follows from Corollary 1.1.12 that (14a) has a unique global minimizer $\hat{x}$. Since X is compact, every sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 1.3.16 must have accumulation points, and since, by Exercise 1.3.17, every accumulation point x^* of this sequence must satisfy $\nabla f(x^*) = 0$, it follows that $\hat{x}$ is the only accumulation point of this sequence, i.e., $x_i \to \hat{x}$, as $i \to \infty$.

(b) Next we turn to the rate of convergence. Expanding the formula used in the Armijo step-size calculation (16b), to second order, we conclude that, for all $\lambda \in [0, \delta/M]$, because $\delta \le M$, $0 \le \lambda \le 1$, $\lambda M \le \delta$, and

$$\begin{aligned} &f(x_i + \lambda[\eta_X(x_i) - x_i]) - f(x_i) - \lambda\alpha\theta_X(x_i) \\ &\le \lambda\left[\langle \nabla f(x_i), \eta_X(x_i) - x_i\rangle + \tfrac{1}{2}\lambda M \|\eta_X(x_i) - x_i\|^2\right] - \lambda\alpha\theta_X(x_i) \\ &\le \lambda(1-\alpha)\theta_X(x_i) \le 0\,. \end{aligned} \tag{18a}$$

Hence we must have that

$$\beta^{k_i} \ge \frac{\beta\delta}{M}\,, \tag{18b}$$

because we must have that $k_i \le \hat{k}$, where $\hat{k}$ is such that $\beta^{\hat{k}} \le \delta/M$ and $\beta^{\hat{k}-1} > \delta/M$. Consequently, from (16b) and (18b), we deduce that

$$f(x_{i+1}) - f(x_i) \le \frac{\beta\alpha\delta}{M}\theta_X(x_i)\,, \tag{18c}$$

for all $i \in \mathbb{N}$.

Next we mimic the development of (1.1.9c). Thus, for any $x \in X$, we have

$$\begin{aligned} f(\hat{x}) - f(x) &\ge \langle \nabla f(x), (\hat{x} - x)\rangle + \frac{m}{2}\|\hat{x} - x\|^2 \\ &= \frac{\delta}{m}\left\{\langle \nabla f(x), \frac{m}{\delta}(\hat{x} - x)\rangle + \tfrac{1}{2}\delta\left\|\frac{m}{\delta}(\hat{x} - x)\right\|^2\right\}. \end{aligned} \tag{18d}$$

Since $m/\delta \le 1$, it follows that if we define x' by $x' = x + (m/\delta)(\hat{x} - x)$, then $x' \in X$, and $x' - x = (m/\delta)(\hat{x} - x)$. Hence, for any $i \in \mathbb{N}$,

$$f(\hat{x}) - f(x_i) \ge \frac{\delta}{m}\theta_X(x_i)\,. \tag{18e}$$

Eliminating $\theta_X(x_i)$ between (18c) and (18e), we conclude that for all $i \in \mathbb{N}$,

$$f(x_{i+1}) - f(x_i) \le \beta\alpha\frac{m}{M}[f(\hat{x}) - f(x_i)]\,. \tag{18f}$$

Subtracting $f(\hat{x})$ from both sides of (18f) and rearranging terms, we conclude that (17a) holds for all $i \in \mathbb{N}$, with $c = 1 - m\beta\alpha/M$. □

Remark 1.3.19.

(a) Relation (17b) shows that $\alpha = 1$ yields the smallest rate constant c for Algorithm 1.3.16. We recall that the best value of α for Algorithm 1.3.3 was $\alpha = \frac{1}{2}$. This discrepancy would not have occurred if we had used $\theta(x_i) = -\frac{1}{2}\|h_i\|^2$ (as defined in (1.1.8g)) instead of $-\|h_i\|^2$ in the step-size rule (3b) and (3c).

(b) If one picks an arbitrary value of $\delta > 0$, and $\delta \notin [m, M]$, then Theorem 1.3.18 remains valid if we replace m by $m' = \min\{m, \delta\}$ and M by $M' = \max\{M, \delta\}$. However, when Assumption 1.3.5 is valid, one can obtain a value of $\delta \in [m, M]$ using the following observation that is based on second-order expansions. Given any $x_0, x_1 \in \mathbb{R}^n$,

$$m \le \frac{f(x_1) - f(x_0) - \langle \nabla f(x_0), x_1 - x_0\rangle}{\|x_1 - x_0\|^2} \le M\,. \tag{19}$$

Hence, after the first iteration of Algorithm 1.3.16, with an arbitrary value of δ, one obtains a $\delta \in [m, M]$, which can, then, be used for all the subsequent iterations.

(c) When X is as in (14b), the optimality function (15a) and corresponding solution point given by (15b) can be evaluated in a finite number of operations using standard quadratic programming subprocedures. □

1.3.4 Notes

The method of steepest descent was proposed in 1847 by Cauchy [Cau.847] and is one of the oldest optimization algorithms known. The step-size rule in Algorithm 1.3.3 was first presented by Armijo in [Arm.66]. It is more efficient than the earlier "two-line" step-size rule proposed by Goldstein [Gol.65, Gol.67]. In [LeP.66] Levitin and Polyak have presented a projected gradient algorithm which differs from Algorithm 1.3.16 only because they used $\delta = 1$ and a different step-size rule.

Our selection of the step-size ε for the numerical computation of approximations to partial derivatives in Algorithm 1.3.13 was not "optimized" to account for machine errors that arise in the evaluation of the function $f(\cdot)$. For a discussion of finite difference techniques for computing approximations to derivatives, as well as guidance in the selection of a step-size ε that results in good approximations in the presence of machine errors in the evaluation of $f(\cdot)$, see Chapter 8 in [GMW.81], as well as the references therein. It should be obvious that when there are unavoidable errors in the evaluation of function and gradient values, all numerical implementations of optimization algorithms, such as those discussed in this section, become "confused" once the gradients are sufficiently small and hence "jam up", i.e., stop yielding reductions in cost.

Since it only converges linearly, the Armijo Gradient Method 1.3.3 is considered to be slow. However, it is extremely robust and hence a useful tool for "globalizing" only locally converging superliner algorithms, as we will see in the following sections.

1.4 Newton's Method

Newton's algorithm is one of the very oldest and best methods for solving many root finding and optimization problems. In its simplest form it converges only if the initial guess is sufficiently close to a solution. We will examine both the simplest *(local)* version as well as *stabilized* versions which have global convergence properties.

1.4.1 The Local Newton Method

Originally, Newton's algorithm was used for solving systems of equations of the form

$$g(x) = 0\,, \tag{1}$$

where $g : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuously differentiable on bounded sets, under the assumption that the Jacobian matrix $g_x(x) \triangleq \partial g(x)/\partial x$ is nonsingular for all x near a solution $\hat{x}$ of (1).

We will use the induced matrix norm $\|g_x(x)\| \triangleq \max_{|y|=1} |g_x(x)y|$. Since we assume that $g_x(\cdot)$ is Lipschitz continuous on bounded sets, it follows that for every bounded set S, there exists a Lipschitz constant $L_S < \infty$ such that

$$\|g_x(x') - g_x(x'')\| \le L_S |x' - x''|\,, \quad \forall\, x'\,, x'' \in S\,. \tag{2a}$$

The brilliant idea behind Newton's algorithm consists of decomposing the nonlinear equation problem (1), for which one is unable to obtain an explicit solution, into an infinite sequence of linear equations, constituting *successive approximations*, for which one can obtain an explicit solution. Thus, given the current approximation $x_i \in \mathbb{R}^n$ to a solution of (1), one linearizes (1) about x_i and constructs the solvable approximating problem

$$g(x_i) + g_x(x_i)(x - x_i) = 0\,, \tag{2b}$$

whose explicit solution

$$x_{i+1} = x_i - g_x(x_i)^{-1} g(x_i)\,, \tag{2c}$$

is the next, and hopefully better, approximation to a solution of (1). In the special case when $g(\cdot)$ is affine, i.e., of the form $g(x) = Ax + b$, given any x_0, its successor, x_1, is a solution of $g(x) = 0$, i.e., the equation is solved in one iteration. As can be seen from Fig.1.4.1, in general, the successive approximation strategy defined by (2c) may or may not be successful. However, as we will now show, it is always successful if initialized sufficiently close to a solution of (1) and, furthermore, it then constructs sequences which converge Q-quadratically (i.e., with quotient rate 2) to that solution.

Theorem 1.4.1. *Consider Problem (1) and suppose that*

(i) $g : \mathbb{R}^n \to \mathbb{R}^n$ *is Lipschitz continuously differentiable on bounded sets, and*

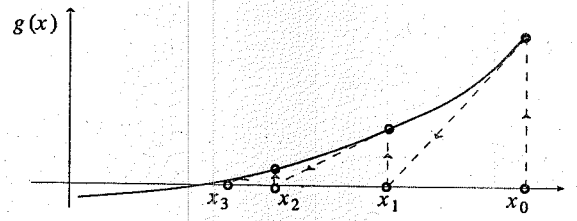
(ii) $\hat{x}$ *is a solution of (1) such that for some* $\rho > 0$, *the Jacobian* $g_x(x)$ *is nonsingular for all* $x \in B(\hat{x}, \rho)$.

Then there exists a $\hat{\rho} \in (0, \rho]$ *such that if* $x_0 \in B(\hat{x}, \hat{\rho})$ *and the sequence* $\{x_i\}_{i=0}^{\infty}$ *is constructed according to the recursion (2c), then* $x_i \to \hat{x}$, *as* $i \to \infty$, Q*-quadratically (with quotient rate 2).*

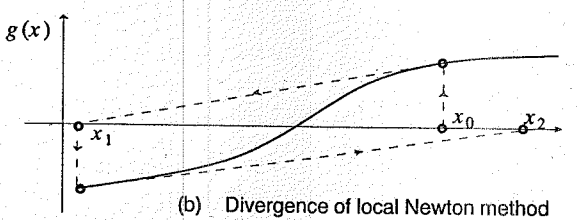
Proof. Let $\rho > 0$ be such that $g_x(x)$ is nonsingular for all $x \in B(\hat{x}, \rho)$. Since $g_x(\cdot)$ is Lipschitz continuous on bounded sets, there exist $b\,,\ L \in (0, \infty)$ such that

$$\|g_x(x)^{-1}\| \le b\,, \quad \forall\, x \in B(\hat{x}, \rho)\,, \tag{3a}$$

and



(a) Convergence of local Newton method



(b) Divergence of local Newton method

Fig. 1.4.1. Local behavior of Newton's method.

$$\|g_x(x') - g_x(x)\| \le L\|x - x'\|, \quad \forall x', x \in B(\hat{x}, \rho). \tag{3b}$$

Suppose that $x_i \in B(\hat{x}, \rho)$. Then, by construction and because $g(\hat{x}) = 0$ by assumption,

$$\begin{aligned} g_x(x_i)(x_{i+1} - x_i) &= -g(x_i) + g(\hat{x}) \\ &= -\int_0^1 g_x(\hat{x} + s(x_i - \hat{x}))\, ds\, (x_i - \hat{x}). \end{aligned} \tag{3c}$$

Hence, since, for $s \in [0,1]$, $\hat{x} + s(x_i - \hat{x}) \in B(\hat{x}, \rho)$, if we add $g_x(x_i)(x_i - \hat{x})$ to both sides of (3c), we find that

$$\begin{aligned} \|x_{i+1} - \hat{x}\| &\le \|g_x(x_i)^{-1}\| \int_0^1 \|g_x(x_i) - g_x(\hat{x} + s(x_i - \hat{x}))\| ds\, \|x_i - \hat{x}\| \\ &\le \tfrac{1}{2} bL\, \|x_i - \hat{x}\|^2. \end{aligned} \tag{3d}$$

Therefore, if

$$\tfrac{1}{2} bL\, \|x_i - \hat{x}\| < 1, \tag{3e}$$

then

$$\|x_{i+1} - \hat{x}\| < \|x_i - \hat{x}\|, \tag{3f}$$

and hence $x_{i+1} \in B(\hat{x}, \rho)$. Therefore, for any $\alpha \in (0,1)$, if we define $\hat{\rho} \triangleq \min\{\rho, 2\alpha/bL\}$, we obtain, by induction, that if $x_0 \in B(\hat{x}, \hat{\rho})$, then the entire sequence $\{x_i\}_{i=0}^{\infty}$ constructed by the Local Newton Algorithm (2c), is well

defined and contained in $B(\hat{x}, \hat{\rho})$ and satisfies the relation (3d). The desired result now follows directly from (3d) and Theorem 1.2.41*(b)*. □

The following alternative result (as well as its more sophisticated manifestation known as the Kantorovitch Theorem, see [KaA.59]), is sometimes found useful.

Exercise 1.4.2. Consider Problem (1). Suppose that *(a)* $g: \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuously differentiable on bounded sets and *(b)* the Jacobian $g_x(x)$ is nonsingular for all $x \in \mathbb{R}^n$. Suppose that (3a) and (3b) hold for all $x \in \mathbb{R}^n$.

Show that, if x_0 is such that

$$\tfrac{1}{2} bL\, \|g_x(x_0)^{-1} g(x_0)\| < 1, \tag{4a}$$

then the sequence $\{x_i\}_{i=0}^{\infty}$, constructed according to the Newtonian recursion (2c), converges Q-quadratically to a point $\hat{x} \in \mathbb{R}^n$ such that $g(\hat{x}) = 0$.

Hint: First show that

$$g_x(x_i)(x_{i+1} - x_i) = -g(x_i) + g_x(x_{i-1})(x_i - x_{i-1}) + g(x_{i-1}). \tag{4b}$$

Then use a first-order expansion of $g(x_i)$ about x_{i-1} to obtain that

$$\|x_{i+1} - x_i\| \le \tfrac{1}{2} bL\, \|x_i - x_{i-1}\|^2. \tag{4c}$$

Finally, make use of Theorem 1.2.42*(b)*. □

An important characteristic of the Local Newton Algorithm, defined by (2c), is that it is *scale-invariant*, i.e., it is invariant under coordinate transformations, as we will now show. Let A be a nonsingular $n \times n$ matrix, and let $x = Az$ define a coordinate change. Then the original equation (1) becomes $\tilde{g}(z) = 0$, where $\tilde{g}: \mathbb{R}^n \to \mathbb{R}^n$ is defined by

$$\tilde{g}(z) = g(Az). \tag{5a}$$

Since $\tilde{g}_z(z) = g_x(Az)A$, it is nonsingular because, by assumption, $g_x(Az)$ is nonsingular. According to (2c), given an initial guess x_0, the sequence constructed by the Local Newton Algorithm for solving the equation $\tilde{g}(z) = 0$ is defined by

$$\begin{aligned} z_{i+1} &= z_i - \tilde{g}_z(z_i)^{-1} \tilde{g}(z_i) \\ &= z_i - A^{-1} g_x(Az_i)^{-1} g(Az_i), \; i = 0, 1, 2, \ldots \end{aligned} \tag{5b}$$

If we multiply both sides of (5b) by A, we obtain

$$Az_{i+1} = Az_i - g_x(Az_i)^{-1} g(Az_i), \; i = 0, 1, 2, \ldots \tag{5c}$$

If we set $x_i = Az_i$, for $i = 0, 1, 2, \ldots$, we find that the x_i satisfy the recursion (2c), i.e., that the change in coordinates has not affected the sequence of iterates.

Next we return to the problem

$$\min_{x \in \mathbb{R}^n} f(x) . \tag{6}$$

Assumption 1.4.3.

(i) The function $f : \mathbb{R}^n \to \mathbb{R}$ in (6) is twice Lipschitz continuously differentiable on bounded sets, i.e., given any bounded set $S \subset \mathbb{R}^n$, there exists an $L_S < \infty$ such that

$$\|H(x') - H(x)\| \le L_S \|x - x'\| , \tag{7a}$$

for all x, $x' \in S$, with

$$H(x) \triangleq f_{xx}(x) . \tag{7b}$$

(ii) We will assume that (6) has a local minimizer $\hat{x}$ satisfying the second-order sufficiency condition in Theorem 1.1.8, and hence that there exist constants $0 < m \le M < \infty$ such that

$$m\|h\|^2 \le \langle h , H(\hat{x})h \rangle \le M\|h\|^2 , \quad \forall\ h \in \mathbb{R}^n . \tag{7c}$$

□

Note that any function that is twice Lipschitz continuously differentiable on bounded sets is also Lipschitz continuous on bounded sets, and so is its gradient.

Exercise 1.4.4. Suppose that H is an $n \times n$, real symmetric matrix and that there exist $0 < m \le M < \infty$ such that, for all $h \in \mathbb{R}^n$,

$$m\|h\|^2 \le \langle h , Hh \rangle \le M\|h\|^2 . \tag{8a}$$

Show that the induce norms of H and of H^{-1} ($\|H\| \triangleq \max_{\|h\| \le 1} \|Hh\|$) satisfy the inequalities

$$\|H\| \le M , \qquad \|H^{-1}\| \le 1/m . \tag{8b}$$

□

The logical extension of the Local Newton Algorithm, defined by (2c), to problem (6) is as follows. Given a current estimate x_i of the local minimizer $\hat{x}$ we expand $f(\cdot)$ to second-order terms about x_i, to obtain the approximation

$$f(x) \approx f(x_i) + \langle \nabla f(x_i) , (x - x_i) \rangle + \tfrac{1}{2} \langle (x - x_i) , H(x_i)(x - x_i) \rangle . \tag{9a}$$

Assuming that $H(x_i)$ is positive-definite, we can compute the minimizer x_{i+1} of the right-hand side of (9a) explicitly, by setting its gradient equal to zero (as per Corollary 1.1.3), i.e., by solving the equation

$$\nabla f(x_i) + H(x_i)(x_{i+1} - x_i) = 0 . \tag{9b}$$

Since $H(x_i)$ must be nonsingular for x_i close enough to $\hat{x}$, (9b) defines the iterative process

$$x_{i+1} = x_i - H(x_i)^{-1} \nabla f(x_i) , \quad i = 0, 1, 2, \ldots \tag{9c}$$

which we recognize as the application of the Local Newton Algorithm in (2c), encountered earlier, to the equation $\nabla f(x) = 0$. Note that because $\nabla f(x) = 0$ holds both for x a local minimizer and for a local maximizer, one will eventually have to introduce a modification to (9c) to eliminate the possibility of converging to a local maximizer.

We can restate (9c) in the form of an algorithm, as follows:

Local Newton Algorithm 1.4.5.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute the *Newton search direction*

$$h_i = -H(x_i)^{-1} \nabla f(x_i) . \tag{10a}$$

Step 2. Set

$$x_{i+1} = x_i + h_i , \tag{10b}$$

replace i by $i + 1$, and go to step 1.

It should be obvious that the Local Newton Algorithm 1.4.5 inherits scale invariance from the Local Newton Algorithm (2c). This is a very nice property which the Armijo Gradient Algorithm 1.3.3 does not possess. Also, when $f(\cdot)$ in (6) is a quadratic function of the form $f(x) = c + \langle g , x \rangle + \tfrac{1}{2}\langle x , Hx \rangle$, then Algorithm 1.4.5 computes a solution to (6) in one iteration. Again, this is an advantage over the Armijo Gradient Algorithm 1.3.3, which does not compute a minimizer for a quadratic function (6) in a finite number of iterations.

The following theorem is an obvious consequence of Theorem 1.4.1.

Theorem 1.4.6. *Consider problem (6), and suppose that Assumption 1.4.3 is satisfied. Then there exists a $\hat{\rho} > 0$ such that, if $x_0 \in B(\hat{x} , \hat{\rho})$ and the sequence $\{x_i\}_{i=0}^{\infty}$ is constructed by the Local Newton Algorithm 1.4.5, then $x_i \to \hat{x}$, as $i \to \infty$, Q-quadratically (with quotient rate 2).*

Proof. Since $H(\cdot)$ is continuous and Assumption 1.4.3 holds, there exist a $\rho > 0$ and an $L < \infty$ such that

$$\frac{m}{2}\|y\|^2 \le \langle y , H(x)y \rangle \le 2M\|y\|^2 , \quad \forall\ x \in B(\hat{x} , \rho) \text{ and } \forall\ y \in \mathbb{R}^n , \tag{11a}$$

and

$$\|H(x') - H(x)\| \leq L\|x - x'\|, \quad \forall\, x', x \in B(\hat{x}, \rho). \tag{11b}$$

Hence, for all $x \in B(\hat{x}, \rho)$, $H(x)$ is nonsingular, and $\|H(x)^{-1}\| \leq 2/m$. The rest of the proof follows directly from the proof of Theorem 1.4.1. □

There are two problems with the Local Newton Algorithm 1.4.5: first, that it converges to a solution of (6) only when initialized with a sufficiently good initial guess x_0; second, that it is basically a root finding, not an optimization algorithm, and hence, when applied to a nonconvex function $f(\cdot)$, it can converge to a local maximizer just as easily as to a local minimizer, since $\nabla f(x) = 0$ at both such points. We will deal with these problems one at a time.

1.4.2 Global Newton Method for Convex Functions

We will now show that the local Newton algorithm can be "globalized", for the case where the function $f : \mathbb{R}^n \to \mathbb{R}$ in (6) is strictly convex, by adding an Armijo type step-size rule to the local algorithm. Furthermore, we will see that the global algorithm converges with Q-quadratic rate.

Assumption 1.4.7.

(i) The function $f : \mathbb{R}^n \to \mathbb{R}$ is twice Lipschitz continuously differentiable on bounded sets.

(ii) There exists an $m > 0$ such that

$$m\|y\|^2 \leq \langle y, H(x)y \rangle, \quad \forall\, x, y \in \mathbb{R}^n. \tag{12}$$

□

In view of Assumption 1.4.7, we recall that the level sets of $f(\cdot)$ are convex and compact (see Proposition 5.2.15). Furthermore, in this case, by Corollary 1.1.12, problem (6) has a unique solution $\hat{x}$.

Newton-Armijo Algorithm 1.4.8.

Parameters. $\alpha \in (0, \frac{1}{2})$, $\beta \in (0, 1)$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute the *Newton search direction*

$$h_i \triangleq -H(x_i)^{-1}\nabla f(x_i). \tag{13a}$$

Stop if $h_i = 0$.

Step 2. Compute the *Armijo step-size*

$$\lambda_i = \max_{k \in \mathbb{N}} \{\beta^k \mid f(x_i + \beta^k h_i) - f(x_i) \leq \alpha\beta^k \langle h_i, \nabla f(x_i) \rangle\}. \tag{13b}$$

Step 3. Set $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i + 1$, and go to Step 1.

Theorem 1.4.9. *Consider problem (6), and suppose that Assumption 1.4.7 is satisfied. If $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by the Newton-Armijo Algorithm 1.4.8, then $x_i \to \hat{x}$, as $i \to \infty$, Q-quadratically, where $\hat{x}$ is the unique minimizer of $f(\cdot)$.*

Proof. The proof is in two parts. First we prove that $x_i \to \hat{x}$, as $i \to \infty$, with $\nabla f(\hat{x}) = 0$. Then we show that this convergence is Q-quadratic.

(a) Clearly, for all $i \in \mathbb{N}$, $h_i \triangleq h(x_i) \triangleq -H(x_i)^{-1}\nabla f(x_i)$. Since *(i)* $h(\cdot)$ is continuous, *(ii)* $h(x) = 0$ if and only if $\nabla f(x) = 0$, and *(iii)* $\langle \nabla f(x), h(x) \rangle < 0$ for all $x \in \mathbb{R}^n$ such that $\nabla f(x) \neq 0$, it follows directly from Theorem 1.2.24a that every accumulation point x^* of the sequence $\{x_i\}_{i=0}^{\infty}$ must satisfy $\nabla f(x^*) = 0$. Now, by Proposition 5.2.15, the level set

$$L_{f(x_0)}(f) \triangleq \{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\} \tag{14a}$$

is compact. Hence $\{x_i\}_{i=0}^{\infty}$ must converge to its set of accumulation points. Since, by Corollary 1.1.12, there is only one point in $L_{f(x_0)}(f)$, viz., $\hat{x}$, that satisfies $\nabla f(\hat{x}) = 0$, we conclude that $x_i \to \hat{x}$, as $i \to \infty$.

(b) To show that $x_i \to \hat{x}$, as $i \to \infty$, Q-quadratically, it is necessary to use the fact that $\alpha \in (0, \frac{1}{2})$. First, let

$$\gamma \triangleq \max\{f(x - \lambda H(x)^{-1}\nabla f(x)) \mid x \in L_{f(x_0)}(f),\ \lambda \in [0, 1]\}. \tag{14b}$$

Clearly, $\gamma \geq f(x_0)$. Since $H(\cdot)$ is Lipschitz continuous on bounded sets, there exists an $L \in (0, \infty)$ such that

$$\|H(x') - H(x'')\| \leq L\|x' - x''\|, \tag{14c}$$

for all $x', x'' \in \{x \mid f(x) \leq \gamma\}$. Next, setting $\lambda = 1$ in the expression for the Armijo step-size rule (see (13b)), upon expansion to second-order terms, we find that for some $s_i \in [0, 1]$,

$$f(x_i + h_i) - f(x_i) - \alpha\langle \nabla f(x_i), h_i \rangle$$

$$= (1 - \alpha)\langle \nabla f(x_i), h_i \rangle + \tfrac{1}{2}\langle h_i, H(x_i + s_i h_i)h_i \rangle$$

$$= -(1-\alpha)\langle h_i, H(x_i)h_i\rangle + \tfrac{1}{2}\langle h_i, H(x_i+s_ih_i)h_i\rangle, \tag{14d}$$

where we have used the fact that $-H(x_i)h_i = \nabla f(x_i)$. Adding and subtracting $\frac{1}{2}\langle h_i, H(x_i), h_i\rangle$ to the right-hand side of (14d), we obtain

$$f(x_i+h_i) - f(x_i) - \alpha\langle \nabla f(x_i), h_i\rangle$$

$$= -(1-\tfrac{1}{2}-\alpha)\langle h_i, H(x_i)h_i\rangle + \tfrac{1}{2}\langle h_i, [H(x_i+s_ih_i) - H(x_i)]h_i\rangle$$

$$\leq \|h_i\|^2\left[-(\tfrac{1}{2}-\alpha)m + \tfrac{1}{2}Ls_i\|h_i\|\right]. \tag{14e}$$

Now, $h_i \to 0$, as $i \to \infty$, because $\|h_i\| \leq \|\nabla f(x_i)\|/m$ and because $\nabla f(x_i) \to 0$, as $i \to \infty$. Since $\alpha \in (0, \frac{1}{2})$, $-(1-\alpha) < 0$ and $s_i \in [0,1]$, it follows that there must exist an i_0 such that, for all $i \geq i_0$,

$$[-(\tfrac{1}{2}-\alpha)m + \tfrac{1}{2}Ls\|h_i\|] \leq 0 \tag{14f}$$

and hence

$$f(x_i+h_i) - f(x_i) - \alpha\langle \nabla f(x_i), h_i\rangle \leq 0, \tag{14g}$$

which shows that $\lambda_i = 1$ for all $i \geq i_0$. Therefore we conclude that the global algorithm degenerates to the local algorithm as the solution $\hat{x}$ is approached. The desired rate of convergence result now follows from Theorem 1.4.6. □

Exercise 1.4.10.

(a) Show that Theorem 1.4.9 remains valid when Assumption 1.4.7 is weakened by replacing *(ii)* by the following hypothesis:

(ii) The Hessian matrix $H(x)$ is positive-definite for all $x \in \mathbb{R}^n$. Furthermore, for every $x_0 \in \mathbb{R}^n$, there exists an $m_0 > 0$ such that

$$m_0\|y\|^2 \leq \langle y, H(x)y\rangle, \quad \forall y \in \mathbb{R}^n, \ \forall x \in L_{f(x_0)}(f), \tag{15a}$$

where $L_{f(x_0)}(f) \triangleq \{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\}$.

(c) Suppose that Q is a symmetric, positive-definite $n \times n$ matrix, and consider the change of variables defined by $z = Qx$, so that problem (6) becomes

$$\min_{z \in \mathbb{R}^n} \tilde{f}(z), \tag{15b}$$

where $\tilde{f}(z) \triangleq f(Q^{-1}z)$. Show that, if the Armijo-Newton Algorithm 1.4.8 is applied to problem (6) from the starting point x_0, and to problem (15b) from the starting point $z_0 = Qx_0$, then it produces sequences $\{x_i\}_{i=0}^{\infty}$ and $\{z_i\}_{i=0}^{\infty}$ such that $z_i = Qx_i$ for all $i \in \mathbb{N}$, i.e., show that the Armijo-Newton Algorithm 1.4.8 is *invariant* under linear transformations based on a symmetric, positive-definite matrix. □

1.4.3 Discrete Newton Method

Let us briefly return to problem (1), with the assumptions stated, and suppose that $\hat{x} \in \mathbb{R}^n$ is such that $g(\hat{x}) = 0$. Consider what happens when we replace the Newtonian iteration formula (2c) by the approximate formula

$$x_{i+1} = x_i - H_i^{-1}g(x_i), \ i = 0,1,2,3,\ldots, \tag{16a}$$

where H_i is an $n \times n$ (hopefully nonsingular) matrix whose jth column $(H_i)_j$ is given by

$$(H_i)_j = \frac{1}{\varepsilon_i}[g(x_i + \varepsilon_i e_j) - g(x_i)], \quad j = 1,2,\ldots,n, \tag{16b}$$

where e_j is the jth column of the $n \times n$ identity matrix.

If we proceed as in (3c), from (16a) we obtain

$$H_i(x_{i+1} - x_i) = -g(x_i) + g(\hat{x})$$

$$= -\int_0^1 g_x(x_i - s(x_i - \hat{x}))\,ds\,(x_i - \hat{x}). \tag{16c}$$

Hence, adding $H_i(x_i - \hat{x})$ to both sides of (16c) and adding and subtracting $g_x(x_i)$, we obtain

$$\|x_{i+1} - \hat{x}\| \leq \|H_i^{-1}\|\int_0^1 \|[H_i - g_x(x_i)] + [g_x(x_i) - g_x(x_i - s(x_i - \hat{x}))]\|ds\,\|x_i - \hat{x}\|$$

$$\leq \|H_i^{-1}\|\int_0^1 (\|H_i - g_x(x_i)\| + sL\|x_i - \hat{x}\|)ds\,\|x_i - \hat{x}\|. \tag{16d}$$

Hence, if there exist m, $L' \in (0,\infty)$ such that $\|H_i^{-1}\| \leq 1/m$, for all $i \in \mathbb{N}$, and

$$\|H_i - g_x(x_i)\| \leq L'\|x_i - \hat{x}\|, \tag{16e}$$

then (16d) implies that, with $L^* \triangleq (L' + \frac{1}{2}L)/m$,

$$\|x_{i+1} - \hat{x}\| \leq L^*\|x_i - \hat{x}\|^2, \ i = 0,1,2,3,\ldots \tag{16f}$$

and hence that if $L^*\|x_0 - \hat{x}\| < 1$, then $x_i \to \hat{x}$, as $i \to \infty$, Q-quadratically.

Now, since $g_x(x)$ is assumed to be nonsingular for all x near $\hat{x}$, there exist $\rho > 0$ and an $m > 0$ such that, for all $x \in B(\hat{x}, \rho)$,

$$\|g_x(x)y\| \geq 2m\|y\|, \ \forall y \in \mathbb{R}^n. \tag{16g}$$

Hence, if $x_i \in B(\hat{x}, \rho)$ and $\|H_i - g_x(x_i)\| \leq m$, then for all $y \in \mathbb{R}^n$,

$$\|H_i y\| = \|H_i y - g_x(x)y + g_x(x)y\|$$

$$\geq \|g_x(x)y\| - \|[H_i - g_x(x_i)]y\| \geq m\|y\|, \tag{16h}$$

(where we have used the fact that $\|a-b\| \geq |\|a\| - \|b\||$) which shows that H_i is nonsingular. Furthermore, since $\|y\| = \|H(H^{-1}y)\| \geq m\|H^{-1}y\|$, for any $y \in \mathbb{R}^n$, it follows that $\|H_i^{-1}\| \leq 1/m$.

Next, since $g(\hat{x}) = 0$ for all $x_i \in B(\hat{x}, \rho)$,

$$\|g(x_i)\| = \left\|\int_0^1 g_x(\hat{x} + s(x_i - \hat{x}))ds\,(x_i - \hat{x})\right\| \leq M\|x_i - \hat{x}\|, \tag{16i}$$

where $M \triangleq \max_{x_i \in B(\hat{x},\rho)} \left\|\int_0^1 g_x(\hat{x} + s(x_i - \hat{x}))ds\right\| < \infty$. Now, by construction, $h_{i,j} = \int_0^1 g_x(x_i + s\varepsilon_i e_j)ds\, e_j$, $j = 1, 2, \ldots, n$, and $g_x(\cdot)$ is Lipschitz continuous on bounded sets, by assumption. Hence there exists a local Lipschitz constant $L'' < \infty$ such that, for all $x_i \in B(\hat{x}, \rho)$,

$$\|H_i - g_x(x_i)\| \leq L''\varepsilon_i. \tag{16j}$$

Therefore, if we set

$$0 < \varepsilon_i \leq \|g(x_i)\|, \tag{16k}$$

we conclude from (16i,j) that (16e) holds with $L' \triangleq L''M$. Now, let $\rho' \triangleq \min\{\rho m/L'\}$. Then it follows from (16e) that $\|H_i - g_x(x_i)\| \leq m$ for all $x_i \in B(\hat{x}, \rho')$ and hence that $\|H^{-1}\| \leq 1/m$ for all $x_i \in B(\hat{x}, \rho')$. Let $L^* \triangleq (L' + \frac{1}{2}L)/m$ (as above), and let $\hat{\rho} = \min\{\rho', 1/L^*\}$. If we begin with an $x_0 \in B(\hat{x}, \hat{\rho})$, then $L^*\|x_0 - \hat{x}\| < 1$. Hence, provided that ε_i satisfies (16k), (16e) and therefore also (16f) will hold for all $i \in \mathbb{N}$. Thus we have proved the following result.

Theorem 1.4.11. *Consider Problem (1) and suppose that*

(i) $g : \mathbb{R}^n \to \mathbb{R}^n$ *is Lipschitz continuously differentiable on bounded sets, and*

(ii) *the Jacobian* $g_x(x)$ *is nonsingular for all* x *near a solution* $\hat{x}$ *of (1).*

Then there exists a $\hat{\rho} > 0$ *such that (a) for all* $x_i \in B(\hat{x}, \hat{\rho})$, x_{i+1} *is well defined by (16a), (16b), and (16k), and (b) if* $x_0 \in B(\hat{x}, \hat{\rho})$, *and the sequence* $\{x_i\}_{i=0}^{\infty}$ *is constructed according to the recursion (16a), (16b), and (16k), then* $x_i \to \hat{x}$, *as* $i \to \infty$, *Q–quadratically (with quotient rate 2).* □

Combining the above results with Algorithm Model 1.2.30, we obtain a globally converging discrete Newton algorithm for solving the minimization problem (6), as follows.

Discrete Newton-Armijo Algorithm 1.4.12.

Parameters. $\alpha \in (0, \frac{1}{2})$, $\beta \in (0, 1)$, $\varepsilon_{-1} > 0$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. If $\nabla f(x_i) = 0$, stop. Else, compute

$$\varepsilon_i = \min\{\varepsilon_{i-1}, \|\nabla f(x_i)\|\}. \tag{17a}$$

Step 2. For $j = 1, 2, \ldots, n$ and $k = j, j+1, \ldots, n$, compute the finite differences

$$(H_i)_{j,k} = \frac{1}{\varepsilon_i}[g^k(x_i + \varepsilon_i e_j) - g^k(x_i)], \quad j = 1, 2, \ldots, n, \tag{17b}$$

with $g(x) \triangleq \nabla f(x)$, set $(H_i)_{k,j} = (H_i)_{j,k}$ for $j = 1, 2, \ldots, n$ and $k = j, j+1, \ldots, n$, and set H_i equal to the $n \times n$ symmetric matrix whose elements are $(H_i)_{j,k}$, $j = 1, 2, \ldots, n$, $k = 1, 2, \ldots, n$.

Step 3. If H_i^{-1} exists, compute the *discrete Newton search direction*

$$h_i \triangleq -H_i^{-1}\nabla f(x_i). \tag{17c}$$

Else, replace ε_i by $\varepsilon_i/2$, and go to Step 2.

Step 4. If

$$\langle \nabla f(x_i), h_i \rangle < 0, \tag{17d}$$

compute the *Armijo step-size*

$$\lambda_i = \max_{k \in \mathbb{N}} \{\beta^k \mid f(x_i + \beta^k h_i) - f(x_i) \leq \alpha\beta^k \langle h_i, \nabla f(x_i)\rangle\}. \tag{17e}$$

Else, replace ε_i by $\varepsilon_i/2$, and go to Step 2.

Step 5. If

$$f(x_i + \lambda_i h_i) - f(x_i) \leq -\varepsilon_i, \tag{17f}$$

set $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i + 1$, and go to Step 1.

Else, replace ε_i by $\varepsilon_i/2$, and go to Step 2.

To obtain reasonable scaling in computing the approximate Hessian by finite differences, it may be desirable to set $\varepsilon_i \leq \min\{\gamma\|x_i\|, \|g(x_i)\|\}$ with $\gamma \leq 1/10$, say.

Exercise 1.4.13. Prove the following result:

Theorem 1.4.14. *Consider problem (6), and suppose that Assumption 1.4.7 is satisfied. If $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by the Discrete Newton-Armijo Algorithm 1.4.12, then $x_i \to \hat{x}$, as $i \to \infty$, Q-quadratically, where $\hat{x}$ is the unique minimizer of $f(\cdot)$.* □

1.4.4 Global Newton Method for General Functions

Both Algorithm Model 1.2.23, with Armijo line search, and the trust region Algorithm Model 1.2.26 provide guidance for constructing "extended" Newton type algorithms that can be used for solving problem (6) under Assumption 1.4.7*(i)* only, i.e., the strict convexity requirement can be dropped. We present one example of such an extension below, based on Algorithm Model 1.2.23, which, in addition to the Armijo step-size parameters $\alpha, \beta \in (0,1)$, uses a bound $\kappa \gg 1$ on the matrix condition number of the Hessian $H(x)$. The bound κ is imposed by the numerical linear algebra computer code used to solve the equation $H(x_i)h = -\nabla f(x_i)$ for the search direction h_i in (18a) below.

Stabilized Newton-Armijo Algorithm 1.4.15.

Parameters. $\alpha \in (0, \frac{1}{2})$, $\beta \in (0,1)$, $\kappa \gg 1$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute the Hessian matrix $H(x_i)$ and its largest and smallest eigenvalues $\lambda_{\max}(x_i)$ and $\lambda_{\min}(x_i)$.

If $\lambda_{\min}(x_i) > 0$ and $\lambda_{\max}(x_i)/\lambda_{\min}(x_i) \le \kappa$, set

$$h_i \triangleq -H(x_i)^{-1}\nabla f(x_i). \tag{18a}$$

Else, set

$$h_i \triangleq -\nabla f(x_i). \tag{18b}$$

Step 2. Compute the *Armijo step-size*

$$\lambda_i = \max_{k \in \mathbb{N}} \{\beta^k \mid f(x_i + \beta^k h_i) - f(x_i) \le \alpha\beta^k \langle h_i, \nabla f(x_i)\rangle\}. \tag{18c}$$

Step 3. Set $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i+1$, and go to Step 1.

Exercise 1.4.16. Use the fact that *(i)* $\|H(x)\| = \lambda_{\max}(x)$, when $H(x) \ge 0$, and *(ii)* $\|H(\cdot)\|$ is continuous, together with the Polak-Sargent-Sebastian Theorem 1.2.24b to prove the following result:

Theorem 1.4.17. *Consider problem (6), and suppose that $f: \mathbb{R}^n \to \mathbb{R}$ is twice Lipschitz continuously differentiable on bounded sets. Then,*

(a) Algorithm 1.4.15 satisfies the hypotheses of Theorem 1.2.24b, and hence, if $\hat{x}$ is an accumulation point of a sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 1.4.15, then $\nabla f(\hat{x}) = 0$; and

(b) if Algorithm 1.4.15 constructs a sequence $\{x_i\}_{i=0}^{\infty}$ which has an accumulation point $\hat{x}$ such that (7c) holds with $0 < m \le M$ and $M/m \le \kappa/2$, then there exists an i_0 such that for all $i \ge i_0$, h_i is determined by (18a), and $\{x_i\}_{i=0}^{\infty}$ converges to $\hat{x}$ Q-quadratically. □

1.4.5 The Iterated Newton Method

As we have seen, under Assumption 1.4.7, the Global Newton Algorithm 1.4.8 converges Q-quadratically, when applied to the minimization problem (6). However, as we will now show, one can improve on its asymptotic efficiency, which is determined by the behavior of the Local Newton Algorithm 1.4.5 and is defined in (1.2.44j) by sacrificing some of its rate of convergence to gain a considerable reduction in work required per iteration.

Suppose that the time needed to compute a partial derivative (such as $\partial f(x)/\partial x^i$, $\partial^2 f(x)/\partial (x^i)^2$) is about the same as that needed to evaluate $f(x)$. Now consider what happens when the Newton-Armijo Algorithm 1.4.8 becomes Q-quadratically convergent, i.e., when it reduces to the Local Newton Algorithm 1.4.5. In this case, it requires one work unit for the Armijo step-size evaluation of $f(x_{i+1})$, n work units for the evaluation of $\nabla f(x_i)$, and $n(n-1)/2$ work units for the evaluation of the symmetric Hessian matrix $H(x_i)$, a total of $W = [1 + n + \frac{1}{2}n(n+1)]$ work units. Assuming that the work of inverting the Hessian is negligible compared to the above evaluations (which may not be true in the case of large matrices), the efficiency E_N of the Newton Algorithm 1.4.8 is given by

$$E_N = \frac{\ln 2}{1 + n + \frac{1}{2}n(n+1)}. \tag{19}$$

Note that the efficiency of the Local Newton Algorithm 1.4.5 decreases rapidly as the dimension of the vector x increases. Now we will show that one can improve matters considerably by computing the Hessian only once every q iterations, with $q > 1$. This results in an algorithm of the form of the Local Newton Algorithm 1.4.5, where $H(x_i)$ is replaced by the matrix H_i, with $H_i = H(x_{jq})$ for $jq \le i < (j+1)q$, and $j = 0, 1, 2, \ldots$.

Notation 1.4.18. *For any real number a, we will denote by $\lfloor a \rfloor$ the largest integer smaller than or equal to a, i.e., $\lfloor a \rfloor \triangleq \max\{a' \in \mathbb{N} \mid a' \le a\}$, so that $\lfloor a \rfloor$ is the integer satisfying the relation $a - 1 < \lfloor a \rfloor \le a$. Thus $\lfloor 1.3 \rfloor = 1$, $\lfloor 2.7 \rfloor = 2, \lfloor 3.9 \rfloor = 3$, etc.* □

Shamanskii Iterated Newton-Armijo Algorithm 1.4.19.

Parameters. $q \in \mathbb{N}$, $\alpha \in (0, \frac{1}{2})$, $\beta \in (0,1)$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute the *iterated Newton search direction*

$$h_i = -H(x_{\lfloor i/q \rfloor q})^{-1} \nabla f(x_i). \tag{20a}$$

Step 2. Compute the *Armijo step-size*

$$\lambda_i = \max_{k \in \mathbb{N}} \{\beta^k \mid f(x_i + \beta^k h_i) - f(x_i) \le \beta^k \alpha \langle h_i, \nabla f(x_i) \rangle\}. \tag{20b}$$

Step 3. Set $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i+1$, and go to step 1.

Theorem 1.4.20. *Consider problem (6), and suppose that Assumption 1.4.7 holds.*

(a) If $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by the Shamanskii Iterated Armijo-Newton Algorithm 1.4.19, then $x_i \to \hat{x}$, as $i \to \infty$, with $\hat{x}$ the global minimizer of $f(\cdot)$.

(b) There is an i_0 such that $\lambda_i = 1$ for all $i \ge i_0$. □

Exercise 1.4.21. Prove Theorem 1.4.20. *Hint*: Use the Polak-Sargent-Sebastian Theorem 1.2.24b to establish part *(a)*, and mimic the proof of Theorem 1.4.9 to establish part *(b)*. □

Next we will use Theorem 1.2.41 to establish the rate of convergence of the local version of the above algorithm.

Theorem 1.4.22. *Consider problem (6), and suppose that Assumption 1.4.7 holds and that $\hat{x}$ is the unique minimizer of $f(\cdot)$. Then there exists a $\hat{\rho} > 0$ such that if $x_0 \in B(\hat{x}, \hat{\rho})$ and $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by the* Local Shamanskii Iterated Newton Algorithm *defined by*

$$x_{i+1} = x_i - H(x_{\lfloor i/q \rfloor q})^{-1} \nabla f(x_i), \quad i = 0, 1, 2, \ldots \tag{21a}$$

where $q \ge 1$ is a given integer, then $x_i \to \hat{x}$, as $i \to \infty$, superlinearly, with root rate r greater than or equal to τ_q, where τ_q is the unique positive root of the equation

$$t^q - t^{q-1} - 1 = 0. \tag{21b}$$

Proof. We will use Theorem 1.2.41 with $p = q - 1$. Referring to Theorem 1.2.41, we see that we must establish relations of the form (1.2.42a) and (1.2.42b) for sequences constructed according to (21a) to deduce that Theorem 1.4.22 is true.

We begin with (1.2.42a). First, it follows from Assumption 1.4.7 that $\|H(x)^{-1}\| \le 1/m$ for all $x \in \mathbb{R}^n$. For all $i \in \mathbb{N}$, let $H_i \triangleq H(x_{\lfloor i/q \rfloor q})$. Let $\rho > 0$ be arbitrary, let $L < \infty$ be a Lipschitz constant for $H(\cdot)$ on $B(\hat{x}, 2\rho)$, and suppose that for all $i \in \mathbb{N}$, $x_i \in B(\hat{x}, \rho)$. Then it follows from (21a), since $\nabla f(\hat{x}) = 0$, that, for any $i \in \mathbb{N}$,

$$H_i(x_{i+1} - x_i) = -\nabla f(x_i) = -\int_0^1 H(\hat{x} + s(x_i - \hat{x}))(x_i - \hat{x})\,ds. \tag{22a}$$

Hence, adding $H_i(x_i - \hat{x})$ to both sides of (22a), we conclude that, for any $i \le q$,

$$H_i(x_{i+1} - \hat{x}) = \int_0^1 [H_i - H(\hat{x} + s(x_i - \hat{x}))]\,ds\,(x_i - \hat{x}). \tag{22b}$$

Since $\|H_i^{-1}\| \le 1/m$ and $H(\cdot)$ is Lipschitz continuous, with some constant $L < \infty$, on $B(\hat{x}, 2\rho)$, we conclude from (22b) that, for all $i \in \mathbb{N}$,

$$\|x_{i+1} - \hat{x}\| \le \frac{L}{m}\|x_i - \hat{x}\| \int_0^1 \|(x_{\lfloor i/q \rfloor q} - \hat{x}) - s(x_i - \hat{x})\|\,ds$$

$$\le \frac{L}{m}\|x_i - \hat{x}\|(\|x_i - \hat{x}\| + \|x_{\lfloor i/q \rfloor q} - \hat{x}\|). \tag{22c}$$

Let $e_i = (L/m)\|x_i - \hat{x}\|$. Then, after multiplying both sides of (22c) by L/m, we find that, for all $i \in \mathbb{N}$,

$$e_{i+1} \le e_i(e_i + e_{\lfloor i/q \rfloor q}). \tag{22d}$$

Now suppose that $\hat{\rho} = 0.9 \min\{\rho, m/2L\}$ and that $x_0 \in B(\hat{x}, \hat{\rho})$. Then $2e_0 \le 0.9$, and hence

$$e_1 \le e_0[e_0 + e_0] = (2e_0)e_0 < e_0. \tag{22e}$$

Now we can start an induction process. Suppose that $e_i \le e_0$ for $i = 0, \ldots, k$. Then it follows from (22d) that

$$e_{k+1} \le e_k(e_k + e_{\lfloor k/q \rfloor q}) \le (2e_0)e_k < e_0. \tag{22f}$$

Hence we conclude that $e_k \le e_0$ for all $k \in \mathbb{N}$ and, therefore, that (22c) is valid

for all $i \in \mathbb{N}$, provided that $x_0 \in B(\hat{x}, \hat{\rho})$. Furthermore, it follows from (22f) that $e_i \to 0$, as $i \to \infty$, monotonically and hence that $\|x_i - \hat{x}\| \to 0$, monotonically, as $i \to \infty$. Hence, since $\lfloor i/q \rfloor \geq i + 1 - q$ for all $i \in \mathbb{N}$, and $q \in \mathbb{N}$ is arbitrary, it follows from (22c) that if $x_0 \in B(\hat{x}, \hat{\rho})$, then, for all $i \in \mathbb{N}$,

$$\|x_{i+1} - \hat{x}\| \leq \frac{L}{m}\|x_i - \hat{x}\|(\|x_i - \hat{x}\| + \|x_{i+1-q} - \hat{x}\|), \tag{22g}$$

which is of the form (1.2.42a), with $p = q - 1$, $\gamma_j = L/m$ for $j = 0$ and $j = q - 1$, and $\gamma_j = 0$, otherwise.

Now $\sum_{j=0}^{q-1} \gamma_j = 2L/m$. Since, by construction of $\hat{\rho}$, we have ensured that $2e_i = (2L/m)\|x_i - \hat{x}\| \leq 0.9$ for all $i \in \mathbb{N}$, we see that (1.2.42b) holds with $\eta = 0.9$. Hence the desired result follows from Theorem 1.2.41. □

Exercise 1.4.23. Consider problem (6), and suppose that Assumption 1.4.7 holds. Show that a slightly tighter bound on the R–rate of convergence of the Local Shamanskii Iterated Newton Algorithm is given by

$$r = (q + 1)^{1/q}. \tag{23}$$

□

It remains to be shown that, for suitable choices of q, the Shamanskii Iterated Armijo-Newton Algorithm 1.4.19 is more efficient than the Armijo-Newton Algorithm 1.4.8, whose efficiency is given in (19). Let E_{IN} denote the efficiency of the Shamanskii Iterated Armijo-Newton Algorithm. For the Shamanskii Iterated Armijo-Newton Algorithm 1.4.19, we define the work per iteration as the *average* work per cycle of q iterations, when the algorithm has become Q–quadratically convergent and the step-size $\lambda_i = 1$ is computed using only one evaluation of $f(\cdot)$. In this case, in a cycle of q iterations, we perform q function evaluations, q gradient evaluations and one Hessian evaluation. If we count the evaluation of $f(\cdot)$ as one unit of work, the evaluation of $\nabla f(x)$ as n units of work, and the evaluation of $H(x)$ as $n(n + 1)/2$ units of work (as we had also assumed for the Armijo-Newton Algorithm), then, in a cycle of q iterations, the Shamanskii Iterated Armijo-Newton Algorithm requires the average work per iteration

$$W = \frac{1}{q}[q + nq + n(n + 1)/2]. \tag{24a}$$

Hence

$$E_{IN} = \frac{\frac{1}{q}\ln(q + 1)}{\frac{1}{q}[q + nq + \tfrac{1}{2}n(n + 1)]} = \frac{\ln(q + 1)}{q(1 + n) + \tfrac{1}{2}n(n + 1)}, \tag{24b}$$

where we have used the slightly tighter rate bound given by (23).

Exercise 1.4.24. Show that $E_{IN} > E_N$ for a suitable choice of q. In particular, examine the case where $n = 50$. What is the best value of q for this case? Show that if the numerical linear algebra involved in computing the search direction h_i is taken into account, the ratio E_{IN}/E_N can become considerably larger. □

Exercise 1.4.25. Develop globally converging versions of Newton's algorithm and of the iterated Newton algorithm based on the trust region Algorithm Model 1.2.26. □

1.4.6 Notes

The local Newton algorithm is probably the oldest root-finding algorithm using derivative information. It is attributed to Sir Isaac Newton, though it is also known as the Newton-Ralphson algorithm. It was described in a treatise written in Latin by Sir Isaac Newton and translated by Mr. Ralphson, published in 1720, see [New.720].

For an in-depth study of the local Newton algorithm, see [KaA.59] and [OrR.70]. The idea of using a step-size rule to globalize Newton's algorithm on convex functions is due to Goldstein [Gol.67], while the local discrete Newton algorithm and modified Newton type algorithms that reuse previous Hessians or their approximations, as in the Shamanskii Iterated Newton Algorithm 1.4.19, was first proposed by Shamanskii [Sha.67].

1.5 Methods of Conjugate Directions

Methods of conjugate directions were first proposed by Hestenes and Stiefel in 1952 (see [HeS.52]) as a technique for solving large systems of linear equations, of the form $Gh = g$, where G is a given $n \times n$ nonsingular matrix and $g \in \mathbb{R}^n$ is a given vector. Their method depended on three facts: *(i)* the solution of the equation $Gh = g$ is the same as the solution of the minimization problem $\min_{h \in \mathbb{R}^n} \|Gh - g\|^2$, *(ii)* when the minimization problem $\min_{h \in \mathbb{R}^n} \|Gh - g\|^2$ is expressed in terms of a G^TG-conjugate basis (to be defined shortly), this problem decomposes into n one-dimensional quadratic minimization problems, and *(iii)* the availability of an iterative formula for constructing, simultaneously, both an orthogonal and a conjugate basis, by using a positive-definite matrix.

Subsequently, methods of conjugate directions were extended to problems of the form $\min_{x \in \mathbb{R}^n} f(x)$, where $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, by Fletcher and Reeves [FlR.64] and Polak and Ribière [PoR.69], among others, with a rationale provided by the fact that the quadratic approximation $\frac{1}{2}\langle x - \hat{x}, f_{xx}(\hat{x})(x - \hat{x})\rangle$ to $f(x)$ is quite accurate near a solution point $\hat{x}$.

1.5.1 Decomposition of Quadratic Functions

In view of the above discussion, we begin by considering the quadratic minimization problem

$$\min_{x \in \mathbb{R}^n} \tfrac{1}{2}\langle x, Hx \rangle + \langle d, x \rangle , \tag{1}$$

where H is an $n \times n$, symmetric, positive-definite matrix.

Definition 1.5.1. *A set of nonzero vectors $\{u_i\}_{i=1}^n$ for $\mathbb{R}^n$ is said to be H-conjugate (or simply* conjugate*) relative to a symmetric, positive-definite, $n \times n$ matrix H, if $\langle u_i, Hu_j \rangle = 0$ for all $i \neq j$.* □

Note that a set of nonzero vectors $\{u_i\}_{i=1}^n$ for $\mathbb{R}^n$ that is H-conjugate, relative to a symmetric, positive-definite, $n \times n$ matrix H is orthogonal with respect to the norm $\|x\|_H \triangleq \langle x, Hx \rangle$. Hence, a set of vectors that is H-conjugate can also be called H-orthogonal.

Our first observation is that a set of vectors $\{u_i\}_{i=1}^n$ for $\mathbb{R}^n$ that is *H-conjugate* relative to a symmetric, positive-definite, $n \times n$ matrix H is linearly independent, i.e., it is a basis for $\mathbb{R}^n$. To see this, let the scalars α_i, $i = 1, 2, \ldots, n$, be such that

$$\sum_{i=1}^n \alpha_i u_i = 0 . \tag{2a}$$

Then, because $\langle u_i, Hu_j \rangle = 0$ for all $i \neq j$, if we take the scalar product of both sides of (2a) with any Hu_j, $j \in \mathbf{n}$, we conclude that $\alpha_j \langle u_j, Hu_j \rangle = 0$. Since $\langle u_j, Hu_j \rangle \neq 0$, because H is positive-definite, $\alpha_j = 0$. Hence it follows that the vectors $\{u_i\}_{i=1}^n$ are linearly independent.

The most obvious conjugate basis that one can associate with an $n \times n$, symmetric, positive-definite matrix H is any orthogonal basis of eigenvectors of H, $\{v_i\}_{i=1}^n$. (We recall that a basis $\{u_i\}_{i=1}^n$, for $\mathbb{R}^n$ is said to be *orthogonal* if $\langle u_i, u_j \rangle = 0$ for all $i \neq j$). Clearly, in the case of orthogonal eigenvectors,

$$\langle v_i, Hv_j \rangle = \lambda_j \langle v_i, v_j \rangle = 0 , \quad \forall\, i \neq j , \tag{2b}$$

where λ_j is the eigenvalue corresponding to the eigenvector v_j.

Sets of orthogonal eigenvectors are not the only examples of H-conjugate bases, as the reader can establish by examining the following "bootstrap" method for constructing H-conjugate bases proposed by Hestenes and Stiefel.

Exercise 1.5.2. Let H be a symmetric, positive-definite, $n \times n$ matrix and suppose that $g_0 \in \mathbb{R}^n$, not an eigenvector of H, is given. Show that if the following process does not stop because $g_i = 0$, then it constructs, *simultaneously*, both an orthogonal basis $\{g_i\}_{i=0}^{n-1}$ and an H-conjugate basis $\{h_i\}_{i=0}^{n-1}$ for $\mathbb{R}^n$:

$$\left.\begin{aligned} &g_0 \text{ given}, \\ &g_{i+1} = g_i + \lambda_i H h_i , \\ &\lambda_i = -\|g_i\|^2 / \langle g_i, Hh_i \rangle \end{aligned}\right\} , \quad i = 0, 1, \ldots, n-1 ; \tag{3a}$$

$$\left.\begin{aligned} &h_0 = -g_0 , \\ &h_{i+1} = -g_{i+1} + \gamma_i h_i \\ &\gamma_i = \langle Hh_i, g_{i+1} \rangle / \langle Hh_i, h_i \rangle \end{aligned}\right\} , \quad i = 0, 1, \ldots, n-1 . \tag{3b}$$

□

Let us return to the function $f : \mathbb{R}^n \to \mathbb{R}$, defined by

$$f(x) \triangleq \tfrac{1}{2}\langle x, Hx \rangle + \langle d, x \rangle , \tag{4}$$

(with H an $n \times n$, symmetric and positive-definite matrix) that appears in (1). Suppose that $\{h_j\}_{j=0}^{n-1}$ is an H-conjugate basis for $\mathbb{R}^n$. Then, given a point $x_0 \in \mathbb{R}^n$, to which we translate the origin, we can express an arbitrary point $x \in \mathbb{R}^n$ in terms of the H-conjugate basis, as follows:

$$x = x_0 - \sum_{j=0}^{n-1} \lambda_j h_j , \tag{5a}$$

(note that we have changed notation and that now λ_j *is not* an eigenvalue). Replacing x in (4) with the expression in (5a), we obtain

$$\begin{aligned} f(x) &= f\Big(x_0 + \sum_{j=0}^{n-1} \lambda_j h_j\Big) \\ &= f(x_0) + \sum_{j=0}^{n-1} \Big[\tfrac{1}{2}\langle h_j, Hh_j \rangle \lambda_j^2 + \langle d + Hx_0, h_j \rangle \lambda_j \Big] . \end{aligned} \tag{5b}$$

We see that the use of the H-conjugate basis decomposes the original optimization problem (1), which is in terms of the n-dimensional variable x, into n problems in terms of the one-dimensional variables λ_j. Clearly, the values $\hat{\lambda}_j$ that minimize (5b) are given by

$$\hat{\lambda}_j = -\frac{\langle d + Hx_0, h_j \rangle}{\langle h_j, Hh_j \rangle} , \tag{6a}$$

and hence the optimal solution to (1) is given by

$$\hat{x} = x_0 + \sum_{j=0}^{n-1} \hat{\lambda}_j h_j . \tag{6b}$$

The global minimizer $\hat{x}$ of (4) can be obtained either by computing the $\hat{\lambda}_j$ in parallel and then forming the sum (6b), or it can be accumulated sequentially, starting from x_0 and using the recursion

$$x_{i+1} = x_i + \hat{\lambda}_i h_i\,,\ i = 0,1,..,n-1\,, \tag{6c}$$

yielding $\hat{x} = x_n$.

Proposition 1.5.3. *Consider problem (1). Let $\{h_j\}_{j=0}^{n-1}$ be a H-conjugate basis for $\mathbb{R}^n$. Then,*

(a) the recursion (6c) is equivalent to the recursion

$$\left.\begin{aligned} &x_0 \text{ given},\\ &x_{i+1} = x_i + \lambda_i h_i\,,\\ &\lambda_i = \arg\min_{\lambda \geq 0} f(x_i + \lambda h_i)\,, \end{aligned}\right\},\ i = 0,1,\ldots,n\,, \tag{7a}$$

where the quadratic function $f(\cdot)$ is defined as in (4), and

(b) for $i = 1,2,\ldots,n$, x_i, constructed as above, minimizes the quadratic function $f(\cdot)$ on the affine manifold

$$M_i \triangleq \{x \in \mathbb{R}^n \mid x = x_0 + \textstyle\sum_{j=1}^{i} y^j h_{j-1},\ y \in \mathbb{R}^i\} \tag{7b}$$

(note that $M_n = \mathbb{R}^n$).

Proof. *(a)* It follows from (7) that

$$x_i = x_0 + \sum_{j=0}^{i-1} \lambda_j h_j\,. \tag{8a}$$

For $i = 0,\ldots,n$, let $g_i \triangleq \nabla f(x_i)$. Since λ_j is computed by exact minimization along the line spanned by the vector h_j, it follows that $\langle g_{i+1}, h_j\rangle = 0$. Therefore, since

$$g_{i+1} = Hx_{i+1} + d = g_0 + \sum_{j=0}^{i-1} \lambda_j H h_j \tag{8b}$$

and the vectors h_j are H-conjugate, we conclude that, for $j = 0,1,\ldots,n-1$,

$$\lambda_j = -\frac{\langle g_0, h_j\rangle}{\langle h_j, Hh_j\rangle} = \hat{\lambda}_j\,. \tag{8c}$$

Consequently, it follows that (7) produces the same sequence of vectors as (6c).

(b) It follows from (8a) that, for any $i \in \{0,\ldots,n-1\}$, $g_i \triangleq \nabla f(x_i)$ is given by

$$g_i = g_0 + \sum_{j=0}^{i-1} \lambda_j H h_j\,. \tag{8d}$$

Hence, since the h_j are H-conjugate, for any $k \in \{0,\ldots,n\}$ and $j \in \{0,\ldots,k-1\}$,

$$\langle g_k, h_j\rangle = \langle g_0, h_j\rangle + \lambda_j \langle h_j, Hh_j\rangle\,. \tag{8e}$$

Since, in view of the step-size calculation in (7), $\langle g_{j+1}, h_j\rangle = 0$ must hold, we conclude from (8e), with $k = j+1$, that $\langle g_0, h_j\rangle + \lambda_j\langle h_j, Hh_j\rangle = 0$ for $j = 0,1,\ldots,i-1$. Hence it follows from (8e) that

$$\langle g_i, h_j\rangle = 0,\ \forall\, j \in \{0,\ldots,i-1\}\,, \tag{8f}$$

which proves that $\langle \nabla f(x_i), x - x_0\rangle = 0$ for all $x \in M_i$, i.e., that x_i satisfies the first-order optimality condition (1.1.5) for the problem $\min_{x \in M_i} f(x)$. Since $f(\cdot)$ is strictly convex, it follows that x_i is the global minimizer $f(\cdot)$ on M_i. □

1.5.2 Methods of Conjugate Gradients

We begin with a *progenitor* conjugate gradient algorithm, based on the recursions (3a), (3b), which is useful *only* for minimizing positive-definite quadratic functions as in (1).

Progenitor Conjugate Gradient Algorithm 1.5.4.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$, and set $h_0 = -g_0 \triangleq -(Hx_0 + d)$.

Step 1. Compute the *step-size*

$$\lambda_i = \arg\min_{\lambda \geq 0} f(x_i + \lambda h_i)\,. \tag{9a}$$

Step 2. *Update:* Set

$$\left.\begin{aligned} &x_{i+1} = x_i + \lambda_i h_i\,,\\ &g_{i+1} \triangleq Hx_{i+1} + d\,,\\ &h_{i+1} = -g_{i+1} + \gamma_i h_i\,, \end{aligned}\right. \tag{9b}$$

with

$$\gamma_i = \frac{\langle Hh_i, g_{i+1}\rangle}{\langle h_i, Hh_i\rangle}\,, \tag{9c}$$

so that

$$\langle h_{i+1}, Hh_i\rangle = 0\,. \tag{9d}$$

Step 3. Replace i by $i+1$, and go to Step 1.

The Progenitor Conjugate Gradient Algorithm 1.5.4 serves as the point of departure in the construction of several algorithms which produce H-conjugate

directions while minimizing a positive-definite quadratic function

$$f(x) = \tfrac{1}{2}\langle x, Hx\rangle + \langle d, x\rangle .$$

These algorithms are called conjugate gradient methods and arise from the fact that the formula for γ_i in (3b) can be replaced with several equivalent alternatives. As a result, when minimizing a quadratic function, these algorithms produce identical search directions, and, since they use the same step-size rule, they construct identical sequences of points x_i. However, when $f(\cdot)$ is not quadratic, their behavior can be quite different.

Theorem 1.5.5. *The Progenitor Conjugate Gradient Algorithm 1.5.4 solves problem (1) in, at most, n iterations.*

Proof. In view of Proposition 1.5.3, (6b) and (6c), we need to show only that the search directions h_i constructed by Algorithm 1.5.4 are H-conjugate, i.e., that $\langle h_i, Hh_j\rangle = 0$ for all $i \neq j$. In the process of establishing this fact, we will show that the g_i, $i = 0, 1, 2, \ldots, n$, constructed by Algorithm 1.5.4, are orthogonal, i.e., that $\langle g_i, g_j\rangle = 0$, for all $i \neq j$.

Our proof proceeds by induction. Since $h_0 = -g_0$,

$$\langle g_1, h_0\rangle = -\langle g_1, g_0\rangle = 0 , \tag{10a}$$

by construction of λ_0. Next, by construction of γ_0,

$$\langle h_1, Hh_0\rangle = 0 . \tag{10b}$$

Hence, suppose that

$$\left.\begin{aligned}\langle g_i, g_j\rangle &= 0\\ \langle h_i, Hh_j\rangle &= 0\end{aligned}\right\}, \quad \forall\ 0 \le i, j \le k \le n-1,\ i \neq j . \tag{10c}$$

First, by construction of λ_i, $\langle g_{i+1}, h_i\rangle = 0$ for all i. Next, for $i = 0, 1, \ldots, n-1$,

$$g_{i+1} = Hx_{i+1} + d = H(x_i + \lambda_i h_i) + d = g_i + \lambda_i Hh_i . \tag{10d}$$

Consequently,

$$\langle g_{i+1}, h_i\rangle = 0 = \langle h_i, g_i\rangle + \lambda_i \langle h_i, Hh_i\rangle , \text{ for } i = 0, 1, \ldots, n-1, \tag{10e}$$

so that

$$\lambda_i = -\frac{\langle h_i, g_i\rangle}{\langle h_i, Hh_i\rangle} , \quad i = 0, 1, \ldots, n-1 . \tag{10f}$$

Thus, from (10d) and (10f), we find that, for $i = 0, 1, \ldots, n-1$,

$$\langle g_i, g_{i+1}\rangle = \langle g_i, g_i\rangle - \frac{\langle h_i, g_i\rangle \langle g_i, Hh_i\rangle}{\langle h_i, Hh_i\rangle} . \tag{10g}$$

Now,

$$\langle h_i, g_i\rangle = \langle -g_i + \gamma_{i-1}h_{i-1}, g_i\rangle = -\langle g_i, g_i\rangle , \tag{10h}$$

and

$$\langle h_i, Hh_i\rangle = \langle -g_i + \gamma_{i-1}h_{i-1}, Hh_i\rangle = -\langle g_i, Hh_i\rangle . \tag{10i}$$

Substituting into (10g), we get

$$\langle g_i, g_{i+1}\rangle = 0, \text{ for } i = 0, 1, \ldots, n-1 , \tag{10j}$$

and hence $\langle g_{k+1}, g_k\rangle = 0$.

Next, for all $0 < i < k$, since $\langle g_i, g_k\rangle = 0$ and $\langle h_i, Hh_k\rangle = 0$, by hypothesis, it follows from (10d) that

$$\begin{aligned}\langle g_{k+1}, g_i\rangle &= \langle g_k + \lambda_k Hh_k, g_i\rangle\\ &= \lambda_k \langle Hh_k, g_i\rangle\\ &= \lambda_k \langle Hh_k, h_i - \gamma_{i-1}h_{i-1}\rangle = 0 .\end{aligned} \tag{10k}$$

Finally, since $-g_0 = h_0$,

$$\begin{aligned}\langle g_{k+1}, g_0\rangle &= \langle g_k + \lambda_k Hh_k\rangle\\ &= \lambda_k \langle Hh_k, g_0\rangle\\ &= -\lambda_k \langle Hh_k, h_0\rangle = 0 .\end{aligned} \tag{10l}$$

Next, for $i = k$, $\langle h_{k+1}, Hh_k\rangle = 0$, by construction of γ_k. For $0 \le i < k$,

$$\begin{aligned}\langle h_{k+1}, Hh_i\rangle &= \langle -g_{k+1} + \gamma_k h_k, Hh_i\rangle\\ &= \langle -g_{k+1}, Hh_i\rangle\\ &= -\langle g_{k+1}, \frac{1}{\lambda_i}(g_{i+1} - g_i)\rangle = 0 ,\end{aligned} \tag{10m}$$

which completes our proof that the vectors $\{g_i\}_{i=0}^{n-1}$ are orthogonal and that the vectors $\{h_i\}_{i=0}^{n-1}$ are H-conjugate. This completes our proof. □

Since it follows from Theorem 1.5.5 that the vectors gradient $\{g_i\}_{i=0}^{n-1}$ are orthogonal and *not conjugate*, the commonly used name "Conjugate Gradient Algorithm" for Algorithm 1.4.1 does not appear to be appropriate at first glance. The reason this name has been adopted is that for $i = 0, 1, \ldots, n-1$, h_i is obtained using the Gram-Schmidt H-orthogonalization procedure on the basis $\{-g_0, \ldots, -g_{n-1}\}$.

1.5.3 Formal Extension to General Functions

Thus, for the quadratic problem (1), the Progenitor Conjugate Gradient Algorithm 1.5.4 requires at most n iterations to produce the result which Newton's method obtains in one iteration. On large quadratic problems, these n iterations require less computing effort than the solution of the linear system of optimality equation $Hx = -d$, which yields the optimal solution $\hat{x} = -H^{-1}d$.

Clearly, we cannot expect extensions of conjugate gradient methods to general problems of the form

$$\min_{x \in \mathbb{R}^n} f(x), \tag{11}$$

with $f : \mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable, to do any better than the Progenitor Conjugate Gradient Algorithm 1.5.4 does on quadratic functions, i.e., to be better than n-step quadratically convergent (which is roughly equivalent to being R-superlinearly convergence with rate $2^{1/n}$). Furthermore, on general problems, the exact minimization step-size rule (9a) consumes a lot of computing time. Hence the use of extensions of conjugate gradient methods on large-scale, general problems of the form (11) can be justified only if the calculation of the Hessian matrix can be avoided.

A *formal* extension of the Progenitor Conjugate Gradient Algorithm 1.5.4 to the solution of (11), would set $\gamma_i = \langle H(x_i)h_i , \nabla f(x_{i+1}) \rangle / \langle H(x_i)h_i , h_i \rangle$ (see (9c)), which involves the Hessian matrix. Hence, to obtain a practical extension of Algorithm 1.5.4, one must begin by developing alternative formulas for γ_i, which are equivalent to (9c) for the quadratic case, but which enable one to apply conjugate gradient algorithms to problem (11) without computing Hessians. We will produce three such formulas.

First, *for the quadratic case*, where $f(x) = \frac{1}{2}\langle x , Hx \rangle + \langle d , x \rangle$, with H a symmetric, positive-definite matrix, we see that, according to (9c),

$$\gamma_i = \frac{\langle Hh_i , g_{i+1} \rangle}{\langle Hh_i , h_i \rangle} . \tag{12}$$

Since, by construction,

$$\lambda_i Hh_i = g_{i+1} - g_i , \tag{13}$$

we conclude, from (12) and (13), that

$$\gamma_i = \frac{\langle g_{i+1} - g_i , g_{i+1} \rangle}{\langle g_{i+1} - g_i , h_i \rangle} = -\frac{\langle g_{i+1} - g_i , g_{i+1} \rangle}{\langle g_i , h_i \rangle} , \tag{14}$$

because $\langle g_{i+1} , h_i \rangle = 0$ by construction of λ_i. Now, by (9b), $h_i = -g_i + \gamma_{i-1}h_{i-1}$, and $\langle g_i , h_{i-1} \rangle = 0$, by construction of λ_{i-1}. Hence, from (14), we obtain the *Polak-Ribière* formula for γ, viz.,

$$\gamma_i^{PR} = \frac{\langle g_{i+1} - g_i , g_{i+1} \rangle}{\|g_i\|^2} . \tag{15a}$$

Formula (15a) was first used in [PoR.69]. When used in the Progenitor Conjugate Gradient Algorithm 1.5.4, with $g_i \triangleq \nabla f(x_i)$, it defines the *Polak-Ribière* method of conjugate gradients for solving (11).

Now, again for the quadratic case, as we have seen in the proof of Theorem 1.5.5, $\langle g_i , g_{i+1} \rangle = 0$ for all i, and hence (15a) can be further reduced to yield the *Fletcher-Reeves* formula for γ,

$$\gamma_i^{FR} = \frac{\|g_{i+1}\|^2}{\|g_i\|^2} , \tag{15b}$$

first used in [FlR.64]. When used in the Progenitor Conjugate Gradient Algorithm, with $g_i \triangleq \nabla f(x_i)$, it defines the Fletcher-Reeves method of conjugate gradients for solving (11). The formula for γ favored by Hestenes and Stiefel [HeS.52] was

$$\gamma_i^{HS} = \frac{\langle g_{i+1} - g_i , g_{i+1} \rangle}{\langle g_{i+1} - g_i , h_i \rangle} , \tag{15c}$$

which reduces to the Polak-Ribière formula under exact line searches.

1.5.4 The Polak-Ribière Conjugate Gradient Algorithm

When we replace (9c) with (15a), we obtain the Polak-Ribière Conjugate Gradient Algorithm for solving (11):

Polak-Ribière Conjugate Gradient Algorithm 1.5.6.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$, $g_0 = \nabla f(x_0)$, and $h_0 = -g_0$.

Step 1. Compute the *step-size*

$$\lambda_i = \arg\min_{\lambda \geq 0} f(x_i + \lambda h_i) . \tag{16a}$$

Step 2. Update: Set

$$\begin{cases} x_{i+1} = x_i + \lambda_i h_i , \\ g_{i+1} = \nabla f(x_{i+1}) , \\ \gamma_i^{PR} = \langle g_{i+1} - g_i , g_{i+1} \rangle / \|g_i\|^2 , \\ h_{i+1} = -g_{i+1} + \gamma_i^{PR} h_i . \end{cases} \tag{16b}$$

Step 3. Replace i by $i+1$, and go to Step 1.

Assumption 1.5.7. *We will suppose that $f : \mathbb{R}^n \to \mathbb{R}$ in (11) is twice continuously differentiable and that there exist $0 < m \le M < \infty$ such that, for all $x , y \in \mathbb{R}^n$,*

$$m\|y\|^2 \le \langle y , H(x)y \rangle \le M\|y\|^2 , \tag{17}$$

where $H(x) \triangleq f_{xx}(x)$. □

Theorem 1.5.8. *Suppose that Assumption 1.5.7 is satisfied and that $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by the Polak-Ribière Algorithm 1.5.6 in solving problem (11). Then,*

(a) for all $i \in \mathbb{N}$,

$$\langle \nabla f(x_i) , h_i \rangle \le -\frac{m}{m+M} \|\nabla f(x_i)\| \|h_i\| , \tag{18}$$

and

(b) the sequence $\{x_i\}_{i=0}^{\infty}$ converges to $\hat{x}$, the unique minimizer of $f(\cdot)$.

Proof. (a) Let $g(x) \triangleq \nabla f(x)$. Since $h_0 = -g_0$, it is obvious that (18) holds for $i = 0$. Hence, suppose that $i \ge 0$. Then we obtain

$$\begin{aligned} g_{i+1} = g(x_{i+1}) &= g(x_i + \lambda_i h_i) \\ &= g_i + \lambda_i \int_0^1 H(x_i + s\lambda_i h_i)\, ds\, h_i . \end{aligned} \tag{19a}$$

Since $\langle g_i , h_{i-1} \rangle = 0$, by construction of λ_i, we conclude from (19a) that

$$\lambda_i = -\frac{\langle h_i , g_i \rangle}{\langle h_i , H_i h_i \rangle} = \frac{\langle g_i , g_i \rangle}{\langle h_i , H_i h_i \rangle} , \tag{19b}$$

where

$$H_i = \int_0^1 H(x_i + s\lambda_i h_i)\, ds . \tag{19c}$$

Hence,

$$\begin{aligned} \gamma_i^{PR} &= \frac{\langle g_{i+1} - g_i , g_{i+1} \rangle}{\|g_i\|^2} \\ &= \lambda_i \frac{\langle H_i h_i , g_{i+1} \rangle}{\|g_i\|^2} \\ &= \frac{\langle H_i h_i , g_{i+1} \rangle}{\langle h_i , H_i h_i \rangle} . \end{aligned} \tag{19d}$$

Therefore,

$$|\gamma_i^{PR}| \le M\|g_{i+1}\| / m\|h_i\| . \tag{19e}$$

Hence,

$$\|h_{i+1}\| \le \|g_{i+1}\| + |\gamma_i^{PR}|\, \|h_i\| \le \|g_{i+1}\|(1 + M/m) . \tag{19f}$$

Finally,

$$\langle g_{i+1} , h_{i+1} \rangle = \langle g_{i+1} , -g_{i+1} + \gamma_i^{PR} h_i \rangle = -\|g_{i+1}\|^2 . \tag{19g}$$

Consequently, making use of (19f), we obtain

$$\frac{\langle g_{i+1} , h_{i+1} \rangle}{\|g_{i+1}\| \|h_{i+1}\|} = \frac{\|g_{i+1}\|}{\|h_{i+1}\|} \le -\frac{1}{1 + M/m} = -\frac{m}{m+M} , \tag{19h}$$

which completes the proof of part *(a)*.

(b) The fact that the sequence $\{x_i\}_{i=0}^{\infty}$ converges to the unique minimizer of $f(\cdot)$ follows from the Wolfe Theorem 1.2.21, where we set $\kappa(x_i) = m/(m+M)$, for all i. □

It is quite easy to show that the Polak-Ribière conjugate gradient algorithm converges at least linearly. However, the bound on its rate of convergence obtained in the theorem below is quite pessimistic since it wrongly suggests that the Polak-Ribière conjugate gradient algorithm may converge more slowly than the method of steepest descent.† The problem lies in the technique used below to establish a lower bound on the rate of convergence.

Theorem 1.5.9. *Suppose that Assumption 1.5.7 is satisfied. If $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by the Polak-Ribière Algorithm 1.5.6 in solving problem (11), then*

† Exercise 10 in Section 8.8 in [Lue.84] suggests that one should be able to establish that $f(x_i - f(\hat{x}) \le 4[(1 - \sqrt{\gamma})/(1 + \sqrt{\gamma})]^{2i} [f(x_0 - f(\hat{x})]$, where $\gamma \triangleq m/M$. This is a much better root rate constant than the one for the method of steepest descent.

(a) $x_i \to \hat{x}$ *as* $i \to \infty$, *with* $\hat{x}$ *the unique minimizer of* $f(\cdot)$, *and*

(b) for all $i \in \mathbb{N}$,

$$f(x_{i+1}) - f(\hat{x}) \leq \delta[f(x_i) - f(\hat{x})], \tag{20a}$$

$$\|x_i - \hat{x}\| \leq \left[\frac{2}{m}[f(x_0) - f(\hat{x})]\right]^{1/2} (\delta^{1/2})^i, \tag{20b}$$

where δ, *the linear rate of convergence constant, is given by*

$$\delta = \left[1 - \rho^2 \frac{m}{M}\right], \tag{20c}$$

with $\rho \triangleq m/(m+M)$.

Proof. *(a)* This part was established in Theorem 1.5.8, and hence we turn to *(b)*. Making use of the second-order expansion formula (5.1.17d) and of (18), we deduce that, for any x_i and $\lambda > 0$,

$$f(x_i + \lambda h_i) - f(x_i) \leq \lambda \langle \nabla f(x_i), h_i \rangle + \lambda^2 \frac{M}{2} \|h_i\|^2$$

$$\leq -\lambda\rho \|\nabla f(x_i)\| \|h_i\| + \lambda^2 \frac{M}{2} \|h_i\|^2, \tag{21a}$$

where $\rho \triangleq m/(m+M)$. The right-hand side of (21a) is minimized by $\lambda = \lambda_i'$ with $\lambda_i' \|h_i\| = \rho \|\nabla f(x_i)\|/M$. Hence we must have that

$$f(x_{i+1}) - f(x_i) \leq f(x_i + \lambda_i' h_i) - f(x_i) \leq -\frac{\rho^2}{2M} \|\nabla f(x_i)\|^2, \tag{21b}$$

for all $i \in \mathbb{N}$. Now, by Lemma 1.3.6, for all $i \in \mathbb{N}$,

$$f(\hat{x}) - f(x_i) \geq -\frac{1}{2m} \|\nabla f(x_i)\|^2. \tag{21c}$$

Substituting for $\|\nabla f(x_i)\|^2$ in (21b) from (21c), we conclude that, for all $i \in \mathbb{N}$,

$$f(x_{i+1}) - f(x_i) \leq \rho^2 \frac{m}{M} [f(\hat{x}) - f(x_i)]. \tag{21d}$$

Subtracting $f(\hat{x}) - f(x_i)$ from both sides of (21d) and rearranging terms, we get (20a).

Since $0 < (1 - \rho^2 m/M) < 1$, it follows from (20a) by recursion that, for all $i \in \mathbb{N}$,

$$0 < f(x_i) - f(\hat{x}) \leq \delta^i [f(x_0) - f(\hat{x})], \tag{21e}$$

with $\delta = 1 - \rho^2 m/M$. Hence, making use of (1.3.5b) we obtain (20b), which completes our proof. □

1.5.5 The Fletcher-Reeves Conjugate Gradient Method

When we replace (9c) with (15b), we obtain the Fletcher-Reeves Conjugate Gradient Algorithm 1.4.10, below, for solving (11). The Fletcher-Reeves method does not appear to satisfy the assumptions of the Wolfe Theorem 1.2.21 and hence seems to show less tolerance to errors in the step-size calculations. However, while the convergence Theorem 1.5.8 seems to limit the Polak-Ribière method to convex functions, as we will soon see, the Fletcher-Reeves method has no such limitation, though on nonconvex functions, the convergence theorem that we get is weaker than the one for the method of steepest descent, in the sense that one can only prove that a bounded sequence $\{x_i\}_{i=0}^{\infty}$, constructed by the Fletcher-Reeves Algorithm 1.5.10, must have at least one accumulation point $\hat{x}$ satisfying $\nabla f(\hat{x}) = 0$. Such a statement does not rule out the possibility that the sequence $\{x_i\}_{i=0}^{\infty}$ also has an accumulation point x^* such that $\nabla f(x^*) \neq 0$.

Fletcher-Reeves Conjugate Gradient Algorithm 1.5.10.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$, $g_0 = \nabla f(x_0)$, and $h_0 = -g_0$.

Step 1. Compute the *step-size*

$$\lambda_i = \arg\min_{\lambda \geq 0} f(x_i + \lambda h_i). \tag{22a}$$

Step 2. Update: Set

$$\begin{cases} x_{i+1} = x_i + \lambda_i h_i, \\ g_{i+1} = \nabla f(x_{i+1}), \\ \gamma_i^{FR} = \|g_{i+1}\|^2 / \|g_i\|^2, \\ h_{i+1} = -g_{i+1} + \gamma_i^{FR} h_i. \end{cases} \tag{22b}$$

Step 3. Replace i by $i+1$, and go to Step 1.

The following result is due to Zoutendijk [Zou.70].

Theorem 1.5.11. *Suppose that*

(i) $x_0 \in \mathbb{R}^n$ *is given,*

(ii) the function $f : \mathbb{R}^n \to \mathbb{R}$ *in (11) is twice continuously differentiable, and*

(iii) the level set $L_{f(x_0)}(f) \triangleq \{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\}$ *is bounded. If* $\{x_i\}_{i=0}^{\infty}$ *is a sequence constructed by the Fletcher-Reeves conjugate gradient method, then*

(a) for $i = 1,2,3,\ldots,$

$$\|h_i\|^2 = \|g_i\|^4\left[\sum_{j=0}^{i}\frac{1}{\|g_j\|^2}\right], \tag{23a}$$

(b) for $i = 1,2,3,\ldots,$

$$\frac{\langle h_i, g_i\rangle}{\|h_i\|\|g_i\|} = -\|g_i\|^{-1}\left[\sum_{j=0}^{i}\frac{1}{\|g_j\|^2}\right]^{-\frac{1}{2}}, \tag{23b}$$

and

(c) the sequence $\{x_i\}_{i=0}^{\infty}$ *must have at least one accumulation point* $\hat{x}$ *such that* $\nabla f(\hat{x}) = 0$, *i.e,* $\underline{\lim}_{i\to\infty}\|g_i\| = 0$.

Proof. (a) We proceed by induction. Since, by the construction of the step-size, $\langle g_{i+1}, h_i\rangle = 0$ for all $i \in \mathbb{R}^n$, we conclude, from (22b), that

$$\|h_{i+1}\|^2 = \|g_{i+1}\|^2 + \frac{\|g_{i+1}\|^4}{\|g_i\|^4}\|h_i\|^2, \ i = 0,1,2,\ldots \tag{24a}$$

and in particular, since $h_0 = -g_0$,

$$\|h_1\|^2 = \|g_1\|^4\left[\sum_{j=0}^{1}\frac{1}{\|g_j\|^2}\right]. \tag{24b}$$

Now suppose that (23a) holds for $i = k$. Then, making use of (24a) and (23a), we find that

$$\begin{aligned}\|h_{k+1}\|^2 &= \|g_{k+1}\|^2 + \frac{\|g_{k+1}\|^4}{\|g_k\|^4}\|h_k\|^2\\ &= \|g_{k+1}\|^2 + \frac{\|g_{k+1}\|^4}{\|g_k\|^4}\|g_k\|^4\left[\sum_{j=0}^{k}\frac{1}{\|g_j\|^2}\right].\end{aligned} \tag{24c}$$

Rearranging terms, we see that (23a) holds for $i = k+1$, which completes this part of the proof.

(b) Next, since by (22b), for all $i \in \mathbb{N}$, $\langle h_i, g_i\rangle = \langle -g_i + \gamma_{i-1}^{FR}h_{i-1}, g_i\rangle$ and $\langle g_i, h_{i-1}\rangle = 0$, it follows that

$$\langle h_i, g_i\rangle = -\|g_i\|^2, \tag{24d}$$

and hence (23b) follows directly from (23a).

(c) Let

$$t_i \triangleq \left[\sum_{j=0}^{i}\frac{1}{\|g_j\|^2}\right]^{-\frac{1}{2}}, \ i = 1,2,3,\ldots, \tag{24e}$$

so that

$$\langle h_i, g_i\rangle = -t_i\|h_i\|, \ i = 1,2,3,\ldots \tag{24f}$$

Now, since by assumption the level set $L_{f(x_0)}(f)$ is bounded, there exists an $M < \infty$ such that $\|H(x)\| \le M$ for all $x \in L_{f(x_0)}(f)$ (where $H(x) = f_{xx}(x)$, as before). Hence, using the second-order expansion formula (5.1.17d), we conclude that (with $s \in [0,1]$),

$$\begin{aligned}f(x_{i+1}) - f(x_i) &\le \min_{\lambda\ge 0}\{\lambda\langle h_i, g_i\rangle + \tfrac{1}{2}\lambda^2\langle h_i, H(x_i - s\lambda h_i)h_i\rangle\}\\ &\le \min_{\lambda\ge 0}\{-\lambda t_i\|h_i\| + \lambda^2\frac{M}{2}\|h_i\|^2\}.\end{aligned} \tag{24g}$$

Since the right-hand side is minimized by $\lambda = \lambda_i'$, where $\lambda_i'\|h_i\| = t_i/M$, we conclude that

$$f(x_{i+1}) - f(x_i) \le -\frac{t_i^2}{2M}, \ i = 0,1,2,\ldots \tag{24h}$$

Now, for the sake of contradiction, suppose that $\underline{\lim}_{i\to\infty}\|g_i\| > 0$. Then there exists an $\alpha > 0$ such that $\|g_i\|^2 \ge \alpha$ for all $i \in \mathbb{N}$, and hence $t_i^2 \ge \alpha/(1+i)$ for all $i \in \mathbb{N}$. Now, for any $k \in \mathbb{N}$,

$$f(x_k) - f(x_0) = \sum_{i=0}^{k-1} f(x_{i+1}) - f(x_i) \le -\frac{1}{2M}\sum_{i=0}^{k-1} t_i^2. \tag{24i}$$

Since $\sum_{i=0}^{k-1} t_i^2 \to \infty$, as $k \to \infty$, we conclude that $f(x_k) \to -\infty$, as $k \to \infty$. However, by assumption, the level set $L_{f(x_0)}(f)$ is bounded and the function $f(\cdot)$ is continuous. Consequently, $f(x)$ is bounded on $L_{f(x_0)}(f)$, which results in a contradiction, concluding the proof of *(c)*. □

Corollary 1.5.12. *Suppose that the Assumption 1.5.7 is satisfied. If* $\{x_i\}_{i=0}^{\infty}$ *is a sequence constructed by the Fletcher-Reeves conjugate gradient method, then* $x_i \to \hat{x}$, *the unique solution of (11).*

Proof. Since, under Assumption 1.5.7, the level sets of $f(\cdot)$ are compact, it is clear that the sequence $\{x_i\}_{i=0}^{\infty}$ must have accumulation points. Since $\hat{x}$ is the only point such that $\nabla f(\hat{x}) = 0$, it follows that $\hat{x}$ must be one of the accumulation points. The fact that $\hat{x}$ is the only accumulation point now follows from the fact that $f(x_i) \to f(\hat{x})$ as $i \to \infty$ (because the cost sequence is monotone decreasing), and the fact that $f(x) > f(\hat{x})$ for all $x \ne \hat{x}$. □

Remark 1.5.13. It is clear from the above theorem that it is distinctly possible in the Fletcher-Reeves method for the angle between the gradient g_i at x_i and the search direction h_i to approach 90° , as $i \to \infty$, while, in the Polak-Ribière method, this angle is well bounded away from 90°. As we have already

mentioned, this fact makes the Polak-Ribière method somewhat less sensitive to numerical errors. □

1.5.6 Partial Conjugate Gradient Methods

We note that even in the quadratic case, the finite convergence of the conjugate gradient methods depends on setting $h_0 = -g_0$. Now, near a minimizer $\hat{x}$ which satisfies the second-order sufficiency condition (Theorem 1.1.8), a function $f(\cdot)$ does have a reasonably good quadratic approximation, and it may be conjectured that, near the minimizer $\hat{x}$, if we reinitialize a conjugate gradient method by setting $h_i = -g_i$, from time to time, we might get better performance than by constructing h_i by one of the standard formulas at each iteration. There are two interesting results dealing with this case that we include without proofs. The Theorem below combines results in [Coh.72] and in [Lue.84] into a single statement. The second result (*(b)* below) is useful only when $M \gg m$.

Theorem 1.5.14. *Suppose that (i) Assumption 1.5.7 is satisfied and (ii) that the Polak-Ribière or Fletcher-Reeves conjugate gradient algorithm is modified so that for a given $k \in \mathbb{N}$, $k \geq 1$, whenever $(i+1)/k$ is an integer, h_{i+1} is constructed according to*

$$h_{i+1} = -g_{i+1}\,, \tag{25a}$$

and according to the original formula

$$h_{i+1} = -g_{i+1} + \gamma_i h_i\,, \tag{25b}$$

otherwise, where $\gamma_i = \gamma_i^{PR}$ or $\gamma_i = \gamma_i^{FR}$, as appropriate.

(a) If $k = n$, then there exist an $i_0 \in \mathbb{N}$ and a constant $c < \infty$ such that, for all $i \geq i_0$,

$$\|x_{n(i+1)} - x_{ni}\| \leq c\,\|x_{ni} - x_{n(i-1)}\|^2\,, \tag{25c}$$

i.e., $x_i \to \hat{x}$, as $i \to \infty$, n-step Q-quadratically, where $\hat{x}$ is the global minimizer of $f(\cdot)$.

(b) If $k < n$, then there exists an $i_0 \in \mathbb{N}$ such that, for all $i \geq i_0$,

$$f(x_{k(i+1)}) - f(\hat{x}) \leq \left[\frac{b-a}{b+a}\right]^2 [f(x_{ki}) - f(\hat{x})]\,, \tag{25d}$$

and

$$\|x_{ki} - \hat{x}\| \leq \left[\frac{2}{m}\right]^{1/2} \left(f(x_0) - f(\hat{x})\right)^{1/2} \left[\frac{b-a}{b+a}\right]^i\,, \tag{25e}$$

where $m = a < b < M$ are such that $(n-k)$ eigenvalues of $H(\hat{x})$ are contained in the interval $[a, b]$ and the remaining k eigenvalues are larger than b.

(Thus the effect of the k worst eigenvalues has been removed.) □

Note that (25e) implies that $x_i \to \hat{x}$, as $i \to \infty$, linearly, with a rate constant $\delta = [(b-a)/(b+a)]^{1/k}$.

1.5.7 Notes

As we have already mentioned, methods of conjugate directions were introduced by Hestenes and Stiefel in 1952 (see [HeS.52]) as a technique for solving large systems of linear equations. Subsequently, methods of conjugate directions were extended to the solution of nonquadratic unconstrained optimization problems by Fletcher and Reeves [FlP.63] and Polak and Ribière [PoR.69]. Since then, the literature on conjugate gradient methods has become very large, though the practical utility of many of the proposed variants is unclear. Overviews of the literature can be found in [Sto.77, NaN.82, GiL.89]. Variants that are important from a practical point of view include those that use scaling and preconditioning, see, e.g., [GiL.89, GiN.92, GiM.79, Sha.78, Lue.84], and inexact line searches, see [Naz.79, Dix.75, Al-.85, GiN.92].

A variation of the restart procedure for partial conjugate gradient methods is given in [Pow.77]. Limited storage versions for very large problems were presented in [Naz.76, NaN.82, Sha.78, BuL.83, KLS.92].

Results on bounds on the linear rate of convergence of conjugate gradient methods, other than the ones presented in this section, can be found in [CrW.72, Pow.76, Sto.77, BaS.77].

At present, variants of the Fletcher-Reeves conjugate gradient method and of the Polak-Ribière conjugate gradient method are the most used, with a preference given to variants of the Polak-Ribière method. On problems of large dimension, when implemented with an inexact line search, their efficiency is considerably better than that of Newton's method, because they do not require computing the Hessian matrix $f_{xx}(x)$. In numerical comparisons, enhanced versions of the Polak-Ribière Algorithm 1.5.6 have been found to be substantially better than the enhanced versions of Fletcher-Reeves Algorithm 1.5.10 (e.g., see the tables in [GiN.92]). This superior numerical performance is explained by Powell [Pow.86] as due to the fact that when the angle between $\nabla f(x_i)$ and h_i tends to 90°, $x_{i+1} \approx x_i$ and $g_{i+1} \approx g_i$ hold. In this case $\gamma_i^{PR} \approx 0$ which causes the Polak-Ribière method to reset itself, while γ_i^{FR} tends to unity. In addition, if h_i is very large relative to g_i, then $h_{i+1} \approx h_i$ holds, and hence a preservation of a bad search direction is the result.

It was also shown by Powell in [Pow.84a] that on a particular nonconvex optimization problem the Polak-Ribière Algorithm 1.5.6 fails to converge. Hence the strict convexity assumption in Theorem 1.5.8 cannot be weakened, if the conclusions of that theorem are to remain valid. It was suggested by Powell in [Pow.86] that, if the the Polak-Ribière Algorithm 1.5.6 were modified by replacing γ_i^{PR} by $\gamma_{+i}^{PR} \triangleq \max\{0, \gamma_i^{PR}\}$, then it would also converge on nonconvex functions. In fact, Gilbert and Nocedal were able to prove the following result in [GiN.92].

Theorem 1.5.15. *Suppose that*

(i) the cost function $f(\cdot)$ in (11) is twice continuously differentiable and that its level sets are bounded, and

(ii) in Algorithm 1.5.5, γ_i^{PR} is replaced by γ_{+i}^{PR} and the step-size λ_i, defined by (16a), is replaced by a step-size λ_i satisfying the following three inequalities:

$$f(x_i + \lambda_i h_i) - f(x_i) \leq \alpha_1 \lambda_i \langle g_i , h_i \rangle , \tag{26a}$$

$$\langle \nabla f(x_i + \lambda_i h_i), h_i \rangle \geq \alpha_2 \langle g_i , h_i \rangle , \tag{26b}$$

$$\langle g_i , h_i \rangle \leq -\alpha_3 \|g_i\|^2 , \tag{26c}$$

where $0 < \alpha_1 < \alpha_2 < 1$ and $\alpha_3 \in (0, 1]$.

If $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by the thus modified Algorithm 1.5.6 in solving problem (11), then $\underline{\lim} \|\nabla f(x_i)\| = 0$. □

The inequalities (26a,b) are known as the Wolfe step-size rule [Wol.69, Wol.71] and can be satisfied by using a sufficiently large β in the Armijo Model Algorithm 1.2.23, with $s = 0$. Next, in view of the fact that

$$h_i = -g_i + \gamma_{+i}^{PR} h_{i-1} , \tag{27a}$$

we conclude that

$$\langle g_i , h_i \rangle = -\|g_i\|^2 + \gamma_{+i}^{PR} \langle g_i , h_{i-1} \rangle . \tag{27b}$$

Hence for (26c) to hold, the second term in (27b) must be small, i.e., λ_{i-1} must be fairly close to a local minimizer of $f(x_{i-1} + \lambda h_{i-1})$. Some sketching shows that when α_1 is close to unity, (26a,b) and (26c) may be inconsistent conditions. Hence α_1 must be chosen small, and a combination of a back stepping procedure, as in the Armijo Model Algorithm 1.2.23, with $s = 0$, and a one-dimensional optimization algorithm (such as the secant method in Section 1.7) must be used in computing a step-size satisfying (26a,b,c). Other appropriate one-dimensional minimization schemes can be found in [Lem.81, MoT.94].

1.6 Quasi-Newton Methods

Quasi-Newton methods are methods of unconstrained optimization which approximate the Newton search direction, usually without evaluating second-order derivatives of the cost function. As a class, they include the Iterated Newton-Armijo Algorithm 1.4.19, the Discrete Newton-Armijo Algorithm 1.4.12, variations of this method, called secant methods, that we will see in this section, and the so-called variable metric methods, based on a secant principle, which use rank-one or rank-two updates to improve a current estimate of the Hessian of the cost function.

We begin this section with an explanation of the variable metric concept, followed by an introduction to secant methods, which are the simplest and most easily analyzed examples of variable metric methods. They also provide insights into the more efficient variable metric methods. One of the secant methods that we will describe can be globally stabilized in the same way as Newton's method and then used to solve both convex and nonconvex problems.

Then we proceed with a description of symmetric, rank-one and rank-two updates for generating approximations to second derivatives. These approximations are much less straightforward and less obvious than secant approximations and require separate treatment of quadratic functions and general convex functions. We conclude this section by showing that an algorithm using the BFGS rank-two updates and an Armijo type step-size rule converges superlinearly on strictly convex functions.

In this section, we will restrict ourselves to problems of the form

$$\min_{x \in \mathbb{R}^n} f(x) , \tag{1}$$

where the function $f(\cdot)$ satisfies Assumption 1.4.7, which was also required for the local Newton method. For convenience, we repeat this assumption here in a slightly stronger form:

Assumption 1.6.1.

(a) The function $f : \mathbb{R}^n \to \mathbb{R}$ is twice Lipschitz continuously differentiable on bounded sets.

(b) There exist $0 < m \leq M < \infty$ such that

$$m\|y\|^2 \leq \langle y , f_{xx}(x) y \rangle \leq M\|y\|^2 , \quad \forall\, x , y \in \mathbb{R}^n . \tag{2}$$

□

We recall that, in view of Assumption 1.6.1, the level sets of $f(\cdot)$ are convex and compact (see Proposition 5.2.15) and hence, by Corollary 1.1.12, problem (1) has a unique solution $\hat{x}$.

1.6.1 The Variable Metric Concept

Consider problem (1), and let Q be a positive-definite, symmetric $n \times n$ matrix. Now suppose that we carry out a change of variables, setting $x = Q^{-1/2} z$. Then, in the new coordinate system, problem (1) becomes

$$\min_{z \in \mathbb{R}^n} F(z), \tag{3a}$$

where the function $F(\cdot)$ is defined by $F(z) \triangleq f(Q^{-1/2} z)$. The steepest descent direction for $F(\cdot)$ at $z = Q^{1/2} x$ is given by $d(z) = -\nabla F(z) = -Q^{-1/2} \nabla f(x)$, which, when transformed back to the original coordinate system, becomes $h(x) = -Q^{-1} \nabla f(x)$. We note that this search direction can also be obtained by solving the problem

$$h(x) = \arg \min_{h \in \mathbb{R}^n} \{ \langle \nabla f(x) , h \rangle + \tfrac{1}{2}\|h\|_Q^2 \} , \tag{3b}$$

where $\|h\|_Q^2 \triangleq \langle h , Qh \rangle$, i.e., by computing the steepest descent direction of $f(\cdot)$ with respect to the Q-norm. It is common to refer to the Q-norm as a *Q-metric*,

since it does define a metric, though the reasons for the preference of the term *metric* over *norm* are not clear.

To simplify notation, we denote again the Hessian of $f(\cdot)$ by $H(\cdot)$, i.e., $H(\cdot) \triangleq f_{xx}(\cdot)$. Now suppose that $H(\hat{x})$ is known, and suppose that we modify the Armijo Gradient Algorithm 1.3.3 so that it computes the search direction according to (3b), with the *fixed* metric determined by $Q = H(\hat{x})$, rather than with $Q = I$, which was used to obtain $h(x) = -\nabla f(x)$ in (1.3.3a). Then the sequence that this modified Armijo Gradient Algorithm constructs in solving (1) satisfies the recursion

$$x_{i+1} = x_i - \lambda_i H(\hat{x})^{-1} \nabla f(x_i), \quad i = 0, 1, 2, 3, \ldots \tag{3c}$$

If we carry out an analysis for the Newton-Armijo Algorithm 1.4.8, similar to that in the proof of Theorem 1.4.9, we find that there must exist an i_0 such that $\lambda_i = 1$ for all $i \geq i_0$, and hence that for all $i \geq i_0$,

$$\begin{aligned} \|x_{i+1} - \hat{x}\| &\leq \|H(\hat{x})^{-1}\| \int_0^1 \|H(\hat{x}) - H(\hat{x} + s(x_i - \hat{x}))\| \, ds \, \|x_i - \hat{x}\| \\ &\leq \frac{L}{2m} \|x_i - \hat{x}\|^2 , \end{aligned} \tag{3d}$$

which implies that $\{x_i\}_{i=0}^{\infty}$ converges Q-quadratically to $\hat{x}$. Thus we see that a change in metric can transform a linearly converging algorithm, such as the Armijo Gradient Algorithm, into a quadratically converging algorithm.

Now, the Hessian $H(\hat{x})$ is not usually known in advance. Therefore Newton's method uses a *variable* metric, defined at x_i by $Q_i = H(x_i)$. We have seen that this variable metric converges to the ideal one fast enough not to affect the rate of convergence of Newton's method.

Clearly, the Discrete Newton-Armijo Algorithm 1.4.12 and the Iterated Newton-Armijo Algorithm 1.4.19 that we saw in Section 1.4 are also variable metric algorithms in the above sense.

Many of the algorithms that we will see in this section make use of the labor saving Sherman-Morrison formula for inverting the sum of a matrix with a dyad.† We state this result below; it can be easily verified by multiplying the proposed inverse by the original matrix.

Proposition 1.6.2. *Suppose that A is a nonsingular, $n \times n$ matrix and that $a, b \in \mathbb{R}^n$ are such that $1 + \langle b, A^{-1}a \rangle \neq 0$. Then*

† A *dyad* is an $n \times n$, rank-one matrix of the form $D = ab^T$, where $a, b \in \mathbb{R}^n$.

$$[A + ab^T]^{-1} = A^{-1} - \frac{1}{1 + \langle b, A^{-1}a \rangle} (A^{-1}a)(b^T A^{-1}) . \tag{4}$$

□

1.6.2 Secant Methods

Secant algorithms are variable metric algorithms characterized by the fact that they use finite differences to obtain an approximation to the Hessian of $f(\cdot)$. Thus, the Discrete Newton-Armijo Algorithm 1.4.12 is a member of this class. The original secant algorithm is about as old as Newton's algorithm and was designed for solving equations of the form

$$g(x) = 0 , \tag{5a}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a Lipschitz continuously differentiable on bounded sets, whose derivative $g_x(x)$ is bounded away from 0 near a solution $\hat{x}$ of (5a). It requires *two* starting points x_0 and x_1 for initialization. Assuming that it has already constructed $x_0, x_1, \ldots, x_{i-1}, x_i$, it constructs x_{i+1} according to the formula

$$x_{i+1} = x_i - \{[g(x_i) - g(x_{i-1})] / (x_i - x_{i-1})\}^{-1} g(x_i) , \tag{5b}$$

i.e., it replaces $g_x(x_i)^{-1}$ in (1.4.2c) by a finite difference approximation or, more precisely, by $g_x(x_{i-1} + s(x_i - x_{i-1})^{-1}$, where $s \in [0, 1]$ (see Fig. 1.6.1). Proceeding as in the analysis of Newton's algorithm and using Theorem 1.2.43, one can show that, if initialized with points sufficiently close to a solution $\hat{x}$, the "scalar" secant algorithm constructs a sequence that converges R-superlinearly to $\hat{x}$, with rate at least equal to $\tau_1 = 1.618$.

We will call the natural "vector" extension of the above "scalar" secant algorithm the classical secant algorithm, as distinct from the more labor intensive "conservative" secant algorithm to be introduced shortly. The classical secant algorithm can be used for solving equations of the form (5a), where $g : \mathbb{R}^n \to \mathbb{R}^n$ is a locally Lipschitz continuously differentiable function whose Jacobian $g_x(x)$ is nonsingular and whose inverse $g_x(x)^{-1}$ is bounded near a solution $\hat{x}$ of (5a). To simplify notation, we will use the definition $H(x) \triangleq g_x(x)$ for all $x \in \mathbb{R}^n$.

The classical secant algorithm requires $n+1$ linearly independent vectors $x_0, \ldots, x_n$ for initialization. Then, given that it has already constructed the points $x_{i-n}, x_{i-n+1}, \ldots, x_{i-1}, x_i$, at which the values $g_j \triangleq g(x_j)$, $j = i-n, i-n+1, \ldots, i$, have been computed, it forms two sets of difference vectors, each consisting of n elements

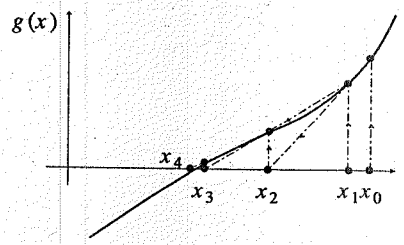


Fig. 1.6.1. Behavior of the scalar secant method.

$$\left.\begin{aligned} \Delta x_j &\triangleq x_{j+1} - x_j\,, \\ \Delta g_j &\triangleq g_{j+1} - g_j\,, \end{aligned}\right\} \tag{5c}$$

$j = i-n\,, i-n+1\,, \ldots, i-1$, and uses them to construct an approximation H_i to $H(x_i)$, which satisfies the relation

$$H_i \Delta x_j = \Delta g_j\,,\ j = i-n, \ldots, 1-1\,, \tag{5d}$$

by solving the equation

$$H_i \Xi_{i-1} = \Gamma_{i-1}\,, \tag{5e}$$

where $\Xi_{i-1} = [\Delta x_{i-n}, \ldots, \Delta x_{i-1}]$ is an $n \times n$ matrix with columns Δx_j and $\Gamma_{i-1} = [\Delta g_{i-n}, \ldots, \Delta g_{i-1}]$ is an $n \times n$ matrix with columns Δg_j, $j = i-n\,, i-n+1\,, \ldots, i-1$. Assuming that Ξ_i and Γ_i are both nonsingular, (5e) determines H_i and H_i^{-1} uniquely:

$$\left.\begin{aligned} H_i &= \Gamma_{i-1}\Xi_{i-1}^{-1}\,, \\ H_i^{-1} &= \Xi_{i-1}\Gamma_{i-1}^{-1}\,. \end{aligned}\right\} \tag{5f}$$

The next point x_{i+1} is now computed according to the Newtonian formula

$$x_{i+1} = x_i - H_i^{-1} g(x_i)\,. \tag{5g}$$

One can use the fact that

$$\Delta g_j = \int_0^1 H(x_j + s\Delta x_j)\, ds\, \Delta x_j \tag{5h}$$

to establish error bounds on $\|H_i - H(x_i)\|$. Assuming that the matrices H_i, remain nonsingular, with inverses that are bounded, it can be shown, roughly in the same way as for the local Iterated Newton Algorithm (1.4.21a) in the proof of Theorem 1.4.22 that, if the classical secant algorithm is initialized with points sufficiently close to a solution $\hat{x}$ of (5a), then the sequence $\{x_i\}_{i=0}^{\infty}$,

constructed by the classical secant algorithm, is well defined and converges R-superlinearly to $\hat{x}$, with rate $r \geq \tau_n$, where τ_n is defined by (1.2.42c) in Theorem 1.2.43.

The main disadvantage of the classical secant algorithm is that the matrices H_i can (and have been observed to) become singular, which causes the algorithm to crash. This fact is also inherited by some of the "rank-one" variable metric algorithms.

We now return to problem (1) and associate the gradient $\nabla f(x)$ with the function $g(x)$ in (5a). The collapse of invertibility of the secant matrices H_i can be avoided, at least near the solution $\hat{x}$ of problem (1), by carrying out an extra gradient evaluation at each iteration, as in the following conservative secant algorithm which is a modification of the Discrete Newton-Armijo Algorithm 1.4.12.

The conservative secant algorithm makes use of the following observations. First, for $j = 1, \ldots, n$, let e_j denote the jth column of the $n \times n$ identity matrix, and let Q be the $n \times n$ matrix defined by

$$Q \triangleq [e_2, e_3, \ldots, e_n, e_1]\,. \tag{6a}$$

If we set $v_0 = e_1$ and define $v_{i+1} = Qv_i$ for $i = 0, 1, 2, \ldots$, then we find that the sequence $\{v_i\}_{i=0}^{\infty}$ cycles through the columns of the identity matrix. Second, given x_i, if we have a current estimate H_{i-1}, of $H(x_i)$, we can update the estimate of one of its columns using the formula

$$H_i' = H_{i-1} + (\eta_i - H_{i-1}v_i)v_i^T\,, \tag{6b}$$

where, for $\varepsilon_i > 0$,

$$\eta_i = \frac{1}{\varepsilon_i}[\nabla f(x_i + \varepsilon_i v_i) - \nabla f(x_i)]\,. \tag{6c}$$

Assuming that $v_i = e_k$, we see that $\eta_i^j = (\partial^2 f^j(x_i + s_{jk}\varepsilon_i e_k)/\partial x^j x^k)$, $j = 1, \ldots, n$, with $s_{jk} \in [0, 1]$, i.e., it is an approximation to the jth element of the kth column of $H(x_i)$.

Now, H_i' is not symmetric. Hence we use $H_i \triangleq \frac{1}{2}[H_i' + H_i'^T]$ as our updated estimate of $H(x_i)$, i.e., we simultaneously update a column and the corresponding row of our estimate. Note that

$$H_i = H_{i-1} + \tfrac{1}{2}v_i(\eta_i^T - v_i^T H_{i-1}) + \tfrac{1}{2}(\eta_i - H_{i-1}v_i)v_i^T\,, \tag{6d}$$

and hence that H_i^{-1} can be obtained from H_{i-1}^{-1} by two applications of the Sherman-Morrison formula (4).

Conservative Secant Algorithm 1.6.3.

Parameters. $\alpha \in (0, \frac{1}{2})$, $\beta, \gamma \in (0,1)$, $\varepsilon_{-1} > 0$, $s_{\max} \gg 1$.

Data. $x_0 \in \mathbb{R}^n$, H_{-1}, a symmetric, positive-definite matrix.

Step 0. Set $i = 0$, $v_0 = e_n$. psc..TS

Step 1. If $\nabla f(x_i) = 0$, stop.
Else compute

$$\varepsilon_i = \min\{\varepsilon_{i-1}, \|\nabla f(x_i)\|\}. \tag{7a}$$

Step 2. Compute the approximate Hessian H_i according to (6b), (6c), and (6d).

Step 3. If H_i^{-1} exists, compute the *secant search direction*

$$h_i' \triangleq -H_i^{-1}\nabla f(x_i), \tag{7b}$$

and go to Step 4.
Else, set $h_i = -\nabla f(x_i)$, and go to Step 5.

Step 4. If $\|h_i'\| \le s_{\max}\|\nabla f(x_i)\|$ and $\langle h_i', \nabla f(x_i)\rangle < 0$, compute

$$\delta_i \triangleq \frac{1}{\|\nabla f(x_i)\|^2}\{\langle h_i', \nabla f(x_i)\rangle + \min\{\gamma, \|\nabla f(x_i)\|^3\}\}_+, .sp{-}0.2 \tag{7c}$$

(where for any $a \in \mathbb{R}$, $a_+ \triangleq \max\{0, a\}$) and set

$$h_i = \frac{1}{1+\delta_i}[h_i' - \delta_i \nabla f(x_i)]. \tag{7d}$$

Else, set $h_i = -\nabla f(x_i)$.

Step 5. Compute the *Armijo step-size*

$$\lambda_i = \max_{k \in \mathbb{N}}\{\beta^k \mid f(x_i + \beta^k h_i) - f(x_i) \le \alpha\beta^k \langle \nabla f(x_i), h_i\rangle\}. \tag{7e}$$

Step 6. Set $x_{i+1} = x_i + \lambda_i h_i$, set $v_{i+1} = Pv_i$, replace i by $i+1$, and go to Step 1.

The algorithm above illustrates an alternative to the global stabilization scheme used in Algorithm 1.4.12. Note that, rather than just switching to the steepest descent direction, when the secant search direction may not satisfy the tests in the Polak-Sargent-Sebastian Theorem 1.2.24b, the above algorithm constructs a search direction that combines both of these directions.

Exercise 1.6.4. Note that the definitions of δ_i in (7c) and of h_i in (7d) ensure that

$$\langle h_i, \nabla f(x_i)\rangle \le -\frac{1}{1+\delta_i}\min\{\gamma, \|\nabla f(x_i)\|^3\}. \tag{8}$$

Use the Polak-Sargent-Sebastian Theorem 1.2.24b, the fact that $\delta_i = 0$ for all i sufficiently large, and the rate of convergence Theorem 1.2.41 to establish the following result.

Theorem 1.6.5. *Suppose that Assumption 1.6.1 is satisfied and that $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by the Conservative Secant Algorithm 1.6.3 in solving problem (1). Then, (a) there is an i_0 such that $\lambda_i = 1$ for all $i \ge i_0$, and (b) $x_i \to \hat{x}$, as $i \to \infty$, R-superlinearly, with root rate τ_n (defined by (1.2.42c)), where $\hat{x}$ is the global minimizer of $f(\cdot)$.* □

Exercise 1.6.6. Use the trust region Algorithm Model 1.2.26 to develop a globally converging conservative secant algorithm for which the conclusions of Theorem 1.6.5 remain valid. □

1.6.3 Symmetric Rank-One Updates

Rank-one variable metric algorithms incorporate the following two properties of the classical secant algorithm. The first is given by (5e), which can be restated in column-by-column format as follows:

$$H_{i+1}\Delta x_j = \Delta g_j, \quad j = i, i-1, \ldots, i-n+1, \tag{9a}$$

$$H_{i+1}^{-1}\Delta g_j = \Delta x_j, \quad j = i, i-1, \ldots, i-n+1. \tag{9b}$$

This property is commonly referred to as the *secant property*.

The second property is the *rank-one update* property, i.e., H_{i+1} is obtained from H_i by adding a dyad (i.e., a rank-one matrix) to H_i. Hence the matrices H_i satisfy a recursion of the form

$$H_{i+1} = H_i + u_i v_i^T, \quad i = 0, 1, 2, \ldots \tag{9c}$$

with $u_i, v_i \in \mathbb{R}^n$.

Now we will show that a recursion of the form (9c) is valid for the the classical secant algorithm. Recalling that, for $j = 1, 2, .., n$, e_j denotes the jth column of the $n \times n$ identity matrix, let

$$P \triangleq [e_n, e_1, e_2, e_3, .., e_{n-1}]. \tag{9d}$$

Then we see that $\tilde{\Xi}_{i+1} \triangleq \Xi_{i+1}P = [\Delta x_i, \Delta x_{i-n+1}, \ldots, \Delta x_{i-1}]$ differs from Ξ_i only by the first column and, similarly, $\tilde{\Gamma}_{i+1} \triangleq \Gamma_{i+1}P$ differs from Γ_i, also, only by the

first column. Thus, $\tilde{\Xi}_{i+1}$ and Ξ_I differ only by a dyad (a rank-one matrix), and similarly, $\tilde{\Gamma}_{i+1}$ and Γ_i differ only by a dyad, as shown below:

$$\left.\begin{aligned}\tilde{\Xi}_{i+1} &= \Xi_i + (\Delta x_i - \Delta x_{i-n})e_1^T\,,\\ \tilde{\Gamma}_{i+1} &= \Gamma_i + (\Delta g_i - \Delta g_{i-n})e_1^T\,.\end{aligned}\right\} \tag{9e}$$

Furthermore,

$$H_{i+1} = \Gamma_{i+1}\Xi_{i+1}^{-1} = \tilde{\Gamma}_{i+1}P^{-1}P\tilde{\Xi}_{i+1}^{-1} = \tilde{\Gamma}_{i+1}\tilde{\Xi}_{i+1}^{-1}\,, \tag{9f}$$

and similarly,[‡]

$$H_{i+1}^{-1} = \tilde{\Xi}_{i+1}\tilde{\Gamma}_{i+1}^{-1}\,. \tag{9g}$$

Next, it follows from (9f) that

$$H_{i+1}\tilde{\Xi}_{i+1} = \tilde{\Gamma}_{i+1}\,. \tag{9h}$$

Now, it follows from (9e) that $\tilde{\Xi}_{i+1} = \Xi_i + \Delta\Xi_i$, and $\tilde{\Gamma}_{i+1} = \Gamma_i + \Delta\Gamma_i$, where $\Delta\Xi_i = (\Delta x_i - \Delta x_{i-n})e_1^T$ and $\Delta\Gamma_i = (\Delta g_i - \Delta g_{i-n})e_1^T$. Hence, if we let $H_{i+1} = H_i + \Delta H_i$, then we conclude from (9h) that

$$(H_i + \Delta H_i)(\Xi_i + \Delta\Xi_i) = (\Gamma_i + \Delta\Gamma_i)\,. \tag{9i}$$

Since $H_i\Xi_i = \Gamma_i$, we conclude from (9i) that

$$\Delta H_i = (-H_i\Delta\Xi_i + \Delta\Gamma_i)\Xi_{i+1}^{-1} = u_i v_i^T\,, \tag{9j}$$

where $u_i = [-H_i(\Delta x_i - \Delta x_{i-n}) + (\Delta g_i - \Delta g_{i-n})]$ and $v_i^T = e_1^T\Xi_{i+1}^{-1}$. Thus we find that

$$H_{i+1} = H_i + u_i v_i^T\,, \tag{9k}$$

i.e., that H_{i+1} differs from H_i only by the rank-one matrix $u_i v_i$.

Similarly, it can be shown that

$$H_{i+1}^{-1} = H_i^{-1} + u_i^* v_i^{*T}\,, \tag{9l}$$

where $u_i^* = [-H_i^{-1}(\Delta g_i - \Delta g_{i-n}) + (\Delta x_i - \Delta x_{i-n})]$ and $v_i^{*T} = e_1^T\Gamma_{i+1}^{-1}$, and we see that H_{i+1}^{-1} differs from H_i^{-1} only by the rank-one matrix $u^*{}_i v^*{}_i$.

Next we will extend the classical secant algorithm to convex optimization problems of the form (1). Since the Hessian matrix $H(x) \triangleq f_{xx}(x)$ is symmetric, it is desirable to approximate it by a symmetric matrix. Hence we will consider the possibility of constructing approximations to Hessian matrices using symmetric, rank-one corrections. Such corrections are usually referred to as *updates*.

[‡] It should be obvious that the Sherman-Morrison formula can be used in conjunction with the equations (9e) and (9f) to obtain an efficient numerical procedure for computing the search directions $H_i^{-1}g(x_i)$ for the classical secant algorithm.

We note that the formula (6d), which corresponds to updating one column and one row of H_i at each iteration, is a rank-two formula. Hence, if there is a satisfactory rank-one update formula, it must be obtained in a way unrelated to updating one row and one column at at time. To develop such a formula, we begin by considering the special case of problem (1) where $f(\cdot)$ is defined by

$$f(x) \triangleq \tfrac{1}{2}\langle x\,, Hx\rangle + \langle d\,, x\rangle\,, \tag{10}$$

with H a symmetric, positive-definite, $n \times n$ matrix and $d \in \mathbb{R}^n$. Thus, suppose that we have a sequence of $n+1$ vectors $x_0, x_1, \ldots, x_n$, and corresponding gradients, $g_i \triangleq \nabla f(x_i) = Hx_i + d$, $i = 0, 1, \ldots, n-1$. Suppose that the vectors Δx_i, $i = 0, 1, \ldots, n-1$, defined by (5c), are linearly independent. Clearly, since H is nonsingular, the corresponding gradient differences Δg_i, defined by (5d), are also linearly independent. We will now see that, in this case, it is possible to reconstruct the matrices H and H^{-1} by a sequential process, involving symmetric, rank-one updates of a matrix, satisfying the two secant properties (9a) and (9k), as follows. Since the matrix H is symmetric, it makes sense to require that the dyad uv^T in (9k) also be symmetric, by setting $u = v = z$.

We start with an arbitrary, symmetric, positive-definite matrix H_0, and, for $i = 0, 1, 2, \ldots, n-1$, we define the symmetric, rank-one update process by

$$H_{i+1} = H_i + z_i z_i^T\,, \tag{11a}$$

where z_i is such that

$$H_{i+1}\Delta x_i = \Delta g_i\,. \tag{11b}$$

It follows directly from (11a) and (11b), that

$$z_i\langle z_i\,, \Delta x_i\rangle = \Delta g_i - H_i\Delta x_i \tag{11c}$$

and hence that

$$\langle z_i\,, \Delta x_i\rangle^2 = \langle \Delta g_i - H_i\Delta x_i\,, \Delta x_i\rangle\,. \tag{11d}$$

Consequently,

$$H_{i+1} = H_i + \frac{1}{\langle \Delta g_i - H_i\Delta x_i\,, \Delta x_i\rangle}(\Delta g_i - H_i\Delta x_i)(\Delta g_i - H_i\Delta x_i)^T\,, \tag{11e}$$

provided that $\langle \Delta g_i - H_i\Delta x_i\,, \Delta x_i\rangle \neq 0$.

Similarly, if we begin with a symmetric, positive-definite matrix G_0 and, for $i = 0, 1, 2, \ldots, n-1$, we define the symmetric, rank-one update process by

$$G_{i+1} = G_i + z_i z_i^T\,, \tag{12a}$$

where z_i is such that

$$G_{i+1}\Delta g_i = \Delta x_i\,, \tag{12b}$$

then, reasoning as above, we conclude that

$$G_{i+1} = G_i + \frac{1}{\langle \Delta x_i - G_i \Delta g_i \, , \Delta g_i \rangle} (\Delta x_i - G_i \Delta g_i)(\Delta x_i - G_i \Delta g_i)^T \, , \quad (12c)$$

provided that $\langle \Delta x_i - G_i \Delta g_i \, , \Delta g_i \rangle \neq 0$.

We will now show that, if these constructions do not break down, then they yield $H_n = H$ and $G_n = G = H^{-1}$.

Theorem 1.6.7. *Suppose that for $i = 0, 1, 2, \ldots, n-1$, the difference vectors Δx_i are linearly independent and that the corresponding gradient differences Δg_i are defined by $\Delta g_i = H \Delta x_i$, with H an $n \times n$, symmetric, positive-definite matrix. Then,*

(a) if for $i = 0, 1, 2, \ldots, n-1$, the matrices H_i are constructed according to (11e) (which implies that for $i = 0, 1, 2, \ldots, n-1$, $\langle \Delta g_i - H_i \Delta x_i \, , \Delta x_i \rangle \neq 0$), with H_0 an arbitrary, symmetric, positive-definite, $n \times n$ matrix, then for $i = 0, 1, 2, \ldots, n-1$, the H_i satisfy the secant relations

$$H_{i+1} \Delta x_j = \Delta g_j \, , \; j = 0, 1, 2, \ldots, i \, , \quad (13a)$$

and $H_n = H$;

(b) if for $i = 0, 1, 2, \ldots, n-1$, the matrices G_i are constructed according to (12c) (which implies that for $i = 0, 1, 2, \ldots, n-1$, $\langle \Delta x_i - G_i \Delta g_i \, , \Delta g_i \rangle \neq 0$), with G_0 an arbitrary, symmetric, positive-definite, $n \times n$ matrix, then, for $i = 0, 1, 2, \ldots, n-1$, the G_i satisfy the secant relations

$$G_{i+1} \Delta g_j = \Delta x_j \, , \; j = 0, 1, 2, \ldots, i \, , \quad (13b)$$

and $G_n = G = H^{-1}$.

Proof. Clearly, given the identity of form, we need only to establish *(a)* or *(b)* and the other one will follow by symmetry. We will prove *(b)*. Thus, for $i = 0$, $G_1 \Delta g_0 = \Delta x_0$, by construction of G_1. Next, proceeding by induction, we assume that (13b) holds for $i = 0, 1, 2, \ldots, k-1$, with $k \leq n-1$. Then, for $i \leq k-1$,

$$G_{k+1} \Delta g_i = G_k \Delta g_i + y_k \langle \Delta x_k - G_k \Delta g_k \, , \Delta g_i \rangle \, , \quad (14a)$$

with y_k defined by

$$y_k = \frac{1}{\langle \Delta g_k \, , \Delta x_k - G_k \Delta g_k \rangle} [\Delta x_k - G_k \Delta g_k] \, . \quad (14b)$$

Since by hypothesis $G_k \Delta g_i = \Delta x_i$, G_k is symmetric, and $\Delta g_k = H \Delta x_k$, we find in (14a) that

$$\begin{aligned} \langle \Delta x_k - G_k \Delta g_k \, , \Delta g_i \rangle &= \langle \Delta x_k \, , \Delta g_i \rangle - \langle \Delta g_k \, , G_k \Delta g_i \rangle \, , \\ &= \langle \Delta x_k \, , H \Delta x_i \rangle - \langle H \Delta x_k \, , \Delta x_i \rangle = 0 \, . \end{aligned} \quad (14c)$$

Consequently,

$$G_{k+1} \Delta g_i = \Delta x_i \, , \text{ for } i = 0, 1, \ldots, k-1 \, . \quad (14d)$$

Since $G_{k+1} \Delta g_k = \Delta x_k$ by construction of G_{k+1}, (13b) follows.

Since the vectors Δx_i are linearly independent, by assumption, and since (13b) holds for $i = n-1$, we see that $G_n \Gamma = \Xi$, where $\Gamma = [\Delta g_0, \ldots, \Delta g_{n-1}]$ and $\Xi = [\Delta x_0, \ldots, \Delta x_{n-1}]$. Since by the secant relation $\Gamma = H \Xi$, it follows that $G_n = G$. Hence the theorem is proved. □

Now suppose that $f(\cdot)$ in (1) is a general function satisfying Assumption 1.6.1, with global minimizer $\hat{x} \in \mathbb{R}^n$ and that $H(\hat{x})$ is positive-definite. Let $\{x_i\}_{i=0}^{n-1}$ be points near $\hat{x}$ such that the difference vectors Δx_i, $i = 0, 1, \ldots, n-1$, are linearly independent and that the corresponding gradient differences Δg_i are defined by $\Delta g_i = \nabla f(x_{i+1}) - \nabla f(x_i)$, $i = 0, 1, \ldots, n-1$. If H_n is constructed using the recursion (11e), with H_0 an arbitrary, symmetric, positive-definite matrix, then Theorem 1.6.7 suggests that H_n should be "close" to $H(\hat{x})$ and hence that the recursion (11e) (or (12c)) should be usable in a secant-like algorithm for minimizing $f(\cdot)$. Note, however, that when $f(\cdot)$ is not a quadratic function, (14c) need not hold, and hence (13a) may hold only for $j = 1$.

Since it is possible for $\langle \Delta x_i \, , \Delta g_i - H_i \Delta x_i \rangle$ or for $\langle \Delta g_i \, , \Delta x_i - G_i \Delta g_i \rangle$ to be zero, which causes the symmetric rank-one constructions to break down, we see that the construction of the matrices H_{i+1} and G_{i+1}, according to (11e) and (12c) shares the advantages and disadvantages of the classical secant method. In addition, note that if H_i is positive-definite and H_{i+1} is constructed according to (11e), then it is not at all certain that H_{i+1} is also positive-definite. A similar statement holds for G_{i+1}, constructed according to (12c). All is not lost, however, since we can extract symmetric rank-two update formulas from (13a) and (13b), which are more complex, but have a *hereditary positive-definiteness property* and which can be used in solving both the quadratic and the general case of (1).

1.6.4 Symmetric Rank-Two Updates

We begin by deriving the Davidon-Fletcher-Powell (DFP) formula. It was the first rank-two formula to be discovered, although not by the route that we are following. If we expand (12c), we get an expression of the form

$$\begin{aligned} G_{i+1} = G_i &+ \beta_i \Delta x_i \Delta x_i^T + \gamma_i (G_i \Delta g_i)(G_i \Delta g_i)^T \\ &+ \delta_i [\Delta x_i (G_i \Delta g_i)^T + (G_i \Delta g_i) \Delta x_i^T] \, , \end{aligned} \quad (15a)$$

where β_i, γ_i, δ_i are coefficients determined from (12c). If we suppress the

nonsymmetrical terms $\Delta x_i (G_i \Delta g_i)^T$ and $(G_i \Delta g_i)\Delta x_i^T$ in (15a), we obtain

$$G_{i+1} = G_i + \beta_i \Delta x_i \Delta x_i^T + \gamma_i (G_i \Delta g_i)(G_i \Delta g_i)^T . \tag{15b}$$

Since we need $G_{i+1}\Delta g_i = \Delta x_i$ to hold, we require that

$$\Delta x_i = G_i \Delta g_i + \beta_i \Delta x_i \langle \Delta x_i , \Delta g_i \rangle + \gamma_i (G_i \Delta g_i) \langle G_i \Delta g_i , \Delta g_i \rangle . \tag{15c}$$

If we set $\beta_i = 1/\langle \Delta x_i , \Delta g_i \rangle$ and $\gamma_i = -1/\langle G_i \Delta g_i , \Delta g_i \rangle$, we find that $G_{i+1}\Delta g_i = \Delta x_i$ holds. With these values of β_i and γ_i, (15b) becomes the symmetric, rank-two, update formula invented by Davidon [Dav.59] and analyzed by Fletcher and Powell [FlP.63]:

$$G_{i+1} = G_i + \frac{1}{\langle \Delta x_i , \Delta g_i \rangle} \Delta x_i \Delta x_i^T - \frac{1}{\langle G_i \Delta g_i , \Delta g_i \rangle} (G_i \Delta g_i)(G_i \Delta g_i)^T \tag{15d}$$

Formula (15d) is by no means the only valid, rank-two, update formula. The literature contains a continuum of formulas extracted from (12c).

If we proceed in a similar manner from formula (11e), we obtain the rank-two, Broyden-Fletcher-Goldfarb-Shanno (BFGS), update formula ([Bro.65, FlW.80, Gol.70, Sha.70]):

$$H_{i+1} = H_i + \frac{1}{\langle \Delta g_i , \Delta x_i \rangle} \Delta g_i \Delta g_i^T - \frac{1}{\langle H_i \Delta x_i , \Delta x_i \rangle} (H_i \Delta x_i)(H_i \Delta x_i)^T \tag{15e}$$

Note that (15e) can be obtained formally from (15d) by replacing G_i by H_i, Δx_i by Δg_i, and Δg_i by Δx_i. It should be obvious to the reader that when H_{i+1} is defined by (15e), given H_i^{-1}, H_{i+1}^{-1} is easily computed by means of two applications of the Sherman-Morrison formula (4). The result is

$$H_{i+1}^{-1} = \left(I - \frac{\Delta x_i \Delta g_i^T}{\langle \Delta x_i , \Delta g_i \rangle} \right) H_i^{-1} \left(I - \frac{\Delta x_i \Delta g_i^T}{\langle \Delta x_i , \Delta g_i \rangle} \right)^T + \frac{\Delta x_i \Delta x_i^T}{\langle \Delta x_i , \Delta g_i \rangle} . \tag{15f}$$

We begin by exhibiting the hereditary positive-definiteness properties of the update formulas (15d) and (15e).

Theorem 1.6.8. *Suppose that H is a symmetric, positive-definite, $n \times n$ matrix, that $\Delta x_i \in \mathbb{R}^n$ is a nonzero vector, and that $\Delta g_i = H \Delta x_i$. If H_i, G_i are symmetric, positive-definite, $n \times n$ matrices, then H_{i+1}, G_{i+1}, given by (15d) and (15e), respectively, are also symmetric, positive-definite matrices.*

Proof. First we note that because H is positive-definite, by assumption, $\langle \Delta g_i , \Delta x_i \rangle = \langle H \Delta x_i , \Delta x_i \rangle > 0$. Similarly, because G_i and H_i are positive-definite by assumption, $\langle \Delta g_i , G_i \Delta g_i \rangle > 0$, and $\Delta x_i , H_i \Delta x_i \rangle > 0$. Hence, G_{i+1} is well defined by (15d), and H_{i+1} is well defined by (15e).

Next, for any $y \in \mathbb{R}^n$, $y \neq 0$,

$$\langle y , G_{i+1} y \rangle = \langle y , G_i y \rangle + \frac{\langle y , \Delta x_i \rangle^2}{\langle \Delta g_i , \Delta x_i \rangle} - \frac{\langle y , G_i \Delta g_i \rangle^2}{\langle \Delta g_i , G_i \Delta g_i \rangle} . \tag{16a}$$

Let $a = G_i^{1/2} y$ and $b = G_i^{1/2} \Delta g_i$. Then (16a) becomes

$$\langle y , G_{i+1} y \rangle = \frac{\|a\|^2 \|b\|^2 - \langle a , b \rangle^2}{\|b\|^2} + \frac{\langle y , \Delta x_i \rangle^2}{\langle \Delta x_i , H \Delta x_i \rangle} . \tag{16b}$$

By the Schwartz inequality,

$$\|a\|^2 \|b\|^2 - \langle a , b \rangle^2 \geq 0 , \tag{16c}$$

and hence $\langle y , G_{i+1} y \rangle \geq 0$. Now suppose that $\langle y , G_{i+1} y \rangle = 0$ for some $y \neq 0$. Then *both terms* in the right-hand side of (16b) must be zero. But $\|a\|^2 \|b\|^2 - \langle a , b \rangle^2 = 0$ implies that $a = \alpha b$ for some $\alpha \in \mathbb{R}$; i.e., that $y = \alpha \Delta g_i$. But then

$$\langle y , \Delta x_i \rangle^2 = \alpha^2 \langle \Delta g_i , \Delta x_i \rangle^2 = \alpha^2 \langle H \Delta x_i , \Delta x_i \rangle^2 > 0 , \tag{16d}$$

which contradicts our assumption that $\langle y , G_{i+1} y \rangle = 0$. Hence G_{i+1} is positive-definite.

Next, for any $y \in \mathbb{R}^n$, $y \neq 0$,

$$\langle y , H_{i+1} y \rangle = \langle y , H_i y \rangle + \frac{\langle y , \Delta g_i \rangle^2}{\langle \Delta x_i , \Delta g_i \rangle} - \frac{\langle y , H_i \Delta x_i \rangle^2}{\langle \Delta x_i , H_i \Delta x_i \rangle} . \tag{16e}$$

If we now define $a = H_i^{1/2} y$ and $b = H_i^{1/2} \Delta x_i$, the fact that H_{i+1} is positive-definite follows by symmetry from the arguments used to establish the fact that G_{i+1} is positive-definite. □

Thus, unlike the rank-one formulas, the DFP and BFGS rank-two formulas do not break down as the computation proceeds.

1.6.5 Finite Convergence on Quadratic Functions

When the DFP and BFGS formulas are incorporated into an algorithm with an exact line search step-size rule, they both lead to the solution of the quadratic problem (10) in, at most, n iterations. However, when the BFGS formula is used in an algorithm either with an exact line search step-size rule or other type of step-size rule, it has been proven much superior in solving the general problem (1). Hence we will present proofs for the DFP method only for the

quadratic case. We will leave the corresponding proofs for the BFGS method as an easy exercise for the reader, and we will deal in full with the BFGS algorithm for the general problem (1).

We recall that optimization algorithms with an exact line search step-size rule are of theoretical interest only, since exact line searches are not implementable, in general. The DFP variable metric method for solving (1), with an exact line search step-size rule, has the following form:

DFP Variable Metric Algorithm 1.6.9.

Data. $x_0 \in \mathbb{R}^n$, G_0, a symmetric, $n \times n$, positive-definite matrix.

Step 0. Set $i = 0$.

Step 1. If $g_i = \nabla f(x_i) = 0$, stop. Else, compute

$$\lambda_i = \arg\min_{\lambda \geq 0} f(x_i - \lambda G_i g_i) . \tag{17a}$$

Step 2. Compute

$$x_{i+1} = x_i - \lambda_i G_i g_i , \tag{17b}$$

$$g_{i+1} = \nabla f(x_{i+1}) , \tag{17c}$$

$$\Delta x_i = x_{i+1} - x_i , \quad \Delta g_i = g_{i+1} - g_i , \tag{17d}$$

$$G_{i+1} = G_i + \frac{1}{\langle \Delta g_i , \Delta x_i \rangle} \Delta x_i \Delta x_i^T - \frac{1}{\langle \Delta g_i , G_i \Delta g_i \rangle} (G_i \Delta g_i)(G_i \Delta g_i)^T . \tag{17e}$$

Step 3. Replace i by $i + 1$, and go to Step 1.

The following theorem shows that the DFP Algorithm 1.6.9 solves a quadratic problem, in $\mathbb{R}^n$, in, at most, n iterations (recall that Newton's method solves such problems in one iteration).

Theorem 1.6.10. *Suppose that $f(x) = \frac{1}{2}\langle x , Hx \rangle + \langle d , x \rangle$ with H a symmetric, positive-definite, $n \times n$, matrix. Then, for x_i, G_i, $i = 0, 1, 2, \ldots$, constructed by Algorithm 1.6.9,*

(a) the secant property holds, i.e.,

$$G_{i+1} \Delta g_k = \Delta x_k , \quad 0 \leq k \leq i \leq n-1 ; \tag{18a}$$

(b) the Δx_i are H conjugate (and hence linearly independent), i.e.,

$$\langle \Delta x_i , H \Delta x_j \rangle = 0 , \quad \forall\, i \neq j , \; 0 \leq i , j \leq n-1 ; \tag{18b}$$

and

(c) x_n is the minimizer of $f(\cdot)$ over $\mathbb{R}^n$.

Proof. We will prove (18a) and (18b) together, by induction. (Recall that $\Delta g_i = H \Delta x_i$ because of the form of $f(\cdot)$ and that $G_{i+1} \Delta g_i = \Delta x_i$ by construction of G_{i+1}). First,

$$\begin{aligned} \langle H \Delta x_0 , \Delta x_1 \rangle &= \langle H \Delta x_0 , -\lambda_1 G_1 g_1 \rangle \\ &= -\lambda_1 \langle H \Delta x_0 , G_1 g_1 \rangle \\ &= -\lambda_1 \langle G_1 \Delta g_0 , g_1 \rangle \\ &= -\lambda_1 \langle \Delta x_0 , g_1 \rangle = 0 , \end{aligned} \tag{19a}$$

because λ_0 was computed by exact minimization, and by the construction of G_1. Hence we see that (18b) holds for $0 \leq i , j \leq 1$.

Next,

$$G_1 \Delta g_0 = \Delta x_0 , \tag{19b}$$

by construction. Hence (18a) holds for $i = k = 0$. Thus we have initialized the proof by induction of (18a) and (18b). Consequently, suppose that (18a) holds for all $0 \leq i \leq l < n$, and that (18b) holds for all $0 \leq i , j \leq l < n$. For any $i \in \{0, 1, \ldots, l\}$, we get, by adding and subtracting terms, that

$$g_{l+1} = g_{i+1} + \sum_{j=i+1}^{l} H \Delta x_j . \tag{19c}$$

Since $\langle \Delta x_i , g_{i+1} \rangle = 0$ by construction of λ_i and $\langle \Delta x_i , H \Delta x_j \rangle = 0$, for $j \neq i$, $i , j \leq l$, by assumption, we find that for any $i \leq l$,

$$\langle \Delta x_i , g_{l+1} \rangle = \langle \Delta x_i , g_{i+1} \rangle + \langle \Delta x_i , \sum_{j=i+1}^{l} H \Delta x_j \rangle = 0 . \tag{19d}$$

Hence, for any $i \in \{0, 1, \ldots, l\}$,

$$\begin{aligned} \langle \Delta x_i , H \Delta x_{l+1} \rangle &= -\lambda_{l+1} \langle \Delta x_i , H G_{l+1} g_{l+1} \rangle \\ &= -\lambda_{l+1} \langle G_{l+1} H \Delta x_i , g_{l+1} \rangle \\ &= -\lambda_{l+1} \langle G_{l+1} \Delta g_i , g_{l+1} \rangle \\ &= -\lambda_{l+1} \langle \Delta x_i , g_{l+1} \rangle = 0 , \end{aligned} \tag{19e}$$

i.e., the vectors $\{\Delta x_i\}_{i=0}^{l+1}$ are H-conjugate for $0 \leq i \leq l+1$.

Next, for $0 \le i \le l$,

$$
\begin{aligned}
G_{l+2}\Delta g_i &= G_{l+2}H\Delta x_i \\
&= G_{l+1}H\Delta x_i + \frac{1}{\langle \Delta x_{l+1}, \Delta g_{l+1}\rangle}\Delta x_{l+1}\langle \Delta x_{l+1}, H\Delta x_i\rangle \\
&\quad - \frac{1}{\langle \Delta g_{l+1}, G_{l+1}\Delta g_{l+1}\rangle}(G_{l+1}\Delta g_{l+1})\langle G_{l+1}\Delta g_{l+1}, H\Delta x_i\rangle \\
&= \Delta x_i - \frac{1}{\langle \Delta g_{l+1}, G_{l+1}\Delta g_{l+1}\rangle}(G_{l+1}\Delta g_{l+1})\langle H\Delta x_{l+1}, G_{l+1}\Delta g_i\rangle \\
&= \Delta x_i - \frac{1}{\langle \Delta g_{l+1}, G_{l+1}\Delta g_{l+1}\rangle}(G_{l+1}\Delta g_{l+1})\langle H\Delta x_{l+1}, \Delta x_i\rangle \\
&= \Delta x_i ,
\end{aligned}
\tag{19f}
$$

and $G_{l+2}\Delta g_{l+1} = \Delta x_{l+1}$ by construction. Hence (18a) holds for $i = l+1$. This completes our proof of (18a) and (18b).

Since (19d) implies that x_{l+1} minimizes $f(x)$ on the $(l+1)$-dimensional affine manifold $M_{l+1} \triangleq \{x \in \mathbb{R}^n \mid x = x_0 + \sum_{j=1}^{l+1} y^j \Delta x_{j-1}\ y \in \mathbb{R}^{l+1}\}$, which is spanned by the linearly independent vectors $\Delta x_0, \Delta x_1, \ldots, \Delta x_l$, part *(c)* of the theorem follows by setting $l = n-1$, since $M_n = \mathbb{R}^n$. Hence our proof is complete. □

This concludes our discussion of the DFP variable metric method.

Next, we turn to the BFGS variable metric method, which we will discuss in two forms: with an exact line search step-size rule and with an Armijo type step-size rule. The version with the exact line search step-size rule is of theoretical interest only, since exact line searches are not implementable, in general.

The BFGS variable metric method for solving (1), with exact line search, has the following form:

BFGS Variable Metric Algorithm 1.6.11.

Data. $x_0 \in \mathbb{R}^n$, H_0, a symmetric, $n \times n$, positive-definite matrix.

Step 0. Set $i = 0$.

Step 1. If $g_i = \nabla f(x_i) = 0$, stop. Else, compute

$$\lambda_i = \arg\min_{\lambda \ge 0} f(x_i - \lambda H_i^{-1} g_i). \tag{20a}$$

Step 2. Compute

$$x_{i+1} = x_i - \lambda_i H_i^{-1} g_i , \tag{20b}$$

$$g_{i+1} = \nabla f(x_{i+1}), \tag{20c}$$

$$\Delta x_i = x_{i+1} - x_i , \quad \Delta g_i = g_{i+1} - g_i , \tag{20d}$$

$$H_{i+1} = H_i + \frac{1}{\langle \Delta g_i , \Delta x_i\rangle}\Delta g_i \Delta g_i^T - \frac{1}{\langle H_i \Delta x_i , \Delta x_i\rangle}(H_i\Delta x_i)(H_i \Delta x_i)^T. \tag{20e}$$

Step 3. Replace i by $i+1$, and go to Step 1.

Exercise 1.6.12. Prove the following result.

Theorem 1.6.13. *Suppose that $f(x) = \frac{1}{2}\langle x , Hx\rangle + \langle d , x\rangle$, with H a symmetric, positive-definite, $n \times n$ matrix. Then, for every x_i and H_i constructed by the BFGS Algorithm 1.6.11,*

(a) the secant property holds, i.e.,

$$H_{i+1}\Delta x_k = \Delta g_k , \quad 0 \le k \le \le i \le n-1 ; \tag{21a}$$

(b) the Δx_i are H-conjugate, i.e.,

$$\langle \Delta x_i , H\Delta x_j\rangle = 0 , \quad \forall\, i \ne j , \ 0 \le i , j \le n-1 ; \tag{21b}$$

and

(c) x_n is the minimizer of $f(\cdot)$ over $\mathbb{R}^n$. □

The fact that both the DFP Variable Metric Algorithm 1.6.9 and the BFGS Variable Metric Algorithm 1.6.11 generate H-conjugate directions in solving (1) when $f(x) = \frac{1}{2}\langle x , Hx\rangle + \langle d , x\rangle$, with H a symmetric, positive-definite, $n \times n$ matrix, raises the question of their relationship to conjugate gradient methods. In fact, it was shown in [Mye.68] that the search direction vectors constructed by the DFP Variable Metric Algorithm 1.6.9 and the Progenitor Conjugate Gradient Algorithm 1.5.4 (and hence also by the Fletcher-Reeves and Polak-Ribière conjugate gradient algorithms) are scalar multiples of each other. To conclude our discussion of the behavior of variable metric methods on quadratic functions, we will show that in solving (1) with

$$f(x) = \tfrac{1}{2}\langle x , Hx\rangle + \langle d , x\rangle ,$$

as above, when initialized with $H_0 = I$, the BFGS Variable Metric Algorithm 1.6.11 and the Progenitor Conjugate Gradient Algorithm 1.5.4 construct identical search direction vectors. Thus we are led to the conclusion that on quadratic functions, the Progenitor Conjugate Gradient Algorithm 1.5.4, the DFP Variable

Metric Algorithm 1.6.9, and the BFGS Variable Metric Algorithm 1.6.1 construct identical sequences in minimizing positive-definite quadratic functions. We follow Nazareth [Naz.94] in our presentation (see, also, [Dix.72]).

Lemma 1.6.14. *Suppose that $H_0 = I$ in Algorithm 1.6.11 and that for $i = 0, 1, \ldots, n-1$, x_i, g_i, and H_i were constructed by Algorithm 1.6.11 in solving (1) with $f(\cdot)$ defined by (10), with H a positive-definite, symmetric $n \times n$ matrix. Let*

$$h_i^B \triangleq -H_i^{-1} g_i \,, \quad i = 0, 1, \ldots, n-1 \,, \tag{22a}$$

be the associated search direction vectors. Then, for $i = 0, 1, \ldots, n-1$,

$$h_i^B \in \text{span}\{ g_0, \ldots, g_i \} \,, \tag{22b}$$

and H_i^{-1} has a dyadic decomposition of the form

$$H_i^{-1} = I + \sum_{j=0}^{k_i} p_{i,j} q_{i,j}^T \,, \tag{22c}$$

where $k_i \leq i$ is an integer which depends on i; $p_{i,j}$, $q_{i,j} \in \mathbb{R}^n$, and

$$p_{i,j} \in \text{span}\{ g_0, \ldots, g_i \} \,, \; 0 \leq j \leq k_i \,. \tag{22d}$$

Proof. We will give a proof by induction. Clearly (22b), (22c), and (22d) are true for $i = 0$. Hence suppose that (22b), (22c), and (22d) are true for some $0 \leq i \leq n-2$. We will show that they must also be true for $i+1$. By direct calculation, it follows from this assumption and (15f) that

$$H_{i+1}^{-1} = I + \sum_{j=0}^{k_{i+1}} p_{i+1,j} q_{i+1,j}^T \tag{23a}$$

with $k_{i+1} \leq i$ and

$$p_{i+1,j} \in \text{span}\{ g_0, \ldots, g_{i+1} \} \,, \; 0 \leq j \leq k_{i+1} \,. \tag{23b}$$

Since $h_{i+1}^B = H_{i+1}^{-1} g_{i+1}$, it follows that $h_{i+1}^B \in \text{span}\{ g_0, \ldots, g_{i+1} \}$, and hence our proof is complete. □

Note that because the vectors $\{ h_j \}_{j=0}^i$ are H-conjugate, they form a basis for the subspace $S_i \triangleq \text{span}\{ g_0, \ldots, g_i \}$. Furthermore, since the number of g_j's spanning S_i is the same as the cardinality of the basis $\{ h_j \}_{j=0}^i$, it follows that the vectors in the set $\{ g_0, \ldots, g_i \}$ are linearly independent.

Theorem 1.6.15. *Consider problem (10) with H a positive-definite, symmetric $n \times n$ matrix. Suppose x_0 is given and that $H_0 = I$ in Algorithm 1.6.11. Then the search direction vectors h_i^B, $i = 0, 1, \ldots, n-1$, defined by (22a), and the search direction vectors h_i^{CG}, $i = 0, 1, \ldots, n-1$, defined by the recursion (1.5.9a-c) in the Progenitor Conjugate Gradient Algorithm 1.5.4 (and hence also by the Fletcher-Reeves and Polak-Ribière algorithms) are identical, i.e.,*

$$h_i^B = h_i^{CG} \,, \; i = 0, 1, \ldots, n-1 \,. \tag{24}$$

Proof. It follows from (22b) that

$$h_i^B = \sum_{j=0}^{i} \alpha_{i,j} g_j \,, \tag{25a}$$

for some $\alpha_{i,j} \in \mathbb{R}$. Similarly, it follows from the recursion in (1.5.9a-c), that

$$h_i^{CG} = \sum_{j=0}^{i} \alpha'_{i,j} g_j \,, \tag{25b}$$

for some $\alpha'_{i,j} \in \mathbb{R}$, and that $\alpha'_{i,i} = -1$. Since $h_i^B \in \text{span}\{ g_i, g_{i-1}, \ldots, g_0 \}$ and $\langle h_i^B, H h_k^B \rangle = 0$ for all $i \neq k$, it follows that h_i is H-conjugate to the subspace S_i spanned by the vectors $\{ g_{i-1}, \ldots, g_0 \}$ and hence is colinear with the projection of g_i on the H-conjugate complement of the subspace S_i. Therefore it follows that the directions of the unit vectors $h_i^B / \|h_i^B\|$ are uniquely defined, that $\alpha_{i,i} \neq 0$ in (25a), and that the coefficients $\alpha_{i,j} / \alpha_{i,i}$ are uniquely defined. Thus, since we also have that $\langle h_i^{CG}, H h_k^{CG} \rangle = 0$ for all $i \neq k$, the same unique relations also define the $\alpha'_{i,j}$, and therefore it follows that $\alpha_{i,j} / \alpha_{i,i} = \alpha'_{i,j}$, $j = 0, 1, \ldots, i$, $i = 0, 1, \ldots, n-1$, i.e., that the vectors h_i^B and h_i^{CG} are colinear.

Since Algorithm 1.6.11 and Algorithm 1.5.4 both compute step-sizes by exact minimization along the search directions, it follows that they construct identical sequences $\{ x_i \}_{i=0}^n$, $\{ g_i \}_{i=0}^n$, $\{ \Delta x_i \}_{i=0}^n$, and $\{ \Delta g_i \}_{i=0}^n$. Hence, it follows from (1.5.10c) that

$$\langle g_i, g_j \rangle = 0 \,, \quad \forall\, i \neq j \,, \; i, j = 0, 1, \ldots, n \,. \tag{25c}$$

Since, for the BFGS Algorithm 1.6.11, $\Delta x_k = \lambda_k h_k^B$, we conclude from (22b) that

$$\langle g_k, \Delta x_i \rangle = 0 \,, \quad 0 \leq i < k \leq n \,. \tag{25d}$$

Next, for convenience, let us assume that $g_i \neq 0$ for all $i < n$, i.e., that early termination does not occur. First we will show by induction that

$$H_i^{-1} g_k = g_k \,, \quad \forall\, 0 \leq i < k \leq n \,. \tag{25e}$$

Clearly, (25e) holds for $i = 0$. Hence suppose that (25e) holds for some $i < n$. Then it follows from (25a), (25d), and (15f) that (25e) also holds for i replaced by $i+1$. Thus, (25e) is true.

Returning to our proof that $h_i^B = h_i^{CG}$ for $i = 0, \ldots, n-1$, we note that by definition, $h_{i+1}^B = -H_{i+1}^{-1} g_{i+1}$, $i = 0, \ldots, n-1$. Hence it follows from (15f), (25a), (25d), and (25e), the fact that h_i^B and h_i^{CG} are colinear, the fact that $\Delta x_i = \lambda_i^B h_i^B = \lambda_i^{CG} h_i^{CG}$, $\Delta g_i = \lambda_i^{CG} H h_i^{CG}$, and (1.5.9b,c) that

$$h_{i+1}^B = -H_i^{-1} g_{i+1} + \frac{\langle \Delta g_i , H_i^{-1} g_{i+1} \rangle}{\langle \Delta g_i , \Delta x_i \rangle} \Delta x_i$$

$$= -g_{i+1} + \frac{\langle \Delta g_i , g_{i+1} \rangle}{\langle \Delta g_i , h_i^B \rangle} h_i^B$$

$$= -g_{i+1} + \frac{\langle \Delta g_i , g_{i+1} \rangle}{\langle \Delta g_i , h_i^{CG} \rangle} h_i^{CG}$$

$$= h_i^{CG}, \quad 1 \le i \le n-1 . \tag{25f}$$

Since $h_0^B = h_0^{CG}$, our proof is complete. □

1.6.6 Global Convergence on Convex Functions

Now we return to the general case of problem (1), for which the BFGS update formula can be combined with the Armijo step-size rule to produce the following algorithm:

BFGS-Armijo Variable Metric Algorithm 1.6.16.

Parameters. $\alpha \in (0, \frac{1}{2})$, $\beta \in (0, 1)$.

Data. $x_0 \in \mathbb{R}^n$, H_0, a symmetric, $n \times n$, positive-definite matrix.

Step 0. Set $i = 0$, and $h_0 = -H_0^{-1} \nabla f(x_0)$.

Step 1. If $g_i = \nabla f(x_i) = 0$, stop.

Else, compute the *Armijo* step-size

$$\lambda_i = \max_{k \in \mathbb{N}} \{ \beta^k \mid f(x_i + \beta^k h_i) - f(x_i) \le \alpha \beta^k \langle \nabla f(x_i) , h_i \rangle \}. \tag{26a}$$

Step 2. Set

$$x_{i+1} = x_i + \lambda_i h_i , \tag{26b}$$

$$g_i = \nabla f(x_{i+1}) , \tag{26c}$$

$$\Delta x_i = x_{i+1} - x_i , \quad \Delta g_i = g_{i+1} - g_i , \tag{26d}$$

$$H_{i+1} = H_i + \frac{\Delta g_i \, \Delta g_i^T}{\langle \Delta g_i , \Delta x_i \rangle} - \frac{(H_i \Delta x_i)(H_i \Delta x_i)^T}{\langle H_i \Delta x_i , \Delta x_i \rangle} , \tag{26e}$$

$$h_{i+1} = -H_{i+1}^{-1} \nabla f(x_i) . \tag{26f}$$

Step 3. Replace i by $i + 1$, and go to Step 1.

We begin by showing that, just like Newton's method, Algorithm 1.6.16 is invariant under coordinate transformations performed using a symmetric, positive-definite matrix. Thus, suppose that Q is a symmetric, positive-definite, $n \times n$ matrix, and let $z = Qx$. Then, in the new coordinates, problem (1) becomes

$$\min_{z \in \mathbb{R}^n} \tilde{f}(z) , \tag{27a}$$

where $\tilde{f}(z) \triangleq f(Q^{-1}z)$. We note that $\hat{z} \triangleq Q\hat{x}$ is the solution to (27a), that $\nabla \tilde{f}(z) = Q^{-1} \nabla f(Q^{-1}z)$, and that $\tilde{H}(z) \triangleq \tilde{f}_{zz}(z) = Q^{-1} H(Q^{-1}z) Q^{-1}$.

To track the performance of Algorithm 1.6.16 on problem (1) in the new coordinate system, we define $z_i \triangleq Qx_i$, $\tilde{g}_i \triangleq \nabla \tilde{f}(z_i)$, $\Delta z_i \triangleq z_{i+1} - z_i$, and $\Delta \tilde{g}_i \triangleq \tilde{g}_{i+1} - \tilde{g}_i$. Then we see that

$$\left.\begin{aligned} \Delta z_i &= Q \Delta x_i , \\ \Delta \tilde{g}_i &= Q^{-1} \Delta g_i . \end{aligned}\right\} \tag{27b}$$

Next, if, for all $i \in \mathbb{N}$, we define the matrices $\tilde{H}_i$ by

$$\tilde{H}_i \triangleq Q^{-1} H_i Q^{-1} , \tag{27c}$$

then we see that the matrices $\tilde{H}_i$ satisfy the recursion

$$\tilde{H}_{i+1} = \tilde{H}_i + \frac{\Delta \tilde{g}_i \, \Delta \tilde{g}_i^T}{\langle \Delta \tilde{g}_i , \Delta z_i \rangle} - \frac{(\tilde{H}_i \Delta z_i)(\tilde{H}_i \Delta z_i)^T}{\langle \tilde{H}_i \Delta z_i , \Delta z_i \rangle} . \tag{27d}$$

Comparing (27d) with (26e), we see that the matrices $\tilde{H}_i$ are those that Algorithm 1.6.16 would construct in solving the transformed problem (27a), assuming that it generated the sequence $\{z_i\}_{i=0}^{\infty}$. If we now set $\tilde{h}_i = -\tilde{H}_i^{-1} \nabla \tilde{f}(z_i)$, then we see that $\tilde{h}_i = -QH_i^{-1}QQ^{-1} \nabla f(Q^{-1}z_i) = Qh_i$.

Next, we consider the step-size λ_i defined by (26a). First we note that because $h_i = Q^{-1}\tilde{h}_i$,

$$\langle \nabla f(x_i) , h_i \rangle = \langle \nabla \tilde{f}(z_i) , \tilde{h}_i \rangle , \tag{27e}$$

and, also, that for any $k \in \mathbb{N}$,

$$f(x_i + \beta^k h_i) = \tilde{f}(z_i + \beta^k \tilde{h}_i). \tag{27f}$$

Hence, if λ_i satisfies the step-size condition (26a) in the original coordinates, then it also satisfies the step-size condition in the new coordinates, and vice versa. Therefore we conclude as follows. If we initialize Algorithm 1.6.16 with x_0 and H_0 and apply it to problem (1), we construct a sequence $\{x_i\}_{i=0}^{\infty}$. If we initialize it with $z_0 = Qx_0$, and $\tilde{H}_0 = Q^{-1}H_0Q^{-1}$ and apply it to (27a), we construct a sequence $\{z_i\}_{i=0}^{\infty}$ which satisfies the relation $z_i = Qx_i$, for all $i \in \mathbb{N}$, i.e., Algorithm 1.6.16 is *scale-invariant*.

Now suppose that $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 1.6.16. We begin with a few observations regarding the quantities Δx_i, Δg_i, and H_{i+1}, defined in (26d) and (26e), respectively. First, (5h) applies. Hence, if we define $H \triangleq \int_0^1 f_{xx}(x_i + s\Delta x_i)ds$, then, because of Assumption 1.6.1, H is a symmetric, positive-definite matrix, and, in addition that $H\Delta x_i = \Delta g_i$ holds. Therefore, referring to Theorem 1.6.8, we see that if H_i in (26e) is symmetric and positive-definite, then so is H_{i+1}, i.e., the hereditary positive-definiteness property is valid for the general class of functions satisfying Assumption 1.6.1. Since we start out with a symmetric, positive-definite matrix H_0, it is clear that the entire sequence of matrices $\{H_i\}_{i=0}^{\infty}$ is positive-definite. Now, for $i \in \mathbb{N}$, let $\bar{\sigma}_i$, $\underline{\sigma}_i$ denote the largest and smallest eigenvalues, respectively, of the matrix H_i, and suppose for the moment that there exist bounds $0 < \sigma_* \le \sigma^* < \infty$, such that $\sigma_* \le \underline{\sigma}_i \le \bar{\sigma}_i \le \sigma^*$ for all $i \in \mathbb{N}$. Then we must have that

$$\langle \nabla f(x_i), h_i \rangle = -\langle \nabla f(x_i), H_i^{-1}\nabla f(x_i) \rangle \le -\frac{1}{\sigma^*}\|\nabla f(x_i)\|^2 \tag{28a}$$

and

$$\frac{1}{\sigma^*}\|\nabla f(x_i)\| \le \|h_i\| \le \|H_i^{-1}\|\|\nabla f(x_i)\| \le \frac{1}{\sigma_*}\|\nabla f(x_i)\|. \tag{28b}$$

Let κ_i denote the cosine of the angle between $-\nabla f(x_i)$ and h_i, i.e.,

$$\kappa_i \triangleq \frac{\langle -\nabla f(x_i), h_i \rangle}{\|\nabla f(x_i)\|\,\|h_i\|}. \tag{28c}$$

Then we see that (28a) and (28b) imply that

$$\kappa_i \ge \frac{\sigma_*}{\sigma^*} \triangleq \kappa_*, \tag{28d}$$

and that the relations (28b) and (28d) imply that

$$\langle \nabla f(x_i), h_i \rangle \le -(\kappa_*/\sigma_*)\|\nabla f(x_i)\|^2.$$

Thus, if either (28a)and (28b) or (28b), (28c), and (28d) could be shown to hold for all $i \in \mathbb{N}$, then it would follow directly from the Polak-Sargent-Sebastian Theorem 1.2.24b and the strict convexity of $f(\cdot)$ that $x_i \to \hat{x}$, as

$i \to \infty$. Eventually, we will establish the existence of such eigenvalue bounds. However, for the time being we will use the fact that, under Assumption 1.6.1, it is sufficient for (28b,c,d) to hold only on an infinite subsequence, for us to be able to prove that $x_i \to \hat{x}$, as $i \to \infty$. Hence we proceed to show that (28b,c,d) are, indeed, satisfied infinitely often.

We recall that for any symmetric, positive-definite $n \times n$ matrix H, with eigenvalues $0 < \sigma^1 \le \sigma^2 \le \ldots \le \sigma^n$, the determinant $\det(H) = \Pi_{j=1}^n \sigma^j$ and the trace $\mathrm{tr}(H) = \sum_{j=1}^n \sigma^j$. Now let us define the function $\psi(H)$, from the space of $n \times n$, symmetric, positive-definite matrices into the reals, by

$$\psi(H) \triangleq \mathrm{tr}(H) - \ln(\det(H)) = \sum_{j=1}^{n} (\sigma^j - \ln \sigma^j), \tag{28e}$$

where $0 < \sigma^1 \le \sigma^2 \le \ldots \le \sigma^n$ are the eigenvalues of H. We see that $\psi(H) > 0$, and that when $\psi(H)$ is not very large, then σ^1 cannot be very small nor can σ^n be very large. Not surprisingly, as we will soon see, the function $\psi(\cdot)$ turns out to be a very useful tool in proofs.

First we need to develop convenient expressions for $\mathrm{tr}(H_{i+1})$ and $\det(H_{i+1})$. Thus, using the fact that, for any rank-one matrix $A = uv^T$, with $u, v \in \mathbb{R}^n$, $\mathrm{tr}(A) = \langle u, v \rangle$, we conclude from (26e) that

$$\mathrm{tr}(H_{i+1}) = \mathrm{tr}(H_i) + \frac{\|\Delta g_i\|^2}{\langle \Delta g_i, \Delta x_i \rangle} - \frac{\|H_i \Delta x_i\|^2}{\langle \Delta x_i, H_i \Delta x_i \rangle}. \tag{28f}$$

□

The next result is more difficult to establish.

Proposition 1.6.17. *Suppose that H_{i+1} is defined as in (26e), then*

$$\det(H_{i+1}) = \det(H_i)\frac{\langle \Delta g_i, \Delta x_i \rangle}{\langle \Delta x_i, H_i \Delta x_i \rangle}. \tag{29}$$

Proof. First we rewrite (26e) as follows:

$$H_{i+1} = H_i\left[I + \frac{H_i^{-1}\Delta g_i\, \Delta g_i^T}{\langle \Delta g_i, \Delta x_i \rangle} - \frac{(\Delta x_i)(H_i \Delta x_i)^T}{\langle H_i \Delta x_i, \Delta x_i \rangle} \right]. \tag{30a}$$

Next, let

$$u_1 \triangleq \frac{H_i^{-1}\Delta g_i}{\langle \Delta g_i, \Delta x_i \rangle^{1/2}}, \tag{30b}$$

$$u_2 \triangleq \frac{\Delta g_i}{\langle \Delta g_i, \Delta x_i \rangle^{1/2}}, \tag{30c}$$

$$u_3 \triangleq \frac{\Delta x_i}{\langle \Delta x_i , H_i \Delta x_i \rangle^{1/2}}, \tag{30d}$$

and

$$u_4 \triangleq -\frac{H_i \Delta x_i}{\langle \Delta x_i , H_i \Delta x_i \rangle^{1/2}}. \tag{30e}$$

Since $\langle u_1, u_2 \rangle > 0$, it follows from the Sherman-Morrison formula (4) that $I + u_1 u_2^T$ is nonsingular, so that $(I + u_1 u_2^T)^{-1} = I - [1/(1 + \langle u_1, u_2 \rangle)] u_1 u_2^T$. Hence (30a) becomes

$$H_{i+1} = H_i(I + u_1 u_2^T + u_3 u_4^T) = H_i(I + u_1 u_2^T)[I + (I + u_1 u_2^T)^{-1} u_3 u_4^T]. \tag{30f}$$

Because for any vectors $v, w \in \mathbb{R}^n$, $\det(I + vw^T) = 1 + \langle v, w \rangle$, we conclude from (30f) that

$$\det(H_{i+1}) = \det(H_i)(1 + \langle u_1, u_2 \rangle)(1 + \langle (I + u_1 u_2^T)^{-1} u_3, u_4 \rangle). \tag{30g}$$

Substituting for u_i, $i = 1$–4, and simplifying, we obtain (29). □

Lemma 1.6.18. *Suppose that Assumption 1.6.1 is satisfied and that the sequences $\{x_i\}_{i=0}^{\infty}$ and $\{H_i\}_{i=0}^{\infty}$ were constructed by the BFGS-Armijo Algorithm 1.6.16 in the process of solving problem (1). Then, for any $r \in (0,1)$, there exist constants κ_*, σ_*, and $\sigma^* \in (0, \infty)$ such that for any $i > 0$, the inequalities*

$$\kappa_j \triangleq \frac{\langle -\nabla f(x_j), h_j \rangle}{\|\nabla f(x_j)\| \|h_j\|} \geq \kappa_* \tag{31a}$$

and

$$\sigma_* \leq \frac{\|\nabla f(x_i)\|}{\|h_i\|} \leq \sigma^* \tag{31b}$$

hold for at least $i^* \triangleq \lfloor ri \rfloor$ † values of $j \in \overline{1\text{–}i}$.

Proof. First we note that because $H_i \Delta x_i = -\lambda_i \nabla f(x_i)$,

$$\kappa_i = \frac{\langle -\nabla f(x_i), \Delta x_i \rangle}{\|\nabla f(x_i)\| \|\Delta x_i\|} = \frac{\langle H_i \Delta x_i, \Delta x_i \rangle}{\|H_i \Delta x_i\| \|\Delta x_i\|}, \tag{32a}$$

and

† We define $\lfloor i \rfloor \triangleq \max\{k \in \mathbb{N} \mid k \leq i\}$, i.e., the largest integer smaller than or equal to ri.

$$\frac{\|\nabla f(x_i)\|}{\|h_i\|} = \frac{\|H_i \Delta x_i\|}{\|\Delta x_i\|}. \tag{32b}$$

Next, for all $i \in \mathbb{N}$, the Rayleigh quotient q_i is defined by

$$q_i \triangleq \frac{\langle \Delta x_i, H_i \Delta x_i \rangle}{\|\Delta x_i\|^2}. \tag{32c}$$

Then we see from (32a) and (32b) that

$$\frac{\|\nabla f(x_i)\|}{\|h_i\|} = \frac{q_i}{\kappa_i}. \tag{32d}$$

Now we can proceed. First note that, because of Assumption 1.6.1, with $\tilde{H}_j \triangleq \int_0^1 H(x_j + s\Delta x_j) ds$,

$$\frac{\langle \Delta g_j, \Delta x_j \rangle}{\|\Delta x_j\|^2} = \frac{\langle \tilde{H}_j \Delta x_j, \Delta x_j \rangle}{\|\Delta x_j\|^2} \geq m > 0, \tag{32e}$$

and

$$\frac{\|\Delta g_j\|^2}{\langle \Delta g_j, \Delta x_j \rangle} = \frac{\langle \tilde{H}_j \Delta x_j, \tilde{H}_j \Delta x_j \rangle}{\langle H_j \Delta x_j, \Delta x_j \rangle} = \frac{\langle (\tilde{H}_j^{1/2} \Delta x_j), \tilde{H}_j (\tilde{H}_j^{1/2} \Delta x_j) \rangle}{\langle \tilde{H}_j^{1/2} \Delta x_j, \tilde{H}_j^{1/2} \Delta x_j \rangle} \leq M. \tag{32f}$$

It follows from (26e), (28e), (28f), (29), (32e) and (32f) that

$$\begin{aligned}
\psi(H_{i+1}) &= \psi(H_i) + \frac{\|\Delta g_i\|^2}{\langle \Delta g_i, \Delta x_i \rangle} - \frac{\|H_i \Delta x_i\|^2}{\langle \Delta x_i, H_i \Delta x_i \rangle} - \ln \frac{\langle \Delta g_i, \Delta x_i \rangle}{\langle \Delta x_i, H_i \Delta x_i \rangle} \\
&= \psi(H_i) + \frac{\|\Delta g_i\|^2}{\langle \Delta g_i, \Delta x_i \rangle} - \left[\frac{\|H_i \Delta x_i\| \|\Delta x_i\|}{\langle \Delta x_i, H_i \Delta x_i \rangle} \right]^2 \frac{\langle \Delta x_i, H_i \Delta x_i \rangle}{\|\Delta x_i\|^2} \\
&\quad - \ln \left[\frac{\langle \Delta g_i, \Delta x_i \rangle}{\|\Delta x_i\|^2} \frac{\|\Delta x_i\|^2}{\langle \Delta x_i, H_i \Delta x_i \rangle} \right] \\
&= \psi(H_i) + \frac{\|\Delta g_i\|^2}{\langle \Delta g_i, \Delta x_i \rangle} - \frac{q_i}{\kappa_i^2} - \ln \frac{\langle \Delta g_i, \Delta x_i \rangle}{\|\Delta x_i\|^2} + \ln q_i \\
&= \psi(H_i) - 1 + \frac{\|\Delta g_i\|^2}{\langle \Delta g_i, \Delta x_i \rangle} - \ln \frac{\langle \Delta g_i, \Delta x_i \rangle}{\|\Delta x_i\|^2} + \ln \kappa_i^2
\end{aligned}$$

$$+\left(1-\frac{q_i}{\kappa_i^2}+\ln\frac{q_i}{\kappa_i^2}\right)$$

$$\leq \psi(H_i)-1+M-\ln m+\ln\kappa_i^2+\left(1-\frac{q_i}{\kappa_i^2}+\ln\frac{q_i}{\kappa_i^2}\right). \tag{32g}$$

Hence,

$$0<\psi(H_i)\leq\psi(H_0)+(M-1-\ln m)i+\sum_{j=0}^{i-1}\left(\ln\kappa_j^2+1-\frac{q_j}{\kappa_j^2}+\ln\frac{q_j}{\kappa_j^2}\right). \tag{32h}$$

For all $j\in\mathbb{N}$, let

$$\zeta_j\triangleq 1-\frac{q_j}{\kappa_j^2}+\ln\frac{q_j}{\kappa_j^2} \tag{32i}$$

and

$$\eta_j\triangleq-\ln\kappa_j^2-\zeta_j, \tag{32j}$$

so that (32h) becomes

$$0<\psi(H_i)\leq\psi(H_0)+(M-1-\ln m)i-\sum_{j=0}^{i-1}\eta_j. \tag{32k}$$

Hence, we must have that

$$\frac{1}{i}\sum_{j=0}^{i-1}\eta_j<\frac{\psi(H_0)}{i}+(M-1-\ln m). \tag{32l}$$

Now suppose that $r\in(0,1)$ is given, and let $i^*\triangleq\lfloor ri\rfloor$. Let the indices $j_k\in\overline{0\text{–}i-1}$, $k=0,1,\ldots,i$, be such that $\eta_{j_0}\leq\eta_{j_1}\leq\eta_{j_2}\leq\ldots\leq\eta_{j_{i-1}}$, and let $J_{i^*}\triangleq\{j_k\in\overline{0\text{–}i-1}\mid k\leq i^*\}$. Since the function $1-t+\ln t\leq 0$, for all $t>0$, it follows that $\zeta_j\leq 0$ for all j, and hence that $\eta_j\geq 0$ for all j. Consequently, since $\eta_{j_k}\geq\eta_{j_{i^*}}$ for all $k>i^*$ and since all $\eta_j\geq 0$,

$$\frac{1}{i}\sum_{j=0}^{i-1}\eta_j\geq\frac{1}{i}\sum_{k=i^*}^{i-1}\eta_j$$

$$\geq\frac{1}{i}\eta_{j_{i^*}}[1+(i-1)-i^*]\geq\eta_{j_{i^*}}(1-r). \tag{32m}$$

It follows from (32l, m) that, for all $j\in J_{i^*}$,

$$\eta_j\leq\eta_{j_{i^*}}<c_0\triangleq\frac{1}{1-r}[\psi(H_0)+M-1-\ln m]. \tag{32n}$$

Therefore, since $\zeta_j\leq 0$, for all $j\in\mathbb{N}$, it follows from (32j) and (32n) that, for all $j\in J_{i^*}$, $-\ln\kappa_j^2\leq c_0$, which implies that $\kappa_j\geq\kappa_*\triangleq e^{-c_0/2}$, i.e., (31a) holds for all $j\in J_{i^*}$.

Similarly, it follows from (32j) and (32n) that $\zeta_j>-c_0$ for all $j\in J_{i^*}$. Now let us examine (32h). Since the function $u(t)\triangleq 1-t+\ln t$, $t>0$, is such that $u(t)\leq 0$ for all t, $u(1)=0$ and $u(t)\to-\infty$, both as $t\to\infty$ and as $t\to 0$, and $\zeta_j\geq-c_0$, it follows that there exist constants $c_1,c_2>0$ such that, for all $j\in J_{i^*}$, $c_1\leq q_j/\kappa_j^2\leq c_2$. If we now use (31a), the fact that $\kappa_i\leq 1$, and set $c_3=\kappa_*^2c_1$, we find that for all $j\in J_{i^*}$,

$$c_3\leq q_j\leq c_2. \tag{32o}$$

Finally, in view of (32d), we conclude from (31a) and (32o) that for all $j\in J_{i^*}$, (31b) holds, with $\sigma_*=c_3/\kappa_*$ and $\sigma^*=c_2/\kappa_*$. □

Now we are ready to show that on twice continuously differentiable, strictly convex functions, the BFGS-Armijo Variable Metric Algorithm 1.6.16 converges at least R–linearly, in cost.

Theorem 1.6.19. *Suppose that Assumption 1.6.1 is satisfied. If $\{x_i\}_{i=0}^{\infty}$, is a sequence constructed by the BFGS-Armijo Variable Metric Algorithm 1.6.16 in the process of solving problem (1), then*

(a) $x_i\to\hat{x}$, *as* $i\to\infty$;

(b) there exist a $C\in(0,\infty)$ *and a* $\delta\in(0,1)$ *such that, for all* $i\in\mathbb{N}$,

$$f(x_i)-f(\hat{x})\leq C\delta^i, \tag{33a}$$

and

(c)

$$\sum_{i=0}^{\infty}\|x_i-\hat{x}\|<\infty. \tag{33b}$$

Proof. *(a)* First we note that $\{f(x_i)\}_{i=0}^{\infty}$, is a monotone decreasing sequence which is bounded from below because the level set $\{x\in\mathbb{R}^n\mid f(x)\leq f(x_0)\}$ is compact. Hence it must converge to some value f^*. Next, by Lemma 1.6.18, there exists an infinite subset $K\subset\mathbb{N}$ such that (31a,b) hold for all $i\in K$. We will now obtain an estimate of the cost decrease which can be inferred from Assumption 1.6.1, (26a), and (31a,b) at all $i\in K$. Thus, for all $i\in K$, using second-order expansions and Assumption 1.6.1, we obtain

$$f(x_i+\lambda h_i)-f(x_i)-\lambda\alpha\langle\nabla f(x_i),h_i\rangle$$

$$\leq\lambda(1-\alpha)\langle\nabla f(x_i),h_i\rangle+\lambda^2\frac{M}{2}\|h_i\|^2$$

$$= -\lambda(1-\alpha)\kappa_i \|\nabla f(x_i)\| \|h_i\| + \lambda^2 \frac{M}{2}\|h_i\|^2$$

$$= \lambda \frac{M}{2}\|h_i\|^2 [-\frac{2\kappa_i(1-\alpha)}{M}\frac{\|\nabla f(x_i)\|}{\|h_i\|} + \lambda]$$

$$\leq \lambda \frac{M}{2}\|h_i\|^2 [-\frac{2\kappa_*\sigma_*(1-\alpha)}{M} + \lambda] \triangleq \phi(\lambda) . \tag{34a}$$

Since $\phi(\lambda) \leq 0$ for all $\lambda \in [0, 2\kappa_*\sigma_*(1-\alpha)/M]$, and $\lambda_i \leq 1$, by construction, it follows that for all $i \in K$, $\lambda_i = \beta^{k_i} \geq \gamma \triangleq \min\{1, 2\kappa_*\sigma_*(1-\alpha)/M\}$. Hence, for all $i \in K$,

$$f(x_{i+1}) - f(x_i) \leq \lambda_i \alpha \langle \nabla f(x_i), h_i \rangle$$

$$\leq -\gamma\alpha\kappa_i \|\nabla f(x_i)\| \|h_i\| \leq -\gamma\alpha \frac{\kappa_*}{\sigma^*}\|\nabla f(x_i)\|^2 . \tag{34b}$$

We conclude from (34b) that $\nabla f(x_i) \to^K 0$, as $i \to \infty$. For suppose that this is not so. Then there must exist an infinite subset $K' \subset K$ and $b > 0$ such that $\|\nabla f(x_i)\| > b$ for all $i \in K'$. Hence, by (34b), for all $i \in K'$, $f(x_{i+1}) - f(x_i) \leq -\gamma\alpha(\kappa_*/\sigma^*)b^2$. Since this implies that $f(x_i) \to -\infty$, as $i \to \infty$, we obtain a contradiction.

It follows that, if x^* is an accumulation point of $\{x_i\}_{i\in K}$, then $\nabla f(x^*) = 0$. However, $\hat{x}$, the global minimizer of $f(\cdot)$, is the only point in $\mathbb{R}^n$ such that $\nabla f(\hat{x}) = 0$. Hence it follows that $x^* = \hat{x}$ and, in addition, that $f(x_i) \to f(\hat{x})$, as $i \to \infty$. Since there is no other point $x' \in \mathbb{R}^n$ such that $f(x') = f(\hat{x})$, it follows that $\hat{x}$ is the only accumulation point of $\{x_i\}_{i=0}^{\infty}$, i.e., that $x_i \to \hat{x}$, as $i \to \infty$.

(b) Making use of (1.3.5a) and (34b), we find that for all $i \in K$,

$$f(x_{i+1}) - f(x_i) \leq -\frac{2m\gamma\alpha\kappa_*}{\sigma^*}[f(x_i) - f(\hat{x})] . \tag{34c}$$

Adding $f(x_i) - f(\hat{x})$ to both sides of (34c), we obtain that for all $i \in K$,

$$f(x_{i+1}) - f(\hat{x}) \leq \left[1 - \frac{2m\gamma\alpha\kappa_*}{\sigma^*}\right][f(x_i) - f(\hat{x})] . \tag{34d}$$

Note that because the sequence $\{f(x_i)\}_{i=0}^{\infty}$ is monotone decreasing, we must have that $0 \leq 1 - 2m\gamma\alpha\kappa_*/\sigma^* < 1$. Now let $r \in (0,1)$ be arbitrary, and let $\delta \triangleq (1 - 2m\gamma\alpha\kappa_*/\sigma^*)^r$, so that $\delta^{1/r} = (1 - 2m\gamma\alpha\kappa_*/\sigma^*)$. Since in view of Lemma 1.6.18, given any $k \in \mathbb{N}$, (34d) must hold for at least $\lfloor r(k+1) \rfloor$ indices $i \leq k$, and the sequence $\{f(x_i)\}_{i=0}^{\infty}$ is monotone decreasing, we conclude from (34d) that for all $k \in \mathbb{N}$,

$$f(x_k) - f(\hat{x}) \leq (\delta^{1/r})^{\lfloor r(k+1) \rfloor} \leq \delta^k (\delta^{(r-1)/r}[f(x_0) - f(\hat{x})]) \triangleq C\delta^k , \tag{34e}$$

because $\lfloor r(k+1) \rfloor \geq rk + r - 1$.

(c) Finally, using (1.3.5b) and (34d), we conclude that

$$\sum_{i=0}^{\infty} \|x_i - \hat{x}\| \leq \frac{2}{m}\sum_{i=0}^{\infty}[f(x_i) - f(\hat{x})] \leq \frac{2C}{m(1-\delta)} , \tag{34f}$$

which completes our proof. □

Next we will show that the BFGS-Armijo Variable Metric Algorithm 1.6.16 converges Q-superlinearly under Assumption 1.6.1. To this end, first we will establish that the search directions constructed by the BFGS-Armijo Variable Metric Algorithm 1.6.16 converge to the Newton directions. Then we will show that $\lambda_i = 1$ for all i sufficiently large, and last, we will show that the local version of the method (using unity step-sizes) is Q-superlinearly convergent.

We recall that the ideal Newton method defined by (3) computes search directions by solving the equation $H(\hat{x})h = -\nabla f(x_i)$, while the BFGS-Armijo Variable Metric Algorithm 1.6.16 computes search directions by solving the equation $H_i h = -\nabla f(x_i)$. Since $\nabla f(x_i) \to 0$, as $i \to \infty$, the quantity

$$\frac{\|[H_i - H(\hat{x})]h_i\|}{\|h_i\|} = \frac{\|[H_i - H(\hat{x})]\Delta x_i\|}{\|\Delta x_i\|}$$

provides a good, normalized measure of the difference between the BFGS and ideal Newton search directions.

Lemma 1.6.20. *Suppose that Assumption 1.6.1 is satisfied and that the sequences $\{x_i\}_{i=0}^{\infty}$ and $\{H_i\}_{i=0}^{\infty}$ were constructed by the BFGS-Armijo Variable Metric Algorithm 1.6.16 in the process of solving problem (1). Then*

$$\lim_{i\to\infty} \frac{\|[H_i - H(\hat{x})]\Delta x_i\|}{\|\Delta x_i\|} = 0 , \tag{35}$$

and there exist $0 < \sigma_* \leq \sigma^* < \infty$ such that, for all $i \in \mathbb{N}$, the eigenvalues of H_i are contained in the interval $[\sigma_*, \sigma^*]$.

Proof. Since Algorithm 1.6.16 is scale-invariant, we can simplify the linear algebra by assuming that Algorithm 1.6.16 has been applied to problem (27a), with $Q \triangleq H(\hat{x})^{1/2}$. In this case, $\tilde{H}(\hat{z}) = I$. To avoid complicating notation, we now rename the variable z as x and drop the tildes on all the quantities associated with problem (27a). Next we observe that, because $H(\cdot)$ is Lipschitz continuous and because $x_i + s\Delta x_i = (1-s)(x_i - \hat{x}) + s(x_{i+1} - \hat{x}) + \hat{x}$, there exists a constant $L < \infty$ such that

$$\frac{\|\Delta g_i - H(\hat{x})\Delta x_i\|}{\|\Delta x_i\|} \leq \int_0^1 \|H(x_i + s\Delta x_i) - H(\hat{x})\|\, ds$$

$$\leq L[\|x_i - \hat{x}\| + \|x_{i+1} - \hat{x}\|]. \tag{36a}$$

Since, by Theorem 1.6.19, $x_i \to \hat{x}$, as $i \to \infty$, it follows that

$$\lim_{i \to \infty} \frac{\|\Delta g_i - H(\hat{x})\Delta x_i\|}{\|\Delta x_i\|} = 0. \tag{36b}$$

If for all $i \in \mathbb{N}$, we define

$$\varepsilon_i \triangleq \frac{\|\Delta g_i - \Delta x_i\|}{\|\Delta x_i\|}, \tag{36c}$$

then, since $H(\hat{x}) = I$, we conclude from (36b) that $\varepsilon_i \to 0$, as $i \to \infty$. Also, $\|\Delta g_i - \Delta x_i\| = \varepsilon_i \|\Delta x_i\|$. Since $|\|\Delta g_i\| - \|\Delta x_i\|| \leq \|\Delta g_i - \Delta x_i\|$, we conclude that

$$(1 - \varepsilon_i)\|\Delta x_i\| \leq \|\Delta g_i\| \leq (1 + \varepsilon_i)\|\Delta x_i\|. \tag{36d}$$

It now follows from (36c, d) that

$$(1 - \varepsilon_i)^2\|\Delta x_i\|^2 - 2\langle \Delta g_i, \Delta x_i \rangle + \|\Delta x_i\|^2 \leq \|\Delta g_i\|^2 - 2\langle \Delta g_i, \Delta x_i \rangle + \|\Delta x_i\|^2$$

$$= \|\Delta g_i - \Delta x_i\|^2 = \varepsilon_i^2\|\Delta x_i\|^2. \tag{36e}$$

Hence, in view of (36e),

$$2\langle \Delta g_i, \Delta x_i \rangle \geq (2 - 2\varepsilon_i)\|\Delta x_i\|^2. \tag{36f}$$

Consequently,

$$\frac{\langle \Delta g_i, \Delta x_i \rangle}{\|\Delta x_i\|^2} \geq 1 - \varepsilon_i. \tag{36g}$$

Next, since $\langle \Delta g_i, \Delta x_i \rangle = \langle \int_0^1 H(x_i + s\Delta x_i) ds\, \Delta x_i, \Delta x_i \rangle > 0$, by using (36d) we obtain

$$\frac{\|\Delta g_i\|^2}{\langle \Delta g_i, \Delta x_i \rangle} \leq (1 + \varepsilon_i)^2 \frac{\|\Delta x_i\|^2}{\langle \Delta g_i, \Delta x_i \rangle}. \tag{36h}$$

Since $\varepsilon_i \to 0$, as $i \to \infty$, it follows from (36g) that $\|\Delta x_i\|^2 / \langle \Delta g_i, \Delta x_i \rangle$ is bounded and hence that there exists a $c \in (3, \infty)$ and an i_0 such that, for all $i \geq i_0$,

$$\frac{\|\Delta g_i\|^2}{\langle \Delta g_i, \Delta x_i \rangle} \leq \frac{(1 + \varepsilon_i)^2}{1 - \varepsilon_i} \leq 1 + c\varepsilon_i. \tag{36i}$$

Clearly, since $\varepsilon_i \to 0$, as $i \to \infty$, there exists an $i_1 \geq i_0$ such that, for all $i \geq i_1$,

$c\varepsilon_i < \frac{1}{2}$, and also $-\ln(1 - \varepsilon_i) \leq 2c\varepsilon_i$. Consequently, making use of (36g), (36i) and the next to last equality in (32g), we find that for all $i \geq i_1$,

$$\psi(H_{i+1}) \leq \psi(H_i) + 3c\varepsilon_i + \ln \kappa_i^2 + \left[1 - \frac{q_i}{\kappa_i^2} + \ln \frac{q_i}{\kappa_i^2}\right], \tag{36j}$$

where κ_i was defined in (31a) and q_i was defined in (32c). Hence for any $k \in \mathbb{N}$, summing (36j) from $i = 0$ to $i = k - 1$ yields

$$\psi(H_k) \leq \psi(H_0) + \sum_{i=0}^{k-1} \left[3c\varepsilon_i + \ln \kappa_i^2 + \left[1 - \frac{q_i}{\kappa_i^2} + \ln \frac{q_i}{\kappa_i^2}\right]\right]. \tag{36k}$$

Now, since, by Assumption 1.6.1, $H(\cdot)$ is Lipschitz continuous and $H(\hat{x}) = I$, it follows from (36a) that there exists a constant $L < \infty$ such that

$$\varepsilon_i = \frac{\|\int_0^1 [H(x_i + s\Delta x_i) - I]\, ds\, \Delta x_i\|}{\|\Delta x_i\|}$$

$$\leq L[\|x_i - \hat{x}\| + \|x_{i+1} - \hat{x}\|]$$

$$\leq 2L \max\{\|x_i - \hat{x}\|, \|x_{i+1} - \hat{x}\|\}. \tag{36l}$$

Since by (33b), $\sum_{i=0}^{\infty} \|x_i - \hat{x}\| < \infty$, it follows from (36l) that $\sum_{i=0}^{\infty} \varepsilon_i < \infty$. Because $\ln \kappa_i^2 \leq 0$ and $(1 - q_i/\kappa_i^2 + \ln q_i/\kappa_i^2) \leq 0$ for all i, we conclude from (36k) that the sequence $\{\psi(H_i)\}_{i=0}^{\infty}$ is bounded from above. Since, by construction, $\psi(H_i)$ is bounded from below by zero, it follows that $\{\psi(H_i)\}_{i=0}^{\infty}$ is a bounded sequence and hence that

$$\ln \kappa_i \to 0, \tag{36m}$$

and

$$\left[1 - \frac{q_i}{\kappa_i^2} + \ln \frac{q_i}{\kappa_i^2}\right] \to 0, \text{ as } i \to \infty. \tag{36n}$$

Hence, since $t = 1$ is the only root of the equation $1 - t + \ln t = 0$, we conclude that

$$\lim_{i \to \infty} \kappa_i = \lim_{i \to \infty} q_i = 1. \tag{36o}$$

Now let

$$\delta_i \triangleq \frac{\|(H_i - I)\Delta x_i\|^2}{\|\Delta x_i\|^2} = \frac{\|H_i \Delta x_i\|^2 - 2\langle \Delta x_i, H_i \Delta x_i \rangle + \|\Delta x_i\|^2}{\|\Delta x_i\|^2}$$

$$= \frac{q_i^2}{\kappa_i^2} - 2q_i + 1 . \tag{36p}$$

It follows directly from (36o) that $\delta_i \to 0$, as $i \to \infty$, which implies (35). Since the sequence $\{\psi(H_i)\}_{i=0}^{\infty}$ is bounded, it follows from the definition of $\psi(\cdot)$ that there exists $0 < \sigma_* \le \sigma^* < \infty$ such that the eigenvalues of the matrices H_i are contained in the interval $[\sigma_*, \sigma^*]$. Since the fact that (35) must hold in the original coordinate system is now obvious, our proof is complete. □

Now we can complete the last two steps of our demonstration that the BFGS Algorithm 1.6.16 converges Q-superlinearly.

Theorem 1.6.21. *Suppose that Assumption 1.6.1 is satisfied. If $\{x_i\}_{i=0}^{\infty}$, is a sequence constructed by the BFGS-Armijo Variable Metric Algorithm 1.6.16 in the process of solving problem (1), then $x_i \to \hat{x}$, as $i \to \infty$ Q-superlinearly.*

Proof. Since we have already established convergence in Theorem 1.6.19, we need to show only that the convergence is Q-superlinear. We begin by showing that there exists an i_0 such that, for all $i \ge i_0$, $\lambda_i = 1$. Thus, consider the step-size test in (26a), where we set $k = 0$. Then, making use of the second-order expansion formula (5.1.17d), we find that for some $s_i \in [0, 1]$

$$f(x_i + h_i) - f(x_i) - \alpha \langle \nabla f(x_i), h_i \rangle$$

$$= (1 - \alpha)\langle \nabla f(x_i), h_i \rangle + \tfrac{1}{2} \langle h_i, H(x_i + s_i h_i) h_i \rangle . \tag{37a}$$

Taking into account that $\nabla f(x_i) = -H_i h_i$ and adding and subtracting $\frac{1}{2}\langle h_i, H_i h_i \rangle$ in (37a), we find that

$$f(x_i + h_i) - f(x_i) - \alpha\langle \nabla f(x_i), h_i \rangle$$

$$= -(\tfrac{1}{2} - \alpha)\langle H_i h_i, h_i \rangle + \tfrac{1}{2}\langle h_i, [H(x_i + s_i h_i) - H_i] h_i \rangle$$

$$\le -\|h_i\|^2 \left[\sigma_*(\tfrac{1}{2} - \alpha) - \|H(x_i + s_i h_i) - H(\hat{x})\| - \frac{\|(H_i - H(\hat{x}))h_i\|}{\|h_i\|} \right], \tag{37b}$$

where we have used the lower bound σ_* on the eigenvalues of H_i, established in Lemma 1.6.20. Now $h_i \to 0$, as $i \to \infty$, because $\nabla f(x_i) \to 0$, as $i \to \infty$ and the eigenvalues of H_i are bounded. Hence, by continuity,

$$\|H(x_i + s_i h_i) - H(\hat{x})\| \to 0 ,$$

as $i \to \infty$. Next, since $\lambda_i h_i = \Delta x_i$, it follows from Lemma 1.6.20 (i.e., from (35)) that $(\|(H_i - H(\hat{x}))h_i\| / \|h_i\|) \to 0$, as $i \to \infty$. Hence there must exist an i_0 such that the right-hand side of (37b) is negative, which implies that $\lambda_i = 1$ for all $i \ge i_0$.

Clearly, since $\nabla f(\hat{x}) = 0$, for all $i \ge i_0$,

$$H_i \Delta x_i = -\nabla f(x_i) + \nabla f(\hat{x})$$

$$= -\int_0^1 H(\hat{x} + s(x_i - \hat{x}))\, ds\, (x_i - \hat{x}) . \tag{37c}$$

Let $\tilde{H}_i \triangleq \int_0^1 H(\hat{x} + s(x_i - \hat{x}))\, ds$. Then we deduce from (37c) that

$$H(\hat{x})(x_{i+1} - x_i) = [H(\hat{x}) - H_i]\Delta x_i - \tilde{H}_i(x_i - \hat{x}) . \tag{37d}$$

Adding $H(\hat{x})(x_i - \hat{x})$ to both sides of (37d), we conclude that

$$H(\hat{x})(x_{i+1} - \hat{x}) = [H(\hat{x}) - \tilde{H}_i](x_i - \hat{x}) + [H(\hat{x}) - H_i]\Delta x_i . \tag{37e}$$

Since by Assumption 1.6.1, $H(\cdot)$ is Lipschitz continuous, there exists an $L < \infty$ such that $\|H(\hat{x}) - \tilde{H}_i\| \le L\|x_i - \hat{x}\|$. It follows that

$$\|x_{i+1} - \hat{x}\| \le \frac{1}{\sigma_*}\left[L\|x_i - \hat{x}\|^2 + \frac{\|[H(\hat{x}) - H_i]\Delta x_i\|}{\|\Delta x_i\|}\|\Delta x_i\| \right] . \tag{37f}$$

Let

$$\nu_i \triangleq (1/\sigma_*)\frac{\|[H(\hat{x}) - H_i]\Delta x_i\|}{\|\Delta x_i\|} . \tag{37g}$$

Then, since

$$\|\Delta x_i\| \le \|x_{i+1} - \hat{x}\| + \|x_i - \hat{x}\| , \tag{37h}$$

we conclude from (37f) that

$$(1 - \nu_i)\|x_{i+1} - \hat{x}\| \le \frac{1}{\sigma_*}[L\|x_i - \hat{x}\| + \sigma_* \nu_i]\|x_i - \hat{x}\| . \tag{37i}$$

Because both $\|x_i - \hat{x}\|$ and ν_i converge to zero, as $i \to \infty$, the desired result follows directly. □

1.6.7 Notes

An earlier version of the Secant Algorithm 1.6.3, using a somewhat more complex globalization technique, was described in [Pol.74]. The globalization technique used in the Secant Algorithm 1.6.3 serves the purpose of illustrating the versatility of the Polak-Sargent-Sebastian theorem as a tool for designing such mechanisms. For a more complete presentation of secant methods, see [OrR.70] and [Bre.72].

The earliest variable metric method was introduced by Davidon [Dav.59]. It was promulgated and clarified by Fletcher and Powell in their seminal paper [FlP.63]. This method has become known as the DFP method. Since that time, many individuals have labored at developing various versions of variable metric methods and proving their convergence and rate of convergence properties. For a survey of this work, see [DeM.77] and [DeS.89], which contain a large number of references. For a unified treatment of conjugate directions and variable metric methods, as well as a survey of

the relevant literature, see the recent monograph [Naz.94].

The symmetric rank-one update that we presented was proposed independently by Broyden [Bro.67], Davidon [Dav.68], Fiacco and McCormick [FiM.68], Murtagh and Sargent [MuS.69] and Wolfe [Wol.68]. Broyden [Bro.70], Fletcher [Fle.70], Goldfarb [Gol.70] and Shanno [Sha.70] are responsible for the BFGS version bearing their initials that has emerged as the most efficient of variable metric methods. Convergence proofs for variable metric methods using trust regions and other inexact line searches were presented in [Pow.76] and [Wer.78].

Our exposition of variable metric methods draws on Fletcher and Powell [FlP.63], the survey paper by Dennis and MoreMorè [DeM.77], and the elegant convergence analysis by Byrd and Nocedal in [ByN.89], which in turn, relies on results in [DeM.74a, Pow.76, Wer.78, DeM.74, BLN.87, BNY.87, Rit.79, Rit.81].

As a practical matter, the BFGS-based variable metric method loses its "hereditary positive-definiteness" property on nonconvex problems, and hence it may fail to construct descent directions. In that case, it is probably easiest to stabilize it using trust region techniques that modify the matrix H_i by adding a multiple of the identity matrix, as in the Levenberg-Marquardt method [Lev.44, Mar. 63].

1.7 One-Dimensional Optimization

We saw that the method of steepest descent as well as the conjugate gradient and variable metric methods require solving one-dimensional minimization problems of the form

$$\min_{\lambda \geq 0} f(x + \lambda h), \tag{1a}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and $\langle \nabla f(x), h \rangle < 0$. We can write (1a) as

$$\min_{\lambda \geq 0} \phi(\lambda), \tag{1b}$$

with $\phi(\lambda) \triangleq f(x + \lambda h)$ and $\phi'(0) < 0$. Note that the derivative $\phi'(\lambda)$ is given by

$$\phi'(\lambda) = \langle \nabla f(x + \lambda h), h \rangle \tag{1c}$$

and that it can be expensive to compute if formula (2a) is used. When high precision is not important, it is much cheaper to evaluate $\phi'(\lambda)$ by finite differences, i.e., by using a formula such as

$$\phi'(\lambda) \approx \frac{\phi(\lambda + \varepsilon) - \phi(\lambda)}{\varepsilon} \tag{1d}$$

or

$$\phi'(\lambda) \approx \frac{\phi(\lambda + \varepsilon) - \phi(\lambda - \varepsilon)}{2\varepsilon}. \tag{1e}$$

We will consider the slightly more general problem

$$\min_{\lambda \in \mathbb{R}} \phi(\lambda), \tag{2}$$

under the assumption that $\phi(\cdot)$ is at least once continuously differentiable.

We will discuss three commonly used methods for solving (2).

1.7.1 Secant Method Based on Cubic Interpolation

The common, garden variety secant method, already mentioned in the preceding section, is a discrete form of Newton's method for solving equations of the form

$$g(\lambda) = 0, \tag{3}$$

where $g : \mathbb{R} \to \mathbb{R}$. Given two points, λ_{i-1} and λ_i, it approximates the derivative $g'(\lambda_i)$ by the finite difference expression

$$\tilde{g}'(\lambda_i, \lambda_{i-1}) \triangleq \frac{g(\lambda_i) - g(\lambda_{i-1})}{\lambda_i - \lambda_{i-1}}, \tag{4a}$$

and computes the next point λ_{i+1} by the formula

$$\lambda_{i+1} = \lambda_i - \tilde{g}'(\lambda_i, \lambda_{i-1})^{-1} g(\lambda_i). \tag{4b}$$

We can establish its local convergence properties by following the pattern used in proving Theorem 1.4.1, as follows.

Theorem 1.7.1. *Consider Problem (3). Suppose that*

(i) $g : \mathbb{R} \to \mathbb{R}$ *is Lipschitz continuously differentiable on bounded sets, and*

(ii) the derivative $g'(\lambda)$ *is not zero for all* λ *near a solution* $\hat{\lambda}$ *of (3).*

Then there exists a $\hat{\rho} > 0$ *such that, if* $\lambda_0 \in B(\hat{\lambda}, \hat{\rho})$ *and the sequence* $\{\lambda_i\}_{i=0}^{\infty}$ *is constructed according to the recursion (4), then* $\lambda_i \to \hat{\lambda}$, *as* $i \to \infty$, *superlinearly, with root rate* $\tau_1 \approx 1.618$.

Proof. By the Mean-Value Theorem 5.1.28*(a)*, for any $\lambda, \lambda' \in \mathbb{R}$, $\tilde{g}'(\lambda, \lambda') = g'(\lambda + s(\lambda' - \lambda))$, with $s \in [0, 1]$, Since $g'(\cdot)$ is continuous, there exist $\rho > 0$ and $b < \infty$ such that, for all $\lambda, \lambda' \in B(\hat{\lambda}, \rho)$, $\tilde{g}'(\lambda, \lambda')$ is nonzero and

$$|\tilde{g}'(\lambda, \lambda')^{-1}| \leq b. \tag{5a}$$

Since $g'(\cdot)$ is Lipschitz continuous on bounded sets, there exists an $L \in (0,\infty)$ such that

$$|g'(\lambda') - g'(\lambda)| \leq L|\lambda - \lambda'|, \quad \forall \lambda', \lambda \in B(\hat{\lambda}, \rho). \tag{5b}$$

Suppose that $\lambda_{i-1}, \lambda_i \in B(\hat{\lambda}, \rho)$, then, by construction and because $g(\hat{\lambda}) = 0$ by assumption, it follows from (4b) (via the expansion formula (5.1.17a)) that

$$\begin{aligned} \tilde{g}'(\lambda_i, \lambda_{i-1})(\lambda_{i+1} - \lambda_i) &= -g(\lambda_i) + g(\hat{\lambda}) \\ &= -\int_0^1 g'(\hat{\lambda} + s(\lambda_i - \hat{\lambda}))\, ds\, (\lambda_i - \hat{\lambda}). \end{aligned} \tag{5c}$$

Hence, adding $\tilde{g}'(\lambda_i, \lambda_{i-1})(\lambda_i - \hat{\lambda}) = \int_0^1 g'(\lambda_{i-1} + s(\lambda_i - \lambda_{i-1}))\, ds\, (\lambda_i - \hat{\lambda})$ to both sides of (5c), we obtain

$$\begin{aligned} |\lambda_{i+1} - \hat{\lambda}| &\leq |\tilde{g}'(\lambda_i)^{-1}| \int_0^1 |g'(\lambda_{i-1} + s(\lambda_i - \lambda_{i-1})) - g'(\hat{\lambda} + s(\lambda_i - \hat{\lambda}))|\, ds\, |\lambda_i - \hat{\lambda}| \\ &\leq \tfrac{1}{2} bL\, |\lambda_i - \hat{\lambda}|\, |\lambda_{i-1} - \hat{\lambda}| \end{aligned} \tag{5d}$$

Therefore, if $\tfrac{1}{2} bL\, |\lambda_{i-1} - \hat{\lambda}| < 1$, then $|\lambda_{i+1} - \hat{\lambda}| < |\lambda_i - \hat{\lambda}|$ and hence $\lambda_{i+1} \in B(\hat{\lambda}, \rho)$. Therefore, if for any $\alpha \in (0,1)$, we define

$$\hat{\rho} \triangleq \min\{\rho, 2\alpha / bL\}, \tag{5e}$$

we find, by induction, that, if $\lambda_0, \lambda_1 \in B(\hat{\lambda}, \hat{\rho})$, then the entire sequence $\{\lambda_i\}_{i=0}^{\infty}$, constructed by the secant method (4b), is well defined, contained in $B(\hat{\lambda}, \hat{\rho})$, and satisfies the relation (5e). The desired result now follows from Theorem 1.2.43. □

Now let us return to problem (2) and assume that $\phi(\cdot)$ is twice Lipschitz continuously differentiable on bounded sets and that, at a local solution $\hat{\lambda}$ of (2), $\phi''(\hat{\lambda}) > 0$. We associate problem (2) with problem (3) by defining $g(\lambda) \triangleq \phi'(\lambda)$. Given $\lambda_i \neq \lambda_{i-1}$, if we allow ourselves to compute not only the two derivatives $\phi'(\lambda_{i-1})$, $\phi'(\lambda_i)$, but also the values $\phi(\lambda_{i-1})$, $\phi(\lambda_i)$ (as will be the case in a descent method for solving (2)), then we have two more pieces of information. These four pieces of information can be used to construct a cubic polynomial, $p_i(\lambda) = a_i^3\lambda^3 + a_i^2\lambda^2 + a_i^1\lambda + a_i^0$, which interpolates this data, i.e., $p_i(\lambda_j) = \phi(\lambda_j)$, and $p'(\lambda_j) = \phi'(\lambda_j)$, for $j = i, i-1$. The coefficients of this cubic polynomial are the solution of the following linear equation:

$$\begin{bmatrix} \lambda_i^3 & \lambda_i^2 & \lambda_i & 1 \\ \lambda_{i-1}^3 & \lambda_{i-1}^2 & \lambda_{i-1} & 1 \\ 3\lambda_i^2 & 2\lambda_i & 1 & 0 \\ 3\lambda_{i-1}^2 & 2\lambda_{i-1} & 1 & 0 \end{bmatrix} \begin{bmatrix} a_i^3 \\ a_i^2 \\ a_i^1 \\ a_i^0 \end{bmatrix} = \begin{bmatrix} \phi(\lambda_i) \\ \phi(\lambda_{i-1}) \\ \phi'(\lambda_i) \\ \phi'(\lambda_{i-1}) \end{bmatrix}. \tag{6a}$$

The matrix in (6a) is a van der Monde matrix and hence known to be nonsingular. There are several closed form expressions for $p_i(\cdot)$. These include the well-known Lagrange interpolation formula

$$\begin{aligned} p_i(\lambda) = {} & \frac{\phi(\lambda_i)}{(\lambda_i - \lambda_{i-1})^3}(\lambda - \lambda_{i-1})^3 + \frac{\phi(\lambda_{i-1})}{(\lambda_{i-1} - \lambda_i)^3}(\lambda - \lambda_i)^3 \\ & + \frac{\phi'(\lambda_i)(\lambda_i - \lambda_{i-1}) - 3\phi(\lambda_i)}{(\lambda_i - \lambda_{i-1})^3}(\lambda - \lambda_{i-1})^2(\lambda - \lambda_i) \\ & + \frac{\phi'(\lambda_{i-1})(\lambda_{i-1} - \lambda_i) - 3\phi(\lambda_{i-1})}{(\lambda_{i-1} - \lambda_i)^3}(\lambda - \lambda_i)^2(\lambda - \lambda_{i-1}), \end{aligned} \tag{6b}$$

as well as the following easily verified formula obtained by solving (6a):

$$p_i(\lambda) = \phi(\lambda_{i-1}) + \phi'(\lambda_{i-1})(\lambda - \lambda_{i-1}) + c_i(\lambda - \lambda_{i-1})^2 + d_i(\lambda - \lambda_{i-1})^2(\lambda - \lambda_i), \tag{6c}$$

where

$$c_i \triangleq \frac{1}{\lambda_i - \lambda_{i-1}}\left[\frac{\phi(\lambda_i) - \phi(\lambda_{i-1})}{\lambda_i - \lambda_{i-1}} - \phi'(\lambda_{i-1})\right] \tag{6d}$$

and

$$d_i \triangleq \frac{1}{(\lambda_i - \lambda_{i-1})^2}\left[\phi'(\lambda_i) - 2\frac{\phi(\lambda_i) - \phi(\lambda_{i-1})}{\lambda_i - \lambda_{i-1}} + \phi'(\lambda_{i-1})\right]. \tag{6e}$$

Given $\lambda_i, \lambda_{i-1} \in \mathbb{R}$, if we define an approximation $\tilde{\phi}''(\lambda_i)$ to $\phi''(\lambda_i)$, using (4a), with $g(\cdot) \triangleq \phi'(\cdot)$, we deduce from the Mean-Value Theorem 5.1.28*(b)* that $\tilde{\phi}''(\lambda_i) = \phi''(\lambda_{i-1} + s_i(\lambda_i - \lambda_{i-1}))$, for some $s_i \in [0,1]$. Hence, since $\phi''(\cdot)$ is Lipschitz continuous on bounded sets by assumption,

$$|\tilde{\phi}''(\lambda_i) - \phi''(\lambda_i)| \leq L|\lambda_i - \lambda_{i-1}|, \tag{6f}$$

for some $L < \infty$. We will now show that, when $\phi(\cdot)$ is four times continuously differentiable, $p_i''(\lambda_i)$ is a much better approximation to $\phi''(\lambda_i)$ than $\tilde{\phi}''(\lambda_i)$. The following result is a special case of Theorem A-1 in [Tra.64].

Lemma 1.7.2. *Suppose that $\phi(\cdot)$ is four times continuously differentiable.*

Then, given any $\lambda_i \neq \lambda_{i-1} \in \mathbb{R}$, *there exists a* $\mu \in [\lambda_{i-1}, \lambda_i]$[†] *such that*

$$p_i''(\lambda_i) - \phi''(\lambda_i) = \frac{2}{4!}\phi^{(4)}(\mu)(\lambda_i - \lambda_{i-1})^2 . \tag{7}$$

Proof. Let $\nu : \mathbb{R} \to \mathbb{R}$ be defined by

$$\nu(\lambda) \triangleq (\lambda - \lambda_i)^2(\lambda - \lambda_{i-1})^2 . \tag{8a}$$

Then

$$\nu''(\lambda_i) = 2(\lambda_i - \lambda_{i-i}) . \tag{8b}$$

Let

$$\sigma \triangleq [\phi''(\lambda_i) - p_i''(\lambda_i)] / \nu''(\lambda_i) , \tag{8c}$$

and let $F : \mathbb{R} \to \mathbb{R}$ be defined by

$$F(\lambda) \triangleq \phi(\lambda) - p_i(\lambda) - \sigma\nu(\lambda) . \tag{8d}$$

Then we see that $F(\lambda_i) = F'(\lambda_i) = F''(\lambda_i) = 0$ (the last by definition of σ), and also $F(\lambda_{i-1}) = F'(\lambda_{i-1}) = 0$, i.e., $F(\cdot)$ has at least five zeros in the interval $[\lambda_i, \lambda_{i-1}]$. It now follows from a straightforward generalization of Rolle's Theorem (see, e.g.,[Apo.60]) that there exists a $\mu \in [\lambda_i, \lambda_{i-1}]$ such that $F^{(4)}(\mu) = 0$. Since $p_i^{(4)}(\mu) = 0$ and $\nu^{(4)}(\mu) = 4!$, it follows that

$$0 = F^{(4)}(\mu) = \phi^{(4)}(\mu) - 0 - \sigma\nu^4(\mu) = \phi^{(4)}(\mu) - \sigma 4! , \tag{8e}$$

which leads to the conclusion that

$$\sigma = [\phi''(\lambda_i) - p_i''(\lambda_i)] / \nu''(\lambda_i) = \phi^{(4)}(\mu) / 4! . \tag{8f}$$

Substituting for $\nu''(\lambda_i)$ from (8b) into (8d), we obtain the desired result. □

In view of the above result, we can expect that the local method defined by

$$\lambda_{i+1} = \lambda_i - p_i''(\lambda_i)^{-1} g(\lambda_i) , \quad i = 0, 1, 2, \ldots , \tag{9}$$

converges faster than the secant method.

Kirjner-Neto-Polak Stabilized Cubic-Secant Algorithm 1.7.3.

Parameters. $\alpha \in (0, \frac{1}{2})$, $\beta \in (0, 1)$, machine precision parameters $\sigma_* \ll 1$, $\sigma^* \gg 1$.

Data. $\lambda_0, \lambda_1 \in \mathbb{R}$.

[†] We denote an interval with end points λ_{i-1}, λ_i by $[\lambda_{i-1}, \lambda_i]$, irrespective of whether $\lambda_{i-1} < \lambda_i$ or not.

Step 0. Set $i = 0$.

Step 1. Compute $p_i''(\lambda_i)$ using formula (6c), which gives

$$p_i''(\lambda_i) = 2c_i + 4d_i(\lambda_i - \lambda_{i-1}) . \tag{10a}$$

If $\sigma_* \leq p_i''(\lambda_i) \leq \sigma^*$, set

$$\eta_i \triangleq -p_i''(\lambda_i)^{-1}\phi'(\lambda_i) . \tag{10b}$$

Else, set

$$\eta_i \triangleq -\phi'(\lambda_i) . \tag{10c}$$

Step 2. Compute the *Armijo step-size*

$$s_i = \max_{k \in \mathbb{N}} \{\beta^k \mid \phi(\lambda_i + \beta^k\eta_i) - \phi(\lambda_i) \leq \alpha\beta^k \eta_i\phi'(\lambda_i)\} . \tag{10d}$$

Step 3. Set $\lambda_{i+1} = \lambda_i + s_i\eta_i$, replace i by $i + 1$, and go to Step 1.

Theorem 1.7.4. *Suppose that* $\phi : \mathbb{R} \to \mathbb{R}$ *in (2) is four times continuously differentiable, that* $\phi'(0) < 0$, *and that there exists an* $M \in (1, \infty)$ *such that* $|\phi''(\lambda)| \leq M$ *for all* $\lambda \in \mathbb{R}$. *Then,*

(a) Algorithm 1.7.3 satisfies the hypotheses of Theorem 1.2.24b, and hence if $\hat{\lambda}$ *is an accumulation point of a sequence* $\{\lambda_i\}_{i=0}^{\infty}$ *constructed by Algorithm 1.7.2, then* $\phi'(\hat{\lambda}) = 0$, *and*

(b) if Algorithm 1.7.3 constructs a sequence $\{\lambda_i\}_{i=0}^{\infty}$ *which has an accumulation point* $\hat{\lambda}$ *satisfying the second-order sufficiency condition (1.1.12a), i.e.,* $\phi''(\hat{\lambda}) \geq m$, *with* $m \geq 3\sigma_*$, *then* $\{\lambda_i\}_{i=0}^{\infty}$ *converges to* $\hat{\lambda}$, *R–quadratically.*

Proof. *(a)* Suppose that $\{\lambda_i\}_{i=0}^{\infty}$ was constructed by Algorithm 1.7.3, and that $\lambda_i \to^K \hat{\lambda}$, as $i \to \infty$, for some $K \subset \mathbb{N}$. Since $\phi'(0) < 0$, $\hat{\lambda} > 0$. First, it follows from (6d) and the second-order expansion formula (5.1.18a) that

$$c_i = \frac{1}{(\lambda_i - \lambda_{i-1})^2}\left[\phi(\lambda_i) - \phi(\lambda_{i-1}) - \phi'(\lambda_{i-1})(\lambda_i - \lambda_{i-1})\right] = \tfrac{1}{2}\phi''(\mu) , \tag{11a}$$

where $\mu \in [\lambda_{i-1}, \lambda_i]$. Similarly, it follows from (6e) and the second-order expansion formula (5.1.17d) that

$$d_i(\lambda_i - \lambda_{i-1}) = \frac{1}{(\lambda_i - \lambda_{i-1})^2}\{\phi(\lambda_{i-1}) - \phi(\lambda_i) - \phi'(\lambda_i)(\lambda_{i-1} - \lambda_i)$$

$$- [\phi(\lambda_i) - \phi(\lambda_{i-1}) - \phi'(\lambda_{i-1})(\lambda_i - \lambda_{i-1})] \Big\}$$

$$= \tfrac{1}{2}\phi''(\mu') + \tfrac{1}{2}\phi''(\mu''), \tag{11b}$$

where $\mu', \mu'' \in [\lambda_{i-1}, \lambda_i]$. Therefore it follows from (10a) that

$$|p_i''(\lambda_i)| = |2c_i + 4d_i(\lambda_i - \lambda_{i-1})| \leq 5M . \tag{11c}$$

We conclude from (11c) and Step 2 in Algorithm 1.7.3 that, for any $i \in \mathbb{N}$,

$$\phi'(\lambda_i)\eta_i \leq -\min\{1, \frac{1}{5M}\}\phi'(\lambda_i)^2 \tag{11d}$$

and

$$|\eta_i| \leq \max\{1, \frac{1}{m}\}|\phi'(\lambda_i)| . \tag{11e}$$

In view of (11d,e), the desired result now follows directly from the Polak-Sargent-Sebastian Theorem 1.2.24.

(b) First we will show that $\lambda_i \to \hat{\lambda}$, as $i \to \infty$. Let $\omega \triangleq \sigma_* m / M \leq 1$, with σ_* as in Algorithm 1.7.3. Clearly, there exists a $\rho > 0$ such that $\phi''(\lambda) \geq m/2$ for all $\lambda \in B(\hat{\lambda}, \rho)$, and hence, since $\phi'(\hat{\lambda}) = 0$, it follows from the second-order expansion formula (5.1.18c) that $\phi(\lambda) - \phi(\hat{\lambda}) \geq \frac{1}{4}m(\lambda - \hat{\lambda})^2$, for all $\lambda \in B(\hat{\lambda}, \rho)$. By assumption, there exists an infinite subset $K \subset \mathbb{N}$ such that $\lambda_i \to^K \hat{\lambda}$, as $i \to \infty$. Since the sequence $\{\phi(\lambda_i)\}_{i=0}^{\infty}$ is monotone decreasing and $\phi(\hat{\lambda})$ is an accumulation point of this sequence, it follows from Proposition 5.1.16 that $\{\phi(\lambda_i)\}_{i=0}^{\infty}$ converges to $\phi(\hat{\lambda})$. Hence, there exists an i_0 such that $\lambda_{i_0} \in B(0, \rho)$ and $\phi(\lambda_i) - \phi(\hat{\lambda}) \leq \rho^2\omega^2 m/16$, for all $i \in \mathbb{N}$, $i \geq i_0$. In view of the above, we must have that $\frac{1}{4}m(\lambda_{i_0} - \hat{\lambda})^2 \leq \phi(\lambda_{i_0}) - \phi(\hat{\lambda}) \leq \rho^2\omega^2 m/16$, which implies that $|\lambda_{i_0} - \hat{\lambda}| \leq \frac{1}{2}\rho\omega < \frac{1}{2}\rho$. Since $\phi'(\lambda_{i_0}) = \phi''(\nu)(\lambda_i - \hat{\lambda})$ for some $\nu \in [\lambda_{i_0}, \hat{\lambda}]$, we must have that $|\phi'(\lambda_{i_0})| \leq \frac{1}{2}M\rho\omega$ and hence it follows from (11e), that $|\eta(\lambda_{i_0})| \leq \frac{1}{2}\rho$. Since the step-size $s_i \leq 1$, we conclude that $\lambda_{i_0+1} \in B(\hat{\lambda}, \rho)$ and, in addition, that $\phi(\lambda_{i_0+1}) - \phi(\hat{\lambda}) \leq \rho^2\omega^2 m/16$. It now follows by induction that $\lambda_i \in B(\hat{\lambda}, \rho)$, for all $i \geq i_0$. Since $B(\hat{\lambda}, \rho)$ is compact, and $\hat{\lambda}$ is the only stationary point in $B(\hat{\lambda}, \rho)$, it follows that $\lambda_i \to \hat{\lambda}$, as $i \to \infty$.

Next we will show that there exists an i_1 such that, for all $i \geq i_1$, the step-size $s_i = 1$. Since by assumption, $\phi''(\hat{\lambda}) \geq 3\sigma_*$, it follows by continuity of $\phi''(\cdot)$ that there exists a $\rho' > 0$ such that, for all $\lambda \in B(\hat{\lambda}, \rho')$ and any $s \in [0, 1]$,

$\phi''(\hat{\lambda} + s(x - \hat{\lambda})) \geq 2\sigma_*$. Now, as we have shown, $\lambda_i \to \hat{\lambda}$, as $i \to \infty$. Hence there exists an $i' \in \mathbb{N}$ such that, for all $i \geq i'$, $\lambda_i \in B(\hat{\lambda}, \rho')$ and, in addition, in view of Lemma 1.7.2, $p_i''(\lambda_i) \geq \sigma_*$. Therefore, by construction in Step 2 of Algorithm 1.7.3, $\phi'(\lambda_i) = -p_i''(\lambda_i)\eta_i$. Hence, using the second-order expansion formula (5.1.17c), we find that, for some $\sigma_i \in [0, 1]$,

$$\begin{aligned}\phi(\lambda_i + \eta_i) - \phi(\lambda_i) - \alpha\eta_i\phi'(\lambda_i) &= (1 - \alpha)\phi'(\lambda_i)\eta_i + \tfrac{1}{2}\phi''(\lambda_i + \sigma_i\eta_i)\eta_i^2 \\ &= -(\tfrac{1}{2} - \alpha)\eta_i^2 p_i''(\lambda_i) \\ &\quad + \tfrac{1}{2}[\phi''(\lambda_i + \sigma_i\eta_i) - p_i''(\lambda_i)]\eta_i^2] . \end{aligned} \tag{11f}$$

If we add and subtract $\frac{1}{2}\eta_i^2\phi''(\lambda_i)$ in the second term in the second line of (11f) and use (7) and the local Lipschitz continuity of $\phi''(\cdot)$, we conclude that there exists an $i'' \geq i'$ and a $C < \infty$ such that, for all $i \geq i_1$,

$$\begin{aligned}\phi(\lambda_i + \eta_i) - f(\lambda_i) - \alpha\eta_i\phi'(\lambda_i) \leq \eta_i^2 \{&-(\tfrac{1}{2} - \alpha)m \\ &+ C[|\eta_i| + (\lambda_i - \lambda_{i-1})^2]\} . \end{aligned} \tag{11g}$$

Since both $\lambda_i - \lambda_{i-1} \to 0$ and $\eta_i \to 0$ as $i \to \infty$, it follows from (11g) that there exists an $i_1 \geq i''$, such that $\phi(\lambda_i + \eta_i) - f(\lambda_i) \leq \alpha\eta_i\phi'(\lambda_i)$ for all $i \geq i_1$, which implies that $\lambda_i = 1$ for all $i \geq i_1$.

From the above and the Mean-Value Theorem 5.1.28*(a)*, it follows that for all $i \geq i_1$,

$$p_i''(\lambda_i)(\lambda_{i+1} - \lambda_i) = -\phi'(\lambda_i) + \phi'(\hat{\lambda}) = -\phi''(\lambda_i + \sigma_i(\lambda_i - \hat{\lambda}))(\lambda_i - \hat{\lambda}) , \tag{11h}$$

with $\sigma_i \in [0, 1]$. Adding and subtracting $\phi''(\lambda_i)(\lambda_i - \hat{\lambda})$ from the left side of (11f) and reordering terms, we find that

$$\begin{aligned}|p_i''(\lambda_i)|\,|\lambda_{i+1} - \hat{\lambda}| \leq [&|\phi''(\lambda_i - \sigma_i(\lambda_i - \hat{\lambda})) - \phi''(\lambda_i)| \\ &+ |\phi''(\lambda_i) - p_i''\lambda_i)|]|\lambda_i - \hat{\lambda}| . \end{aligned} \tag{11i}$$

For all $i \geq i_1$, let $e_i \triangleq |\lambda_i - \hat{\lambda}|$. Clearly $e_i \to 0$, as $i \to \infty$, and hence we can assume that i_1 is large enough to ensure that $e_i \leq 1$, for all $i \geq i_1$. Since $\phi''(\lambda) \leq M$ for all $\lambda \in \mathbb{R}$, by assumption, it follows from (11i), Lemma 1.7.2, and the fact that $p_i''(\lambda_i) \geq m$, for all $i \geq i_1$, that for all $i \geq i_1$ and some $C < \infty$,

$$\begin{aligned} e_{i+1} = |\lambda_{i+1} - \hat{\lambda}| &\leq C(|\lambda_i - \hat{\lambda}|^2 + |\lambda_i - \hat{\lambda} + \hat{\lambda} - \lambda_{i-1}|^2|\lambda_i - \hat{\lambda}|) \\ &\leq C(e_i^2 + e_{i-1}^2 e_i) , \end{aligned} \tag{11j}$$

where we have used the fact that $e_i^3 \leq e_i^2$. To complete the proof we proceed by

induction. Let $a > 0$ be such that $C(a^2 + a) < 1$. Let $i_2 \geq i_1$ be such that, for all $i \geq i_2$, $e_i \leq a/4$. Then $e_{i_2+l} \leq a(\frac{1}{2})^{2^l}$, for $l = 0, 1$. Suppose that $e_{i_2+l} \leq a(\frac{1}{2})^{2^l}$, for $l = 0, 1, \ldots, n$. It follows from (11j) that

$$\begin{aligned} e_{i_2+n+1} &\leq C(e_{i_2+n}^2 + e_{i_2+n-1}^2 e_{i_2+n}) \\ &= C[a^2(\tfrac{1}{2})^{2^{i_2+n+1}+1} + a(\tfrac{1}{2})^{2^{i_2+n+1}} a^2(\tfrac{1}{2})^{2^{i_2+n+1}}] \\ &\leq Ca(\tfrac{1}{2})^{2^{i_2+n+1}+1}(a + a^2) \leq a(\tfrac{1}{2})^{2^{n+1}}, \end{aligned} \tag{11k}$$

which completes the proof of the theorem. □

Referring to [KiP.91], we see that, if we assume that $\phi(\cdot)$ is six times continuously differentiable, and Algorithm 1.7.2 constructs a sequence $\{\lambda_i\}_{i=0}^{\infty}$ which has an accumulation point $\hat{\lambda}$ satisfying the second-order sufficiency condition (1.1.12a), then $\{\lambda_i\}_{i=0}^{\infty}$ converges to $\hat{\lambda}$ Q-quadratically (and hence also R-quadratically).

The next two methods do not require computing derivatives of $\phi(\cdot)$, but they can be used only when the function $\phi(\cdot)$ is unimodal. To avoid any ambiguity, for our purposes we define unimodal functions as follows.

Definition 1.7.5. *We say that a continuously differentiable function* $\phi : \mathbb{R} \to \mathbb{R}$ *is* unimodal *if there is a unique point* $\hat{\lambda} \in \mathbb{R}$ *such that* $\phi'(\hat{\lambda}) = 0$, *which is also the unique global minimizer of* $\phi(\cdot)$. □

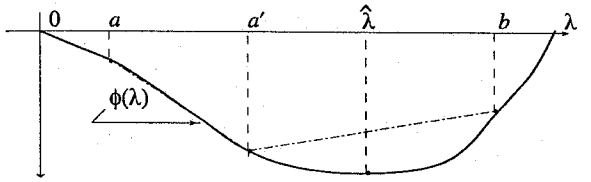
1.7.2 The Golden Section Search

We begin with the golden section search method for finding the minimizer of a unimodal function. The golden section search method does not require that the function $\phi(\cdot)$ be differentiable. However, we will present it under the simplifying assumption that $\phi(\cdot)$ is differentiable. The golden section search method is based on three observations.

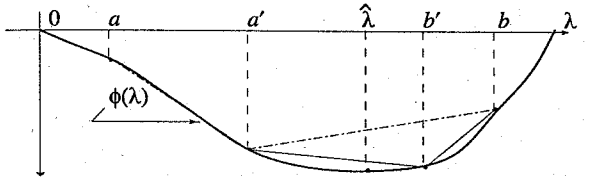
(a) First (see Fig. 1.7.1a), suppose that we have an interval $[a, b]$ containing a point a' such that $\phi(a') \leq \min\{\phi(a), \phi(b)\}$. Then it follows from the Mean Value Theorem 5.1.28*(a)* that there is a $\lambda_1 \in (a, a']$ such that $\phi'(\lambda_1) \leq 0$ and a $\lambda_2 \in [a', b)$ such that $\phi'(\lambda_2) \geq 0$. Therefore, it follows from the continuity of $\phi'(\cdot)$ that the interval $[\lambda_1, \lambda_2]$ must contain a zero of $\phi'(\cdot)$. Since, by assumption, $\hat{\lambda}$, the unique minimizer of $\phi(\cdot)$, is also the unique zero of $\phi'(\cdot)$, $\hat{\lambda} \in [\lambda_1, \lambda_2] \subset [a, b]$, i.e., the interval $[a, b]$ brackets $\hat{\lambda}$.

(b) Second, with a, b as above (see Fig. 1.7.1b), let $b' > a'$ be another point in $[a, b]$. If $\phi(a') > \phi(b')$, then $\phi(b') < \min\{\phi(a'), \phi(b)\}$ and hence, by the same arguments as above, we conclude that $[a', b]$ is a smaller interval containing $\hat{\lambda}$. If $\phi(a') \leq \phi(b')$, then $\phi(a') \leq \min\{\phi(a), \phi(b')\}$, and hence, by the same arguments as above, we conclude that $[a, b']$ is a smaller interval containing $\hat{\lambda}$.

(c) Third, the process of reducing the bracket $[a, b]$, containing the global minimizer $\hat{\lambda}$, can be made more efficient as follows. Let $a_0 = a$, $b_0 = b$ and suppose that we wish to construct a sequence of nested intervals $[a_{i+1}, b_{i+1}] \subset [a_i, b_i] \subset [a, b]$, $i = 0, 1, 2, 3 \ldots$, such that either $\phi(a_{i+1}) \leq \min\{\phi(a_i), \phi(b_i)\}$ or $\phi(b_{i+1}) \leq \min\{\phi(a_i), \phi(b_i)\}$, so that each of these intervals also contains the global minimizer $\hat{\lambda}$. In keeping with the preceding discussion, we assume that we construct two points $a_i' < b_i' \in (a_i, b_i)$



(a) Construction of an initial bracket



(b) Construction of a reduced bracket

Fig. 1.7.1. Construction for golden section algorithm.

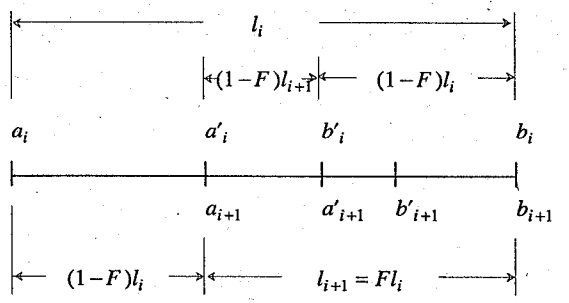


Figure 1.7.2. Division of interval in golden section construction.

at each stage and hence that either $[a_{i+1}, b_{i+1}] = [a_i, b_i']$ or $[a_{i+1}, b_{i+1}] = [a_i', b_i]$ holds. Now, the computing work would be considerably reduced if we could reuse one of the two points a_i', b_i'. It turns out that if the points a_i', b_i' are placed *symmetrically*, we can ensure that either $b_i' = a'_{i+1}$ or $a_i' = b_i'$+1 holds, as shown in Fig. 1.7.2, by requiring that, for some $F \in (0, 1)$, with $l_i \triangleq b_i - a_i$,

$$l_{i+1} = Fl_i\,, \tag{12a}$$

$$l_{i+1} - (1-F)l_i \;=\; (1-F)l_{i+1} \tag{12b}$$

is satisfied at each stage. Eliminating l_i and l_{i+1} from (12a) and (12b), we obtain

$$F^2 + F - 1 = 0\,. \tag{12c}$$

Hence

$$F = \tfrac{1}{2}(5^{1/2} - 1) \approx 0.61804\,. \tag{12d}$$

Thus, with this construction, the length of the interval containing $\hat{\lambda}$ shrinks linearly to zero, with rate constant F, i.e.,

$$l_i = F^i l_0 \approx (0.61804)^i l_0\,. \tag{12e}$$

Hence the precision with which $\hat{\lambda}$ is identified increases very rapidly with i, e.g., $l_{15} = 0.00073\, l_0$, so that if $l_0 = 1$, then $\hat{\lambda} = (a_{15} + b_{15})/2 \pm 0.000365$.

Since we must associate an evaluation of the function $\phi(\cdot)$ with each placement of an a_i' or b'_i, we can compare the efficiency of the golden section search with any other "two point" scheme. It should be clear that no matter how we place two additional points in the interval $[a_i, b_i]$, of length l_i, the next interval will have length $l_{i+1} \geq \tfrac{1}{2} l_i$. Hence, a limit on the efficiency of any two point scheme which uses two function evaluations per iteration is $-\tfrac{1}{2} \ln \tfrac{1}{2} = 0.3466$, whereas the efficiency of the golden section search, which uses only one function evaluation per iteration, is $-\ln 0.61804 = 0.4812$, which is considerably better. We now state the golden section search formally.

Golden Section Algorithm 1.7.6.

Step 0. Compute a bracket $[a_0, b_0]$ containing $\hat{\lambda}$, the minimizer of $\phi(\lambda)$, and set $i = 0$.

Step 1. Set $l_i = b_i - a_i$, and compute

$$a_i' = a_i + l_i(1-F)\,, \tag{13a}$$

$$b_i' = b_i - l_i(1-F)\,. \tag{13b}$$

Step 2. If $\phi(b_i') \leq \min\{\phi(a_i'), \phi(b_i)\}$, set $a_{i+1} = a_i', b_{i+1} = b_i$. Else, set $a_{i+1} = a_i$ and $b_{i+1} = b_i'$.

Step 3. Replace i by $i + 1$, and go to Step 1.

Since $\phi'(0) < 0$, in the applications that interest us, we can construct an initial bracket $[a_0, b_0]$, containing the minimizer $\hat{\lambda}$, by using the Armijo step-Size construction defined in (1.2.20a), which depends on two parameters, α, $\beta \in (0, 1)$, to compute an integer k such that $\phi(\beta^k) \leq \alpha\beta^k\phi'(0) < \phi(\beta^{k-1})$. There is a good chance that $\phi(\beta^k) \leq \min\{\phi(0), \phi(\beta^{k-1})\}$. If not, one can continue to reduce the integer k until a bracketing interval is constructed. This results in an exponential scanning of the real line. Alternatively, one can evaluate $\phi(\lambda_i)$ for $\lambda_i = \lambda_0 + i\Delta$, with $\Delta > 0$ and $i = 1, 2\ldots$, until three consecutive values are obtained such that $\phi(\lambda_{i-1}) \leq \min\{\phi(\lambda_i), \phi(\lambda_{i-2})\}$, which ensures that $\lambda_{i-2} \leq \hat{\lambda} \leq \lambda_i$. The main drawback of this second scheme is that it requires intuition in selecting an efficient value for Δ. A small Δ produces a small interval containing $\hat{\lambda}$ but it may take many function evaluations to find it, whereas a large Δ produces a large interval containing $\hat{\lambda}$, but requires few function evaluations to find it.

1.7.3 Method of Sequential Quadratic Interpolations

Suppose that $\phi: \mathbb{R} \to \mathbb{R}$ is a unimodal function and that the triplet of real numbers $\zeta_i = (\zeta_i^1, \zeta_i^2, \zeta_i^3)$ (i.e., $\zeta_i \in \mathbb{R}^3$) is such that $\zeta_i^1 < \zeta_i^2 < \zeta_i^3$ and $\phi(\zeta_i^2) \leq \min\{\phi(\zeta^1), \phi(\zeta_i^3)\}$ hold, so that the interval $[\zeta_i^1, \zeta_i^3]$ contains $\hat{\lambda}$, the unique minimizer of $\phi(\cdot)$. There are several methods of sequential quadratic interpolations. They all begin by constructing an interpolating quadratic $q(\lambda; \zeta_i) = a_2\lambda^2 + a_1\lambda + a_0$ whose coefficients are determined by the equation

$$\begin{bmatrix} (\zeta_i^1)^2 & \zeta_i^1 & 1 \\ (\zeta_i^2)^2 & \zeta_i^2 & 1 \\ (\zeta_i^3)^2 & \zeta_i^3 & 1 \end{bmatrix} \begin{bmatrix} a_2 \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} \phi(\zeta_i^1) \\ \phi(\zeta_i^2) \\ \phi(\zeta_i^3) \end{bmatrix}. \tag{14a}$$

Because $\zeta_i^1 < \zeta_i^2 < \zeta_i^3$, the van der Monde matrix in (14a) is nonsingular. Note that the polynomial $q(\lambda, \zeta_i)$ can also be expressed in closed form using the Lagrange formula

$$q(\lambda;\zeta_i) = \phi(\zeta_i^1)\frac{(\lambda-\zeta_i^2)(\lambda-\zeta_i^3)}{(\zeta_i^1-\zeta_i^2)(\zeta_i^1-\zeta_i^3)} + \phi(\zeta_i^2)\frac{(\lambda-\zeta_i^1)(\lambda-\zeta_i^3)}{(\zeta_i^2-\zeta_i^1)(\zeta_i^2-\zeta_i^3)}$$

$$+ \phi(\zeta_i^3)\frac{(\lambda-\zeta_i^1)(\lambda-\zeta_i^2)}{(\zeta_i^3-\zeta_i^1)(\zeta_i^3-\zeta_i^2)}. \tag{14b}$$

The sequential quadratic interpolation methods compute the minimizer of $q(\cdot;\zeta_i)$, $\lambda_i \triangleq \arg\min q(\lambda;\zeta_i)$, which, by inspection, is in the interval $[\zeta_i^1, \zeta_i^3]$. Finally, they select a new triplet ζ_{i+1} out of the available quadruplet $\zeta_i^1, \zeta_i^2, \zeta_i^3, \lambda_i$. It is shown in [KoO.68] that when ζ_{i+1} is defined by the update rule $\zeta_{i+1} = (\zeta_i^2, \zeta_i^3, \lambda_i)$ and the initial triplet ζ_0 is sufficiently close to $\hat{\lambda}$, the sequence $\{\lambda_i\}_{i=0}^{\infty}$ converges superlinearly to $\hat{\lambda}$ with root rate $r \approx 1.3$. Unfortunately, it is not clear how to stabilize this method so as to make it globally convergent. Therefore, we will describe an alternative proposed by Luenberger [Lue.84] which is globally convergent, but which has not been shown to retain the superlinear rate of convergence of the simple method. In practice, however, the Luenberger version performs quite well. In spirit, the Luenberger version of the method of sequential quadratic interpolation is quite close to that of the Golden Section Algorithm 1.7.6, but it takes more space to explain and justify.

First, we define the set of *admissible triplets* $T \subset \mathbb{R}^3$ as the set of vectors $\zeta \in \mathbb{R}^3$ which define an interval $[\zeta^1, \zeta^3]$ that contains the minimizer $\hat{\lambda}$. Hence we see that

$$\begin{aligned} T \triangleq\ & \{\zeta \in \mathbb{R}^3 \mid \zeta^1 < \zeta^2 < \zeta^3,\ \phi(\zeta^2) \le \min\{\phi(\zeta^1), \phi(\zeta^3)\}\} \\ & \cup \{\zeta \in \mathbb{R}^3 \mid \zeta^1 = \zeta^2 < \zeta^3,\ \phi'(\zeta^1) \le 0,\ \phi(\zeta^3) \ge \phi(\zeta^1)\} \\ & \cup \{\zeta \in \mathbb{R}^3 \mid \zeta^1 < \zeta^2 = \zeta^3,\ \phi'(\zeta^3) \ge 0,\ \phi(\zeta^1) \ge \phi(\zeta^3)\} \\ & \cup \{\zeta \in \mathbb{R}^3 \mid \zeta^1 = \zeta^2 = \zeta^3 = \hat{\lambda}\}. \end{aligned} \tag{15}$$

Lemma 1.7.7.

(a) For every $\zeta \in T$, $\zeta^1 \le \hat{\lambda} \le \zeta^3$ holds.

(b) The set T defined by (15) is closed.

Proof. *(a)* This part follows directly from the definition of T and the Mean-Value Theorem 5.1.28*(a)*.

(b) Suppose that $\{\zeta_i\}_{i=0}^{\infty} \subset T$ is such that $\zeta_i \to \zeta_*$, as $i \to \infty$. Since $\zeta_i^1 \le \zeta_i^2 \le \zeta_i^3$ holds for all i, we must have that $\zeta_*^1 \le \zeta_*^2 \le \zeta_*^3$. We now consider the only four possibilities.

(i) Suppose that $\zeta_*^1 < \zeta_*^2 < \zeta_*^3$. Then there must exist an i_0 such that, for all $i \ge i_0$, $\zeta_i^1 < \zeta_i^2 < \zeta_i^3$ and hence, for all $i \ge i_0$, $\phi(\zeta_i^2) \le \min\{\phi(\zeta_i^1), \phi(\zeta_i^3)\}$. It now follows from the continuity of $\phi(\cdot)$ that $\phi(\zeta_*^2) \le \min\{\phi(\zeta_*^1), \phi(\zeta_*^3)\}$ and hence that $\zeta_* \in T$.

(ii) Next, suppose that $\zeta_*^1 < \zeta_*^2 = \zeta_*^3$. Then we need to consider two subcases:

(α) There is an infinite subsequence $\{\zeta_i\}_{i \in K}$ such that $\zeta_i^1 < \zeta_i^2 = \zeta_i^3$. In this case, $\phi'(\zeta_i^3) \ge 0$ and $\phi(\zeta_i^1) \ge \phi(\zeta_i^3)$ for all $i \in K$, and hence, since $\zeta_i \to^K \zeta_*$, as $i \to \infty$, it follows from the continuity of $\phi'(\cdot)$ and $\phi(\cdot)$ that $\phi'(\zeta_*^3) \ge 0$ and $\phi(\zeta_*^1) \ge \phi(\zeta_*^3)$, so that $\zeta_* \in T$.

(β) There exists an i_0 such that, for all $i \ge i_0$, $\zeta_i^1 < \zeta_i^2 < \zeta_i^3$. Hence, for all $i \ge i_0$, we must have that $\phi(\zeta_i^2) \le \phi(\zeta_i^1)$ and $[\phi(\zeta_i^3) - \phi(\zeta_i^2)] / [\zeta_i^3 - \zeta_i^2] \ge 0$. Therefore, in the limit, since $\zeta_i^2 \to \zeta_*^3$ and $\zeta_i^3 \to \zeta_*^3$, as $i \to \infty$, $\phi(\zeta_*^3) \le \phi(\zeta_*^1)$ and $\phi'(\zeta_*^3) \ge 0$, i.e., $\zeta_* \in T$.

(iii) The case where $\zeta_*^1 = \zeta_*^2 < \zeta_*^3$ follows by symmetry from *(ii)*.

(iv) Finally, consider the case where $\zeta_*^1 = \zeta_*^2 = \zeta_*^3$. Since $\zeta_i^1 \le \hat{\lambda} \le \zeta_i^3$ must hold, for all $i \in \mathbb{N}$, it is clear that $\zeta_*^j = \hat{\lambda}$ for $j = 1, 2, 3$ and hence that $\zeta_* \in T$.

Therefore we conclude that T is closed. □

Lemma 1.7.8. *For any $\zeta \in T$, consider the polynominal*

$$q(\lambda;\zeta) = a_2(\zeta)\lambda^2 + a_1(\zeta)\lambda + a_0(\zeta), \tag{16}$$

whose coefficients are defined as follows:

(i) *When $\zeta^1 < \zeta^2 < \zeta^3$, $a_0(\zeta), a_1(\zeta), a_2(\zeta)$ are defined by (14a) (with $\zeta_i \triangleq \zeta$).*

(ii) *When $\zeta^1 < \zeta^2 = \zeta^3$, $a_0(\zeta), a_1(\zeta), a_2(\zeta)$ are defined by*

$$\begin{bmatrix} (\zeta^1)^2 & \zeta^1 & 1 \\ (\zeta^3)^2 & \zeta^3 & 1 \\ 2\zeta^3 & 1 & 0 \end{bmatrix} \begin{bmatrix} a_2 \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} \phi(\zeta^1) \\ \phi(\zeta^3) \\ \phi'(\zeta^3) \end{bmatrix}. \tag{17a}$$

(iii) *When $\zeta^1 = \zeta^2 < \zeta^3$, $a_0(\zeta), a_1(\zeta), a_2(\zeta)$ are defined by*

$$\begin{bmatrix} (\zeta^1)^2 & \zeta^1 & 1 \\ (\zeta^3)^2 & \zeta^3 & 1 \\ 2\zeta^1 & 1 & 0 \end{bmatrix} \begin{bmatrix} a_2 \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} \phi(\zeta^1) \\ \phi(\zeta^3) \\ \phi'(\zeta^1) \end{bmatrix} \tag{17b}$$

(note that the matrices in (17a), (17b) are nonsingular).

(iv) *Finally, when $\zeta^1 = \zeta^2 = \zeta^3 = \hat{\lambda}$, we define $q(\lambda;\zeta) \equiv \phi(\hat{\lambda})$, i.e., $a_0(\zeta) = \phi(\hat{\lambda})$, $a_1(\zeta) = a_2(\zeta) = 0$.*

Then,

(a) the coefficients $a_0(\cdot), a_1(\cdot), a_2(\cdot)$ *of* $q(\lambda;\cdot)$ *are continuous on* T,

(b) the minimizer $\lambda^*(\zeta)$, *of* $q(\cdot;\zeta)$, *is continuous in* ζ *on* T, *and*

(c) the minimizer $\lambda^*(\zeta)$, *of* $q(\cdot;\zeta)$, *satisfies* $\lambda^*(\zeta) \in [\zeta^1, \zeta^3]$.

Proof. *(a)* Suppose that $\{\zeta_i\}_{i=0}^{\infty}$ is such that $\zeta_i \to \zeta_*$, as $i \to \infty$. We must consider a number of cases.

(i) Suppose that $\zeta_*^1 < \zeta_*^2 < \zeta_*^3$. Then there has to be an i_0 such that $\zeta_i^1 < \zeta_i^2 < \zeta_i^3$ for all $i \geq i_0$, and hence the coefficients $a_2(\zeta_i)$, $a_1(\zeta_i)$, $a_0(\zeta_i)$ are determined by (14a). Since the matrix in the left-hand side of (14a) is nonsingular and continuous, and since the vector in the right-hand side in (14a) is continuous, it follows that $a_2(\zeta_i) \to a_2(\zeta_*)$, $a_1(\zeta_i) \to a_1(\zeta_*)$, and $a_0(\zeta_i) \to a_0(\zeta_*)$, as $i \to \infty$.

(ii) Suppose that $\zeta_*^1 < \zeta_*^2 = \zeta_*^3$. If there exists an infinite subsequence $\{\zeta_i\}_{i \in K}$ such that $\zeta_i^2 = \zeta_i^3$ for all $i \in K$, then the continuity of the coefficients of $q(\cdot;\zeta)$ follows from (17a), because of the continuity and nonsingularity of the matrix in the left-hand side and the continuity of the vector in the right-hand side of this equation. If such a subsequence does not exist, then there must be an i_0 such that $\zeta_i^1 < \zeta_i^2 < \zeta_i^3$ for all $i \geq i_0$, and hence the last equation in (14a) can be replaced by the result of subtracting the last equation from the second one, i.e., by

$$(\zeta_i^2 - \zeta_i^3)(\zeta_i^2 + \zeta_i^3)a_2 + (\zeta_i^2 - \zeta_i^3)a_1 = \phi(\zeta_i^2) - \phi(\zeta_i^3), \tag{18a}$$

which can be rewritten as

$$(\zeta_i^2 + \zeta_i^3)a_2 + a_1 = \frac{\phi(\zeta_i^2) - \phi(\zeta_i^3)}{\zeta_i^2 - \zeta_i^3}. \tag{18b}$$

Hence (14a) can be replaced by

$$\begin{bmatrix} (\zeta_i^1)^2 & \zeta_i^1 & 1 \\ (\zeta_i^2)^2 & \zeta_i^2 & 1 \\ \zeta_i^2 + \zeta_i^3 & 1 & 0 \end{bmatrix} \begin{bmatrix} a_2 \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} \phi(\zeta_i^1) \\ \phi(\zeta_i^3) \\ [\phi(\zeta_i^2) - \phi(\zeta_i^3)] / [\zeta_i^2 - \zeta_i^3] \end{bmatrix}. \tag{18c}$$

It now follows from the continuity and nonsingularity of the matrix in the left-hand side of (18c) and the continuity of the vector in the right-hand side in (18c) that $a_2(\zeta_i) \to a_2(\zeta_*)$, $a_1(\zeta_i) \to a_1(\zeta_*)$, and $a_0(\zeta_i) \to a_0(\zeta_*)$, as $i \to \infty$, with $(a_2(\zeta_*), a_1(\zeta_*), a_0(\zeta_*))$, defined by (17a).

Since all other cases follow in a similar manner, we conclude that the coefficients $a_2(\cdot)$, $a_1(\cdot)$, and $a_0(\cdot)$ of $q(\lambda;\zeta)$ are continuous.

(b) The minimizer $\lambda^*(\zeta)$ of $q(\lambda;\zeta)$ has the property that $\lambda^*(\zeta) \in [\zeta^1, \zeta^3]$ and is given by

$$\lambda^*(\zeta) = -a_1(\zeta)/2a_2(\zeta), \tag{18d}$$

when $a_2(\zeta) \neq 0$. Now $a_2(\zeta) \neq 0$ for all $\zeta \in T$, with the exception of the point $\zeta = (\hat{\lambda}, \hat{\lambda}, \hat{\lambda})$, because $\phi(\cdot)$ is unimodal, and $\lambda^*(\hat{\lambda}) = \hat{\lambda}$. Hence $\lambda^*(\cdot)$ is continuous because $a_2(\cdot)$ and $a_1(\cdot)$ are continuous.

(c) Since $q(\zeta^2;\zeta) \leq \min\{q(\zeta^1;\zeta), q(\zeta^3;\zeta)\}$, and $q(\cdot;\zeta)$ is convex, it follows that $\lambda^*(\zeta) \in [\zeta^1, \zeta^3]$. □

We need two more quantities to define an algorithm. We begin by defining the candidate triplets that might replace a $\zeta \in T$ and define a smaller interval containing $\hat{\lambda}$ by

$$u_1(\zeta) = (\zeta^1, \lambda^*(\zeta), \zeta^2), \tag{19a}$$

$$u_2(\zeta) = (\zeta^2, \lambda^*(\zeta), \zeta^3), \tag{19b}$$

$$u_3(\zeta) = (\lambda^*(\zeta), \zeta^2, \zeta^3), \tag{19c}$$

and

$$u_4(\zeta) = (\zeta^1, \zeta^2, \lambda^*(\zeta)), \tag{19d}$$

where $\lambda^*(\zeta)$ denotes the minimizer of $q(\lambda, \zeta)$. We now define the set of *admissible replacement triplets* by

$$A(\zeta) \triangleq T \cap \{u_1(\zeta), u_2(\zeta), u_3(\zeta), u_4(\zeta)\}. \tag{19e}$$

Second, we define the *surrogate cost* function $c : \mathbb{R}^3 \to \mathbb{R}$ by

$$c(\zeta) \triangleq \phi(\zeta^1) + \phi(\zeta^2) + \phi(\zeta^3). \tag{19f}$$

Finally, to get in line with the algorithm model theory in Section 1.2, we define the set of "stationary" points S by

$$S \triangleq \{\zeta \in T \mid \phi'(\zeta^1) = 0, \text{ or } \phi'(\zeta^2) = 0, \text{ or } \phi'(\zeta^3) = 0\}. \tag{20}$$

Lemma 1.7.9.

(a) For every $\zeta \in T$, *the set* $A(\zeta)$, *defined by (19e), is nonempty.*

(b) The set valued map $A(\cdot)$ *is outer semicontinuous.*

(c) For every $\zeta \in T \setminus S \triangleq \{\zeta \in T \mid \zeta \notin S\}$ *such that* $\zeta^1 < \zeta^2 < \zeta^3$, $c(y) < c(\zeta)$ *for all* $y \in A(\zeta)$.

(d) For every $\zeta \in T \setminus S$ *such that* $\zeta^1 = \zeta^2 < \zeta^3$ *or* $\zeta^1 < \zeta^2 = \zeta^3$, *there exists a* $y \in A(\zeta)$ *such that* $c(y) < c(\zeta)$.

Proof. *(a)* Let $\zeta \in T$ be arbitrary. Suppose $\lambda^*(\zeta) \in [\zeta^1, \zeta^2]$. Then $A(\zeta)$ is empty if and only if both $u_1(\zeta)$ and $u_3(\zeta)$ are not in IT, i.e., if and only if

$$\phi(\lambda^*(\zeta)) > \min\{\phi(\zeta^1), \phi(\zeta^2)\} = \phi(\zeta^2), \tag{21a}$$

and also

$$\phi(\zeta^2) > \min\{\phi(\lambda^*(\zeta)), \phi(\zeta^3)\} \geq \min\{\phi(\lambda^*(\zeta)), \phi(\zeta^2)\}. \tag{21b}$$

Since the latter implies that $\phi(\zeta^2) > \phi(\lambda^*(\zeta))$ we have a contradiction. The case where $\lambda^*(\zeta) \in [\zeta^2, \zeta^3]$ follows by symmetry.

(b) Suppose that $\{\zeta_i\}_{i=0}^{\infty} \subset T$ is such that $\zeta_i \to \zeta_*$ and that there exist a $\zeta_* \in T$, an infinite subset $K \subset \mathbb{N}$, and $\zeta_{i+1} \in A(\zeta_i)$ for all $i \in K$, such that $\zeta_{i+1} \to^K \zeta_*$, as $i \to \infty$. Then there must exist a $k \in \{1,2,3,4\}$ and an infinite subset $K' \subset K$ such that $\zeta_{i+1} = u_k(\zeta_i)$ for all $i \in K'$. Since $u_k(\cdot)$ is continuous, $u_k(\zeta_i) \to u_k(\zeta_*) = \zeta_*$, as $i \to \infty$, and since T is closed, $u_k(\zeta_{**}) \in T$. Hence $\zeta_* = u_k(\zeta_*) \in A(\zeta_*)$, which shows that $A(\cdot)$ is upper semicontinuous.

(c) Suppose that $\zeta \in T \setminus D$ and $\zeta^1 < \zeta^2 < \zeta^3$. Then $\lambda^*(\zeta) \in (\zeta^1, \zeta^3)$. Since the situation is symmetric with respect to ζ^1 and ζ^3, it suffices to consider the case where $\lambda^*(\zeta) \in (\zeta^1, \zeta^2]$. In this case *(i)* $\phi(\zeta^2) < \phi(\zeta^3)$, because when $\phi(\zeta^2) = \phi(\zeta^3)$, $\lambda^*(\zeta) = \frac{1}{2}(\zeta^2 + \zeta^3)$, and *(ii)* only $u_1(\zeta)$ and $u_3(\zeta)$ can be in $A(\zeta)$. So, suppose that $A(\zeta) = \{u_1(\zeta)\}$. Then, because $u_3(\zeta) \notin A(\zeta)$, $\phi(\lambda^*(\zeta)) \leq \phi(\zeta^2)$ and $\phi(\zeta^2) < \phi(\zeta^3)$, we must have that

$$c(u_1(\zeta)) = \phi(\zeta^1) + \phi(\lambda^*(\zeta)) + \phi(\zeta^2)$$

$$< \phi(\zeta^1) + \phi(\zeta^2) + \phi(\zeta^3) = c(\zeta). \tag{21c}$$

Next, suppose that $A(\zeta) = \{u_3(\zeta)\}$. Then, since $u_1(\zeta) \notin A(\zeta)$, we must have that $\phi(\zeta^2) \leq \phi(\lambda^*(\zeta))$. Also, $\phi(\lambda^*(\zeta)) < \phi(\zeta^1)$ must hold, because, otherwise, we would have a local maximum in $[\zeta^1, \zeta^2]$, contradicting the unimodality of $\phi(\cdot)$. Hence, in this case,

$$c(u_3(\zeta)) = \phi(\lambda^*(\zeta)) + \phi(\zeta^2) + \phi(\zeta^3)$$

$$< \phi(\zeta^1) + \phi(\zeta^2) + \phi(\zeta^3) = c(\zeta). \tag{21d}$$

Finally, suppose that $A(\zeta) = \{u_1(\zeta), u_3(\zeta)\}$. Then we must have that *(i)* $\phi(\zeta^2) < \phi(\zeta^1)$ by assumption, *(ii)* $\lambda^*(\zeta) < \zeta^2$, and *(iii)* $\phi(\lambda^*(\zeta)) = \phi(\zeta^2)$ because, otherwise, we would have a contradiction of the unimodality of $\phi(\cdot)$. Hence, since $\phi(\zeta^2) < \min\{\phi(\zeta^1), \phi(\zeta^3)\}$, it follows that both $c(u_1(\zeta)) < c(\zeta)$ and $c(u_3(\zeta)) < c(\zeta)$. This exhausts all the possibilities.

(d) Suppose that $\zeta \in T \setminus D$ and that either $\zeta^1 = \zeta^2 < \zeta^3$ or $\zeta^1 < \zeta^2 = \zeta^3$. Clearly, because of symmetry, we need to consider only one of these two cases. Thus suppose that $\zeta^1 = \zeta^2 < \zeta^3$, so that $\phi(\zeta^3) \geq \phi(\zeta^1)$. In this case only $u_2(\zeta)$ and $u_4(\zeta)$ can be in $A(\zeta)$. Suppose that $A(\zeta) = \{u_2(\zeta)\}$. Then, because $u_4(\zeta) \notin A(\zeta)$, we must have that $\phi(\lambda^*(\zeta)) < \phi(\zeta^1)$, and hence

$$c(u_2(\zeta)) = \phi(\zeta^2) + \phi(\lambda^*(\zeta)) + \phi(\zeta^3) < c(\zeta). \tag{21e}$$

Next, suppose that $A(\zeta) = \{u_4(\zeta)\}$. Then we must have that $\phi(\lambda^*(\zeta)) < \phi(\zeta^3)$, for otherwise $\phi(\cdot)$ would have two stationary points in $[\zeta^1, \zeta^3]$, which is impossible because $\phi(\cdot)$ is unimodal. Hence,

$$c(u_4(\zeta)) = \phi(\zeta^1) + \phi(\zeta^2) + \phi(\lambda^*(\zeta)) < c(\zeta). \tag{21f}$$

Finally, suppose that $A(\zeta) = \{u_2(\zeta), u_4(\zeta)\}$. Then we must have that $\phi(\zeta^1) = \phi(\zeta^2) = \phi(\lambda^*(\zeta))$, and also $\phi(\zeta^2) < \phi(\zeta^3)$, for otherwise $\phi(\cdot)$ would have two stationary points in $[\zeta^1, \zeta^3]$. Hence we see that $c(u_4(\zeta)) < c(\zeta)$. Since this exhausts all the possibilities, our proof is complete. □

Now, we can state formally an algorithm for one-dimensional minimization via sequential quadratic interpolation.

Luenberger SQI Algorithm 1.7.10.

Data. $\zeta_0 \in T$.

Step 0. Set $i = 0$.

Step 1. Compute the coefficients of the quadratic interpolating polynomial $q(\cdot\,; \zeta_i)$, using (14a), (17a), and (17b), as appropriate.

Step 2. Compute $\lambda(\zeta_i) = \arg\min_{\lambda \in \mathbb{R}} q(\lambda\,; \zeta_i)$.

If $\lambda(\zeta_i) = \zeta_i^1$ or $\lambda(\zeta_i) = \zeta_i^3$, stop.

Else, construct the set $A(\zeta_i)$ according to (19e).

Step 3. Compute

$$\zeta_{i+1} \in \arg\min\{c(\zeta) \mid \zeta \in A(\zeta_i)\}. \tag{22}$$

Step 4. Replace i by $i + i$, and go to Step 1.

Theorem 1.7.11. *Suppose that $\{\zeta_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 1.7.10 in minimizing a continuously differentiable and unimodal function $\phi : \mathbb{R} \to \mathbb{R}$. Then $\zeta_i \to \hat{\zeta}$, as $i \to \infty$, with $\hat{\zeta} \in S$.*

Proof. First we observe that the sequence $\{\zeta_i^1\}_{i=0}^{\infty}$ is monotone increasing and bounded from above by ζ_0^3, while the sequence $\{\zeta_i^3\}_{i=0}^{\infty}$ is monotone decreasing and bounded from below by ζ_0^1. Hence, by the Monotone Sequences Proposition 5.1.16, both of these sequences converge, with $\zeta_i^1 \to \hat{\zeta}^1$ and $\zeta_i^3 \to \hat{\zeta}^3$, as $i \to \infty$. Consequently, if $\hat{\zeta}^1 = \hat{\zeta}^3$, then, because T is closed, it follows, trivially, that $\hat{\zeta} \in S$. We must consider two nontrivial cases.

Case 1. Suppose that $\{\zeta_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 1.7.10 and that it has an accumulation point $\hat{\zeta}$ such that $\hat{\zeta}^1 < \hat{\zeta}^2 < \hat{\zeta}^3$. For the sake of contradiction, suppose that $\hat{\zeta} \notin S$. Now, by construction, the sequence $\{c(\zeta_i)\}_{i=0}^{\infty}$ is monotone decreasing, and $c(\cdot)$ is continuous. Hence $c(\hat{\zeta})$ is an accumulation point of $\{c(\zeta_i)\}_{i=0}^{\infty}$ and, therefore, by Proposition 5.1.16, $c(\zeta_i) \to c(\hat{\zeta})$, as $i \to \infty$. Since $\hat{\zeta} \notin S$, it follows from Lemma 1.7.9*(c,d)* that $2\delta \triangleq \max_{y \in A(\hat{\zeta})} c(y) - c(\hat{\zeta}) < 0$. Let $K \subset \mathbb{N}$ be such that $\zeta_i \to^K \hat{\zeta}$, as $i \to \infty$; clearly there is a ζ_* and an infinite subset $K' \subset K$ such that $\zeta_{i+1} \to^{K'} \zeta_*$. Since $A(\cdot)$ is upper semicontinuous, by Lemma 1.7.9*(b)*, it follows that $\zeta_* \in A(\hat{\zeta})$ and hence that $c(\zeta_*) - c(\hat{\zeta}) \leq 2\delta < 0$. It now follows from the continuity of $c(\cdot)$ that there exists an i_0 such that, for all $i \in K'$, with $i \geq i_0$, $c(\zeta_{i+1}) - c(\zeta_i) \leq \delta$, which contradicts the convergence of the monotone decreasing sequence $\{c(\zeta_i)\}_{i=0}^{\infty}$. Therefore we conclude that $\hat{\zeta} \in S$. Since $\hat{\zeta}^1 < \hat{\zeta}^2 < \hat{\zeta}^3$, it follows that $\hat{\zeta}^2 = \hat{\lambda}$, the unique minimizer of $\phi(\cdot)$.

Now suppose that ζ_{**} is another accumulation point of $\{\zeta_i\}_{i=0}^{\infty}$. Then we must have that *(i)* $c(\zeta_{**}) = c(\hat{\zeta})$, and *(ii)* $\hat{\zeta}^1 = \zeta_{**}^1$ and $\hat{\zeta}^3 = \zeta_{**}^3$. It follows that $\phi(\hat{\zeta}^2) = \phi(\zeta_{**}^2)$ and therefore, because $\hat{\zeta}^2 = \hat{\lambda}$, that $\hat{\zeta}^2 = \zeta_{**}^2$, i.e., that $\hat{\zeta} = \zeta_{**}$. Thus in the case considered, the sequence $\{\zeta_i\}_{i=0}^{\infty}$ converges to its unique accumulation point $\hat{\zeta}$.

Case 2. Next, suppose that Case 1 does not apply. Then the sequence $\{\zeta_i\}_{i=0}^{\infty}$ constructed by Algorithm 1.7.10 can have only two accumulation points: $\zeta_* = (\hat{\zeta}^1, \hat{\zeta}^1, \hat{\zeta}^3)$ and $\zeta_{**} = (\hat{\zeta}^1, \hat{\zeta}^3, \hat{\zeta}^3)$. We will now show that only one of these points is an accumulation point of $\{\zeta_i\}_{i=0}^{\infty}$. For suppose that both ζ_* and ζ_{**} are accumulation points. Then they are both in T, and hence we must have that $\phi'(\hat{\zeta}^1) \leq 0$ and $\phi(\hat{\zeta}^3) \geq \phi(\hat{\zeta}^1)$, and, also, that $\phi'(\hat{\zeta}^3) \geq 0$ and $\phi(\hat{\zeta}^1) \geq \phi(\hat{\zeta}^3)$. Thus, $\phi(\hat{\zeta}^1) = \phi(\hat{\zeta}^3)$ must hold. Furthermore, because $\phi(\cdot)$ is unimodal, we must have that $\phi'(\hat{\zeta}^1) < 0$ and $\phi'(\hat{\zeta}^3) > 0$. Consequently, we must have that

$$\hat{\zeta}^1 < \lambda(\zeta_*) < \hat{\zeta}^3 \tag{23a}$$

and also

$$\hat{\zeta}^1 < \lambda(\zeta_{**}) < \hat{\zeta}^3 . \tag{23b}$$

Now, in the case under consideration, by construction, the only accumulation points that the sequence $\{\lambda(\zeta_i)\}_{i=0}^{\infty}$ can have are $\hat{\zeta}^1$ and $\hat{\zeta}^3$. However, by continuity of $\lambda(\cdot)$, the accumulation points of $\{\lambda(\zeta_i)\}_{i=0}^{\infty}$ are $\lambda(\zeta_*)$ and $\lambda(\zeta_{**})$. Since this contradicts (23a), (23b), we conclude that $\{\zeta_i\}_{i=0}^{\infty}$ must converge to either ζ_* or to ζ_{**}. Without loss of generality we may assume that $\{\zeta_i\}_{i=0}^{\infty}$ converges to ζ_*. In this case, we will show that $\{\lambda(\zeta_i)\}_{i=0}^{\infty}$ converges to $\hat{\zeta}^1$. For the sake of contradiction, suppose that $\lambda(\zeta_*) = \hat{\zeta}^3$. Since $q'(\lambda(\zeta_*), \zeta_*) = 0$ and $\hat{\zeta}^1 < \hat{\zeta}^3$, it follows that $\phi(\hat{\zeta}^1) = q(\hat{\zeta}^1, \zeta_*) > q(\hat{\zeta}^3, \zeta_*) = \phi(\hat{\zeta}^3)$, which contradicts the fact that $\zeta_* \in T$. Hence we conclude that $\lambda(\zeta_*) = \zeta_*^1$, and, since in this case $\phi'(\hat{\zeta}^1) = q'(\lambda(\zeta_*), \zeta_*) = 0$, that $\zeta_* \in S$, which completes our proof. □

1.7.4 Notes

For other one-dimensional optimization algorithms based on cubic or quadratic interpolation, see, e.g., [FlP.63, Tam.76, Tam.79, Hag.89]. The "unsafeguarded" cubic interpolation method in [FlP.63] can be shown to be locally quadratically convergent on sufficiently smooth functions (see [Tam.76]), but it may not be globally convergent. In Hager's algorithm [Hag.89], cubic interpolation is "safeguarded" by a bracketing scheme, which guarantees global, quadratic convergence, provided the algorithm is initialized with a bracket containing a local minimizer. Secant type algorithms for one-dimensional minimization can also be found in [KlP.72, Bre.72]. Finite difference and interpolation schemes require careful numerical implementation, to avoid errors caused by finite arithmetic operations. For details, see e.g., [Bre.72, FMM.77, GMW.81].

Nonpolynomial interpolation schemes for one-dimensional optimization are discussed in [BaB.82], a superlinearly converging algorithm for minimizing locally Lipschitz continuous functions can be found in [Mif.84], and a thorough discussion of the golden section and the Fibonnaci section search is presented in [Kie.53].

1.8 Newton's Method for Equations and Inequalities

To conclude this chapter, we will present a version of Newton's method that can be used for solving systems of equations and inequalities of the form

$$f(x) \leq 0 , \tag{1a}$$

$$g(x) = 0 , \tag{1b}$$

where the functions $f : \mathbb{R}^n \to \mathbb{R}^q$ and $g : \mathbb{R}^n \to \mathbb{R}^r$, with $r \leq n$, are Lipschitz continuously differentiable on bounded sets, and for any vector $y \in \mathbb{R}^q$, $y \leq 0$ is to be understood as $y^j \leq 0$, for all $j \in \mathbf{q}$. Similarly, $y < 0$ is to be understood as $y^j < 0$, for all $j \in \mathbf{q}$.

1.8.1 Mangasarian-Fromowitz Constraint Qualification

We will see that the version of Newton's method to be described is well defined in a neighborhood of points satisfying the Mangasarian-Fromowitz constraint qualification (MFCQ).

Definition 1.8.1. *(a) Let $f:\mathbb{R}^n \to \mathbb{R}^q$ and $g:\mathbb{R}^n \to \mathbb{R}^r$, with $r \leq n$, be continuously differentiable functions, and let $\hat{x} \in \mathbb{R}^n$ be such that $f(\hat{x}) \leq 0$ and $g(\hat{x}) = 0$. We will say that the functions $f(\cdot)$, $g(\cdot)$ satisfy the* Mangasarian-Fromowitz constraint qualification *(MFCQ) at $\hat{x}$ if $g_x(\hat{x})$ has maximum row rank and there exists a vector $\hat{h} \in \mathbb{R}^n$ such that*

$$\langle \nabla f^j(\hat{x}), \hat{h} \rangle < 0, \quad \forall\, j \in \mathbf{q}_A(\hat{x}), \tag{2a}$$

$$g_x(\hat{x})\hat{h} = 0, \tag{2b}$$

where, for any $x \in \mathbb{R}^n$, we define

$$\mathbf{q}_A(x) \triangleq \{ j \in \mathbf{q} \mid f^j(x) \geq 0 \}. \tag{2c}$$

(b) Let F be a $s \times n$ matrix and G an $r \times n$ matrix. We will say that the matrices F and G satisfy the Robinson PLI condition *(positive linear independence) if there is no nonzero pair of vectors $\mu \in \mathbb{R}^s$ and $\zeta \in \mathbb{R}^r$ such that*

$$F^T\mu + G^T\zeta = 0, \quad \mu \geq 0, \tag{2d}$$

i.e., (2d) implies that $\mu = 0$ and $\zeta = 0$. □

Note that (2d) implies that the matrix G has maximum row rank, and that $0 \notin \text{co}\{f_j\}$, where the f_j, $j \in \mathbf{s}$, are the columns of F^T.

The Mangasarian-Fromowitz constraint qualification is an assumption on the solvability of a system of equations and inequalities, which turns out to be closely related to the Robinson positive linear independent assumption (PLI), as we will now show (think of F below as a matrix whose rows are $\nabla f^j(\hat{x})^T$, $j \in \mathbf{q}_A(\hat{x})$ and that $G = g_x(\hat{x})$).

Proposition 1.8.2. *Let F be an $s \times n$ matrix and G an $r \times n$ matrix. Then, given any vectors $d_1 \in \mathbb{R}^s$ and $d_2 \in \mathbb{R}^r$, there exists a vector $h \in \mathbb{R}^n$ such that*

$$Fh < d_1, \tag{3a}$$

and

$$Gh = d_2, \tag{3b}$$

if and only if F and G satisfy the PLI condition, i.e., (2d) implies that $\mu = 0$ and $\zeta = 0$.

Proof. We begin by observing that the claim of Proposition 1.8.2 is equivalent to the claim that only one of the following two statements is true:

(i) Given any vectors $d_1 \in \mathbb{R}^s$ and $d_2 \in \mathbb{R}^r$, there exists a vector $h \in \mathbb{R}^n$ such that (3a,b) holds.

(ii) There exist a $\mu \in \mathbb{R}^r$ and a $\zeta \in \mathbb{R}^s$ such that $(\mu, \zeta) \neq 0$ and (2d) holds.

Our proof is based on contraposition. We will show that *(i)* is true if and only if *(ii)* is false. Let $E \triangleq (F^T, G^T)^T$ (an $(s+r) \times n$ matrix) and let

$$S \triangleq \{ y \in \mathbb{R}^s \times \mathbb{R}^r \mid y = Eh,\ h \in \mathbb{R}^n \}. \tag{4a}$$

Suppose that *(i)* is false, i.e., that there exist a $d_1 \in \mathbb{R}^s$ and a $d_2 \in \mathbb{R}^r$ such that (3a,b) does not have a solution. Therefore the set S can be separated from the set $(\overset{\circ}{\mathbb{R}}{}^s_- + d_1) \times \{d_2\}$, where

$$\overset{\circ}{\mathbb{R}}{}^s_- \triangleq \{ z \in \mathbb{R}^s \mid z^j < 0, j \in \mathbf{s} \}, \tag{4b}$$

i.e., there exists a nonzero vector $v = (\mu, \zeta) \in \mathbb{R}^s \times \mathbb{R}^r$ such that

$$\langle v, y \rangle \geq 0, \quad \forall\, y \in S, \tag{4c}$$

$$\langle v, y \rangle \leq 0, \quad \forall\, y \in (\overset{\circ}{\mathbb{R}}{}^s_- + d_1) \times \{d_2\}. \tag{4d}$$

Since (4c) holds if and only if

$$\langle \mu, Fx \rangle + \langle \zeta, Gx \rangle \geq 0, \forall\, x \in \mathbb{R}^n, \tag{4e}$$

we see that $F^T\mu + G^T\zeta = 0$ must hold. Now, for $j \in \mathbf{s}$, let

$$y_j(\alpha) = (d_1 - \alpha e_j, d_2), \tag{4f}$$

with $\alpha > 0$ and e_j, the jth column of the $s \times s$ identity matrix. Then we see that $y_j(\alpha) \in (\overset{\circ}{\mathbb{R}}{}^s_- + d_1) \times \{d_2\}$. Substituting into (4d) and dividing by α, we obtain

$$-\langle \mu, e_j \rangle + \frac{1}{\alpha}(\langle \mu, d_1 \rangle + \langle \zeta, d_2 \rangle) \leq 0, \quad \forall\, \alpha > 0,\ j \in \mathbf{s}. \tag{4g}$$

Letting $\alpha \to \infty$, we conclude that $\mu^j \geq 0$. Since (4g) holds for all $j \in \mathbf{s}$, we conclude that there exists a $(\mu, \zeta) \neq 0$ satisfying (2d). It now follows by contraposition that if *(ii)* is false, then *(i)* is true.

Next suppose that *(ii)* is true. Then there exists a nonzero vector $v = (\mu, \zeta)$ such that (2d) holds, and hence (4c) and (4d) hold for $d_1 = 0$, $d_2 = 0$, which implies that $\overset{\circ}{\mathbb{R}}{}^s_- \times \{0\}$ can be separated from the set S. In turn, this implies that (3a,b) does not have a solution for $d_1 = 0$ and $d_2 = 0$, and hence that *(i)* is false, completing our proof. □

The following result will be required in establishing the rate of convergence of the local Newton method that we will shortly introduce.

Theorem 1.8.3. *Consider the system of equations and inequalities (1a,b). Suppose that the functions $f(\cdot)$, $g(\cdot)$ are continuously differentiable, that $\hat{x} \in \mathbb{R}^n$ is such that $f(\hat{x}) \leq 0$, $g(\hat{x}) = 0$, and that $\mathrm{q}_A(\hat{x}) = \{j_1, j_2, \ldots, j_s\}$. Then the functions $f(\cdot)$ and $g(\cdot)$ satisfy MFCQ at $\hat{x}$ if and only if there exists a $\hat{\rho} > 0$ such that, for every pair of bounded functions $d_1 : B(\hat{x}, \hat{\rho}) \to \mathbb{R}^s$, $d_2 : B(\hat{x}, \hat{\rho}) \to \mathbb{R}^r$, there exists a bounded function $h : B(\hat{x}, \hat{\rho}) \to \mathbb{R}^n$ such that, for all $x \in B(\hat{x}, \hat{\rho})$,*

$$\langle \nabla f^{j_k}(x), h(x) \rangle \leq d_1^k(x), \quad \forall\, j_k \in \mathrm{q}_A(\hat{x}), \tag{5a}$$

and

$$g_x(x)h(x) = d_2(x). \tag{5b}$$

Proof. $\Leftarrow$ Suppose that (5a,b) hold for all $x \in B(\hat{x}, \hat{\rho})$, for some $\hat{\rho} > 0$. Setting $d_1^k(x) \equiv -1$, $k \in \mathrm{s}$, and $d_2(x) \equiv 0$, we find that the corresponding vector $h(\hat{x})$ satisfies (2a,b). The fact that $g_x(\hat{x})$ has maximum row rank follows from the fact that (5b) has a solution at $x = \hat{x}$ for arbitrary $d_2(\hat{x})$ if and only if the columns of $g_x(\hat{x})$ span $\mathbb{R}^r$.

$\Rightarrow$ Now suppose that MFCQ is satisfied at $\hat{x}$. Since $g_x(\hat{x})$ has maximum row rank and $r \leq n$, we can choose an $(n-r) \times n$ matrix W whose rows together with the rows of $g_x(\hat{x})$ form a basis for $\mathbb{R}^n$. Let the $n \times n$ matrix valued function $A : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ be defined by

$$A(x) \triangleq \begin{bmatrix} W \\ g_x(x) \end{bmatrix}. \tag{6a}$$

Then $A(\hat{x})$ is nonsingular and there exists a $\rho > 0$ such that $A(x)^{-1}$ exists and is bounded for all $x \in B(\hat{x}, \rho)$. By assumption, there exists a vector $\hat{h} \in \mathbb{R}^n$ such that (2a,b) are satisfied. Now let $y : B(\hat{x}, \rho) \to \mathbb{R}^n$ be defined by $y(x) \triangleq A(x)^{-1}A(\hat{x})\hat{h}$, so that $y(\hat{x}) = \hat{h}$ and $g_x(x)y(x) = 0$ for all $x \in B(\hat{x}, \rho)$. Let

$$\gamma \triangleq \max_{j \in \mathrm{q}_A(\hat{x})} \langle \nabla f^j(\hat{x}), \hat{h} \rangle < 0. \tag{6b}$$

Clearly, $y(\cdot)$ is continuous. Hence, by continuity of max functions (see Corollary 5.4.2), there exists a $\hat{\rho} \in (0, \rho)$ such that

$$\max_{x \in B(\hat{x}, \hat{\rho})} \max_{j \in \mathrm{q}_A(\hat{x})} \langle \nabla f^j(x), y(x) \rangle \leq \tfrac{1}{2}\gamma < 0. \tag{6c}$$

Now let $d_1 : B(\hat{x}, \hat{\rho}) \to \mathbb{R}^s$ and $d_2 : B(\hat{x}, \hat{\rho}) \to \mathbb{R}^r$ be any given pair of bounded functions, let $\bar{d}(x) = (0, d_2(x)) \in \mathbb{R}^{s+r}$, and, for $x \in B(\hat{x}, \hat{\rho})$, let

$z(x) \triangleq A(x)^{-1}\bar{d}(x)$. Clearly, the function $z(\cdot)$ is bounded on $B(\hat{x}, \hat{\rho})$, and, by construction, $g_x(x)z(x) = d_2(x)$. Let $h : B(\hat{x}, \hat{\rho}) \to \mathbb{R}^n$ be defined by

$$h(x) \triangleq \beta(x)y(x) + z(x), \tag{6d}$$

where $\beta(x) = \max\{0, 2(\delta(x) - \alpha(x))/\gamma\}$, with

$$\delta(x) \triangleq -\max_{k \in \mathrm{s}} |d_1^k(x)| \tag{6e}$$

and

$$\alpha(x) \triangleq \max_{j \in \mathrm{q}_A(\hat{x})} \langle \nabla f^j(x), z(x) \rangle. \tag{6f}$$

Since both $\alpha(x)$ and $\delta(x)$ are bounded on $B(\hat{x}, \hat{\rho})$, because $d_1(x)$ and $d_2(x)$ are bounded, it follows that $\beta(x)$ is bounded on $B(\hat{x}, \hat{\rho})$, and hence $h(x)$ is bounded on $B(\hat{x}, \hat{\rho})$ and satisfies (5a,b). □

Corollary 1.8.4. *Consider the system of equations and inequalities (1a,b). Suppose that the functions $f(\cdot)$, $g(\cdot)$ are continuously differentiable and that $\hat{x} \in \mathbb{R}^n$ is such that $f(\hat{x}) \leq 0$ and $g(\hat{x}) = 0$ and that $\mathrm{q}_A(\hat{x}) = \{j_1, j_2, \ldots, j_s\}$. Furthermore, suppose that the functions $f(\cdot)$, $g(\cdot)$ satisfy MFCQ at $\hat{x}$. Then there exist a $\hat{\rho} > 0$ and a $K \in (0, \infty)$ such that, for any pair of bounded functions $d_1 : B(\hat{x}, \hat{\rho}) \to \mathbb{R}^s$, $d_2 : B(\hat{x}, \hat{\rho}) \to \mathbb{R}^r$, there exists a function $h : B(\hat{x}, \hat{\rho}) \to \mathbb{R}^n$ such that (5a,b) are satisfied for all $x \in B(\hat{x}, \hat{\rho})$ and*

$$\|h(x)\| \leq K \max\{\|d_1(x)\|, \|d_2(x)\|\}, \quad \forall\, x \in B(\hat{x}, \hat{\rho}). \tag{7}$$

Proof. First, by Theorem 1.8.3, there exists a $\hat{\rho} > 0$ such that (5a,b) has a solution for arbitrary, but bounded functions $d_1(\cdot)$, $d_2(\cdot)$. Next, an examination of the second part of the proof of Theorem 1.8.3, leads to the conclusion that the bound on the solution $h(x)$ of (5a,b) depends only on the bounds on norms $\|d_1(x)\|$ and $\|d_2(x)\|$ and not on the direction of these vectors (recall that $(1/\sqrt{n})\|d\| \leq \|d\|_\infty \leq \|d\|$). Hence there exists a $K \in (0, \infty)$ such that, whenever $\max\{\|d_1(x)\|, \|d_2(x)\|\} \leq 1$, there is a solution $h(x)$ of (5a,b) such that $\|h(x)\| \leq K$. The desired result now follows from the linearity of (5a,b). □

1.8.2 The Local Newton Algorithm

We are finally ready to start discussing Newton's method for solving (1a,b). The results presented in Section 1.4 suggest that, given a point x_i, its successor x_{i+1} should satisfy the system of equations and inequalities

$$f(x_i) + f_x(x_i)(x_{i+1} - x_i) \le 0, \tag{8a}$$

$$g(x_i) + g_x(x_i)(x_{i+1} - x_i) = 0. \tag{8b}$$

Since the solution of (8a,b) need not be unique, one possibility is to proceed as follows:

Local Newton Algorithm 1.8.5.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute the *search direction*

$$h_i = -\arg\min \{ \|h\|^2 \mid f_x(x_i)h \le -f(x_i),\ g_x(x_i)h = -g(x_i) \}. \tag{9a}$$

Step 2. Set

$$x_{i+1} = x_i + h_i, \tag{9b}$$

replace i by $i+1$, and go to step 1.

Our first concern is to establish that the search direction finding problem (9a) has a solution near a point $\hat{x}$ satisfying (1a,b).

Lemma 1.8.6. *Consider the system of equations and inequalities (1a,b). Suppose that the functions $f(\cdot)$, $g(\cdot)$ are continuously differentiable and that $\hat{x} \in \mathbb{R}^n$ is such that (a) $f(\hat{x}) \le 0$ and $g(\hat{x}) = 0$ and (b) the functions $f(\cdot)$ and $g(\cdot)$ satisfy MFCQ at $\hat{x}$. Then there exists a $\hat{\rho} > 0$ such that, for all $x \in B(\hat{x}, \hat{\rho})$ the Newton search direction problem*

$$\min \{ \|h\|^2 \mid f_x(x)h \le -f(x),\ g_x(x)h = -g(x) \} \tag{10a}$$

has a solution $\hat{h}(x)$.

Furthermore, there exists a $K \in (0, \infty)$ such that, for all $x \in B(\hat{x}, \hat{\rho})$,

$$\|\hat{h}(x)\| \le K \max \{ \|f_A(x)\|, \|g(x)\| \}, \tag{10b}$$

where $f_A(x) \triangleq (f^{j_1}(x), f^{j_2}(x), \ldots, f^{j_s}(x))$ and $\{j_1, j_2, \ldots, j_s\} = \mathbf{q}_A(\hat{x})$.

Proof. According to Corollary 1.8.4, there exist a $\rho > 0$ and a $K < \infty$ such that, for $d_1(x) \triangleq -f_A(x)$, $d_2(x) \triangleq -g(x)$, and $x \in B(\hat{x}, \rho)$, (5a,b) has a solution $h(x)$ satisfying

$$\|h(x)\| \le K \max \{ \|f_A(x)\|, \|g(x)\| \}. \tag{11a}$$

Clearly, if $f_A(x) = f(x)$, i.e., $\mathbf{q}_A(\hat{x}) = \mathbf{q}$, then $h(x)$ satisfies the constraints in (10a), and we can set $\hat{\rho} = \rho$. Hence suppose that $\mathbf{q}_A(\hat{x}) \ne \mathbf{q}$. Let $\mathbf{q}_A(\hat{x})^c \triangleq \{ j \in \mathbf{q} \mid j \notin \mathbf{q}_A(\hat{x}) \}$, and let

$$\alpha = \max_{j \in \mathbf{q}_A(\hat{x})^c} f^j(\hat{x}) < 0, \tag{11b}$$

and

$$\beta = \max_{j \in \mathbf{q}_A(\hat{x})^c} \max_{x \in B(\hat{x}, \rho)} \|\nabla f^j(x)\|. \tag{11c}$$

By continuity, there exists a $\hat{\rho} \in (0, \rho]$ such that, for all $x \in B(\hat{x}, \hat{\rho})$,

$$\max_{j \in \mathbf{q}_A(\hat{x})^c} f^j(x) \le \alpha/2, \tag{11d}$$

and

$$\|\nabla f^j(x)\| \|h(x)\| \le \beta K \max \{ \|f_A(x)\|, \|g(x)\| \} \le -\alpha/2,\ j \in \mathbf{q}_A(\hat{x})^c. \tag{11e}$$

Hence, for all $x \in B(\hat{x}, \hat{\rho})$ and $j \in \mathbf{q}_A(\hat{x})^c$,

$$f^j(x) + \langle \nabla f^j(x), h(x) \rangle \le f^j(x) + \|\nabla f^j(x)\| \|h(x)\| \le 0. \tag{11f}$$

Thus $h(x)$ satisfies the constraints in (10a). Since the convex problem (10a) must have a unique solution whenever there exists a vector satisfying its constraints, $\hat{h}(x)$ exists, and, since $\|\hat{h}(x)\| \le \|h(x)\|$, we see that (10b) holds. □

Now we are ready to establish the convergence properties of Algorithm 1.8.5.

Theorem 1.8.7. *Consider the system of equations and inequalities (1a,b). Suppose that the functions $f(\cdot)$, $g(\cdot)$ are Lipschitz continuously differentiable on bounded sets and that $\hat{x} \in \mathbb{R}^n$ is such that*

(i) $f(\hat{x}) \le 0$ *and* $g(\hat{x}) = 0$ *and*

(ii) *the functions $f(\cdot)$, $g(\cdot)$ satisfy MFCQ at $\hat{x}$.*

Let $\hat{\rho} > 0$ be such that, for all $x \in B(\hat{x}, \hat{\rho})$ and any $b_1 \in \mathbb{R}^q$ and $b_2 \in \mathbb{R}^r$, the system

$$f_x(x)h \le b_1,\ g_x(x)h = b_2 \tag{12a}$$

has a solution h^* such that

$$\|h^*\| \le K \max \{ \|b_{A1}\|, \|b_2\| \}, \tag{12b}$$

where $b_{A1} = (b_1^{j_1}, \ldots, b_1^{j_s})$, with $\{j_1, \ldots, j_s\} = \mathbf{q}_A(\hat{x})$ and $K < \infty$.

Under these assumptions, there exists a $\hat{K} \in (0, \infty)$ such that, if $x_0 \in B(\hat{x}, \hat{\rho}/2)$ and h_0, as determined by (9a), satisfies $\|h_0\| < \hat{K}$, then the

sequence $\{x_i\}_{i=0}^{\infty}$, constructed by Algorithm 1.8.5 remains in $B(\hat{x}, \hat{\rho})$ and converges R–quadratically to a point $x^* \in B(\hat{x}, \hat{\rho})$ which solves (1a,b).

Proof. First, it follows from Lemma 1.8.6 that the stipulated $\hat{\rho} > 0$ exists. Let $L < \infty$ be a common Lipschitz constant for $f_x(\cdot)$ and $g_x(\cdot)$ on $B(\hat{x}, \hat{\rho})$.

Next, suppose that the sequence $\{x_i\}_{i=0}^{\infty}$, constructed by Algorithm 1.8.5, in fact, remains in $B(\hat{x}, \hat{\rho})$, i.e., the sequence $\{x_i\}_{i=0}^{\infty}$ is well defined. Then, for any integer $i \geq 1$, the system

$$\left.\begin{aligned} f_x(x_i)h &\leq -f(x_i) + f(x_{i-1}) + f_x(x_{i-1})(x_i - x_{i-1}) \\ g_x(x_i)h &= -g(x_i) + g(x_{i-1}) + g_x(x_{i-1})(x_i - x_{i-1}) \end{aligned}\right\} \tag{13a}$$

has a solution h_i^*, which, by (12a,b), satisfies

$$\|h_i^*\| \leq K \max \left\{ \|-f_A(x_i) + f_A(x_{i-1}) + f_{Ax}(x_{i-1})(x_i - x_{i-1})\|, \right.$$

$$\left. \|g(x_i) + g(x_{i-1}) + g_x(x_{i-1})(x_i - x_{i-1})\| \right\}. \tag{13b}$$

In addition, because, by construction, $f(x_{i-1}) + f_x(x_{i-1})(x_i - x_{i-1}) \leq 0$ and $g(x_{i-1}) + g_x(x_{i-1})(x_i - x_{i-1}) = 0$, h_i^* satisfies the constraints in (9a). Hence we must have that

$$\|x_{i+1} - x_i\| = \|h_i\| \leq \|h_i^*\|. \tag{13c}$$

Using the first-order expansion formula (5.1.18a) and the Lipschitz continuity of the derivatives, we conclude that

$$\begin{aligned} &\|-f_A(x_i) + f_A(x_{i-1}) + f_{Ax}(x_{i-1})(x_i - x_{i-1})\| \\ &\quad = \left\| \int_0^1 [f_{Ax}(x_{i-1}) - f_{Ax}(x_{i-1} + s(x_i - x_{i-1}))] ds\, (x_i - x_{i-1}) \right\| \\ &\quad \leq \tfrac{1}{2} L \|x_i - x_{i-1}\|^2. \end{aligned} \tag{13d}$$

Similarly, we can show that

$$\|-g(x_i) + g(x_{i-1}) + g_x(x_{i-1})(x_i - x_{i-1})\| \leq \tfrac{1}{2} L \|x_i - x_{i-1}\|^2. \tag{13e}$$

Hence we must have that

$$\|x_{i+1} - x_i\| \leq \tfrac{1}{2} KL \|x_i - x_{i-1}\|^2, \quad \forall\, i \in \mathbf{N}. \tag{13f}$$

At this point, we make use of Theorem 1.2.42*(b)*. Let $c = \frac{1}{2}KL$, and let $\delta = c\|h_0\|$. Then, by Theorem 1.2.42*(b)* (with $M = 1/c$), if $c\|h_0\| < 1$, then $\{x_i\}_{i=0}^{\infty}$ converges at least R–quadratically to a point x^*, and, by (1.2.40e), with $k = 0$ and $j = i$,

$$\|x_i - x_0\| \leq \frac{\delta}{c(1-\delta)}, \quad \forall\, i \in \mathbf{N}. \tag{13g}$$

If x_0 is sufficiently close to $\hat{x}$ to ensure that $x_0 \in B(\hat{x}, \hat{\rho}/2)$, $c\|h_0\| < 1$, and $\delta/c(1-\delta) < \hat{\rho}/2$, then the sequence $\{x_i\}_{i=0}^{\infty}$ remains in $B(\hat{x}, \hat{\rho})$ and converges to a point $x^* \in B(\hat{x}, \hat{\rho})$. Since this implies that $h_i \to 0$, as $i \to \infty$, it follows from (9a) that x^* satisfies (1a,b). Hence we see that the theorem is true with

$$\hat{K} = \min \left\{ \frac{1}{c}, \frac{\hat{\rho}}{(2 + c\hat{\rho})} \right\}. \tag{13h}$$

This completes our proof. □

Theorem 1.8.7 establishes the behavior of Newton's method near a solution of a system of equations and inequalities. It is also possible to establish a result of the following form for a system of equations.

Theorem 1.8.8. *Suppose that $g : \mathbb{R}^n \to \mathbb{R}^r$, $r \leq n$, is Lipschitz continuously differentiable on bounded sets and that $g_x(x)$ has maximum (row) rank for all $x \in \mathbb{R}^n$. Let $x_0 \in \mathbb{R}^n$, $\rho \in (0, \infty)$, let $L < \infty$ be a Lipschitz constant for $g_x(\cdot)$ on $B(x_0, \rho)$, and let $K \in (0, \infty)$ be such that $\|g_x(x)[g_x(x)g_x(x)^T]^{-1}\| \leq K$ for all $x \in B(x_0, \rho)$. Consider the sequence $\{x_i\}_{i=0}^{\infty}$, constructed by Algorithm 1.8.5 (with the f terms suppressed in (9a)). If*

$$\|x_1 - x_0\| \leq \min \left\{ \frac{2}{KL}, \frac{\rho}{2 + \frac{1}{2}KL\rho} \right\}, \tag{14a}$$

then $\{x_i\}_{i=0}^{\infty}$ converges to a point $\hat{x}$ such that $g(\hat{x}) = 0$ and

$$\|\hat{x} - x_0\| \leq \frac{\|x_1 - x_0\|}{1 - \frac{1}{2}KL\|x_1 - x_0\|}. \tag{14b}$$

□

Theorem 1.8.8 has the following useful consequence.

Corollary 1.8.9. *Suppose that $g : \mathbb{R}^n \to \mathbb{R}^r$, $r \leq n$, is Lipschitz continuously differentiable on bounded sets and that $g_x(x)$ has maximum (row) rank for all $x \in \mathbb{R}^n$. Then, for every bounded subset $S \subset \mathbb{R}^n$, there exist constants $K', K'' \in (0, \infty)$ such that, if $x_0 \in S$ and $\|g(x_0)\| \leq K'$, then there exists an $\hat{x} \in \mathbb{R}^n$ such that $g(\hat{x}) = 0$ and $\|\hat{x} - x_0\| \leq K''\|g(x_0)\|$.* □

Exercise 1.8.10. Prove Theorem 1.8.8 and Corollary 1.8.9. □

1.8.3 Global Newton Method

For $j \in \mathbf{q}$, let $f^j(x)_+ \triangleq \max\{0, f^j(x)\}$, let $f(x)_+ \triangleq (f^1(x)_+, \ldots, f^q(x)_+)$, and let $f^0 : \mathbb{R}^n \to \mathbb{R}$ be defined by

$$f^0(x) \triangleq \tfrac{1}{2}\|g(x)\|^2 + \tfrac{1}{2}\|f(x)_+\|^2 . \tag{15a}$$

Then we see that $f^0(\cdot)$ is a continuously differentiable function whose minimizers are solutions of (1a,b), at which its value is zero. Hence, one way of solving (1a,b) is by minimizing $f^0(\cdot)$, using one of the algorithms that we have encountered in the preceding sections. Next,

$$\nabla f^0(x) = g_x(x)^T g(x) + f_x(x)^T f(x)_+ . \tag{15b}$$

We note that if the matrices $g_x(x)$ and $f_x(x)$ satisfy the PLI condition (see (2d) in Definition 1.8.1) for all $x \in \mathbb{R}^n$, then $\nabla f^0(x) = 0$ if and only if x is a solution to (1a,b). Next, suppose that at $x \in \mathbb{R}^n$, the Newton search direction finding problem (10a) has a solution $h_N(x)$. Then we find that

$$\begin{aligned}\langle \nabla f^0(x), h_N(x)\rangle &= \langle g_x(x)^T g(x), h_N(x)\rangle + \langle f_x(x)^T f(x)_+, h_N(x)\rangle \\ &= \langle g(x), g_x(x)h_N(x)\rangle + \langle f(x)_+, f_x(x)h_N(x)\rangle . \end{aligned} \tag{15c}$$

Now, by construction, $g_x(x)h_N(x) = -g(x)$, and $f_x(x)h_N(x) \leq -f(x)$. Hence we see that

$$\langle \nabla f^0(x), h_N(x)\rangle \leq -(\|g(x)\|^2 + \|f(x)_+\|^2) . \tag{15d}$$

The relations (15b) and (15d) form a basis for constructing a global version of Newton's method, based on the Polak-Sargent-Sebastian Theorem 1.2.24b, for solving (1a,b). Furthermore, when there are no inequalities present and $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$, then $h_N(x) = -g_x(x)^T[g_x(x)g_x(x)^T]^{-1}g(x)$, and hence, since, by assumption, $g(\cdot)$ is continuously differentiable, $h_N(x)$ is a continuous descent direction for $f(\cdot)$.

Exercise 1.8.11. Use the Polak-Sargent-Sebastian Theorem 1.2.24b, the relations (15b) and (15d), and the example set by Algorithm 1.4.15 to construct a globally converging version of Newton's method for solving systems of equations and inequalities such as (1a,b), which, under the assumptions in Theorem 1.8.7, reverts to the Local Newton Algorithm 1.8.5 near a solution of (1a,b). □

1.8.4 Notes

The presentation in this section is based on results in [MaF.67, HaM.79] and [Rob.72]. The Mangasarian-Fromowitz constraint qualification was introduced in [MaF.67]. The Local Newton Algorithm 1.8.5 was proposed, independently, by Pshenichnyi [Psh.70] and Robinson [Rob.72], who also introduced the PLI concept in [Rob.72, Rob.74]. For classical results on Newton's method, see [KaA.59].

Chapter 2

Finite Min-Max and Constrained Optimization

We devote this chapter to optimality conditions and algorithms for solving three classes of progressively more difficult optimization problems: min-max problems of the form

$$\textbf{MMP} \qquad \min_{x \in \mathbb{R}^n} \max_{j \in \mathbf{q}} f^j(x) , \tag{0a}$$

inequality constrained optimization problems of the form

$$\textbf{ICP} \qquad \min\{f^0(x) \mid f^j(x) \leq 0,\ j \in \mathbf{q}\} , \tag{0b}$$

and inequality and equality constrained optimization problems of the form

$$\textbf{IECP} \qquad \min\{f^0(x) \mid f^j(x) \leq 0,\ j \in \mathbf{q},\ g^k(x) = 0,\ k \in \mathbf{r}\} , \tag{0c}$$

where the constraint functions $f^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, and $g^k : \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbf{r}$ are continuously differentiable, while the cost function $f^0 : \mathbb{R}^n \to \mathbb{R}$ can be either continuously differentiable on $\mathbb{R}^n$ or a max function of the form

$$f^0(x) = \max_{k \in \mathbf{p}} c^k(x) , \tag{0d}$$

with the $c^k : \mathbb{R}^n \to \mathbb{R}$ continuously differentiable.

In passing, we note that the problem **MMP** is equivalent to the problem

$$\min_{(x, x^{n+1}) \in \mathbb{R}^{n+1}} \{x^{n+1} \mid f^j(x) - x^{n+1} \leq 0,\ j \in \mathbf{q}\} , \tag{0e}$$

which is of the form **ICP**. Later, we will also see that, given a problem of the form **ICP**, one can construct a problem of the form **MMP** that is locally equivalent to it. These observations indicate that min-max problems are not only important in their own right, but that they also provide a natural bridge from unconstrained optimization problems to constrained optimization

problems. Hence we will deal with them first. In fact, we will obtain optimality conditions and some algorithms for constrained optimization problems directly from those for min-max problems.

As in Section 1.1, we will again present necessary optimality conditions in three equivalent forms. The first form is the most basic. It expresses the fact that to first or second-order, the cost must increase at feasible points sufficiently close to a local minimizer. The second form is a consequence of the first form. For first-order conditions, it consists of an equation involving gradients with coefficients that are called multipliers. In the case of second-order conditions, a quadratic inequality involving the multipliers and second derivatives is added to the first-order conditions. The easiest way to verify whether optimality conditions are satisfied in either the first or second form is to set up an auxiliary optimization problem with quadratic cost and affine inequality and equality constraints. The value of this auxiliary optimization problem is a function of x, the point at which the verification is being carried out. We will call this value function an *optimality function*. We will see that optimality functions, as defined by us, are nonpositive-valued and are zero only at points x satisfying the first (and hence also the second) form of optimality conditions. Thus, the third form of optimality conditions is in the form of zeros of an optimality function. Our favorite optimality functions are based on a strictly convex, first-order local model for an optimization problem and have three important advantages: *(i)* they are continuous, *(ii)* their evaluation yields a continuous cost descent direction, and *(iii)* their value can be used to compute upper and lower bounds on the minimum value that is being sought.

Our presentation will make constant use of the mathematical material in the Sections 5.1 - 5.5, and the reader is therefore advised to read these sections before proceeding any further.

2.1 Optimality Conditions for Min-Max

We begin by developing first- and second-order optimality conditions for the min-max problem **MMP** which can be rewritten as

$$\min_{x \in \mathbb{R}^n} \psi(x), \tag{1a}$$

with

$$\psi(x) \triangleq \max_{j \in \mathbf{q}} f^j(x) \tag{1b}$$

and the functions $f^j : \mathbb{R}^n \to \mathbb{R}$ continuously differentiable. We recall that *necessary conditions* must be satisfied by any local minimizer and that a point satisfying necessary conditions is called a *stationary* point. *Sufficient conditions* imply that a point is a local minimizer. Unlike the situation in the unconstrained

minimization of continuously differentiable functions, we will see that convex max functions are not the only ones for which first-order optimality conditions can be both necessary and sufficient.

2.1.1 First-Order Conditions

First we observe that Definition 1.1.1, defining a local minimizer and a strict local minimizer, applies to problem (1a,b). Next, in view of Corollary 5.4.6, we see that the function $\psi(\cdot)$ is directionally differentiable and that its directional derivative at a point x, in the direction h, is given by the formula

$$d\psi(x;h) = \max_{j \in \hat{\mathbf{q}}(x)} \langle \nabla f^j(x), h \rangle, \tag{2a}$$

where

$$\hat{\mathbf{q}}(x) \triangleq \{ j \in \mathbf{q} \mid f^j(x) = \psi(x) \}. \tag{2b}$$

We begin with the most fundamental first-order optimality condition.

Theorem 2.1.1.

(a) Suppose that the functions $f^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, in (1b) are continuously differentiable and that $\hat{x}$ is a local minimizer of $\psi(\cdot)$. Then

$$d\psi(\hat{x};h) \geq 0, \quad \forall\, h \in \mathbb{R}^n. \tag{3a}$$

(b) Suppose that $\hat{x} \in \mathbb{R}^n$ is such that

$$d\psi(\hat{x};h) > 0, \quad \forall\, h \in \mathbb{R}^n,\ h \neq 0. \tag{3b}$$

Then $\hat{x}$ is a strict local minimizer of $\psi(\cdot)$.

Proof. *(a)* This part follows directly from Theorem 1.1.2.

(b) For the sake of contradiction, suppose that (3b) holds and that $\hat{x}$ is not a strict local minimizer. Then there must exist a sequence $\{x_i\}_{i=0}^{\infty}$ such that $x_i \to \hat{x}$, as $i \to \infty$, and $\psi(x_i) \leq \psi(\hat{x})$ for all $i \in \mathbb{N}$. Hence, using the Mean-Value Theorem 5.1.28*(a)*, we obtain that for all $i \in \mathbb{N}$, there exist a $s_i^j \in [0,1]$, $j \in \mathbf{q}$, such that

$$0 \geq \psi(x_i) - \psi(\hat{x}) = \max_{j \in \mathbf{q}} f^j(\hat{x}) - \psi(\hat{x}) + \langle \nabla f^j(\hat{x} + s_i^j(x_i - \hat{x})), x_i - \hat{x} \rangle. \tag{4a}$$

Since $\hat{\mathbf{q}}(\hat{x}) \subset \mathbf{q}$, it now follows that

$$0 \geq \psi(x_i) - \psi(\hat{x}) \geq \max_{j \in \hat{\mathbf{q}}(\hat{x})} \langle \nabla f^j(\hat{x} + s_i^j(x_i - \hat{x})), x_i - \hat{x} \rangle. \tag{4b}$$

Now, for all $i \in \mathbb{N}$, let $h_i \triangleq (x_i - \hat{x}) / \|x_i - \hat{x}\|$. Then there must exist an infinite

subset $K \subset \mathbb{N}$ and a unit vector $\hat{h}$ such that $h_i \to^K \hat{h}$, as $i \to \infty$. Upon dividing both sides of (4b) by $\|x_i - \hat{x}\|$ and taking the limit for $i \in K$, with $i \to \infty$, we conclude that $d\psi(\hat{x}\,;\hat{h}) \le 0$. Since this contradicts (3b), we are done. □

Corollary 2.1.2. *Suppose that the functions $f^j : \mathbb{R}^n \to \mathbb{R}$ in (1b), $j \in \mathbf{q}$, are convex and continuously differentiable. Then $\hat{x}$ is a global minimizer of $\psi(\cdot)$ if and only if (3a) holds.*

Proof. Clearly, in view of Theorem 2.1.1, we need to show only that if $\hat{x}$ is such that (3a) holds, then $\hat{x}$ is a global minimizer. Making use of (5.2.5) and (3a), we find that, for any $x \in \mathbb{R}^n$, $\psi(x) - \psi(\hat{x}) \ge d\psi(\hat{x}\,; x - \hat{x}) \ge 0$. Hence we are done. □

Next we derive first-order optimality conditions in the second form: an equation involving gradients and multipliers.

Theorem 2.1.3. *Suppose that the functions $f^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, in (1b) are continuously differentiable. Then,*

(a) the relation (3a) holds at $\hat{x} \in \mathbb{R}^n$ if and only if

$$0 \in \partial\psi(\hat{x}), \tag{5a}$$

where, by (5.4.7b), the subgradient $\partial\psi(\hat{x})$ is given by

$$\partial\psi(\hat{x}) \triangleq \operatorname*{co}_{j \in \hat{\mathbf{q}}(\hat{x})} \{\nabla f^j(\hat{x})\}, \tag{5b}$$

and

(b) the relation (3b) holds if and only if $0 \in \operatorname{int}[\partial\psi(\hat{x})]$.

Proof. *(a)* For any $x \in \mathbb{R}^n$, let the point in $\partial\psi(x)$ nearest to the origin in $\mathbb{R}^n$ be denoted by $Nr\,[\partial\psi(x)]$, i.e.,

$$Nr\,[\partial\psi(x)] \triangleq \arg\min \{\|\xi\| \mid \xi \in \partial\psi(x)\}. \tag{6a}$$

$\Rightarrow$ Suppose that (3a) holds, but (5a) is not true. Let $\hat{h} = -Nr\,[\partial\psi(\hat{x})]$. Then we must have that $\hat{h} \neq 0$ and hence, by (5.4.7c), that

$$d\psi(\hat{x}\,,\hat{h}) = \max_{\xi \in \partial\psi(\hat{x})} \langle \xi, \hat{h} \rangle = -\|\hat{h}\|^2 < 0, \tag{6b}$$

which contradicts (3a).

$\Leftarrow$ Next suppose that (5a) holds, but that there is an $\hat{h} \in \mathbb{R}^n$ such that

$$d\psi(\hat{x}\,,\hat{h}) = \max_{\xi \in \partial\psi(\hat{x})} \langle \xi, \hat{h} \rangle < 0. \tag{6c}$$

Then we see that set $\partial\psi(\hat{x})$ lies strictly to one side of the hyperplane $\{x \in \mathbb{R}^n \mid \langle x\,, \hat{h} \rangle = 0\}$ passing through the origin, and hence we have a contradiction of (5a).

(b)

$\Rightarrow$ For the sake of contradiction, suppose that (3b) holds and that $0 \notin \operatorname{int}[\partial\psi(\hat{x})]$. Then, since $\partial\psi(\hat{x})$ is convex, it can be separated from the origin, i.e., there exists $h \neq 0$ such that $\langle h\,, \xi \rangle \le 0$ for all $\xi \in \partial\psi(\hat{x})$. Since this implies that $d\psi(\hat{x}\,; h) \le 0$, we have a contradiction.

$\Leftarrow$ Suppose that $0 \in \operatorname{int}[\partial\psi(\hat{x})]$. Then there exists an $\alpha > 0$ such that for any nonzero vector $h \in \mathbb{R}^n$, $\alpha h / \|h\| \in \partial\psi(\hat{x})$. Hence

$$d\psi(\hat{x}\,; h) = \max_{\xi \in \partial\psi(\hat{x})} \langle \xi, h \rangle \ge \alpha \frac{\langle h\,, h \rangle}{\|h\|} = \alpha\|h\| > 0. \tag{6d}$$

□

We now derive an alternative way of stating (5a), which does not involve the *active index set* $\hat{\mathbf{q}}(\hat{x})$.

Theorem 2.1.4. *Suppose that the functions $f^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, in (1b) are continuously differentiable. Then a point $\hat{x}$ satisfies (5a) if and only if there exists a multiplier vector $\hat{\mu}$ in the unit simplex*

$$\Sigma_q \triangleq \{\mu \in \mathbb{R}^q \mid \mu^j \ge 0,\ j \in \mathbf{q},\ \textstyle\sum_{j=1}^q \mu^j = 1\}, \tag{7a}$$

such that

$$\sum_{j=1}^{q} \hat{\mu}^j \nabla f^j(\hat{x}) = 0 \tag{7b}$$

and

$$\sum_{j=1}^{q} \hat{\mu}^j [\psi(\hat{x}) - f^j(\hat{x})] = 0. \tag{7c}$$

Proof. $\Rightarrow$ Suppose that $\hat{x}$ satisfies (5a). Then there must exist multipliers $\hat{\mu}^j \in [0\,,1]$, $j \in \hat{\mathbf{q}}(\hat{x})$, such that

$$\sum_{j \in \hat{\mathbf{q}}(\hat{x})} \hat{\mu}^j = 1, \tag{8a}$$

$$\sum_{j \in \hat{\mathbf{q}}(\hat{x})} \hat{\mu}^j \nabla f^j(\hat{x}) = 0, \tag{8b}$$

and

$$\sum_{j \in \hat{\mathbf{q}}(\hat{x})} \hat{\mu}^j [\psi(\hat{x}) - f^j(\hat{x})] = 0 \tag{8c}$$

(because $\psi(\hat{x}) - f^j(\hat{x}) = 0$ for all $j \in \hat{\mathbf{q}}(\hat{x})$). Hence, if we set $\hat{\mu}^j = 0$ for all $j \notin \hat{\mathbf{q}}(\hat{x})$, we see that (7b) and (7c) are satisfied.

$\Longleftarrow$ Now suppose that (7b) and (7c) hold. Since $\psi(\hat{x}) - f^j(\hat{x}) \geq 0$ for all $j \in \mathbf{q}$, (7c) implies that $\hat{\mu}^j = 0$ for all $j \notin \hat{\mathbf{q}}(\hat{x})$. Hence (7b) is equivalent to (5a), which completes our proof. □

We will refer to the multipliers $\hat{\mu}^j$ in (7b) and (7c) as *Danskin-Demyanov* multipliers.

Theorem 2.1.5. *Suppose that the functions $f^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, in (1b) are convex and continuously differentiable. If if $\hat{x} \in \mathbb{R}^n$ is such that for some $\hat{\mu} \in \Sigma_q$ (7b), (7c) are satisfied, then $\hat{x}$ is a global minimizer for (1a,b).*

Proof. Since (7b) and (7c) imply that (3a) holds, it follows from (5.2.5a) that, for all $x \in \mathbb{R}^n$, $\psi(x) - \psi(\hat{x}) \geq d\psi(\hat{x}; x - \hat{x}) \geq 0$, which concludes our proof. □

2.1.2 Optimality Functions

Now let us consider how we would go about verifying numerically whether (3a) or (5a) or (7b,c) are satisfied at a point $\hat{x}$. Let us begin with (3a). If we define the nonpositive-valued function $\theta_1 : \mathbb{R}^n \to \mathbb{R}$ by

$$\theta_1(x) \triangleq \min_{\|h\|_\infty \leq 1} \max_{j \in \hat{\mathbf{q}}(x)} \langle \nabla f^j(x), h \rangle, \tag{9a}$$

then we see that (3a) holds if and only if $\theta_1(\hat{x}) = 0$. Expression (9a) can be recast as a linear program and, hence, can be evaluated in a finite number of operations. The function $\theta_1(\cdot)$ is u.s.c. This can be shown as follows. Recall that $d\psi(\cdot\,;\cdot)$ is u.s.c. Suppose that $x_i \to \hat{x}$ as $i \to \infty$, and that $\theta_1(\hat{x}) = d\psi(\hat{x}; \hat{h})$. Then $\theta(x_i) \leq d\psi(x_i; \hat{h})$ must hold for all $i \in \mathbb{N}$, and therefore $\overline{\lim}\, \theta_1(x_i) \leq \overline{\lim}\, d\psi(x_i; \hat{h}) \leq d\psi(\hat{x}; \hat{h}) = \theta_1(\hat{x})$.

Furthermore, at any $x \in \mathbb{R}^n$ such that $\theta_1(x) < 0$, the search direction

$$h_1(x) \triangleq \arg \min_{\|h\|_\infty \leq 1} \max_{j \in \hat{\mathbf{q}}(x)} \langle \nabla f^j(x), h \rangle \tag{9b}$$

is obviously a descent direction function for $\psi(\cdot)$ at x. However, because $h_1(\cdot)$ is not continuous, there is a distinct possibility that a steepest descent or Armijo type algorithm for solving (1a,b), using the search direction function $h_1(\cdot)$, may converge to points that do not satisfy first-order optimality conditions.

Next, to determine if (5a) is satisfied, we can define the nonpositive-valued function $\theta_2 : \mathbb{R}^n \to \mathbb{R}$ by

$$\theta_2(x) \triangleq -\tfrac{1}{2} \min_{\xi \in \partial\psi(x)} \|\xi\|^2$$

$$= -\min \{ \tfrac{1}{2} \| \sum_{j \in \hat{\mathbf{q}}(x)} \mu^j \nabla f^j(x) \|^2 \mid \sum_{j \in \hat{\mathbf{q}}(x)} \mu^j = 1, \mu^j \geq 0, j \in \hat{\mathbf{q}}(x) \}. \tag{9c}$$

Clearly, $\theta_2(x) = 0$ if and only if (5a) is satisfied. Since (9c) is a quadratic program, it can be evaluated in a finite number of operations. It follows from Theorem 5.4.1 that $\theta_2(\cdot)$ is u.s.c.

The reason for the seemingly arbitrary $-\frac{1}{2}$ term in (9c) is that, in this form, $\theta_2(\cdot)$ is an extension of the optimality function $\theta(\cdot)$, defined for differentiable functions in (1.1.8f), as we will show now. The exact extension of $\theta(\cdot)$ in (1.1.8f) is the function $\theta'_2 : \mathbb{R}^n \to \mathbb{R}$, defined by

$$\theta'_2(x) \triangleq \min_{h \in \mathbb{R}^n} d\psi(x; h) + \tfrac{1}{2}\|h\|^2$$

$$= \min_{h \in \mathbb{R}^n} \max_{\xi \in \partial\psi(x)} \langle \xi, h \rangle + \tfrac{1}{2}\|h\|^2. \tag{9d}$$

Making use of Corollary 5.5.6, we conclude that

$$\min_{h \in \mathbb{R}^n} \max_{\xi \in \partial\psi(x)} \langle \xi, h \rangle + \tfrac{1}{2}\|h\|^2 = \max_{\xi \in \partial\psi(x)} \min_{h \in \mathbb{R}^n} \langle \xi, h \rangle + \tfrac{1}{2}\|h\|^2. \tag{9e}$$

Since, in the right-hand side of (9e), the minimization with respect to h is unconstrained, we conclude that its solution h_ξ is given by

$$h_\xi = -\xi, \tag{9f}$$

and hence that

$$\theta'_2(x) = \min_{h \in \mathbb{R}^n} \max_{\xi \in \partial\psi(x)} \langle \xi, h \rangle + \tfrac{1}{2}\|h\|^2 = \max_{\xi \in \partial\psi(x)} -\tfrac{1}{2}\|\xi\|^2$$

$$= -\min_{\xi \in \partial\psi(x)} \tfrac{1}{2}\|\xi\|^2 = -\tfrac{1}{2}\|Nr\,[\partial\psi(x)]\|^2 = \theta_2(x), \tag{9g}$$

where $Nr\,[\cdot]$ was defined in (6a). It now follows from Corollary 5.5.6 and (9f,g) that $(-Nr\,[\partial\psi(x)], Nr\,[\partial\psi(x)])$ is the solution pair for the min-max problem (9d) and, hence, that

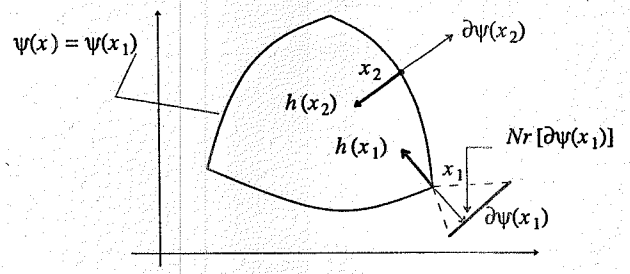


Fig. 2.1.1. The descent direction $h(x) = -Nr[\partial\psi(x)]$.

$$h_2(x) = -Nr[\partial\psi(x)] = \arg\min_{h \in \mathbb{R}^n} \{ \max_{\xi \in \partial\psi(x)} \langle \xi, h \rangle + \tfrac{1}{2}\|h\|^2 \}. \tag{9h}$$

Clearly, whenever $\theta_2(x) < 0$, $d\psi(x; h(x)) < 0$, i.e., that $h_2(x)$ is a descent direction for $\psi(\cdot)$ at x. Unfortunately it is not continuous, because $\partial\psi(x)$ is only o.s.c. since the active function index set $\hat{\mathbf{q}}(x)$ used in the definition of $\partial\psi(x)$ is only o.s.c. by Corollary 5.4.4 and, hence, can change abruptly. This fact is brought out graphically in Fig. 2.1.1. Therefore, when used in a steepest descent type algorithm for solving (1a,b), the search direction $h_2(x) = -Nr[\partial\psi(x)]$ may cause the algorithm to jam, i.e., converge to points that do not satisfy first-order optimality conditions.

Therefore, we will develop an optimality function for problem (1a,b), based on a quadratic, first-order local model for $\psi(\cdot)$, which does yield continuous descent directions. First, as we have already done in (1.1.8a), given a point $x \in \mathbb{R}^n$, for any $j \in \mathbf{q}$ and $h \in \mathbb{R}^n$, we define the first-order, quadratic local model, $\tilde{f}^j(x, \cdot): \mathbb{R}^n \to \mathbb{R}$, for $f^j(\cdot)$, by

$$\tilde{f}^j(x, x') \triangleq f^j(x) + \langle \nabla f^j(x), x' - x \rangle + \tfrac{1}{2}\delta\|x' - x\|^2, \tag{9i}$$

where $\delta > 0$ is an estimate of the average of the eigenvalues of the matrices $f^j_{xx}(x)$ over a subset of $\mathbb{R}^n$ within which one expects to be computing. For the convex case, when these eigenvalues are assumed to lie in an interval $[m, M]$ with $0 < m \leq M < \infty$, such an estimate can be computed using formula (1.3.19), with $f(\cdot)$ replaced by $f^j(\cdot)$, $j \in \mathbf{q}$. In the absence of any information, we set $\delta = 1$. Using this model, $f^j(x + h)$ is approximated by

$$\tilde{f}^j(x, x + h) = f^j(x) + \langle \nabla f^j(x), h \rangle + \tfrac{1}{2}\delta\|h\|^2, \tag{9j}$$

and it should be obvious that there exists a $C < \infty$ such that

$$\overline{\lim_{h \to 0}} \frac{|f^j(x + h) - \tilde{f}^j(x, x + h)|}{\|h\|} \leq C, \tag{9k}$$

which shows that $\tilde{f}^j(x, x + h)$ is a first-order approximation to $f^j(x + h)$.

The level sets of $\tilde{f}^j(x, \cdot)$ are balls centered at $x - \nabla f^j(x)$ (see Fig. 2.1.2). We now define $\tilde{\psi}(x, \cdot): \mathbb{R}^n \to \mathbb{R}$, the piecewise quadratic, first-order local model for $\psi(\cdot)$, by

$$\tilde{\psi}(x, x') \triangleq \max_{j \in \mathbf{q}} \tilde{f}^j(x, x'), \tag{9l}$$

which approximates the value $\psi(x + h)$ by the $\tilde{\psi}(x, x + h)$. Note that because $\tilde{\psi}(x, \cdot)$ is a max of convex functions, it follows from Proposition 5.2.16 that it is a convex. Figure 2.1.2 gives a geometric interpretation of this discussion for the case where $q = 2$. Finally, we define the optimality function $\theta: \mathbb{R}^n \to \mathbb{R}$ and the associated search direction function $h: \mathbb{R}^n \to \mathbb{R}^n$ by

$$\theta(x) \triangleq \min_{h \in \mathbb{R}^n} \tilde{\psi}(x, x + h) - \psi(x)$$

$$= \min_{h \in \mathbb{R}^n} \max_{j \in \mathbf{q}} \{ f^j(x) - \psi(x) + \langle \nabla f^j(x), h \rangle + \tfrac{1}{2}\delta\|h\|^2 \} \tag{9m}$$

and

$$h(x) \triangleq \arg\min_{h \in \mathbb{R}^n} \max_{j \in \mathbf{q}} \{ f^j(x) - \psi(x) + \langle \nabla f^j(x), h \rangle + \tfrac{1}{2}\delta\|h\|^2 \}. \tag{9n}$$

Note that (9l) is a more natural analog of (1.1.8b) than (9d) because it retains more of the structure of the function $\psi(\cdot)$, while still based on a first-order expansion of $\psi(x)$.

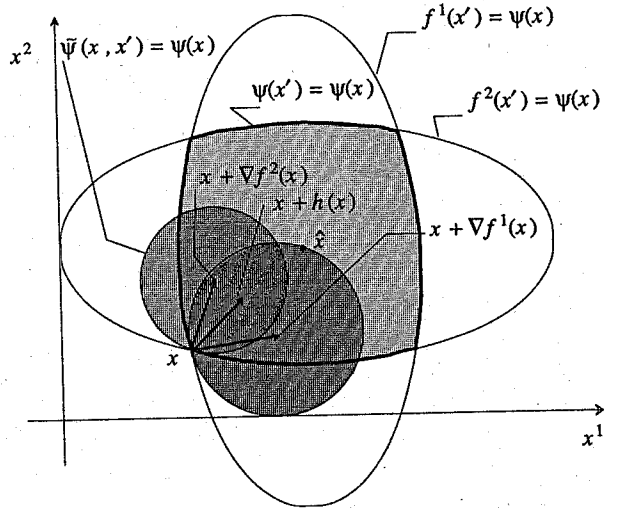


Fig. 2.1.2. A geometric interpretation of $\theta(x)$.

The quantity $\psi(x)+\theta(x)$ can be viewed as a first-order estimate of a local minimum value and $x+h(x)$ can be viewed as a first-order estimate of a local minimizer. In fact, by proceeding as in the proof of Proposition 1.1.6*(e)* (see also (2.4.9f)), one can show that when the functions $f^j(\cdot)$, $j \in \mathbf{q}$, are twice continuously differentiable, and there exist $0 < m \le M < \infty$ such that $m\|h\|^2 \le \langle h, f^j_{xx}(x)h\rangle \le M\|h\|^2$ for all $x, h \in \mathbb{R}^n$ and $j \in \mathbf{q}$, then for any global minimizer $\hat{x}$ of $\psi(\cdot)$ and any $x \in \mathbb{R}^n$,

$$\theta(x)/m \le \psi(\hat{x}) - \psi(x) \le \theta(x)/M\,. \tag{9o}$$

Theorem 2.1.6. *Consider the functions $\theta(\cdot)$ and $h(\cdot)$ defined by (9m) and (9n). Then the following statements hold:*

(a) For all $x \in \mathbb{R}^n$,

$$\theta(x) \le 0\,. \tag{10a}$$

(b) For all $x \in \mathbb{R}^n$,

$$d\psi(x; h(x)) \le \theta(x) - \tfrac{1}{2}\delta\|h(x)\|^2 \le \theta(x)\,. \tag{10b}$$

(c) Alternative expressions for $\theta(x)$ and $h(x)$ are given by[†]

$$\theta(x) = -\min_{\mu \in \Sigma_q} \{\sum_{j=1}^{q} \mu^j[\psi(x) - f^j(x)] + \frac{1}{2\delta}\|\sum_{j=1}^{q}\mu^j\nabla f^j(x)\|^2\} \tag{10c}$$

and

$$h(x) = -\frac{1}{\delta}\sum_{j=1}^{q}\mu_x^j\nabla f^j(x)\,, \tag{10d}$$

where μ_x is any solution of (10c).

Equivalently, $\theta(x)$ and $h(x)$ can be expressed in the form[‡]

$$\theta(x) = -\min_{\bar{\xi} \in \bar{G}\psi(x)} \{\xi^0 + \frac{1}{2\delta}\|\xi\|^2\}\,, \tag{10e}$$

$$\bar{h}(x) \triangleq (h^0(x), h(x)) = -\arg\min_{\bar{\xi} \in \bar{G}\psi(x)} \{\xi^0 + \frac{1}{2\delta}\|\xi\|^2\}\,, \tag{10f}$$

where $\bar{G}\psi(x) \subset \mathbb{R}^{n+1}$ is a set with elements denoted by $\bar{\xi} = (\xi^0, \xi)$, where $\xi^0 \in \mathbb{R}$, $\xi \in \mathbb{R}^n$ and is defined by

[†] The form (10c) is also suggested by (7b) and (7c).

[‡] Note that the scalar component $h^0(x)$ of $\bar{h}(x)$ gets discarded by us.

$$\bar{G}\psi(x) \triangleq \operatorname*{co}_{j \in \mathbf{q}}\left\{\begin{bmatrix}\psi(x) - f^j(x)\\ \nabla f^j(x)\end{bmatrix}\right\}. \tag{10g}$$

(d) For any $x \in \mathbb{R}^n$, $0 \in \partial\psi(x)$ if and only if $0 \in \bar{G}\psi(x)$ if and only if $\theta(x) = 0$.

(e) The function $\theta : \mathbb{R}^n \to \mathbb{R}$ is continuous.

(f) The function $h(\cdot)$ maps $\mathbb{R}^n$ into $\mathbb{R}^n$, i.e., it is point-valued and continuous.

Proof. *(a)* Let

$$\omega(h) \triangleq \max_{j \in \mathbf{q}} \{f^j(x) - \psi(x) + \langle \nabla f^j(x), h\rangle + \tfrac{1}{2}\delta\|h\|^2\}. \tag{11a}$$

Since $\omega(0) = 0$ and $\theta(x) \le \omega(0)$, the desired result follows.

(b) From (9m) and (9n) we obtain

$$\theta(x) = \max_{j \in \mathbf{q}} \{f^j(x) - \psi(x) + \langle \nabla f^j(x), h(x)\rangle + \tfrac{1}{2}\delta\|h(x)\|^2\}. \tag{11b}$$

Hence, since $f^j(x) - \psi(x) = 0$, for every $j \in \hat{\mathbf{q}}(x)$, we have that

$$d\psi(x; h(x)) = \max_{j \in \hat{\mathbf{q}}(x)} \langle \nabla f^j(x), h(x)\rangle \le \theta(x) - \tfrac{1}{2}\delta\|h(x)\|^2 \le \theta(x)\,. \tag{11c}$$

(c) Next, using (9l) and the fact that the maximum over a set of scalars is equal to the maximum over their convex hull, we find that

$$\theta(x) = \min_{h \in \mathbb{R}^n}\max_{\mu \in \Sigma_q} \{\sum_{j=1}^{q}\mu^j[f^j(x) - \psi(x)] + \sum_{j=1}^{q}\mu^j\langle \nabla f^j(x), h\rangle + \tfrac{1}{2}\delta\|h\|^2\}. \tag{11d}$$

Applying Corollary 5.5.6 to (11d), we conclude that

$$\theta(x) = \max_{\mu \in \Sigma_q}\min_{h \in \mathbb{R}^n} \{\sum_{j=1}^{q}\mu^j[f^j(x) - \psi(x)] + \sum_{j=1}^{q}\mu^j\langle \nabla f^j(x), h\rangle + \tfrac{1}{2}\delta\|h\|^2\}. \tag{11e}$$

Now consider the function

$$v(\mu) \triangleq \min_{h \in \mathbb{R}^n} \{\sum_{j=1}^{q}\mu^j[f^j(x) - \psi(x)] + \sum_{j=1}^{q}\mu^j\langle \nabla f^j(x), h\rangle + \tfrac{1}{2}\delta\|h\|^2\}. \tag{11f}$$

Solving the unconstrained minimization problem in (11f) for h in terms of μ, we find that

$$\delta h = -\sum_{j=1}^{q}\mu^j\nabla f^j(x)\,, \tag{11g}$$

and, hence, that

$$v(\mu) = \sum_{j=1}^{q}\mu^j[f^j(x) - \psi(x)] - \frac{1}{2\delta}\|\sum_{j=1}^{q}\mu^j\nabla f^j(x)\|^2\,. \tag{11h}$$

Substituting back into (11e), we obtain (10c).

The expression (10e) follows from (10c) by inspection, while (10d) follows from (11g) and Corollary 5.5.6.

(d) Since $\bar{\xi} = (\xi^0, \xi) \in \bar{G}\psi(x)$ implies that $\xi^0 \geq 0$, it follows that $0 \in \partial\psi(x)$ if and only if $0 \in G\psi(x)$ and also, from (10e), that $\theta(x) = 0$ if and only if $0 \in \bar{G}\psi(x)$. Hence *(d)* is proved.

(e) The continuity of $\theta(\cdot)$ follows from Corollary 5.4.2 and the form of (10c).

(f) First we will show that $\bar{h}(\cdot)$ is point-valued and continuous. For the sake of contradiction, suppose that $\bar{\xi}_* \neq \bar{\xi}_{**}$ are two distinct minimizers for problem (10e). Then we must have that

$$\xi_*^0 + |\xi_*|^2/2\delta = \xi_{**}^0 + |\xi_{**}|^2/2\delta\,. \tag{11i}$$

Suppose that $\xi_* \neq \xi_{**}$. Then, for any $\lambda \in (0,1)$, if we define $\bar{\xi}_\lambda \triangleq \lambda\bar{\xi}_* + (1-\lambda)\bar{\xi}_{**}$, we find that

$$\xi_\lambda^0 + |\xi_\lambda|^2/2\delta < \lambda(\xi_*^0 + |\xi_*|^2/2\delta) + (1-\lambda)(\xi_{**}^0 + |\xi_{**}|^2/2\delta)$$

$$= \xi_*^0 + |\xi_*|^2/2\delta\,. \tag{11j}$$

Since this is clearly impossible because $|\cdot|^2$ is strictly convex, we must assume that $\xi_* = \xi_{**}$. Therefore, because $\xi_*^0 + |\xi_*|^2/2\delta = \xi_{**}^0 + |\xi_{**}|^2/2\delta$, we conclude that $\xi_*^0 = \xi_{**}^0$ and, hence, that $\bar{\xi}_* = \bar{\xi}_{**}$, which implies that $\bar{h}(x)$ is a singleton, i.e., that $\bar{h}(\cdot)$ is point-valued.

To establish the continuity of $h(\cdot)$, we note that, by Corollary 5.3.9, $\bar{G}\psi(\cdot)$ is continuous and, hence, by Theorem 5.4.3, $\{\bar{h}(\cdot)\}$ is an outer semicontinuous set-valued function. However, since it is a singleton, it follows that $\bar{h}(\cdot)$ is continuous and, hence, that $h(\cdot)$ is continuous. □

2.1.3 Second-Order Conditions

Before attempting to develop second-order optimality conditions for problem (1a,b), we must pause to examine the shape of the epigraph of $\psi(\cdot)$. Clearly, it can have ridges, and the minimum value may well lie on a ridge, as in Figures 2.1.3a,b.

Suppose that a local minimizer $\hat{x}$ is such that $(\hat{x}, \psi(\hat{x}))$ is a vertex of the epigraph of $\psi(\cdot)$, as in Figure 2.1.3b, which also shows the projection of the ridges of the epigraph of $\psi(\cdot)$ onto the level sets of $\psi(\cdot)$ for this case (with $q = 3$ and $n = 2$). Then we are in the situation corresponding to (3b) and second-order effects are dominated by first-order effects. Consequently, in this case there can be no second-order optimality conditions.

Another possibility is that $(\hat{x}, \psi(\hat{x}))$ lies on a smooth ridge of the epigraph of $\psi(\cdot)$, as in Figure 2.1.3a, which also shows the projection of the ridge of the epigraph of $\psi(\cdot)$ onto the level sets of $\psi(\cdot)$ for this case (with $q = 2$ and $n = 2$).

The projection of a smooth ridge of the epigraph of $\psi(\cdot)$ is a smooth manifold and the restriction of $\psi(\cdot)$ to this projection is a continuously differentiable function that assumes a local minimum value at $\hat{x}$ and, hence, second-order conditions become meaningful.

Referring to Fig. 2.1.3a, we see that the projection $\boldsymbol{R}$ of a smooth ridge of the epigraph of $\psi(\cdot)$, containing a local minimizer $\hat{x}$ of $\psi(\cdot)$, can be described by the following expression:

$$\boldsymbol{R} \triangleq \{x \in \mathbb{R}^n \mid \hat{\mathbf{q}}(x) = \hat{\mathbf{q}}(\hat{x})\}\,.$$

The development of expressions for smooth curves in $\boldsymbol{R}$ will be facilitated by the following results.

Definition 2.1.7. *(a) The vectors $g_j \in \mathbb{R}^n$, $j \in \mathbf{p}$, are said to be* affinely independent *if the $n+1$-dimensional vectors $(1, g_j)$, $j \in \mathbf{p}$, are linearly independent.*

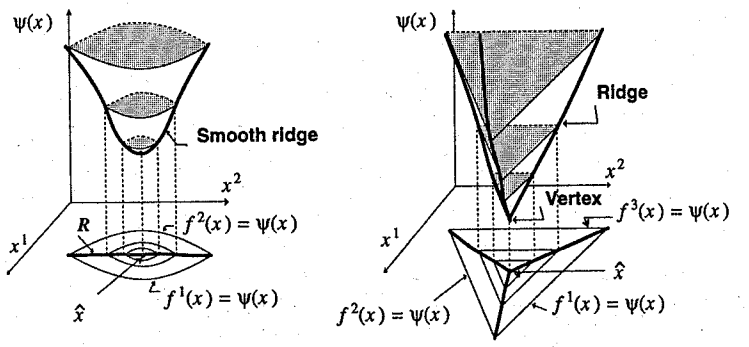
(b) The affine hull *of the set $S \triangleq \{g_j\}_{j=1}^p$ is defined by*

$$\text{aff}\, S \triangleq \{x \in \mathbb{R}^n \mid x = \sum_{j=1}^p \alpha^j g_j\,,\ \sum_{j=1}^p \alpha^j = 1\}. \tag{12}$$

□

The following two exercises give the most obvious properties of affinely independent sets of vectors.

Exercise 2.1.8. Suppose that $S = \{g_j\}_{j=1}^p$, with $g_j \in \mathbb{R}^n$, and that $x \in \mathbb{R}^n$. Show that the vectors in the set $S \cup \{x\}$ are affinely independent if and only if



(a) Solution on smooth ridge (b) Solution at a vertex

Fig. 2.1.3. Geometry of the min-max problem.

the vectors in the set S are affinely independent and $x \notin \text{aff}\, S$. □

Exercise 2.1.9. Show that the vectors in the set $S = \{g_j\}_{j=1}^p$, with $g_j \in \mathbb{R}^n$, are affinely independent if and only if, for every $x \in \text{aff}\, S$, there exist unique (α^j, g_j), $j \in \mathbf{p}$, such that $\sum_{j=1}^p \alpha^j = 1$ and $\sum_{j=1}^p \alpha^j g_j = x$. □

Proposition 2.1.10. *The vectors $g_j \in \mathbb{R}^n$, $j \in \mathbf{p}$, are affinely independent if and only if the vectors $(g_j - g_p)$, $j \in \mathbf{p-1}$ are linearly independent.*

Proof. We will give a proof by contraposition. The vectors $(g_j - g_p)$, $j \in \mathbf{p-1}$, are linearly dependent if and only if there exists a set of coefficients α^j, not all zero, such that

$$\sum_{j=1}^{p-1} \alpha^j g_j - (\sum_{j=1}^{p-1} \alpha^j) g_p = 0. \tag{13a}$$

Let $\alpha^p = -\sum_{j=1}^{p-1} \alpha^j$. Then (13a) can be rewritten in the form

$$\left.\begin{aligned} \sum_{j=1^p} \alpha^j &= 0, \\ \sum_{j=1^p} \alpha^j g_j &= 0, \end{aligned}\right\} \tag{13b}$$

which holds if and only if the g_j, $j \in \mathbf{p}$, are not affinely independent. □

Since the affine independence of a set of vectors $\{g_j\}_{j=1}^p$ is preserved when the vectors are reordered, it should be obvious that the vector g_p, above, was singled out only for notational convenience. The following result will also be useful.

Proposition 2.1.11. *Suppose that the vectors $g_j \in \mathbb{R}^n$, $j \in \mathbf{p}$, are such that for some nonzero $\alpha \in \mathbb{R}^p$,*

$$\sum_{j \in \mathbf{p}} \alpha^j g_j = 0. \tag{14a}$$

Then

$$H_1 \triangleq \{h \in \mathbb{R}^n \mid \langle g_j, h \rangle = 0,\ j \in \mathbf{p}\}$$

$$= H_2 \triangleq \{h \in \mathbb{R}^n \mid \langle g_j - g_p, h \rangle = 0,\ j \in \mathbf{p-1}\}. \tag{14b}$$

Furthermore, if the vectors g_j, $j \in \mathbf{p-1}$, are linearly independent, then the vectors g_j, $j \in \mathbf{p}$, are affinely independent, and

$$H_1 = H_3 \triangleq \{h \in \mathbb{R}^n \mid \langle g_j, h \rangle = 0,\ j \in \mathbf{p-1}\}. \tag{14c}$$

Proof. Clearly, $H_1 \subset H_2$, hence we need to show only that $H_2 \subset H_1$. Suppose that $h \in H_2$. Then we must have that

$$\langle g_j, h \rangle = \langle g_p, h \rangle,\ \forall\, j \in \mathbf{p}. \tag{15a}$$

Now, by assumption, there exists an $\alpha \in \mathbb{R}^p$ such that (14a) holds. Without loss of generality, we can assume that $\sum_{j=1}^p \alpha^j = 1$. Hence, from (15a),

$$0 = \sum_{j=1}^p \alpha^j \langle g_j, h \rangle = \langle g_p, h \rangle. \tag{15b}$$

The fact that $h \in H_1$ now follows from (15a) and (15b).

Next, to obtain a contradiction, suppose that the vectors g_j, $j \in \mathbf{p-1}$, are linearly independent, but the vectors g_j, $j \in \mathbf{p}$, are not affinely independent. Then there exist coefficients β^j, not all zero, such that

$$\sum_{j \in \mathbf{p}} \beta^j \begin{bmatrix} 1 \\ g_j \end{bmatrix} = 0. \tag{15c}$$

Now, by assumption, there exists an $\alpha \in \mathbb{R}^p$, such that (14a) holds, and we can assume, without loss of generality, that $\sum_{j=1}^p \alpha^j = 1$. Since the vectors g_j, $j \in \mathbf{p-1}$, are linearly independent, α is unique. Therefore, it follows that $\beta^j = \gamma\alpha^j$, for all $j \in \mathbf{p}$ and some $\gamma \neq 0$. Clearly, in this case, $\sum_{j \in \mathbf{p}} \beta^j = 0$ cannot hold, as in (15c), and hence we have a contradiction, which proves that the vectors g_j, $j \in \mathbf{p}$, are affinely independent.

Finally, if the vectors g_j, $j \in \mathbf{p-1}$ are linearly independent, then the vector α satisfying (14a), and normalized to satisfy $\sum_{j=1}^p \alpha^j = 1$, is unique, and $\alpha^p \neq 0$. It now follows that

$$g_p = -\frac{1}{\alpha^p} \sum_{j=1}^p \alpha^j g_j \tag{15d}$$

and hence that any $h \in H_3$ is also in H_1. Since, by inspection, $H_1 \subset H_3$, our proof is complete. □

Next, we can associate with problem (1a,b) the *Lagrangian* function $L : \mathbb{R}^n \times \Sigma_q \to \mathbb{R}$, defined by

$$L(x, \mu) \triangleq \sum_{j=1}^q \mu^j f^j(x). \tag{16}$$

We note that if $\hat{x}$ is a local minimizer for (1a,b) and $\hat{\mu}$ is an associated Danskin-Demyanov multiplier, then, in view of (7c), $L(\hat{x}, \hat{\mu}) = \psi(\hat{x})$. Furthermore, $L(x, \hat{\mu}) = \psi(x)$ for all $x \in \boldsymbol{R}$, because $\hat{\mu}^j = 0$ for all $j \notin \hat{\mathbf{q}}(\hat{x})$ and $\sum_{j=1}^q \hat{\mu}^j = 1$. Hence $\hat{x}$ must also be a local minimizer of the restriction of $L(\cdot, \hat{\mu})$ to $\boldsymbol{R}$. Not surprisingly, then, our second-order optimality conditions will be expressed in terms of the Hessian of $L(\cdot, \hat{\mu})$.

Theorem 2.1.12. *Suppose that the functions $f^j(\cdot)$, $j \in \mathbf{q}$, in (1b) are twice continuously differentiable, that $\hat{x}$ is a local minimizer for (1a,b), that $\hat{\mu} \in \Sigma_q$ is an associated Danskin-Demyanov multiplier, satisfying (7b,c), and that the vectors $\{\nabla f^j(\hat{x})\}_{j \in \hat{\mathbf{q}}(\hat{x})}$ are affinely independent. Let*

$$\mathcal{H}(\hat{x}) \triangleq \{h \in \mathbb{R}^n \mid \langle \nabla f^j(\hat{x}), h \rangle = 0,\ j \in \hat{\mathbf{q}}(\hat{x})\}. \tag{17a}$$

Then for all $h \in \mathcal{H}(\hat{x})$,

$$d\psi(\hat{x}; h) = 0, \tag{17b}$$

and

$$\langle h, L_{xx}(\hat{x}, \hat{\mu})h \rangle \geq 0. \tag{17c}$$

Proof. First, since $d\psi(\hat{x}; h) = \max_{j \in \hat{\mathbf{q}}(\hat{x})} \langle \nabla f^j(\hat{x}), h \rangle$, (17b) follows directly from (17a). Next, suppose that $\hat{\mathbf{q}}(\hat{x}) = \{j_1, j_2, \ldots, j_p\}$. Then, by Proposition 2.1.10, the vectors $\{\nabla f^{j_k}(\hat{x}) - \nabla f^{j_p}(\hat{x})\}_{k=1}^{p-1}$ are linearly independent, and by Proposition 2.1.11,

$$\mathcal{H}(\hat{x}) = \{h \in \mathbb{R}^n \mid \langle \nabla f^{j_k}(\hat{x}) - \nabla f^{j_p}(\hat{x}), h \rangle = 0,\ k \in \mathbf{p-1}\}. \tag{18a}$$

It now follows from Corollary 5.1.34 that for every $h \in \mathcal{H}(\hat{x})$, there exists a $t_h > 0$ and a twice continuously differentiable function $s : [0, t_h] \to \mathbb{R}^n$ such that $s(0) = \hat{x}$, $\dot{s}(0) = h$ (where the dot denotes differentiation with respect to t), and $f^{j_k}(s(t)) - f^{j_p}(s(t)) = 0$ for all $t \in [0, t_h]$ and $j_k \in \hat{\mathbf{q}}(\hat{x})$. Clearly, for all $t \in [0, t_h]$, and $j_k \in \hat{\mathbf{q}}(\hat{x})$, the values $f^{j_k}(s(t))$ are all equal. Since, by Corollary 5.4.4, there exists a $\hat{\rho} > 0$ such that, for all $x \in B(\hat{x}, \hat{\rho})$, $\hat{\mathbf{q}}(x) \subset \hat{\mathbf{q}}(\hat{x})$, it follows that there exists a $t'_h \in (0, t_h]$ such that $\psi(s(t)) = f^{j_k}(s(t))$ for at least one $j_k \in \hat{\mathbf{q}}(\hat{x})$, which implies that $f^{j_k}(s(t)) = \psi(s(t))$ for all $j_k \in \hat{\mathbf{q}}(\hat{x})$, i.e., that $s(t) \in \boldsymbol{R}$ for all $t \in [0, t'_h]$. It follows that $L(s(t), \hat{\mu}) = \psi(s(t))$ for all $t \in [0, t'_h]$ and hence that there exists a $t''_h \in (0, t'_h]$ such that $L(s(t), \hat{\mu}) \geq \psi(\hat{x})$ for all $t \in [0, t''_h]$. Since, by inspection, the function $v(t) \triangleq L(s(t), \hat{\mu})$ is twice continuously differentiable on $[0, t'_h]$ and since $v(t) \geq v(0)$ for all $t \in [0, t''_h]$, we see that for all $t \in [0, t''_h]$,

$$0 \leq v(t) - v(0) = t\dot{v}(0) + \tfrac{1}{2}t^2\ddot{v}(0) + o(t^2), \tag{18b}$$

where $o(t^2)/t^2 \to 0$, as $t \to 0$. Now, it follows from the chain rule and (16) that $\dot{v}(0) = \langle \sum_{j=1}^q \hat{\mu}^j \nabla f^j(\hat{x}), h \rangle = 0$ and, also, that

$$\ddot{v}(0) = \langle \textstyle\sum_{j=1}^q \hat{\mu}^j \nabla f^j(\hat{x}), \ddot{s}(0) \rangle + \langle h, L_{xx}(\hat{x}, \hat{\mu})h \rangle = \langle h, L_{xx}(\hat{x}, \hat{\mu})h \rangle, \tag{18c}$$

because $\sum_{j=1}^q \hat{\mu}^j \nabla f^j(\hat{x}) = 0$. Substituting into (18b) and dividing by $t^2/2$, we conclude that

$$\langle h, L_{xx}(\hat{x}, \hat{\mu})h \rangle + 2o(t^2)/t^2 \geq 0. \tag{18d}$$

Letting $t \to 0$, we obtain the desired result. □

Next we develop second-order sufficient conditions.

Theorem 2.1.13. *Suppose that the functions $f^j(\cdot)$, $j \in \mathbf{q}$, in (1b) are twice continuously differentiable and that $\hat{x} \in \mathbb{R}^n$ together with $\hat{\mu} \in \Sigma_q$ satisfy the first-order optimality conditions (7b,c). Let*

$$\hat{\mathbf{q}}_+(\hat{x}, \hat{\mu}) \triangleq \{j \in \hat{\mathbf{q}}(\hat{x}) \mid \hat{\mu}^j > 0\}, \tag{19a}$$

$$\hat{\mathbf{q}}_+^c(\hat{x}, \hat{\mu}) \triangleq \{j \in \hat{\mathbf{q}}(\hat{x}) \mid j \notin \hat{\mathbf{q}}_+(\hat{x}, \hat{\mu})\}, \tag{19b}$$

and

$$\mathcal{H}'(\hat{x}) \triangleq \{h \in \mathbb{R}^n \mid \langle \nabla f^j(\hat{x}), h \rangle = 0,\ j \in \hat{\mathbf{q}}_+(\hat{x}, \hat{\mu}),$$

$$\langle \nabla f^j(\hat{x}), h \rangle \leq 0,\ j \in \hat{\mathbf{q}}_+^c(\hat{x}, \hat{\mu})\}. \tag{19c}$$

If there exists an $m > 0$ such that

$$\langle h, L_{xx}(\hat{x}, \hat{\mu})h \rangle \geq m\|h\|^2,\ \forall h \in \mathcal{H}'(\hat{x}), \tag{19d}$$

then $\hat{x}$ is a strict local minimizer for (1a,b).

Proof. Suppose that $\hat{x}$ is not a strict local minimizer for (1a,b). Then there exists a sequence $\{x_i\}_{i=0}^{\infty}$, such that $x_i \to \hat{x}$, as $i \to \infty$ and $\psi(x_i) \leq \psi(\hat{x})$ for all $i \in \mathbb{N}$. Let $\delta x_i \triangleq x_i - \hat{x}$ and $h_i \triangleq \delta x_i / \|\delta x_i\|$. Since the vectors h_i are all of unit norm, the sequence $\{h_i\}_{i=0}^{\infty}$ must have accumulation points. Without loss of generality we may assume that $h_i \to \hat{h}$, as $i \to \infty$ (with $\|\hat{h}\| = 1$). We need to consider two cases.

Case 1. First suppose that $\hat{h} \in \mathcal{H}'(\hat{x})$. Then, since $\|\hat{h}\| = 1$, $\langle \hat{h}, L_{xx}(\hat{x}, \hat{\mu})\hat{h} \rangle \geq m$, and hence, since $x_i \to \hat{x}$ and $h_i \to \hat{h}$ as $i \to \infty$, and $L_{xx}(\cdot, \hat{\mu})$ is continuous, there exists an i_0 such that, for all $i \geq i_0$,

$$\langle h_i, L_{xx}(\hat{x}, \hat{\mu})h_i \rangle - m \geq -\frac{m}{2}, \tag{20a}$$

and, for all $s \in [0, 1]$, $\|L_{xx}(\hat{x} + s\,\delta x_i, \hat{\mu}) - L_{xx}(\hat{x}, \hat{\mu})\| \leq m/4$. Hence, for all

$i \geq i_0$ and $s \in [0,1]$,

$$\langle h_i, L_{xx}(\hat{x} + s\delta x_i, \hat{\mu})h_i \rangle \geq \langle h_i, L_{xx}(\hat{x}, \hat{\mu})h_i \rangle - |\langle h_i, [L_{xx}(\hat{x} + s\delta x_i, \hat{\mu}) - L_{xx}(\hat{x}, \hat{\mu})]h_i \rangle|$$

$$\geq \frac{m}{2} - \frac{m}{4} = \frac{m}{4}. \tag{20b}$$

Using second-order expansions and (20b), we find that

$$\begin{aligned}
\psi(x_i) - \psi(\hat{x}) &= \max_{j \in \mathbf{q}} f^j(x_i) - \psi(\hat{x}) \\
&= \max_{\mu \in \Sigma_q} \sum_{j=1}^{q} \mu^j [f^j(x_i) - \psi(\hat{x})] \\
&= \max_{\mu \in \Sigma_q} \Big\{ \sum_{j=1}^{q} \mu^j [f^j(\hat{x}) - \psi(\hat{x})] + \langle \sum_{j=1}^{q} \mu^j \nabla f^j(\hat{x}), \delta x_i \rangle \\
&\quad + \int_0^1 (1-s) \langle \delta x_i, \sum_{j=1}^{q} \mu^j f^j_{xx}(\hat{x} + s\delta x_i)\delta x_i \rangle \, ds \Big\} \\
&\geq \sum_{j=1}^{q} \hat{\mu}^j [f^j(\hat{x}) - \psi(\hat{x})] + \langle \sum_{j=1}^{q} \hat{\mu}^j \nabla f^j(\hat{x}), \delta x_i \rangle + \frac{m}{8} \|\delta x_i\|^2. \qquad (20c)
\end{aligned}$$

Since the first and second terms in the last line of (20c) are zero by assumption, it follows from (20c) that

$$\psi(x_i) - \psi(\hat{x}) \geq \frac{m}{8} \|\delta x_i\|^2, \tag{20d}$$

for all $i \geq i_0$, a contradiction.

Case 2. Hence we must consider the only other alternative, viz., that $\hat{h} \notin \mathcal{H}'(\hat{x})$. Since, by assumption, $\psi(x_i) \leq \psi(\hat{x})$ for all $i \in \mathbb{N}$, it follows that $f^j(x_i) \leq \psi(\hat{x})$ for all $j \in \mathbf{q}$, and all $i \in \mathbb{N}$. Consequently, since, for all $j \in \hat{\mathbf{q}}(\hat{x})$, $f^j(\hat{x}) = \psi(\hat{x})$, it follows from the Mean-Value Theorem 5.1.28*(a)* that there exist $s_i^j \in (0,1)$ such that for all $j \in \hat{\mathbf{q}}(\hat{x})$,

$$0 \geq \psi(x_i) - \psi(\hat{x}) \geq f^j(x_i) - f^j(\hat{x}) = \langle \nabla f^j(\hat{x} + s_i^j \delta x_i), \delta x_i \rangle, \tag{20e}$$

Since $s_i^j \delta x_i \to 0$, as $i \to \infty$, if we divide by $\|\delta x_i\|$ and let $i \to \infty$, we conclude from (20e) that

$$\langle \nabla f^j(\hat{x}), \hat{h} \rangle \leq 0, \quad \forall\, j \in \hat{\mathbf{q}}(\hat{x}). \tag{20f}$$

Now, the fact that $\hat{h} \notin \mathcal{H}'(\hat{x})$ implies that either *(a)* there exists a $k \in \hat{\mathbf{q}}_+(\hat{x}, \hat{\mu})$ such that $\langle \nabla f^k(\hat{x}), \hat{h} \rangle \neq 0$ or *(b)* there exists an $l \in \hat{\mathbf{q}}_+^c(\hat{x}, \hat{\mu})$ such that $\langle \nabla f^l(\hat{x}), \hat{h} \rangle > 0$. Clearly, in view of (20f), *(b)* cannot hold. Hence we must assume that *(a)* holds, which, because of (20f), implies that $\langle \nabla f^k(\hat{x}), \hat{h} \rangle < 0$. Since, by assumption, $\hat{\mu}^k > 0$, using (7b) we obtain

$$0 > \langle \nabla f^k(\hat{x}), \hat{h} \rangle = - \sum_{\substack{j \in \hat{\mathbf{q}}(\hat{x}) \\ j \neq k}} (\hat{\mu}^j / \hat{\mu}^k) \langle \nabla f^j(\hat{x}), \hat{h} \rangle, \tag{20g}$$

which implies that for at least one $j \in \hat{\mathbf{q}}(\hat{x})$, $\langle \nabla f^j(\hat{x}), \hat{h} \rangle > 0$, contradicting (20f). Hence $\hat{x}$ is a strict local minimizer for **MMP**. □

2.1.4 Notes

First-order optimality conditions for min-max problems, such as Theorems 2.1.1, 2.1.3, and 2.1.4, were first established, independently, by Danskin [Dan.69] and Demyanov [Dem.66], see also [Dem.67, Dem.68, DeM.74, DeV.81]. Second-order conditions can be found in [DeM.74, DeV.81] and in [Han.78]. A systematic presentation of optimality functions was first given in [Pol.87].

2.2 Optimality Conditions for Constrained Optimization

First we will develop optimality conditions for the problem **ICP** and then for the problem **IECP**, defined in (2.1.0b) and (2.1.0c), respectively. We define the inequality constraint violation function $\psi : \mathbb{R}^n \to \mathbb{R}$ by

$$\psi(x) \triangleq \max_{j \in \mathbf{q}} f^j(x), \tag{1a}$$

and we define the equality constraint violation function $g : \mathbb{R}^n \to \mathbb{R}^r$ by

$$g(x) \triangleq (g^1(x), \ldots, g^r(x)), \tag{1b}$$

where the functions $f^j(\cdot)$ and $g^k(\cdot)$ are as in **ICP** and **IECP** (see (2.1.0b,c)).

2.2.1 First-Order Optimality Conditions for ICP

We now turn to the problem

$$\textbf{ICP} \qquad \min \{ f^0(x) \mid f^j(x) \leq 0,\ j \in \mathbf{q} \}, \tag{2a}$$

where we will assume that

$$f^0(x) \triangleq \max_{k \in \mathbf{p}} c^k(x) \tag{2b}$$

and that the functions $f^j, c^k : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, $k \in \mathbf{p}$, are at least once

continuously differentiable. Clearly, a more compact way of restating problem o2a) is

$$\min \{ f^0(x) \mid \psi(x) \leq 0 \}. \tag{2c}$$

We will show that the local minimizers of **ICP** are also local minimizers of a certain max function, and hence, by extension of the results in the preceding section, we will establish first-order optimality conditions for the problem **ICP** in three equivalent forms: a basic directional derivative form, a derived multiplier form, and an optimality function which provides a computational means for verifying whether or not the optimality conditions in the first two forms are satisfied.

Definition 2.2.1. *Let* $X_I \triangleq \{ x \in \mathbb{R}^n \mid \psi(x) \leq 0 \}$.

(a) We will say that $\hat{x} \in X_I$ *is a* local minimizer *for* ICP *if there exists a* $\rho > 0$ *such that* $f^0(x) \geq f^0(\hat{x})$ *for all* $x \in X_I \cap B(\hat{x}, \rho)$.

(b) We will say that $\hat{x}$ *is a* strict local minimizer *if* $f^0(x) > f^0(\hat{x})$ *for all* $x \in X_I \cap B(\hat{x}, \rho), x \neq \hat{x}$. □

We will now develop a min-max problem that is locally equivalent to the problem **ICP**. Suppose that $\hat{x}$ is a local minimizer for (2a,b), and let $\hat{F} : \mathbb{R}^n \to \mathbb{R}$ be defined by

$$\hat{F}(x) \triangleq \max \{ f^0(x) - f^0(\hat{x}), \psi(x) \}$$

$$= \max \{ \max_{k \in \mathbf{p}} [c^k(x) - f^0(\hat{x})], \max_{j \in \mathbf{q}} f^j(x) \}. \tag{3}$$

Theorem 2.2.2. *Consider the problem* **ICP** *in (2a,b). Suppose that the functions* $c^k(\cdot)$, $k \in \mathbf{p}$ *and* $f^j(\cdot)$, $j \in \mathbf{q}$, *are continuous and that* $\hat{x}$ *is a local minimizer for* **ICP**. *Then* $\hat{x}$ *is a local minimizer for* $\hat{F}(\cdot)$.

Proof. First, since $\psi(\hat{x}) \leq 0$, by assumption, we see that $\hat{F}(\hat{x}) = 0$. Next, let $\hat{\rho} > 0$ be the radius associated with $\hat{x}$. Then, for all $x \in B(\hat{x}, \hat{\rho})$, if $\psi(x) \geq 0$, then $\hat{F}(x) \geq 0$, and if $\psi(x) \leq 0$, then $f^0(x) - f^0(\hat{x}) \geq 0$, which also implies that $\hat{F}(x) \geq 0$. Hence $\hat{x}$ is a local minimizer for $\hat{F}(\cdot)$. □

The converse of Theorem 2.2.2 is not always true. Thus, suppose that $\hat{x}$ is a local minimizer of the function $\hat{F}(\cdot)$ and that $\psi(\hat{x}) = 0$. Then, for any x near $\hat{x}$ such that $\psi(x) = 0$, we will have $\hat{F}(x) = \hat{F}(\hat{x}) = 0$, as expected, but this does not exclude the possibility that $f^0(x) < f^0(\hat{x})$, i.e., that $\hat{x}$ is not a local minimizer of **ICP**. Because of this, the following theorem includes a restriction, $0 \notin \partial\psi(\hat{x})$ (sometimes referred to as a *constraint qualification*), on the problem **ICP**.

Theorem 2.2.3. *Suppose that the functions* $c^k(\cdot)$, $k \in \mathbf{p}$, *are continuous and that the functions* $f^j(\cdot)$, $j \in \mathbf{q}$, *in (2a,b) are continuously differentiable. If* $\hat{x}$ *is a local minimizer of* $\hat{F}(\cdot)$ *and either* $\psi(\hat{x}) < 0$ *or* $\psi(\hat{x}) = 0$ *and* $0 \notin \partial\psi(\hat{x})$, *then* $\hat{x}$ *is a local minimizer for* **ICP** in (2a,b).

Proof. First suppose that $\psi(\hat{x}) < 0$. Since $\psi(\cdot)$ is continuous, and $\hat{F}(\hat{x}) = 0$, and $\hat{x}$ is a local minimizer of $\hat{F}(\cdot)$, there exists a $\hat{\rho} > 0$ such that, for all $x \in B(\hat{x}, \hat{\rho})$, $\hat{F}(x) \geq 0$ and $\psi(x) < 0$. It now follows by inspection that, for all $x \in B(\hat{x}, \hat{\rho})$, $f^0(x) - f^0(\hat{x}) \geq 0$, which shows that $\hat{x}$ is a local minimizer for **ICP**.

Next suppose that $\psi(\hat{x}) = 0$. Since $\hat{x}$ is a local minimizer of $\hat{F}(\cdot)$, it follows that there exists a $\hat{\rho} > 0$ such that for all $x \in B(\hat{x}, \hat{\rho})$, $\hat{F}(x) \geq 0$. Also, since $0 \notin \partial\psi(\hat{x})$, by assumption, and since $\partial\psi(\cdot)$ is outer semicontinuous by Corollary 5.4.6, it follows that there exists a $\rho^* \in (0, \hat{\rho}/2]$ such that $0 \notin \partial\psi(x)$ for all $x \in B(\hat{x}, \rho^*)$. Now, if $x \in B(\hat{x}, \rho^*)$ is such that $\psi(x) < 0$, then $\hat{F}(x) \geq 0$ implies that $f^0(x) - f^0(\hat{x}) \geq 0$. If $x \in B(\hat{x}, \rho^*)$ is such that $\psi(x) = 0$, then, because $0 \notin \partial\psi(x)$, there exists, by Theorem 2.1.3, a vector $h \in \mathbb{R}^n$ such that $d\psi(\hat{x}; h) < 0$. Consequently, there exists a sequence of points $x_i = x + \lambda_i h \in B(\hat{x}, \hat{\rho})$, $i \in \mathbb{N}$, (with $\lambda_i \to 0$, as $i \to \infty$) converging to x, such that $\psi(x_i) < 0$ for all i, which implies that $\hat{F}(x_i) = f^0(x_i) - f^0(\hat{x}) \geq 0$ for all i. It now follows from the continuity of $f^0(\cdot)$ that $f^0(x) - f^0(\hat{x}) \geq 0$. Hence we conclude that $\hat{x}$ is a local minimizer for **ICP**. □

In view of Theorems 2.2.2 and 2.2.3, it should be clear that we can obtain both necessary and sufficient conditions of optimality for the inequality constrained problem **ICP** in (2a,b) by interpreting those for min-max problem (2.1.1a,b). In particular, Theorem 2.1.1 implies that if $\hat{x}$ is a local minimizer for **ICP**, then

$$d\hat{F}(\hat{x}; h) \geq 0, \quad \forall h \in \mathbb{R}^n, \tag{4a}$$

or, equivalently,

$$\min_{h \in \mathbb{R}^n} \max_{\mu^0 \in \mathcal{M}} \mu^0 df^0(\hat{x}; h) + (1 - \mu^0) d\psi(\hat{x}; h) \geq 0, \quad \forall h \in \mathbb{R}^n, \tag{4b}$$

where $\mathcal{M} \triangleq \{ \mu^0 \in [0, 1] \mid (1 - \mu^0)\psi(\hat{x}) = 0 \}$ (note that $\mu^0 = 1$ when $\psi(\hat{x}) < 0$).

It now follows from Corollary 5.5.3 that there exists a $\hat{\mu}^0 \in \mathcal{M}$ such that (4b) holds if and only if

$$\min_{h \in \mathbb{R}^n} \hat{\mu}^0 df^0(\hat{x}\,;h) + (1-\hat{\mu}^0) d\psi(\hat{x}\,;h) \geq 0\,, \quad \forall\, h \in \mathbb{R}^n\,. \tag{4c}$$

To deduce an expanded result of the form of Theorem 2.1.4 from these statements, we begin by defining a $q+1$-dimensional unit simplex whose components are numbered from 0 to q. Hence, for any integer $q \geq 1$, we let

$$\Sigma_q^0 \triangleq \{(\mu^0, \mu^1, \ldots, \mu^q) \mid \sum_{j=0}^{q} \mu^j = 1\,;\ \mu^j \geq 0\,, j = 0, 1, \ldots, q\,\}. \tag{4d}$$

If we define the functions $f_*^j(\cdot)$, $j \in \mathbf{p+q}$, as follows:

$$f_*^j(x) \triangleq c^j(x) - f^0(\hat{x})\,, \ \forall\, j \in \mathbf{p}\,, \tag{4e}$$

$$f_*^{j+p}(x) \triangleq f^j(x)\,, \ \forall\, j \in \mathbf{q}\,, \tag{4f}$$

then we obtain the now familiar form

$$\hat{F}(x) = \max_{j \in \mathbf{p+q}} f_*^j(x)\,, \tag{4g}$$

and hence as we will show, it follows from Theorem 2.1.4 that the following result is true:

Theorem 2.2.4. *Consider the problem* **ICP**, *in (2a,b). Suppose that the functions $f^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, in (2a) and the functions $c^k : \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbf{p}$, in (2b) are continuously differentiable. If $\hat{x}$ is a local minimizer for* **ICP**, *then there exist multiplier vectors $\hat{\mu} \in \Sigma_q^0$ and $\hat{\nu} \in \Sigma_p$ such that*

$$\hat{\mu}^0 \Big[\sum_{k=1}^{p} \hat{\nu}^k \nabla c^k(\hat{x})\Big] + \sum_{j=1}^{q} \hat{\mu}^j \nabla f^j(\hat{x}) = 0 \tag{5a}$$

and

$$\hat{\mu}^0 \Big[\sum_{k=1}^{p} \hat{\nu}^k [c^k(\hat{x}) - f^0(\hat{x})]\Big] + \sum_{j=1}^{q} \hat{\mu}^j f^j(\hat{x}) = 0\,. \tag{5b}$$

Alternatively, restating (5a) and (5b) in terms of subgradients, we obtain

$$0 \in \{\hat{\mu}^0 \partial f^0(\hat{x}) + \sum_{j=1}^{q} \hat{\mu}^j \partial f^j(\hat{x})\}\,, \tag{5c}$$

and

$$\sum_{j=1}^{q} \hat{\mu}^j f^j(\hat{x}) = 0\,. \tag{5d}$$

Proof. With the functions $f_*^j(\cdot)$ defined as in (4e,f) and $\hat{F}(\cdot)$ defined as in (4g), it follows from Theorem 2.1.4 that there exists a multiplier vector $\hat{\mu}_* \in \Sigma_{p+q}$ such that

$$\sum_{j=1}^{p} \mu_*^j \nabla f_*^j(\hat{x}) + \sum_{j=p+1}^{p+q} \mu_*^j \nabla f_*^j(\hat{x}) = 0 \tag{6a}$$

and

$$\sum_{j=1}^{p} \mu_*^j [\hat{F}(\hat{x}) - f_*^j(\hat{x})] + \sum_{j=p+1}^{p+q} \mu_*^j [\hat{F}(\hat{x}) - f_*^j(\hat{x})] = 0\,. \tag{6b}$$

Taking into account the definitions of the $f_*^j(\cdot)$, the fact that $\hat{F}(\hat{x}) = 0$, and setting $\hat{\mu}^0 \triangleq \sum_{j=1}^{p} \mu_*^j$, $\hat{\nu}^k \triangleq \mu_*^k/\hat{\mu}^0$, $k \in \mathbf{p}$, and $\hat{\mu}^j \triangleq \hat{\mu}_*^{j+p}$, $j \in \mathbf{q}$, we find that (6a) and (6b) transform into (5a) and (5b), respectively. At this point, the relations (5c) and (5d) are obvious. □

When $\hat{\mu}^0 \geq 0$ in (6a) and (6b), these equations are referred to as *generalized Fritz John* conditions. In that case the multipliers $\hat{\mu}^j$ in (6a) and (6b) are referred to as *Fritz John* multipliers. When (6a) and (6b) hold with $\hat{\mu}^0 > 0$, (6a) and (6b) are referred to as *generalized Karush-Kuhn-Tucker* conditions, and the normalized multipliers, $\hat{\eta}^j \triangleq \hat{\mu}^j/\hat{\mu}^0$, $j \in \mathbf{q}$, are called *Karush-Kuhn-Tucker* (KKT) multipliers. A doublet $(\hat{x}\,, \hat{\eta})$ consisting of a *feasible point* $\hat{x}$ (i.e., a point satisfying $f^j(\hat{x}) \leq 0$, $j \in \mathbf{q}$, and $g(\hat{x}) = 0$) and an associated KKT multiplier $\hat{\eta}$ is called a *Karush-Kuhn-Tucker doublet.*

The qualification *generalized* is removed from "the generalized Karush-Kuhn-Tucker conditions" when the cost function $f^0(\cdot)$ is differentiable.

Note that when $\hat{\mu}^0 = 0$, the above conditions are degenerate in the sense that they contain no information supplied by the cost function. It takes little effort to convince oneself that if $\psi(\hat{x}) = 0$ and $0 \in \partial\psi(\hat{x})$, then one can satisfy (6a) and (6b) with $\hat{\mu}^0 = 0$ because, in this case, $\hat{x}$ also satisfies the necessary condition for a local minimizer of $\psi(\cdot)$. The next result is an obvious consequence of Theorem 2.2.4 and the fact that when $p=1$, then $\hat{\nu}^1 = 1$ in (6a).

Corollary 2.2.5. *Consider the special case of problem* **ICP** *in (2a,b), where $p = 1$, i.e., $f^0(x) \equiv c^1(x)$, so that the functions $f^j(\cdot)$, $j = 0, 1, \ldots, q$, are all continuously differentiable. If $\hat{x}$ is a local minimizer for this special case of (2a,b), then there exists a $\hat{\mu} \in \Sigma_q^0$ such that*

$$\sum_{j=0}^{q} \hat{\mu}^j \nabla f^j(\hat{x}) = 0 \tag{7a}$$

and

$$\sum_{j=1}^{q} \hat{\mu}^j f^j(\hat{x}) = 0\,. \tag{7b}$$

□

Corollary 2.2.6. *Consider the problem* **ICP** *in (2a,b). Suppose that the functions* $c^k, f^j : \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbf{p}$, $j \in \mathbf{q}$, *in (2a,b) are convex and continuously differentiable. Suppose that* $\hat{x} \in \mathbb{R}^n$ *is such that* $\psi(\hat{x}) \leq 0$, *that* $0 \notin \partial\psi(\hat{x})$ *if* $\psi(\hat{x}) < 0$, *and that there exist* $\hat{\mu} \in \Sigma_q^0$, $\hat{\nu} \in \Sigma_p$ *are such that (6a), (6b) hold, then* $\hat{x}$ *is a global minimizer of problem* **ICP**. □

Exercise 2.2.7. *(a)* Suppose that the functions $f^j(\cdot)$, in (2a), are continuously differentiable, and consider the feasible set $F \triangleq \{x \in \mathbb{R}^n \mid \psi(x) \leq 0\}$ for the problem **ICP**. Show that, if, for all $x \in F$ such that $\psi(x) = 0$, $0 \notin \partial\psi(x)$, then

(i) $\operatorname{int} F = \{x \in \mathbb{R}^n \mid \psi(x) < 0\}$,

(ii) $\operatorname{int} F$ is nonempty, and

(iii) F is equal to the closure of $\operatorname{int} F$.

(b) Consider (5a). Show that if $\psi(\hat{x}) = 0$ and $0 \notin \partial\psi(\hat{x})$, then $\hat{\mu}^0 > 0$ must hold in (5a).

(c) Prove Corollary 2.2.6. *Hint*: Use Theorem 2.1.5 and Theorem 2.2.3. □

2.2.2 An Optimality Function for ICP

Our next task is to adapt the optimality function $\theta(\cdot)$ defined in (2.1.9a) and the search direction function $h(\cdot)$ defined in (2.1.9b) to the problem **ICP** in (2a,b). Since we cannot use the function $\hat{F}(\cdot)$ defined in (3) because it requires the advance knowledge of a local minimizer, we approximate it by the function $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$F(z,x) \triangleq \max\{f^0(x) - f^0(z) - \gamma\psi(z)_+, \psi(x) - \psi(z)_+\}$$

$$= \max\left\{\max_{k \in \mathbf{p}}\{c^k(x) - f^0(z) - \gamma\psi(z)_+\}, \max_{j \in \mathbf{q}}\{f^j(x) - \psi(z)_+\}\right\}, \tag{8a}$$

where $\gamma > 0$, and for any $z \in \mathbb{R}^n$,

$$\psi(z)_+ \triangleq \max\{0, \psi(z)\}. \tag{8b}$$

As we will see later, the terms $\gamma\psi(z)_+$ and $\psi(z)_+$ are required for the construction of certain phase I - phase II methods of feasible directions algorithms. Note

that if $\hat{x}$ is a local minimizer of **ICP** in (2a,b), then $\psi(\hat{x}) \leq 0$ and hence $F(\hat{x}, x) = \hat{F}(x)$ for all $x \in \mathbb{R}^n$.

We will require the directional derivatives and subgradients of $F(\cdot,\cdot)$, with respect to the second argument. We will denote these by $d_2F(\cdot,\cdot;\cdot)$, and $\partial_2 F(\cdot,\cdot)$. Thus, by definition,

$$d_2F(z,x;h) \triangleq \lim_{t \downarrow 0} \frac{F(z, x+th) - F(x,z)}{t}, \tag{8c}$$

and

$$\partial_2 F(z,x) \triangleq \{\xi \in \mathbb{R}^n \mid \langle \xi, h\rangle \leq d_2F(z,x;h),\ \forall h \in \mathbb{R}^n\}. \tag{8d}$$

Let $\delta > 0$ be given. Then we observe that

$$\tilde{F}(x, x+h) \triangleq \max\left\{\max_{k \in \mathbf{p}}\{c^k(x) - f^0(x) - \gamma\psi(x)_+ + \langle \nabla c^k(x), h\rangle\},\right.$$

$$\left.\max_{j \in \mathbf{q}}\{f^j(x) - \psi(x)_+ + \langle \nabla f^j(x), h\rangle\}\right\} + \tfrac{1}{2}\delta\|h\|^2 \tag{8e}$$

is a first-order, convex (in h) approximation to $F(x, x+h)$, and hence, taking into account that $F(x,x) = 0$ for all $x \in \mathbb{R}^n$, we obtain the following definitions from (2.1.9a) and (2.1.9b):

$$\theta(x) \triangleq \min_{h \in \mathbb{R}^n} \tilde{F}(x, x+h)$$

$$= \min_{h \in \mathbb{R}^n}\left\{\max\left\{\max_{k \in \mathbf{p}}\{c^k(x) - f^0(x) - \gamma\psi(x)_+ + \langle \nabla c^k(x), h\rangle,\right.\right.$$

$$\left.\left.\max_{j \in \mathbf{q}}\{f^j(x) - \psi(x)_+ + \langle \nabla f^j(x), h\rangle\}\right\} + \tfrac{1}{2}\delta\|h\|^2\right\}, \tag{8f}$$

and

$$h(x) \triangleq \arg\min_{h \in \mathbb{R}^n}\left\{\max\left\{\max_{j \in \mathbf{p}}\{c^k(x) - f^0(x) - \gamma\psi(x)_+ + \langle \nabla c^k(x), h\rangle\},\right.\right.$$

$$\left.\left.\max_{j \in \mathbf{q}}\{f^j(x) - \psi(x)_+ + \langle \nabla f^j(x), h\rangle\}\right\} + \tfrac{1}{2}\delta\|h\|^2\right\}. \tag{8g}$$

Next we get the following extension of Theorem 2.1.6:

Theorem 2.2.8. *Suppose that the functions* $c^k(\cdot)$, $k \in \mathbf{p}$, *in (2b) and the functions* $f^j(\cdot)$, $j \in \mathbf{q}$, *in (2a) are all continuously differentiable. Consider the functions* $\theta(\cdot)$ *and* $h(\cdot)$ *defined by (8f) and (8g). Then the following statements hold:*

(a) For all $x \in \mathbb{R}^n$,

$$\theta(x) \le 0 . \tag{9a}$$

(b) For all $x \in \mathbb{R}^n$,

$$d_2F(x, x; h(x)) \le \theta(x) - \tfrac{1}{2}\delta\|h(x)\|^2 \le \theta(x) . \tag{9b}$$

(c) For all $x \in \mathbb{R}^n$,

$$\psi(x) - \psi(x)_+ + d\psi(x; h(x)) \le \theta(x) - \tfrac{1}{2}\delta\|h(x)\|^2 \le \theta(x) . \tag{9c}$$

(d) For all $x \in \mathbb{R}^n$,

$$-\gamma\psi(x)_+ + df^0(x; h(x)) \le \theta(x) - \tfrac{1}{2}\delta\|h(x)\|^2 \le \theta(x) . \tag{9d}$$

(e) Alternative expressions for $\theta(x)$ *and* $h(x)$ *are given by*

$$\theta(x) = -\min_{\substack{\mu \in \Sigma_q^0 \\ \nu \in \Sigma_p}} \Big\{ \mu^0 \sum_{k=1}^{p} \nu^k [f^0(x) - c^k(x) + \gamma\psi(x)_+] + \sum_{j=1}^{q} \mu^j [\psi(x)_+ - f^j(x)] + \frac{1}{2\delta}\|\mu^0 \sum_{k=1}^{p} \nu^k \nabla c^k(x) + \sum_{j=1}^{q} \mu^j \nabla f^j(x)\|^2 \Big\} \tag{9e}$$

and

$$h(x) = -\frac{1}{\delta}\Big[\mu_x^0 \sum_{k=1}^{p} \nu_x^k \nabla c^k(x) + \sum_{j=1}^{q} \mu_x^j \nabla f^j(x)\Big], \tag{9f}$$

where (μ_x, ν_x) *is any solution of (9e);*

equivalently, $\theta(x)$ *and* $h(x)$ *can be expressed in the form*

$$\theta(x) = -\min_{\bar{\xi} \in \bar{GF}(x)} \{\xi^0 + \frac{1}{2\delta}\|\xi\|^2\}, \tag{9g}$$

and

$$\bar{h}(x) = (h^0(x), h(x)) = -\arg\min_{\bar{\xi} \in \bar{GF}(x)} \{\xi^0 + \frac{1}{2\delta}\|\xi\|^2\}, \tag{9h}$$

where $\bar{GF}(x) \subset \mathbb{R}^{n+1}$ *has elements denoted by* $\bar{\xi} = (\xi^0, \xi)$, *with* $\xi^0 \in \mathbb{R}$, $\xi \in \mathbb{R}^n$, *and is defined by*

$$\bar{GF}(x) \triangleq \operatorname*{co}_{\substack{k \in \mathbf{p} \\ j \in \mathbf{q}}} \left\{ \begin{bmatrix} f^0(x) - c^k(x) + \gamma\psi(x)_+ \\ \nabla c^k(x) \end{bmatrix}, \begin{bmatrix} \psi(x)_+ - f^j(x) \\ \nabla f^j(x) \end{bmatrix} \right\}. \tag{9i}$$

(f) For any $\hat{x} \in \mathbb{R}^n$ *such that* $\psi(\hat{x}) \le 0$, *there exist* $\hat{\nu} \in \Sigma_p$, $\hat{\mu} \in \Sigma_q^0$ *such that (5a) and (5b) hold if and only if* $\theta(\hat{x}) = 0$. *More generally, for any* $\hat{x} \in \mathbb{R}^n$, $0 \in \partial_2 F(\hat{x}, \hat{x})$ *if and only if* $\theta(\hat{x}) = 0$.

(g) Both $\theta(\cdot)$ *and* $h(\cdot)$ *are continuous.* □

Exercise 2.2.9. *(a)* Verify that Theorem 2.2.8 is correctly derived.

(b) Consider the search direction $h(x)$ defined by (9f). Show that

(i) when $x \in \mathbb{R}^n$ is such that $\psi(x) \ge 0$, then $h(x)$ is a descent direction for $\psi(\cdot)$ at x, and

(ii) when $x \in \mathbb{R}^n$ is such that $\psi(x) \le 0$, then $h(x)$ is a descent direction for $f^0(\cdot)$ at x with the property that one can make a small displacement along $h(x)$ without violating the constraints in (2a,b), i.e., for some $\lambda_x > 0$, $\psi(x + \lambda h(x)) \le 0$ for all $\lambda \in [0, \lambda_x]$. □

2.2.3 Second-Order Conditions for ICP

Next, referring to Theorem 2.2.2, it is clear that we can use Theorem 2.1.12 to obtain a second-order necessary condition of optimality for the problem **ICP** in (2a,b). First we note that, according to (2.1.16), given a local minimizer $\hat{x}$ for **ICP**, we must associate, with the function $\hat{F}(\cdot)$ defined in (3), the Lagrangian

$$L(x, \mu, \nu) \triangleq \mu^0\Big(\sum_{k=1}^{p} \nu^k [c^k(x) - f^0(\hat{x})]\Big) + \sum_{j=1}^{q} \mu^j f^j(x), \tag{10}$$

where $\mu \in \Sigma_q^0$ and $\nu \in \Sigma_p$. Next we note that the constant term $f^0(\hat{x})$ cannot show up in the inequality (2.1.17c). Hence we may as well drop it from the definition of $L(x, \mu, \nu)$ and replace the above definition by

$$L(x, \mu, \nu) \triangleq \mu^0\Big[\sum_{k=1}^{p} \nu^k c^k(x)\Big] + \sum_{j=1}^{q} \mu^j f^j(x) . \tag{11}$$

Next, for any $x \in \mathbb{R}^n$, let

$$\mathbf{p}_{\hat{F}}(x) \triangleq \{k \in \mathbf{p} \mid c^k(x) - f^0(\hat{x}) = \hat{F}(x)\}, \tag{12a}$$

and

$$\mathbf{q}_{\hat{F}}(x) \triangleq \{j \in \mathbf{q} \mid f^j(x) = \hat{F}(x)\}. \tag{12b}$$

Then we see that the maximizing index set for the function $\hat{F}(x)$ is the union of the sets $\mathbf{p}_{\hat{F}}(x)$ and $\mathbf{q}_{\hat{F}}(x)$. Furthermore, at $x = \hat{x}$, a local minimizer for **ICP** in (2a,b), because $\hat{F}(\hat{x}) = 0$, $\mathbf{p}_{\hat{F}}(\hat{x}) = \hat{\mathbf{p}}(\hat{x})$ and $\mathbf{q}_{\hat{F}}(\hat{x}) = \mathbf{q}_A(\hat{x})$, where, in our usual notation,

$$\hat{\mathbf{p}}(\hat{x}) \triangleq \{k \in \mathbf{p} \mid c^k(\hat{x}) = f^0(\hat{x})\}, \tag{12c}$$

$$\mathbf{q}_A(\hat{x}) \triangleq \{ j \in \mathbf{q} \mid f^j(\hat{x}) \geq 0 \}. \tag{12d}$$

Obviously, $\hat{\mathbf{p}}(\hat{x})$ must contain at least one index; however, $\mathbf{q}_A(\hat{x})$ may be empty[†].

We now see that in terms of the above Lagrangian and these index sets, Theorem 2.1.12 assumes the following form for problem **ICP** in (2a,b):

Theorem 2.2.10. *Consider the problem* **ICP** *in (2a,b). Suppose that the functions $f^j(\cdot)$, $j \in \mathbf{q}$, in (2a) and the functions $c^k(\cdot)$, $k \in \mathbf{p}$, in (2b) are twice continuously differentiable, that $\hat{x}$ is a local minimizer for* **ICP**, *that $(\hat{\mu}, \hat{\nu})$ is an associated multiplier that satisfies (5a,b), and that the vectors $\nabla c^k(\hat{x})$, $j \in \hat{\mathbf{p}}(\hat{x})$, together with the vectors $\nabla f^j(\hat{x})$, $j \in \mathbf{q}_A(\hat{x})$, are affinely independent. Let*

$$\mathcal{H}_I(\hat{x}) \triangleq \{ h \in \mathbb{R}^n \mid \langle \nabla f^j(\hat{x}), h \rangle = 0,\ j \in \mathbf{q}_A(\hat{x}),$$

$$\langle \nabla c^k(\hat{x}), h \rangle = 0,\ k \in \hat{\mathbf{p}}(\hat{x}) \}. \tag{13a}$$

Then, for all $h \in \mathcal{H}_I(\hat{x})$,

$$df^0(\hat{x}; h) = 0, \tag{13b}$$

and

$$\langle h, L_{xx}(\hat{x}, \hat{\mu}, \hat{\nu}) h \rangle \geq 0. \tag{13c}$$

□

An important special case of Theorem 2.2.10 arises when $p = 1$.

Exercise 2.2.11. Use Proposition 2.1.11 to prove the following result:

Corollary 2.2.12. *Consider the problem* **ICP** *in (2a). Suppose that the functions $f^0(\cdot)$ and $f^j(\cdot)$, $j \in \mathbf{q}$, in (2a) are twice continuously differentiable, that $\hat{x}$ is a local minimizer for* **ICP**, *that $\hat{\mu} \in \Sigma_q^0$ is an associated multiplier (i.e., it satisfies (7a,b)), and that the vectors $\nabla f^j(\hat{x})$, $j \in \mathbf{q}_A(\hat{x})$, are linearly independent. Let*

$$\mathcal{H}_I(\hat{x}) \triangleq \{ h \in \mathbb{R}^n \mid \langle \nabla f^j(\hat{x}), h \rangle = 0,\ j \in \mathbf{q}_A(\hat{x}) \}. \tag{14a}$$

Then, for all $h \in \mathcal{H}_I(\hat{x})$,

[†] Although the case where $f^j(\hat{x}) > 0$ is not relevant at this point, it is convenient do define $\mathbf{q}_A(\hat{x})$ as in (12d) to make it useful in subsequent situations as well.

$$df^0(\hat{x}; h) = 0, \tag{14b}$$

and, with the Lagrangian defined by

$$L(x, \mu) \triangleq \textstyle\sum_{j=0}^{q} \mu^j f^j(x), \tag{14c}$$

$$\langle h, L_{xx}(\hat{x}, \hat{\mu}) h \rangle \geq 0. \tag{14d}$$

□

Note that the term $\langle \nabla f^0(\hat{x}), h \rangle = 0$ does not appear in (14a), contrary to what one would expect from an examination of (13a). The reason for this is that it is made redundant by (7a), since $\hat{\mu}^0 > 0$ must hold, because the vectors $\nabla f^j(\hat{x})$, $j \in \mathbf{q}_A(\hat{x})$, are linearly independent by assumption.

We can interpret Theorem 2.1.13 to obtain a second-order sufficient condition for a point $\hat{x}$ to be a local minimizer for the problem $\min_{x \in \mathbb{R}^n} \hat{F}(x)$. However, as we have shown, in general, a local minimizer of $\hat{F}(\cdot)$ need not be a local minimizer of **ICP** in (2a,b). Hence, to obtain a sufficient second-order condition for **ICP** from that for the problem $\min_{x \in \mathbb{R}^n} \hat{F}(x)$, as done below, we need to include the condition $0 \notin \partial\psi(\hat{x})$ if $\psi(\hat{x}) = 0$, in Theorem 2.2.3, which ensures that, if $\hat{x}$ is a local minimizer of $\hat{F}(\cdot)$, then it is also a local minimizer for **ICP**. Below, we use the notation $\hat{\mathbf{p}} \setminus \hat{\mathbf{p}}_+$ to denote the set $\{ k \in \hat{\mathbf{p}} \mid k \notin \hat{\mathbf{p}}_+ \}$, etc.

Theorem 2.2.13. *Consider the problem* **ICP** *in (2a,b). Suppose that the functions $f^j(\cdot)$, $j \in \mathbf{q}$, in (2a) and the functions $c^k(\cdot)$, $k \in \mathbf{p}$, in (2b) are twice continuously differentiable, that $\hat{x} \in \mathbb{R}^n$ is such that $\psi(\hat{x}) \leq 0$, and that $0 \notin \partial\psi(\hat{x})$ if $\psi(\hat{x}) = 0$. Furthermore, suppose that $\hat{x}$ together with $\hat{\mu} \in \Sigma_q^0$ and $\hat{\nu} \in \Sigma_p$ satisfy the first-order optimality conditions (5a) and (5b). Let*

$$\mathbf{q}_{A+}(\hat{x}, \hat{\mu}, \hat{\nu}) \triangleq \{ j \in \mathbf{q}_A(\hat{x}) \mid \hat{\mu}^j > 0 \}, \tag{15a}$$

$$\hat{\mathbf{p}}_+(\hat{x}, \hat{\mu}, \hat{\nu}) \triangleq \{ k \in \hat{\mathbf{p}}(\hat{x}) \mid \hat{\nu}^k > 0 \}, \tag{15b}$$

$$\mathcal{H}'_I(\hat{x}) \triangleq \{ h \in \mathbb{R}^n \mid \langle \nabla c^k(\hat{x}), h \rangle = 0,\ k \in \hat{\mathbf{p}}_+(\hat{x}, \hat{\mu}, \hat{\nu}),$$

$$\langle \nabla c^k(\hat{x}), h \rangle \leq 0,\ k \in \hat{\mathbf{p}}(\hat{x}) \setminus \hat{\mathbf{p}}_+(\hat{x}, \hat{\mu}, \hat{\nu}),$$

$$\langle \nabla f^j(\hat{x}), h \rangle = 0,\ j \in \mathbf{q}_{A+}(\hat{x}, \hat{\mu}, \hat{\nu}),$$

$$\langle \nabla f^j(\hat{x}), h \rangle \leq 0,\ j \in \hat{\mathbf{q}}(\hat{x}) \setminus \mathbf{q}_{A+}(\hat{x}, \hat{\mu}, \hat{\nu}) \}. \tag{15c}$$

Let the Lagrangian $L(x, \mu, \nu)$ be defined as in (11). If there exists an $m > 0$ such that

$$\langle h, L_{xx}(\hat{x}, \hat{\mu}, \hat{\nu}) h \rangle \geq m \|h\|^2,\ \forall h \in \mathcal{H}_I'(\hat{x}), \tag{15d}$$

then $\hat{x}$ is a strict local minimizer for **ICP**.

Proof. First we note that (5a) and (5b), together with (15a) - (15d) constitute a sufficient condition for $\hat{x}$ to be a local minimizer of $\hat{F}(\cdot)$. Next, because $0 \notin \partial\psi(\hat{x})$, by assumption, we conclude from Theorem 2.2.3 that $\hat{x}$ is a local minimizer for **ICP**. □

The expression (15c) is very clear, but notationally cumbersome. The following equivalent expression may be preferred:

$$\mathcal{H}_I'(\hat{x}) \triangleq \{ h \in \mathbb{R}^n \mid$$

$$\hat{\nu}^k \langle \nabla c^k(\hat{x}), h \rangle = 0,\ (1 - \hat{\nu}^k) \langle \nabla c^k(\hat{x}), h \rangle \leq 0,\ k \in \hat{\mathbf{p}}(\hat{x}),$$

$$\hat{\mu}^j \langle \nabla f^j(\hat{x}), h \rangle = 0,\ (1 - \hat{\mu}^j) \langle \nabla f^j(\hat{x}), h \rangle \leq 0,\ j \in \mathbf{q}_A(\hat{x}) \}. \tag{16}$$

Again we get an important special case when $p = 1$, so that $c^1(\cdot) = f^0(\cdot)$, and the Lagrangian assumes the simplified form $L(x, \mu) = \sum_{j=0}^{q} \mu^j f^j(x)$, because in this case $\nu^1 = 1$.

Corollary 2.2.14. *Consider the problem* **ICP** *in (2a). Suppose that the functions $f^0(\cdot)$ and $f^j(\cdot)$, $j \in \mathbf{q}$, in (2a) are twice continuously differentiable, that $\hat{x} \in \mathbb{R}^n$ is such that $\psi(\hat{x}) \leq 0$, and $0 \notin \partial\psi(\hat{x})$ if $\psi(\hat{x}) = 0$. Furthermore, suppose that $\hat{x}$ together with $\hat{\mu} \in \Sigma_q^0$ satisfy the first-order conditions (7a,b). Let*

$$\mathcal{H}_I'(\hat{x}) \triangleq \{ h \in \mathbb{R}^n \mid \langle \nabla f^j(\hat{x}), h \rangle = 0,\ j \in \mathbf{q}_{A+}(\hat{x}, \hat{\mu}, \hat{\nu}),$$

$$\langle \nabla f^j(\hat{x}), h \rangle \leq 0,\ j \in \mathbf{q}_A(\hat{x}) \setminus \mathbf{q}_{A+}(\hat{x}, \hat{\mu}, \hat{\nu}) \}. \tag{17a}$$

Let the Lagrangian be defined by

$$L(x, \mu) \triangleq \sum_{j=0}^{q} \mu^j f^j(x). \tag{17b}$$

If there exists an $m > 0$ such that

$$\langle h, L_{xx}(\hat{x}, \hat{\mu}) h \rangle \geq m \|h\|^2,\ \forall h \in \mathcal{H}_I'(\hat{x}), \tag{17c}$$

then $\hat{x}$ is a strict local minimizer for (2). □

Note that the assumptions in Corollary 2.2.14 ensure that $\hat{\mu}^0 > 0$ and, hence, that the condition $\langle \nabla f^0(\hat{x}), h \rangle = 0$ that appears to be missing from (17a) is, in fact, implied by the equations in (17a) and (7a).

2.2.4 First-Order Optimality Conditions for IECP

Finally, we turn to the problem

$$\textbf{IECP} \qquad \min \{ f^0(x) \mid f^j(x) \leq 0,\ j \in \mathbf{q},\ g^l(x) = 0,\ l \in \mathbf{r} \}, \tag{18}$$

where we will assume that $f^0(\cdot)$ is defined as in (2b) and that the functions $f^j, c^k, g^l : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, $k \in \mathbf{p}$, $l \in \mathbf{r}$, are at least once continuously differentiable. The functions $g^l(\cdot)$ obviously define a function $g : \mathbb{R}^n \to \mathbb{R}^r$, with $g(x) \triangleq (g^1(x), g^2(x), \ldots, g^r(x))$. To simplify matters, we will assume that the matrix $g_x(x) \triangleq \partial g(x)/\partial x$ is of maximum row rank at all the local minimizers of **IECP**.

Definition 2.2.15. *Let $X_{IE} \triangleq \{ x \in \mathbb{R}^n \mid \psi(x) \leq 0, g(x) = 0 \}$. We will say that $\hat{x} \in X_{IE}$ is a local minimizer for* **IECP** *if there exists a $\rho > 0$ such that $f^0(x) \geq f^0(\hat{x})$ for all $x \in X_{IE} \cap B(\hat{x}, \rho)$. We will say that $\hat{x}$ is a* strict local minimizer *if $f^0(x) > f^0(\hat{x})$ for all $x \in X_{IE} \cap B(\hat{x}, \rho)$.* □

To obtain optimality conditions for **IECP**, we will transcribe it into a simpler, locally equivalent problem, following the pattern set for the problem **ICP**. However, to avoid the introduction of concepts from differential geometry, we will state our assumptions in terms of sequences, rather than in the more elegant, but equivalent terms of relative interiors.

Theorem 2.2.16. *(a) Suppose that $\hat{x}$ is a local minimizer for the problem* **IECP** *in (18), then $\hat{x}$ is a local minimizer for the problem*

$$\min \{ \hat{F}(x) \mid g(x) = 0 \}, \tag{19}$$

where $\hat{F}(\cdot)$ is defined as in (3).

(b) Suppose that, for any $x \in \mathbb{R}^n$ such that both $\psi(x) = 0$ and $g(x) = 0$, there exists a sequence of vectors $\{x_i\}_{i=0}^{\infty}$ converging to x, such that $\psi(x_i) < 0$ and $g(x_i) = 0$ for all $i \in \mathbb{N}$. If $\hat{x}$ is a local minimizer for (19) such that $\psi(\hat{x}) \leq 0$, then $\hat{x}$ is a local minimizer for (18). □

Exercise 2.2.17. Prove Theorem 2.2.17. *Hint*: Mimic the proof of Theorem 2.2.3. □

Next, we will need the following technical result.

Proposition 2.2.18. *Suppose that C is a convex, compact set in $\mathbb{R}^n$ and that H is a subspace of $\mathbb{R}^n$. Then*

$$\max_{\xi \in C} \langle \xi, h \rangle \geq 0, \quad \forall\, h \in H \tag{20a}$$

if and only if

$$C \cap H^{\perp} \neq \emptyset, \tag{20b}$$

where $H^{\perp}$ is the orthogonal complement of H in $\mathbb{R}^n$.

Proof. $\Rightarrow$ We give a proof by contraposition. Suppose that $C \cap H^{\perp} = \emptyset$. Then C can be separated strictly from $H^{\perp}$, i.e., there exists an $h \neq 0$ such that $\langle v, h \rangle = 0$ for all $v \in H^{\perp}$ (i.e., $h \in H$), and $\langle \xi, h \rangle < 0$ for all $\xi \in C$, which shows that (20a) does not hold.

$\Leftarrow$ Suppose that there is a vector $g \in C \cap H^{\perp}$. Then, for any $h \in H$, $\max_{\xi \in C} \langle \xi, h \rangle \geq \langle g, h \rangle = 0$, i.e., (20a) holds. □

Theorem 2.2.19. *Consider problem* **IECP** *in (18). Suppose that $f^0(\cdot)$ is defined as in (2b), and that the functions $c^k : \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbf{p}$, in (2b) and the functions $f^j, g^l : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, $l \in \mathbf{r}$, in (18) are at least once continuously differentiable. If $\hat{x}$ is a local minimizer for* **IECP** *and the vectors $\nabla g^l(\hat{x})$, $l \in \mathbf{r}$, are linearly independent, then*

(a)
$$d\hat{F}(\hat{x}; h) \geq 0, \quad \forall\, h \in \mathcal{H}_E(\hat{x}), \tag{21a}$$

where

$$\mathcal{H}_E(\hat{x}) \triangleq \{h \in \mathbb{R}^n \mid g_x(\hat{x})h = 0\}; \tag{21b}$$

(b) there exist multipliers $\hat{\mu} \in \Sigma_q^0$, $\hat{\nu} \in \Sigma_p$, and $\hat{\zeta} \in \mathbb{R}^r$ such that

$$\hat{\mu}^0\Big[\sum_{k=1}^{p} \hat{\nu}^k \nabla c^k(\hat{x})\Big] + \sum_{j=1}^{q} \hat{\mu}^j \nabla f^j(\hat{x}) + \sum_{l=1}^{r} \hat{\zeta}^l \nabla g^l(\hat{x}) = 0, \tag{21c}$$

and

$$\sum_{k=1}^{p} \hat{\nu}^k [c^k(\hat{x}) - f^0(\hat{x})] + \sum_{j=1}^{q} \hat{\mu}^j f^j(\hat{x}) = 0. \tag{21d}$$

Proof. *(a)* Since $\hat{x}$ is a local minimizer for **IECP**, it must also be a local minimizer for the problem in (19). Hence, for the sake of contradiction, suppose that there is a vector $h \in \mathcal{H}_E(\hat{x})$ such that $d\hat{F}(\hat{x}; h) < 0$. Since the matrix $g_x(\hat{x})$ is of maximum row rank, it follows from Corollary 5.1.34 that there exists a $t_h > 0$ and a continuously differentiable function $s : [0, t_h] \to \mathbb{R}^n$ such that $s(0) = \hat{x}$, $\dot{s}(0) = h$, and $g(s(t)) = 0$ for all $t \in [0, t_h]$. Let $\sigma : [0, t_h] \to \mathbb{R}$ be defined by $\sigma(t) = \hat{F}(s(t))$. Then, by Theorem 5.4.12, the directional derivative $d\sigma(0; 1) = d\hat{F}(\hat{x}; h) < 0$, and hence there exists a $t' \in (0, t_h]$ such that $\sigma(t) < \sigma(0)$ for all $t \in (0, t')$. Consequently, for any $t \in (0, t')$, $g(s(t)) = 0$ and $\hat{F}(s(t)) < \hat{F}(\hat{x}) = 0$, which contradicts the fact that $\hat{x}$ is a local minimizer for the problem (19).

(b) Now, by Corollary 5.4.6, *(i)* for any $h \in \mathbb{R}^n$,

$$d\hat{F}(\hat{x}; h) = \max_{\xi \in \partial \hat{F}(\hat{x})} \langle \xi, h \rangle, \tag{21e}$$

and *(ii)*, using the definitions (12c) and (12d),

$$\partial \hat{F}(\hat{x}) = \operatorname*{co}_{\substack{j \in \mathbf{q}_A(\hat{x}) \\ k \in \hat{\mathbf{p}}(\hat{x})}} \{\nabla f^j(\hat{x}), \nabla c^k(\hat{x})\}. \tag{21f}$$

It now follows from (21a), (21b), and Proposition 2.2.18 that

$$\partial \hat{F}(\hat{x}) \cap \mathcal{H}_E^{\perp}(\hat{x}) \neq \emptyset. \tag{21g}$$

Since the vectors $\nabla g^l(\hat{x})$, $l \in \mathbf{r}$, form a basis for $\mathcal{H}_E(\hat{x})^{\perp}$, it follows that there exist $\hat{\mu} \in \Sigma_q^0$, $\hat{\nu} \in \Sigma_p$ and $\hat{\zeta} \in \mathbb{R}^r$ such that (21c) and (21d) hold, with (21d) ensuring that only the vectors appearing in the expression (21f) for $\partial \hat{F}(\hat{x})$ have nonzero coefficients in (21c). □

The assumption in Theorem 2.2.19, that the vectors $\nabla g^l(\hat{x})$, $l \in \mathbf{r}$, are linearly independent, eliminates certain degeneracies. It is convenient, but not essential, in deriving optimality conditions, as we can see from the following generalization of Theorem 2.2.19.

Theorem 2.2.19a. *Consider problem* **IECP** *in (18). Suppose that $f^0(\cdot)$ is defined as in (2b), and that the functions c^k, $k \in \mathbf{p}$, in (2b), $f^j, g^l : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, $l \in \mathbf{r}$, in (18) are at least once continuously differentiable. Then there exist multipliers $\hat{\mu}^0 \geq 0$, $\hat{\mu} \in \mathbb{R}^q$, with $\hat{\mu} \geq 0$, $\hat{\nu} \in \Sigma_p$, and $\hat{\zeta} \in \mathbb{R}^r$, with $\hat{\mu}^0$, $\hat{\mu}$, and $\hat{\zeta}$, not all zero, such that*

$$\hat{\mu}^0\Big[\sum_{k=1}^{p} \hat{\nu}^k \nabla c^k(\hat{x})\Big] + \sum_{j=1}^{q} \hat{\mu}^j \nabla f^j(\hat{x}) + \sum_{l=1}^{r} \hat{\zeta}^l \nabla g^l(\hat{x}) = 0, \tag{22a}$$

and

$$\sum_{k=1}^{p} \hat{\nu}^k [c^k(\hat{x}) - f^0(\hat{x})] + \sum_{j=1}^{q} \hat{\mu}^j f^j(\hat{x}) = 0. \tag{22b}$$

Proof. Let $\hat{\rho} > 0$ be the radius associated with $\hat{x}$. First, consider the problem

IECP′ $$\min \{ f^0(x) + \|x - \hat{x}\|^2 \mid f(x) \le 0,\ g(x) = 0 \}, \tag{22c}$$

where $f(x) = (f^1(x), \ldots, f^q(x))$ and $g(x) = (g^1(x), \ldots, g^k(x))$. Clearly, $\hat{x}$ is also a local minimizer for **IECP′** and, in addition, for all $x \neq \hat{x}$ in $B(\hat{x}, \hat{\rho})$ that are feasible for **IECP′**,

$$f^0(x) + \|x - \hat{x}\|^2 > f^0(\hat{x}), \tag{22d}$$

i.e., $\hat{x}$ is the *only* local minimizer of **IECP′** in the ball $B(\hat{x}, \hat{\rho})$.

Now, for $i \in \mathbb{N}$, consider the family of inequality constrained problems

$\mathbf{P}_i$ $$\min \{ f^0(x) + \|x - \hat{x}\|^2 \mid f(x) \le 0, \|g(x)\|_\infty \le \varepsilon_i,\ \|x - \hat{x}\|^2 \le \hat{\rho} \}, \tag{22e}$$

with $\varepsilon_i > 0$, for all i and $\varepsilon_i \to 0$, as $i \to \infty$. Let x_i denote the solution of $\mathbf{P}_i$. Then, for every $\varepsilon \ge 0$, $\hat{x}$ is feasible for the problem $\mathbf{P}_i$. Since the sequence $\{x_i\}_{i \in \mathbb{N}} \subset B(\hat{x}, \hat{\rho})$ is bounded, it must have at least one accumulation point, say x^*, with $x^* \in B(\hat{x}, \hat{\rho})$. Clearly, $f(x^*) \le 0$ and $g(x^*) = 0$. Furthermore, since $f^0(x_i) \le f^0(\hat{x})$ must hold for all $i \in \mathbb{N}$,

$$f^0(x^*) \le f^0(\hat{x}), \tag{22f}$$

and hence x^* is a local minimizer for **IECP′**. Since $\hat{x}$ is the only local minimizer of **IECP′** in $B(\hat{x}, \hat{\rho})$, we must have that $x^* = \hat{x}$.

Now, since x_i is a local minimizer of $\mathbf{P}_i$, it follows from Theorem 2.2.4 that there exist multiplier vectors $\nu_i \in \Sigma_p$ and $\bar{\mu}_i \in \Sigma^0_{q+2r}$, with $\bar{\mu}_i = (\mu_i^0, \mu_i, \zeta_{i+}, \zeta_{i-})$, such that $\mu_i^0 \in \mathbb{R}$, $\mu_i \in \mathbb{R}^m$, $\zeta_{i+}, \zeta_{i-} \in \mathbb{R}^r$,† and, with $c(x) = (c^1(x), \ldots, c^p(x))$ and $\mathbf{1}_p = (1, \ldots, 1) \in \mathbb{R}^p$,

$$\mu_i^0 [c_x(x_i)^T \nu_i + 2(x_i - \hat{x})] + f_x(x_i)^T \mu_i + g_x(x_i)^T \zeta_{i+} - g_x(x_i)^T \zeta_{i-} = 0, \tag{22g}$$

$$\langle \nu_i, [c(x_i) - f^0(x_i)\mathbf{1}_p] \rangle + \langle \mu_i, f(x_i) \rangle + \langle \zeta_{i+}, [g(x_i) - \varepsilon_i \mathbf{1}_r] \rangle + \langle \zeta_{i-}, [-g(x_i) - \varepsilon_i \mathbf{1}_r] \rangle = 0. \tag{22h}$$

Since Σ_p and Σ^0_{q+2r} are both compact, there must be an infinite subset $K \subset \mathbb{N}$ such that $\nu_i \xrightarrow{K} \hat{\nu} \in \Sigma_p$ and $\bar{\mu}_i \xrightarrow{K} \bar{\mu} \in \Sigma^0_{q+2r}$, as $i \to \infty$, with $\bar{\mu} = (\hat{\mu}^0, \hat{\mu}, \hat{\zeta}_+, \hat{\zeta}_-)$. In view of (22g,h) and the fact that $g(\hat{x}) = 0$ and $\lim \varepsilon_i = 0$, the multipliers $(\hat{\nu}, \bar{\mu})$ satisfy the following relations

† Note that the multiplier associated with the constraint $\|x - \hat{x}\|^2 \le \hat{\rho}$ can be assumed to be zero because this inequality is slack for all i sufficiently large.

$$\hat{\mu}^0 c_x(\hat{x})^T \hat{\nu} + f_x(\hat{x})^T \hat{\mu} + g_x(\hat{x})^T \hat{\zeta} = 0, \tag{22i}$$

where $\hat{\zeta} = \hat{\zeta}_+ - \hat{\zeta}_-$, and

$$\langle \hat{\nu}, [c(\hat{x}) - f^0(\hat{x})\mathbf{1}_p] \rangle + \langle \hat{\mu}, f(\hat{x}) \rangle = 0, \tag{22j}$$

which are a compact form of (22a,b)

It remains to show that not all the multipliers $\hat{\mu}^0$, $\hat{\mu}$, and $\hat{\zeta}$ are zero in (22h). For the sake of contradiction, suppose that all the coefficients $\hat{\mu}^0$, $\hat{\mu}$, and $\hat{\zeta}$ are zero in (22h). Since $\bar{\mu} \in \Sigma^0_{q+2r}$, it follows that $(\hat{\zeta}_+, \hat{\zeta}_-) \in \Sigma_{2r}$. Next, since $\hat{\zeta} = 0$, $\hat{\zeta}_+ - \hat{\zeta}_- = 0$ and hence $\hat{\zeta}_+ = \hat{\zeta}_-$. However, since for all $i \in K$, either ζ_{i+}^l or ζ_{i-}^l is zero, or both, it follows that for all $l \in \mathbf{r}$, either $\hat{\zeta}_+^l = 0$ or $\hat{\zeta}_-^l = 0$ and hence, since $\hat{\zeta}_+ = \hat{\zeta}_-$, that $\hat{\zeta}_+ = \hat{\zeta}_- = 0$, which contradicts the fact that $(\hat{\zeta}_+, \hat{\zeta}_-) \in \Sigma_{2r}$, completing our proof. □

Note that when the matrices $f_x(\hat{x})$ and $g_x(\hat{x})$ fail to satisfy the Robinson PLI condition, (21c,d) can be satisfied with $\hat{\mu}^0 = 0$. When the possibility exists that $\hat{\mu}^0 = 0$ in (21c,d) or in (22a,b), these equations are referred to as *generalized Fritz John* conditions. In that case, the multipliers $\hat{\mu}^j$ in (21c,d) and (22a,b) are referred to as *Fritz John* multipliers. When (21c,d) (or (22a,b)) hold with $\hat{\mu}^0 > 0$, (21c,d) (or (22a,b)) are referred to as *generalized Karush-Kuhn-Tucker* conditions, and the normalized multipliers, $\hat{\eta}^j \triangleq \hat{\mu}^j / \hat{\mu}^0$, $j \in \mathbf{q}$, *and* $\hat{\xi}^j \triangleq \hat{\zeta}^l / \hat{\mu}^0$, $l \in \mathbf{r}$, are called *Karush-Kuhn-Tucker (KKT) multipliers*. When there are no inequality constraints in the problem **IECP**, the multipliers $\hat{\xi}^l$, $l \in \mathbf{r}$, are called *Lagrange multipliers*, and (21c,d) are called *generalized Lagrange conditions*. The qualification *generalized* is removed when the cost function $f^0(\cdot)$ is differentiable.

Note that when $\hat{\mu}^0 = 0$, the above optimality conditions are degenerate in the sense that they contain no information supplied by the cost function.

To deal with an optimization problem that has only equality constraints, define $f^j(x) \equiv 0$, for all $j \in \mathbf{q}$, in (18). Then, in (21c), we must set $\hat{\mu}^0 = 1$ and $\hat{\mu}^j = 0$ for all $j \in \mathbf{q}$.

Finally, there are two special cases of Theorem 2.2.19a that are worth stating separately. These correspond to $p = 1$ and hence $\hat{\nu}^1 = 1$ in (22a,b).

Corollary 2.2.20. *(a) Consider the problem*

ECP $$\min \{ f^0(x) \mid g(x) = 0 \}, \tag{23a}$$

with $f^0 : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^l$ continuously differentiable. If $\hat{x}$ is a local minimizer for **ECP** *in (23a), then there exists a multiplier $(\hat{\mu}, \hat{\zeta}) \in \mathbb{R} \times \mathbb{R}^l$ such that $\hat{\mu}^0 \geq 0$ and*

$$\hat{\mu}^0 \nabla f^0(\hat{x}) + \sum_{l=1}^{r} \nabla g^l(\hat{x}) \hat{\zeta}^l = 0 . \tag{23b}$$

Furthermore, when $g_x(\hat{x})$ has maximum row rank, $\hat{\mu}$ can be taken to be 1.†

(b) Consider the problem

$$\min \{ f^0(x) \mid f^j(x) \leq 0, \ j \in \mathbf{q}, \ g(x) = 0 \}, \tag{23c}$$

with $f^j : \mathbb{R}^n \to \mathbb{R}$, $j = 0, 1, \ldots, q$, and $g : \mathbb{R}^n \to \mathbb{R}^l$ continuously differentiable. If $\hat{x}$ is a local minimizer for (23c), then there exist multipliers $\hat{\mu}^0 \geq 0$, $\hat{\mu} \in \mathbb{R}^q$, with $\hat{\mu} \geq 0$, and $\hat{\zeta} \in \mathbb{R}^n$ such that

$$\sum_{j=0}^{q} \hat{\mu}^j \nabla f^j(\hat{x}) + \sum_{l=1}^{r} \hat{\zeta}^l \nabla g^l(\hat{x}) = 0 \tag{23d}$$

and

$$\sum_{j=1}^{q} \hat{\mu}^j f^j(\hat{x}) = 0 . \tag{23e}$$

Furthermore, when $g_x(\hat{x})$ has maximum rank, $(\hat{\mu}^0, \hat{\mu}) \in \Sigma_q^0$, and when $f_x(\hat{x})$ and $g_x(\hat{x})$ satisfy the Robinson PLI condition, $\hat{\mu}^0 > 0$. □

We can also get a first-order sufficient condition, analogous to Corollary 2.2.6, for the case where the cost function and inequality constraint functions are convex and the equality constraint function is *affine*, i.e., there exist vectors $a_j \in \mathbb{R}^n$ and scalars b_j such that $g^j(x) = \langle a_j, x \rangle + b_j$ for all $x \in \mathbb{R}^n$ and $j \in \mathbf{r}$.

Corollary 2.2.21. *Consider problem* **IECP** *in (18). Suppose that $f^0(\cdot)$ is defined as in (2b), that the functions f^j, c^k, $j \in \mathbf{q}$, $k \in \mathbf{p}$, in (18) and (2b), respectively, are all convex and at least once continuously differentiable, that the functions $g^l : \mathbb{R}^n \to \mathbb{R}$, $l \in \mathbf{r}$, in (18) are all affine, and that the matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$. Furthermore, suppose that, for any $x \in \mathbb{R}^n$ such that both $\psi(x) = 0$ and $g(x) = 0$, there exists a sequence of vectors $\{x_i\}_{i=0}^{\infty}$ converging to x, such that $\psi(x_i) < 0$ and $g(x_i) = 0$ for all $i \in \mathbb{N}$. If $\hat{x} \in \mathbb{R}^n$ is such that $\psi(\hat{x}) \leq 0$, $g(\hat{x}) = 0$, and (21c,d) are satisfied, then $\hat{x}$ is a local (and hence also a global) minimizer for the problem* **IECP** *in (18).* □

† When $\hat{\mu}^0 = 1$, equation (23b) is usually referred to as the Lagrange condition, and $\hat{\zeta}$ is called a Lagrange multiplier.

Exercise 2.2.22. Prove Corollary 2.2.21. *Hint*: Make use of Theorem 2.2.16 and of Theorem 2.1.5. □

The following result will be needed in showing that the optimality function to be proposed for the problem **IECP** is continuous.

Exercise 2.2.23. Suppose that $g : \mathbb{R}^n \to \mathbb{R}^r$, with $r < n$, is continuously differentiable and that the matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$. Let the set valued function $\mathcal{H}_E(\cdot)$ be defined by

$$\mathcal{H}_E(x) \triangleq \{ h \in \mathbb{R}^n \mid g_x(x) h = 0 \}. \tag{24}$$

Show that $\mathcal{H}_E(\cdot)$ is continuous in the sense of Definition 5.3.3. □

Theorem 2.2.24. *Consider problem* **IECP** *in (18), with $f^0(\cdot)$ defined as in (2b). Suppose that the functions $c^k(\cdot)$, $k \in \mathbf{p}$, in (2b), $f^j(\cdot)$, $j = 0, \ldots, q$, and $g(\cdot)$, in (18) are all continuously differentiable and that the matrix $g_x(x)$ is of maximum row rank for all $x \in \mathbb{R}^n$. For any $x \in \mathbb{R}^n$, let $\mathcal{H}_E(x)$ be defined as in (24), let $\gamma, \delta > 0$, and let the optimality function $\theta : \mathbb{R}^n \to \mathbb{R}$ be defined by*

$$\theta(x) \triangleq \min_{h \in \mathcal{H}_E(x)} \Big\{ \max \Big\{ \max_{k \in \mathbf{p}} \{ c^k(x) - f^0(x) - \gamma\psi(x)_+ + \langle \nabla c^k(x), h \rangle \}, \max_{j \in \mathbf{q}} \{ f^j(x) - \psi(x)_+ + \langle \nabla f^j(x), h \rangle \} \Big\} + \tfrac{1}{2}\delta \|h\|^2 \Big\}, \tag{25a}$$

Then, (a) $\theta(x) \leq 0$ for all $x \in \mathbb{R}^n$; (b) $\theta(\cdot)$ is continuous; and (c) for any *$x \in \mathbb{R}^n$ such that $\psi(x) \leq 0$ and $g(x) = 0$, $\theta(x) = 0$ if and only if there exist multipliers $\hat{\mu} \in \Sigma_q^0$, $\hat{\mu}_0 \in \Sigma_p$, and $\hat{\zeta} \in \mathbb{R}^r$ such that (20c) and (20d) are satisfied.*

Furthermore, an alternative expression for $\theta(x)$ is given by

$$\theta(x) = - \min_{\substack{\mu \in \Sigma_q^0 \\ \nu \in \Sigma_p \\ \zeta \in \mathbb{R}^r}} \Big\{ -\mu^0 \Big(\sum_{k=1}^{p} \nu^k [c^k(x) - f^0(x) - \gamma\psi(x)_+] \Big) - \sum_{j=1}^{q} \mu^j [f^j(x) - \psi(x)_+]$$

$$+\frac{1}{2\delta}\|\mu^0\Big[\sum_{k=1}^{p}\nu^k\nabla c^k(x)\Big]+\sum_{j=1}^{q}\mu^j\nabla f^j(x)+\sum_{l=1}^{r}\zeta^l\nabla g^l(x)\|^2\Big\}. \tag{25b}$$

□

Exercise 2.2.25. Use Corollary 5.4.2 and Corollary 5.5.3 to prove Theorem 2.2.24 and to show that (21c,d) hold if and only if there exist a $\hat{\mu}^0 \in [0,1]$ and a $\hat{\zeta} \in \mathbb{R}^r$ such that

$$\hat{\mu}^0 df^0(\hat{x};h)+(1-\hat{\mu}^0)d\psi(\hat{x};h)+\langle g_x(\hat{x})^T\hat{\zeta},h\rangle \geq 0, \quad \forall h \in \mathbb{R}^n. \tag{26}$$

□

2.2.5 Second-Order Optimality Conditions for IECP

The extension of the second-order conditions from the problem **ICP** to the problem **IECP** is quite straightforward, since, instead of examining the behavior of the function $\hat{F}(\cdot)$ along the projected ridge of its graph (see Fig. 2.1.3),

$$R = \{x \in \mathbb{R}^n \mid \mathbf{q}_{\hat{F}}(x) = \mathbf{q}_A(\hat{x}),\ \hat{\mathbf{p}}(x) = \hat{\mathbf{p}}(\hat{x})\},$$

we must examine the behavior of the function $\hat{F}(\cdot)$ along the intersection of R with the set of points satisfying the equality constraints, i.e., $\{x \in \mathbb{R}^n \mid g(x) = 0\}$. We must also extend the definition of the Lagrangian (11) to the problem **IECP** by redefining it as follows:

$$L(x,\mu,\nu,\zeta) \triangleq \mu^0\Big(\sum_{k=1}^{p}\nu^k[c^k(x)-f^0(\hat{x})]\Big)+\sum_{j=1}^{q}\mu^j f^j(x)+\sum_{l=1}^{r}\zeta^l g^l(x), \tag{27a}$$

where $\mu \in \Sigma_q^0$, $\nu \in \Sigma_p$, and $\zeta \in \mathbb{R}^r$. The form (27a) is a natural one to use in proofs based on the function $\hat{F}(\cdot)$. However, as was also the case with the problem **ICP**, our second-order conditions assume a more elegant and familiar form if we suppress the constant term $f^0(\hat{x})$ which does not show up in any expressions in our theorems. Hence we may replace (27a) by

$$L(x,\mu,\nu,\zeta) \triangleq \mu^0\sum_{k=1}^{p}\nu^k c^k(x)+\sum_{j=1}^{q}\mu^j f^j(x)+\sum_{l=1}^{r}\zeta^l g^l(x). \tag{27b}$$

Finally, we retain the definitions (12c,d).

In obtaining second-order necessary conditions for the problem **IECP**, we have two alternatives. The first consists of combining the cost and inequality constraint functions into a function $\hat{F}(\cdot)$, as we did for first-order conditions, and using Theorem 2.2.16 to obtain a substitute problem, resulting in the assumptions in *(iii)(a)* in Theorem 2.2.26, below. The second alternative is to combine the active inequality functions $f^j(\cdot)$, $j \in \mathbf{q}_A(\hat{x})$, with the equality constraint functions $g^l(\cdot)$, resulting in a different substitute problem which leads to the assumptions in *(iii)(b)* in Theorem 2.2.26 below. The end result differs only in the assumptions needed, as we will now see.

Theorem 2.2.26. *Consider problem* **IECP** *in (18), with $f^0(\cdot)$ defined by (2b). Suppose that*

(i) the functions $c^k(\cdot)$, $k \in \mathbf{p}$, in (2b), and the functions $f^j(\cdot)$, $j \in \mathbf{q}$, $g^l(\cdot)$, $l \in \mathbf{r}$, in (18) are twice continuously differentiable,

(ii) $\hat{x}$ is a local minimizer for **IECP** *(defined by (18) and (2b)), and $(\hat{\mu},\hat{\nu},\hat{\zeta})$ is a triplet of associated multipliers that satisfies (21c,d),*

(iii) for some $k^* \in \hat{\mathbf{p}}(\hat{x})$, either

(a) the vectors $\nabla c^k(\hat{x})$, $k \in \hat{\mathbf{p}}(\hat{x})$, together with the vectors $\nabla f^j(\hat{x})$, $j \in \mathbf{q}_A(\hat{x})$, and the vectors $[\nabla g^l(\hat{x})+\nabla c^{k^*}(\hat{x})]$, $l \in \mathbf{r}$, are affinely independent, or

(b) the vectors $[\nabla c^k(\hat{x})-\nabla c^{k^*}(\hat{x})]$, $k \in \hat{\mathbf{p}}(\hat{x})$, $k \neq k^*$, together with the vectors $\nabla f^j(\hat{x})$, $j \in \mathbf{q}_A(\hat{x})$, and the vectors $\nabla g^l(\hat{x})$, $l \in \mathbf{r}$, are linearly independent.

Let

$$\mathcal{H}_{IE}(\hat{x}) \triangleq \{h \in \mathbb{R}^n \mid \langle \nabla f^j(\hat{x}),h\rangle = 0,\ j \in \mathbf{q}_A(\hat{x}),$$

$$\langle \nabla g^l(\hat{x}),h\rangle = 0,\ l \in \mathbf{r},$$

$$\langle \nabla c^k(\hat{x}),h\rangle = 0,\ k \in \hat{\mathbf{p}}(\hat{x})\}. \tag{28a}$$

Then for all $h \in \mathcal{H}_{IE}(\hat{x})$,

$$df^0(\hat{x};h) = 0, \tag{28b}$$

and

$$\langle h, L_{xx}(\hat{x},\hat{\mu},\hat{\nu},\hat{\zeta})h\rangle \geq 0. \tag{28c}$$

Proof. First we suppose that condition *(iii)(a)* holds, and let $\hat{F}(\cdot)$ be defined as in (3). We will use Theorem 2.2.16 and proceed essentially as in the proof of Theorem 2.1.12. First, it follows from Proposition 2.1.10 that the vectors $\nabla c^k(\hat{x})-\nabla c^{k^*}(\hat{x})$, $k \in \hat{\mathbf{p}}(\hat{x})$, $k \neq k^*$, together with the vectors $\nabla f^j(\hat{x})-\nabla c^{k^*}(\hat{x})$, $j \in \mathbf{q}_A(\hat{x})$ and the vectors $\nabla g^l(\hat{x})$, $l \in \mathbf{r}$, are linearly independent. Let

$$\mathcal{H}^*_{IE}(\hat{x}) \triangleq \{h \in \mathbb{R}^n \mid \langle \nabla c^k(\hat{x})-\nabla c^{k^*}(\hat{x}),h\rangle = 0,\ k \in \hat{\mathbf{p}}(\hat{x}),\ k \neq k^*,$$

$$\langle \nabla f^j(\hat{x}) - \nabla c^{k^*}(\hat{x}), h \rangle = 0,\ j \in \mathbf{q}_A(\hat{x}),$$

$$\langle \nabla g^l(\hat{x}), h \rangle = 0,\ l \in \mathbf{r} \}. \tag{29a}$$

It follows by inspection that $\mathcal{H}_{IE}(\hat{x}) \subset \mathcal{H}^*_{IE}(\hat{x})$. Now suppose that $h \in \mathcal{H}^*_{IE}(\hat{x})$. Then we must have that

$$\langle \Big(\hat{\mu}^0 \sum_{k \in \hat{\mathbf{p}}(\hat{x})} \hat{\nu}^k [\nabla c^k(\hat{x}) - \nabla c^{k^*}(\hat{x})] + \sum_{j \in \mathbf{q}_A(\hat{x})} [\nabla f^j(\hat{x}) - \nabla c^{k^*}(\hat{x})]$$

$$+ \sum_{l \in \mathbf{r}} \hat{\zeta}^l \nabla g^l(\hat{x})\Big), h \rangle = 0. \tag{29b}$$

It now follows from (21c,d) and (29b) that $\langle \nabla c^{k^*}(\hat{x}), h \rangle = 0$ and hence that $h \in \mathcal{H}_{IE}(\hat{x})$. Therefore, $\mathcal{H}_{IE}(\hat{x}) = \mathcal{H}^*_{IE}(\hat{x})$.

Next, it follows from Corollary 5.1.34 that for every $h \in \mathcal{H}^*_{IE}(\hat{x})$ and hence also in $\mathcal{H}_{IE}(\hat{x})$, there exists a $t_h > 0$ and a twice continuously differentiable function $s : [0, t_h] \to \mathbb{R}^n$ such that $s(0) = \hat{x}$, $\dot{s}(0) = h$, and for all $t \in [0, t_h]$,

$$\left.\begin{aligned} c^k(s(t)) - c^{k^*}(s(t)) &= 0, \\ f^j(s(t)) = c^{k^*}(s(t)) - c^{k^*}(\hat{x}), \\ g(s(t)) &= 0. \end{aligned}\right\} \tag{29c}$$

Furthermore, using Corollary 5.4.4, we conclude that there is a $t'_h \in (0, t_h]$ such that for all $t \in [0, t'_h]$, $\mathbf{p}_{\hat{F}}(s(t)) = \hat{\mathbf{p}}(\hat{x})$, $\mathbf{q}_{\hat{F}}(s(t)) = \mathbf{q}_A(\hat{x})$ (defined in (12a,b)), and hence that for all $t \in [0, t'_h]$, using the definition (27a), $L(s(t), \hat{\mu}, \hat{\nu}, \hat{\zeta}) = \hat{F}(s(t))$. Therefore there exists a $t''_h \in (0, t'_h]$ such that $L(s(t), \hat{\mu}, \hat{\nu}, \hat{\zeta}) \geq L(\hat{x}, \hat{\mu}, \hat{\nu}, \hat{\zeta}) = \hat{F}(\hat{x})$, for all $t \in [0, t''_h]$, because $\hat{x}$ is a local minimizer for $\hat{F}(\cdot)$. Since $v(t) \triangleq L(s(t), \hat{\mu}, \hat{\nu}, \hat{\zeta})$ is twice continuously differentiable, and since $\dot{v}(0) = 0$ because $\nabla_x L(\hat{x}, \hat{\mu}, \hat{\nu}, \hat{\zeta}) = 0$, by (21c,d), in view of the expansion (2.1.18b), we must have that $\ddot{v}(0) \geq 0$. Evaluating $\ddot{v}(0)$ we now obtain (28c).

Next suppose that condition *(iii)(b)* holds. Let $\rho > 0$ be such that $f^j(x) \leq 0$ for all $j \notin \mathbf{q}_A(\hat{x})$ and $x \in \text{int}\, B(\hat{x}, \rho)$. We note that since $\hat{x}$ is a local minimizer for the problem **IECP** (18), it must also be a local minimizer for the problem

$$\min \{ f^0(x) \mid f^j(x) = 0,\ j \in \mathbf{q}_A(\hat{x}), g(x) = 0,$$

$$x \in \text{int}\, B(\hat{x}, \rho) \}, \tag{29d}$$

whose feasible set is contained in that of (18). We will use this fact.

First, it follows from a trivial modification of Proposition 2.1.11 that $\mathcal{H}_{IE}(\hat{x}) = \mathcal{H}^{**}_{IE}(\hat{x})$, where

$$\mathcal{H}^{**}_{IE}(\hat{x}) \triangleq \{ h \in \mathbb{R}^n \mid \langle \nabla f^j(\hat{x}), h \rangle = 0,\ j \in \mathbf{q}_A(\hat{x}),$$

$$\langle \nabla g^l(\hat{x}), h \rangle = 0,\ l \in \mathbf{r},$$

$$\langle \nabla c^k(\hat{x}) - \nabla c^{k^*}(\hat{x}), h \rangle = 0,\ k \in \hat{\mathbf{p}}(\hat{x}) \}. \tag{29e}$$

Reasoning as above, we conclude, using Corollary 5.1.34 and Corollary 5.4.4, that for every $h \in \mathcal{H}_{IE}(\hat{x})$, there exists a $t'_h > 0$ and a twice continuously differentiable function $s : [0, t_h] \to \mathbb{R}^n$ such that $s(0) = \hat{x}$, $\dot{s}(0) = h$ and for all $t \in [0, t'_h]$, $s(t) \in B(\hat{x}, \rho)$, $\hat{\mathbf{p}}(s(t)) = \hat{\mathbf{p}}(\hat{x})$, $f^j(s(t))0$ for all $j \in \mathbf{q}_A(\hat{x})$, and $g(s(t)) = 0$. Therefore, using the definition (27b), we conclude that, for all $t \in [0, t'_h]$, $L(s(t), \hat{\mu}, \hat{\nu}, \hat{\zeta}) = \hat{\mu}^0 f^0(s(t))$. Since, by assumption, the gradients $\nabla f^j(\hat{x})$, $j \in \mathbf{q}_A(\hat{x})$ together with the gradients $\nabla g^l(\hat{x})$, $l \in \mathbf{r}$, are linearly independent, it follows from (21c,d) that $\hat{\mu}^0 > 0$. Therefore, since $\hat{x}$ is a local minimizer for (29d), there exists a $t''_h \in (0, t'_h]$ such that $\hat{\mu} f^0(s(t)) = L(s(t), \hat{\mu}, \hat{\nu}, \hat{\zeta}) \geq L(\hat{x}, \hat{\mu}, \hat{\nu}, \hat{\zeta}) = \hat{\mu}^0 f^0(\hat{x})$. Since $v(t) \triangleq L(s(t), \hat{\mu}, \hat{\nu}, \hat{\zeta})$ is twice continuously differentiable, we conclude from (21c,d) that $\dot{v}(0) = 0$, and hence (using the expansion (2.1.18b)) that $\ddot{v}(0) \geq 0$. Evaluating $\ddot{v}(0)$ we now obtain (28c). □

When there are no inequality constraints in problem **IECP** (18), so that it reduces to problem **ECP** in (23a), Theorem 2.2.26 reduces to the following form:

Theorem 2.2.26a. *Consider the problem* **ECP**, *defined in (23a), with* $f^0(\cdot)$ *defined by (2b). Suppose that*

(i) the functions $c^k(\cdot)$, $k \in \mathbf{p}$, *in (2b), and* $g(\cdot)$, *in (23a) are twice continuously differentiable,*

(ii) $\hat{x}$ *is a local minimizer for* **ECP** *and* $(\hat{\nu}, \hat{\zeta})$ *is set of associated multipliers that satisfies (23b), and*

(iii) for some $k^* \in \hat{\mathbf{p}}(\hat{x})$, *either*

(a) the vectors $\nabla c^k(\hat{x})$, $k \in \hat{\mathbf{p}}(\hat{x})$, *together with the vectors* $[\nabla g^l(\hat{x}) + \nabla c^{k^*}(\hat{x})]$, $l \in \mathbf{r}$, *are affinely independent, or*

(b) the vectors $[\nabla c^k(\hat{x}) - \nabla c^{k^*}(\hat{x})]$, $k \in \hat{\mathbf{p}}(\hat{x})$, $k \neq k^*$, *together with the vectors* $\nabla g^l(\hat{x})$, $l \in \mathbf{r}$, *are linearly independent.*†

Let

$$\mathcal{H}^*_E(\hat{x}) \triangleq \{h \in \mathbb{R}^n \mid \langle \nabla c^k(\hat{x}), h\rangle = 0,\ k \in \hat{\mathbf{p}}(\hat{x}),\ g_x(\hat{x})h = 0\}, \tag{30a}$$

and let the Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^r \to \mathbb{R}$ *be defined by*

$$L_{xx}(x, \nu, \zeta) \triangleq \langle \nu, c(x)\rangle + \langle \zeta, g(x)\rangle, \tag{30b}$$

where $c(x) \triangleq (c^1(x), \ldots, c^p(x))$. *Then, for all* $h \in \mathcal{H}^*_E(\hat{x})$,

$$df^0(\hat{x}; h) = 0, \tag{30c}$$

and

$$\langle h, L_{xx}(\hat{x}, \hat{\nu}, \hat{\zeta})h\rangle \geq 0. \tag{30d}$$

□

The easiest way to see that Theorem 2.2.26a is a special case of Theorem 2.2.26 is to assume that in (18), $f^j(x) \equiv -1$, for all $j \in \mathbf{q}$, which ensures that there are no active inequality constraints and hence that the set $\mathbf{q}_A(\hat{x})$ is empty.

An important special case of Theorem 2.2.26, with assumption *(iii)(b)*, arises when $p = 1$. In this case, since the set $\mathbf{p} = \{1\}$, assumption *(iii)(b)* reduces to the assumption that the gradients $\nabla f^j(\hat{x})$, $j \in \mathbf{q}_A(\hat{x})$, together with the gradients $\nabla g^l(\hat{x})$, $l \in \mathbf{r}$, are linearly independent. Furthermore, in this case, the expression (27b), for the Lagrangian, simplifies to $L(x, \mu, \zeta) = \sum_{j=0}^q \mu^j f^j(x) + \sum_{l=1}^r \zeta^l g^l(x)$.

Corollary 2.2.27. *Consider the problem* **IECP** *in (18). Suppose that*

(i) the functions $f^0(\cdot)$, $f^j(\cdot)$, $j \in \mathbf{q}$, *and* $g(\cdot)$ *are twice continuously differentiable,*

(ii) $\hat{x}$ *is a local minimizer for (18),*

(iii) $\hat{\mu} \in \Sigma^0_q$, *with* $\hat{\mu}^0 > 0$, *and* $\hat{\zeta} \in \mathbb{R}^r$ *are associated multipliers, i.e., they satisfy (23c,d), and*

(iv) the vectors $\nabla f^j(\hat{x})$, $j \in \mathbf{q}_A(\hat{x})$, *together with the vectors* $\nabla g^l(\hat{x})$, $l \in \mathbf{r}$, *are linearly independent.*

† Note that it follows from Proposition 2.1.10 that the assumption in part *(a)* implies that the vectors $\nabla g^l(\hat{x})$, $l \in \mathbf{r}$, are required to be linearly independent. It follows by inspection that the assumption in part *(b)* implies that the vectors $\nabla g^l(\hat{x})$, $l \in \mathbf{r}$, are required to be linearly independent.

Let

$$\mathcal{H}_{IE}(\hat{x}) \triangleq \{h \in \mathbb{R}^n \mid \langle \nabla f^j(\hat{x}), h\rangle = 0,\ j \in \mathbf{q}_A(\hat{x}),$$

$$\langle \nabla g^l(\hat{x}), h\rangle = 0,\ l \in \mathbf{r}\}. \tag{31a}$$

Then, for all $h \in \mathcal{H}_{IE}(\hat{x})$,

$$df^0(\hat{x}; h) = 0, \tag{31b}$$

and, with the Lagrangian defined by

$$L(x, \mu, \zeta) \triangleq \sum_{j=0}^q \mu^j f^j(x) + \langle \zeta, g(x)\rangle, \tag{31c}$$

$$\langle h, L_{xx}(\hat{x}, \hat{\mu}, \hat{\zeta})h\rangle \geq 0. \tag{31d}$$

□

In the case of $p = 1$, Theorem 2.2.26a assumes the following form:

Corollary 2.2.27a. *Consider the problem* **ECP** *(23a). Suppose that*

(i) the functions $f^0(\cdot)$ *and* $g(\cdot)$ *are twice continuously differentiable,*

(ii) $\hat{x}$ *is a local minimizer for* **ECP**,

(iii) $\hat{\zeta} \in \mathbb{R}^r$ *is an associated multiplier, i.e., it satisfies (23b), and*

(iv) the matrix $g_x(\hat{x})$ *has maximum row rank.*

Let $\mathcal{H}_E(\hat{x})$ *be defined as in (24). Then, for all* $h \in \mathcal{H}_E(\hat{x})$,

$$df^0(\hat{x}; h) = 0, \tag{32a}$$

and, with the Lagrangian defined by

$$L(x, \zeta) \triangleq f^0(x) + \langle \zeta, g(x)\rangle, \tag{32b}$$

$$\langle h, L_{xx}(\hat{x}, \hat{\zeta})h\rangle \geq 0. \tag{32c}$$

□

Exercise 2.2.28. Prove Corollaries 2.2.27 and 2.2.27a. □

Next we obtain the following second-order sufficient condition for the problem **IECP** (18).

Theorem 2.2.29. *Consider the problem* **IECP** *in (18), with* $f^0(\cdot)$ *defined by (2b). Suppose that*

(i) the functions $c^k(cdot)$, $k \in \mathbf{p}$, *in (2b), the functions* $f^j(\cdot)$, $j \in \mathbf{q}$, *and the functions* $g^l(\cdot)$, $l \in \mathbf{r}$, *in (18) are twice continuously differentiable,*

(ii) $\hat{x} \in \mathbb{R}^n$ *together with* $\hat{\mu} \in \Sigma^0_q$, $\hat{\nu} \in \Sigma_p$, *and* $\hat{\zeta} \in \mathbb{R}^r$ *satisfy the first-order*

optimality conditions (21c) and (21d),

(iii) $\psi(\hat{x}) \le 0$ *and* $g(\hat{x}) = 0$, *and*

(iv) $\hat{\mu}^0 > 0$.

Let $\mathbf{q}_{A+}(\hat{x}, \hat{\mu}, \hat{\nu}, \hat{\zeta})$, $\hat{\mathbf{p}}_+(\hat{x}, \hat{\mu}, \hat{\nu}, \hat{\zeta})$, *and* $\mathcal{H}'_{IE}(\hat{x})$ *be defined by*

$$\mathbf{q}_{A+}(\hat{x}, \hat{\mu}, \hat{\nu}, \hat{\zeta}) \triangleq \{ j \in \mathbf{q} \mid \hat{\mu}^j > 0 \}, \tag{33a}$$

$$\hat{\mathbf{p}}_+(\hat{x}, \hat{\mu}, \hat{\nu}, \hat{\zeta}) \triangleq \{ k \in \mathbf{p} \mid \hat{\nu}^k > 0 \}, \tag{33b}$$

and

$$\begin{aligned} \mathcal{H}_{IE}'(\hat{x}) \triangleq \{ h \in \mathbb{R}^n \mid & \langle \nabla f^j(\hat{x}), h \rangle = 0, \ j \in \mathbf{q}_{A+}(\hat{x}, \hat{\mu}, \hat{\nu}); \\ & \langle \nabla f^j(\hat{x}), h \rangle \le 0, \ j \in \mathbf{q}_A(\hat{x}) \setminus \mathbf{q}_{A+}(\hat{x}, \hat{\mu}, \hat{\nu}); \\ & \langle \nabla c^k(\hat{x}), h \rangle = 0, \ k \in \hat{\mathbf{p}}_+(\hat{x}, \hat{\mu}, \hat{\nu}); \\ & \langle \nabla c^k(\hat{x}), h \rangle \le 0, \ k \in \hat{\mathbf{p}}(\hat{x}) \setminus \hat{\mathbf{p}}_+(\hat{x}, \hat{\mu}, \hat{\nu}); \\ & \langle \nabla g^l(\hat{x}), h \rangle = 0, \ l \in \mathbf{r} \}. \end{aligned} \tag{33c}$$

Let the Lagrangian $L(x, \mu, \nu, \zeta)$ *be defined by (27a). If there exists an* $m > 0$ *such that*

$$\langle h, L_{xx}(\hat{x}, \hat{\mu}, \hat{\nu}, \hat{\zeta}) h \rangle \ge m \|h\|^2, \ \forall h \in \mathcal{H}'_{IE}(\hat{x}), \tag{33d}$$

then $\hat{x}$ *is a strict local minimizer for (18).*

Proof. We will follow the pattern established in the proof of Theorem 2.1.13. Thus, suppose that $\hat{x}$ is not a strict local minimizer for (18). Then there exists a sequence $\{x_i\}_{i=0}^{\infty}$, such that $x_i \to \hat{x}$, as $i \to \infty$, $\psi(x_i) \le 0$, $g(x_i) = 0$, and $f^0(x_i) \le f^0(\hat{x})$ for all $i \in \mathbb{N}$. Let $\delta x_i \triangleq x_i - \hat{x}$ and $h_i \triangleq \delta x_i / \|\delta x_i\|$. Since the vectors h_i are all of unit norm, the sequence $\{h_i\}_{i=0}^{\infty}$ must have accumulation points. Without loss of generality, we may assume that $h_i \to \hat{h}$, as $i \to \infty$ (with $\|\hat{h}\| = 1$). We need to consider two cases.

Case 1. First suppose that $\hat{h} \in \mathcal{H}_{IE}'(\hat{x})$. Then

$$\langle \hat{h}, L_{xx}(\hat{x}, \hat{\mu}, \hat{\nu}, \hat{\zeta}) \hat{h} \rangle \ge m \|\hat{h}\|^2, \tag{34a}$$

and hence, since $x_i \to \hat{x}$ and $h_i \to \hat{h}$, as $i \to \infty$, and $L_{xx}(\cdot, \hat{\mu}, \hat{\nu}, \hat{\zeta})$ is continuous, there exists an i_0 such that, for all $i \ge i_0$,

$$\langle h_i, L_{xx}(\hat{x} + s\delta x_i, \hat{\mu}, \hat{\nu}, \hat{\zeta}) h_i \rangle \ge \frac{m}{2} \|h_i\|^2, \ \forall i \ge i_0, \ \forall s \in [0,1]. \tag{34b}$$

Using second-order expansions (cf. (2.1.20c)), we find that

$$\begin{aligned} f^0(x_i) - f^0(\hat{x}) &= \max_{j \in \mathbf{p}} c^j(x_i) - f^0(\hat{x}) \\ &= \max_{\nu \in \Sigma_p} \sum_{k=1}^{p} \nu^k [c^k(x_i) - f^0(\hat{x})] \\ &= \max_{\nu \in \Sigma_p} \left\{ \sum_{k=1}^{p} \nu^k [c^k(\hat{x}) - f^0(\hat{x})] + \langle \sum_{k=1}^{p} \nu^k \nabla c^k(\hat{x}), \delta x_i \rangle \right. \\ &\quad \left. + \int_0^1 (1-s) \langle \delta x_i, \sum_{k=1}^{p} \nu^k c_{xx}^k(\hat{x} + s\delta x_i) \delta x_i \rangle \, ds \right\} \\ &\ge \left\{ \sum_{k=1}^{p} \hat{\nu}^k [c^k(\hat{x}) - f^0(\hat{x})] + \langle \sum_{k=1}^{p} \hat{\nu}^k \nabla c^k(\hat{x}), \delta x_i \rangle \right. \\ &\quad \left. + \int_0^1 (1-s) \langle \delta x_i, \sum_{k=1}^{p} \hat{\nu}^k c_{xx}^k(\hat{x} + s\delta x_i) \delta x_i \rangle \, ds \right\}. \end{aligned} \tag{34c}$$

Similarly, taking into account the fact that $\sum_{j=1}^{q} \hat{\mu}^j = 1 - \hat{\mu}^0$ and noting that because $\hat{\mu}^0 > 0$, $\hat{\mu}^0 = 1$ when $\psi(\hat{x}) < 0$, we find that

$$\begin{aligned} (1 - \hat{\mu}^0) \psi(x_i) \ge & \left\{ \sum_{j=1}^{p} \hat{\mu}^j f^j(\hat{x}) + \langle \sum_{j=1}^{q} \hat{\mu}^j \nabla f^j(\hat{x}), \delta x_i \rangle \right. \\ & \left. + \int_0^1 (1-s) \langle \delta x_i, \sum_{j=1}^{p} \hat{\mu}^j f_{xx}^j(\hat{x} + s\delta x_i) \delta x_i \rangle \, ds \right\}. \end{aligned} \tag{34d}$$

Finally, because $g(\hat{x}) = 0$, for $l \in \mathbf{r}$,

$$0 = g^l(x_i) = \langle \nabla g^l(\hat{x}), \delta x_i \rangle + \int_0^1 (1-s) \langle \delta x_i, g_{xx}^l(\hat{x} + s\delta x_i) \delta x_i \rangle \, ds. \tag{34e}$$

Now, for all $i \in \mathbb{N}$, we must have that

$$\hat{\mu}^0 [f^0(x_i) - f^0(\hat{x})] \ge \hat{\mu}^0 [f^0(x_i) - f^0(\hat{x})] + (1 - \hat{\mu}^0) \psi(x_i) + \sum_{l=1}^{r} \zeta^l g^l(x_i). \tag{34f}$$

Using (21c,d), (27a), and (34b-34f), we now find that, for all $i \ge i_0$,

$$\hat{\mu}^0 [f^0(x_i) - f^0(\hat{x})] \ge \int_0^1 (1-s) \langle \delta x_i, L_{xx}(\hat{x} + s_i \delta x_i, \hat{\mu}, \hat{\nu}, \hat{\zeta}) \delta x_i \rangle \, ds$$

$$\geq \frac{m}{2}\|\delta x_i\|^2 . \tag{34g}$$

Since $\hat{\mu}^0 > 0$ by assumption, we have a contradiction.

Case 2. Hence we must consider the only other alternative, viz., that $\hat{h} \notin \mathcal{H}_{IE}'(\hat{x})$. Since, by assumption, $f^0(x_i) \leq f^0(\hat{x})$ for all $i \in \mathbb{N}$, it follows that $c^k(x_i) \leq f^0(\hat{x})$ for all $k \in \mathbf{p}$ and all $i \in \mathbb{N}$. Consequently, since for all $k \in \hat{\mathbf{p}}(\hat{x})$, $c^k(\hat{x}) = f^0(\hat{x})$, it follows from the Mean-Value Theorem 5.1.28*(a)*, that there exist $s_i^k \in (0,1)$ such that, for all $i \in \mathbb{N}$ and $k \in \hat{p}(\hat{x})$,

$$c^k(x_i) - c^k(\hat{x}) = \langle \nabla c^k(\hat{x} + s_i^k \delta x_i), \delta x_i \rangle \leq 0 . \tag{34h}$$

Since $s_i^k \delta x_i \to 0$, as $i \to \infty$, letting $i \to \infty$, we conclude from (34g) that

$$\langle \nabla c^k(\hat{x}), \hat{h} \rangle \leq 0, \quad \forall k \in \hat{\mathbf{p}}(\hat{x}) . \tag{34i}$$

Similar reasoning shows that

$$\langle \nabla f^j(\hat{x}), \hat{h} \rangle \leq 0, \quad \forall j \in \mathbf{q}_A(\hat{x}) \tag{34j}$$

and

$$\langle \nabla g^l(\hat{x}), \hat{h} \rangle = 0, \quad \forall l \in \mathbf{r} . \tag{34k}$$

But (34h-34k) and the fact that $\hat{h} \notin \mathcal{H}_{IE}'(\hat{x})$ imply that there exists a $k^* \in \hat{\mathbf{p}}_+(\hat{x}, \hat{\mu}, \hat{\nu}, \hat{\zeta})$ such that $\hat{\nu}^{k^*} \langle \nabla c^{k^*}(\hat{x}), \hat{h} \rangle < 0$, or a $j^* \in \mathbf{q}_{A+}(\hat{x}, \hat{\mu}, \hat{\nu}, \hat{\zeta})$ such that $\hat{\mu}^{j^*} \langle \nabla f^{j^*}(\hat{x}), \hat{h} \rangle < 0$. Since $\hat{\mu}^0 \neq 0$ by assumption, if $\langle \nabla c^{k^*}(\hat{x}), \hat{h} \rangle < 0$, with $k^* \in \hat{\mathbf{p}}_+(\hat{x}, \hat{\mu}, \hat{\nu})$, then, using (21c) we find that

$$\begin{aligned} 0 &> \hat{\mu}^0 \hat{\nu}^{k^*} \langle \nabla c^{k^*}(\hat{x}), \hat{h} \rangle \\ &= -\langle \hat{\mu}^0 \Big(\sum_{\substack{k=1 \\ k \neq k^*}}^{p} \hat{\nu}^k \nabla c^k(\hat{x}) + \sum_{j=1}^{q} \hat{\mu}^j \nabla f^j(\hat{x}) + \sum_{l=1}^{r} \hat{\zeta}^l \nabla g^l(\hat{x}) \Big), \hat{h} \rangle \geq 0 , \end{aligned} \tag{34l}$$

which is impossible. A similar contradiction arises if $\hat{\mu}^{j^*} \langle \nabla f^{j^*}(\hat{x}), \hat{h} \rangle < 0$. Hence $\hat{x}$ is a strict local minimizer for **IECP**. □

Again the reader may prefer to use the following, equivalent, somewhat less cumbersome expression for $\mathcal{H}_{IE}'(\hat{x})$:

$$\begin{aligned} \mathcal{H}_{IE}'(\hat{x}) \triangleq \{ h \in \mathbb{R}^n \mid & \hat{\mu}^j \langle \nabla f^j(\hat{x}), h \rangle = 0, \\ & (1 - \hat{\mu}^j) \langle \nabla f^j(\hat{x}), h \rangle \leq 0, \ j \in \mathbf{q}_A(\hat{x}); \\ & \hat{\nu}^k \langle \nabla c^k(\hat{x}), h \rangle = 0, \ (1 - \hat{\nu}^k) \langle \nabla c^k(\hat{x}), h \rangle \leq 0, \ k \in \hat{\mathbf{p}}(\hat{x}); \\ & \langle \nabla g^l(\hat{x}), h \rangle = 0, \ l \in \mathbf{r} \}. \end{aligned} \tag{35}$$

Note that when (21c,d) hold with $\hat{\mu}^0 = 0$, then because the vectors $\nabla g^l(\hat{x})$, $l \in \mathbf{r}$, are linearly independent, $\hat{\mu} \neq 0$. It then follows from (21d) that there is at least one $j \in \mathbf{q}$ such that $f^j(\hat{x}) = 0$, and hence that $\psi(\hat{x}) = 0$. Also, in this case, if (33d) holds, then it implies that $\hat{x}$ is a strict local minimizer for the problem $\min \{ \psi(x) \mid g(x) = 0 \}$. Consequently, there exists a $\hat{\rho} > 0$ such that, for all $x \in \{ x \in B(\hat{x}, \hat{\rho}) \mid g(x) = 0 \}$, $\psi(x) > 0$, i.e., $\hat{x}$ is an isolated feasible point, and hence, automatically, a local minimizer for (18).

Again we get an important special case when $p = 1$, so that $f^0(\cdot) = c^1(\cdot)$, and the Lagrangian assumes the simplified form

$$L(x, \mu, \zeta) \triangleq \sum_{j=0}^{q} \mu^j f^j(x) + \sum_{l=1}^{r} \zeta^l g^l(x) . \tag{36}$$

Similarly, since $\hat{\nu} = 1$ in this case, it can also be dropped as an argument in the definition of $\mathbf{q}_{A+}$.

Corollary 2.2.30. *Consider the problem* **IECP** *in (18). Suppose that*

(i) the functions $f^0(\cdot)$, $f^j(\cdot)$, $j \in \mathbf{p}$, and $g^l(\cdot)$, $l \in \mathbf{r}$, in (18) are twice continuously differentiable,

(ii) $\hat{x} \in \mathbb{R}^n$ together with $\hat{\mu} \in \Sigma_q^0$, such that $\hat{\mu}^0 > 0$, and $\hat{\zeta} \in \mathbb{R}^r$ satisfy the first-order conditions (23d,e),

(iii) $\psi(\hat{x}) \leq 0$ and $g(\hat{x}) = 0$, and

(iv) the vectors $\nabla g^l(\hat{x})$, $l \in \mathbf{r}$, are linearly independent. Let

$$\begin{aligned} \mathcal{H}_{IE}'(\hat{x}) \triangleq \{ h \in \mathbb{R}^n \mid & \langle \nabla f^j(\hat{x}), h \rangle = 0, \ j \in \mathbf{q}_{A+}(\hat{x}, \hat{\mu}, \hat{\zeta}); \\ & \langle \nabla g^l(\hat{x}), h \rangle = 0, \ l \in \mathbf{r} \}, \end{aligned} \tag{37a}$$

where

$$\mathbf{q}_{A+}(\hat{x}, \hat{\mu}, \hat{\zeta}) \triangleq \{ j \in \mathbf{q} \mid \hat{\mu}^j > 0 \}, \tag{37b}$$

i.e., it is defined as in (33a), with $\hat{\nu} = 1$. Let $L(x, \mu, \zeta)$ be defined as in (36). If there exists an $m > 0$ such that

$$\langle h, L_{xx}(\hat{x}, \hat{\mu}, \hat{\zeta}) h \rangle \geq m \|h\|^2, \quad \forall h \in \mathcal{H}_{IE}'(\hat{x}), \tag{37c}$$

then $\hat{x}$ is a strict local minimizer for **IECP**. □

In the special case when there are no inequality constraints in (18), as in problem (23a), Corollary 2.2.30 assumes the following form:

Corollary 2.2.31. *Consider the problem* **ECP** *in (23a). Suppose that*

(i) the functions $f^0(\cdot)$ and $g^l(\cdot)$, $l \in \mathbf{r}$, in (23a) are twice continuously differentiable,

(ii) $\hat{x} \in \mathbb{R}^n$ together with and $\hat{\zeta} \in \mathbb{R}^r$ satisfy the first-order conditions (23b),

(iii) $g(\hat{x}) = 0$, and

(iv) the matrix $g_x(\hat{x})$ has maximum row rank.

Let

$$\mathcal{H}_E(\hat{x}) \triangleq \{h \in \mathbb{R}^n \mid g_x(\hat{x})h = 0\}, \tag{38a}$$

Let the Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}$ be defined by

$$L(x, \zeta) \triangleq f^0(x) + \langle \zeta, g(x) \rangle. \tag{38b}$$

If there exists an $m > 0$ such that

$$\langle h, L_{xx}(\hat{x}, \hat{\zeta})h \rangle \geq m\|h\|^2, \quad \forall h \in \mathcal{H}_E(\hat{x}), \tag{38c}$$

then $\hat{x}$ is a strict local minimizer for **ECP**. □

The easiest way to see that Corollary 2.2.31 follows directly from Corollary 2.2.30 is to set $f^j(x) \equiv -1$ in (18), for all $j \in \mathbf{q}$.

Exercise 2.2.32. Prove Theorem 2.2.29 and Corollary 2.2.30. □

2.2.6 Notes

First-order, multiplier type optimality conditions for inequality constrained problems, subject to a constraint qualification, were first established by Karush in an unpublished and long unknown M.Sc. dissertation [Kar.39]. The constraint qualification ensured a nonzero multiplier on the gradient of the cost function. These conditions were independently discovered by Kuhn and Tucker [KuT.51] and by John [Joh.48], who did not use a constraint qualification, and hence did not assert that the multiplier associated with the cost function was nonzero. First-order optimality conditions of the John type for equality and inequality constrained problems, were first established by Canon, Cullum, and Polak [CCP.66] and, under a constraint qualification, by Mangasarian and Fromowitz [MaF.67]. A beautiful exposition of the role of multipliers in the characterization of optimality for problems with equality, inequality and abstract constraints (of the form $x \in X$) and their relationship to the existence of exact penalty functions can be found in [Roc.83], which reviews and interprets a large number of recent results.

The transformation technique used in this section for proving first-order optimality conditions can be traced to the Bui-Trong-Liu and Huard method of centers [BuH.66, Hua.67, Hua.68], was pointed out to the author by Francis Clarke, and is quite different from that appearing in earlier works.

The optimality functions in this section were first presented formally in [Pol.87]. However, they already had appeared in the methods of centers and feasible directions analyzed in [PiP.72, PiP.73], and other papers of the author.

Second-order conditions for equality and inequality constrained problems were first established by McCormick [McC.67]. The second-order necessary condition in Theorem 2.2.26, under assumption *(iii)(a)*, appears to be new. The version of second-order sufficient conditions presented in this book was taken from [HaM.79].

2.3 Algorithm Models and Convergence Conditions II

The algorithm models presented in Section 1.2 are based on the assumption that one has a feasible initial point $x_0 \in X$, where X is the constraint set in (1.2.1b), and that one has a set-valued algorithm function $A(\cdot)$ mapping the constraint set into itself. In this case, convergence can be assured by requiring that a *monotone uniform descent* (MUD) property, involving only the cost function and the algorithm function, i.e., every every nonstationary point x has a neighborhood from which the algorithm decreases the cost by a guaranteed amount. When $X = \mathbb{R}^n$ or when X is a "simple" compact, convex set, e.g., a polyhedron or a ball centered at the origin, it is fairly easy to find such an initial point, and it is frequently possible to construct algorithm functions that map the constraint set into itself. However, in the case of the problems **ICP** (i.e., (2.2.2a), (2.2.2b)), and **IECP** (i.e., (2.2.18), (2.2.2b)), the constraint set is defined by systems of inequalities or by systems of inequalities equations and inequalities that need not be convex. Hence finding an initial feasible point may be quite difficult. Furthermore, in the presence of nonlinear equality constraints, it may not be possible to construct algorithm functions that map the constraint set into itself. Consequently, extensions of the theory presented in Section 1.2 are required.

2.3.1 Algorithm Models for ICP

We begin with an algorithm model for **ICP**, which can be written compactly as follows:

$$\textbf{ICP} \qquad \min\{f^0(x) \mid \psi(x) \leq 0\}, \tag{1a}$$

where $f^0, \psi : \mathbb{R}^n \to \mathbb{R}$ are continuous functions. Thus we see that (1a) is of the form (1.2.1b), with the set $X \triangleq \{x \in \mathbb{R}^n \mid \psi(x) \leq 0\}$.

We assume that we have a continuous optimality function $\theta : \mathbb{R}^n \to \mathbb{R}$ (e.g., as in (2.2.8f))). Since it is possible for an optimality function for problems of the form (1a) to have zeros outside the *feasible set*

$$X \triangleq \{x \in \mathbb{R}^n \mid \psi(x) \leq 0\} \tag{1b}$$

and since the algorithms under consideration are not restricted to the set X, we need two concepts that were first introduced in Section 1.2. First we define the set of *quasi-stationary points* $\mathcal{QS}$ by

$$\mathcal{QS} \triangleq \{x \in \mathbb{R}^n \mid \theta(x) = 0\}, \tag{1c}$$

and then we define the set of *stationary points* S by

$$S \triangleq \{x \in \mathbb{R}^n \mid \psi(x) \leq 0,\ \theta(x) = 0\}. \tag{1d}$$

We will see that we can ensure only that phase I - phase II methods of centers construct points in $\mathcal{QS}$. Hence it is desirable to formulate problems of the form (1a) so that $\mathcal{QS} = S$. The requirement that $\mathcal{QS} = S$, i.e., that $\theta(\cdot)$ have no zeros outside the feasible set X, is called a *constraint qualification*.

Our first algorithm model, for finding points in $\mathcal{QS}$, defined by (1c), captures the essential properties of phase I - phase II methods of centers. These can be of two kinds: the first uses the "unified", i.e., artificial, cost function $F(\cdot,\cdot)$, defined in (2.2.8a) to ensure that an appropriate MUD (*monotone uniform descent*) property is satisfied. The second uses the constraint violation function $\psi(\cdot)$ as the cost function to ensure that a MUD property outside of the feasible set X and then uses the cost function $f^0(\cdot)$ to ensure that a MUD property inside of the feasible set X.

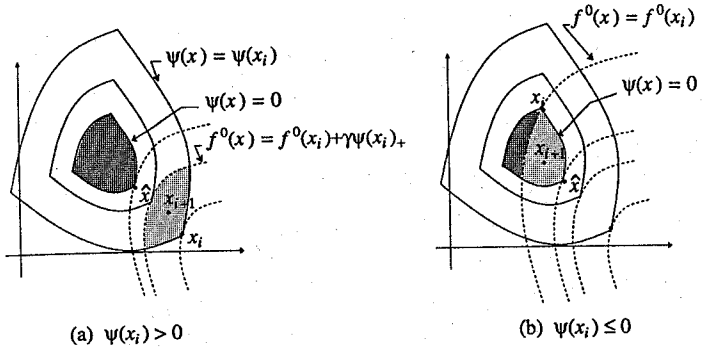
The designation *phase I - phase II* is derived from the simplex algorithm terminology and is used to indicate that an algorithm for solving (1a) can be initialized with *any* point $x_0 \in \mathbb{R}^n$.

Both kinds of phase I - phase II methods of centers use an *algorithm function* $A : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ whose definition, directly or indirectly, involves level sets of the function $F(\cdot,\cdot)$, defined in (2.2.8a), i.e., the function $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$F(x',x) \triangleq \max\{f^0(x) - f^0(x') - \gamma\psi(x')_+, \psi(x) - \psi(x')_+\}, \tag{2}$$

where $\gamma > 0$ is a preselected parameter. All phase I - phase II methods of centers have in common the fact that, given a point x_i, they construct a point x_{i+1} in the interior (i.e., "in the center") of the level set $\{x \mid F(x_i,x) \leq 0\}$, as shown in Fig. 2.3.1. They differ from one another by the formula used to define a "center" of $\{x \mid F(x_i,x) \leq 0\}$.

The following algorithm model can be used in the analysis and construction of the two kinds of methods of centers mentioned above. Note its similarity to Algorithm Model 1.2.11.



(a) $\psi(x_i) > 0$ (b) $\psi(x_i) \leq 0$

Fig. 2.3.1. A phase I - phase II method of centers algorithm.

Algorithm Model 2.3.1.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute a $x_{i+1} \in A(x_i)$.

Step 2. If $\theta(x_{i+1}) = 0$, stop.
Else, replace i by $i+1$, and go to Step 1.

The following theorem applies to phase I - phase II methods of centers based on the "unified" (one-cost) descent function $F(\cdot,\cdot)$, examples of which can be found in [PoH.91a, PYM.93].

Theorem 2.3.2. *Suppose that, for $\gamma > 0$, $F(\cdot,\cdot)$ is defined by (2) and that, for every $x \notin \mathcal{QS}$, there exist a $\rho_x > 0$ and a $\delta_x > 0$ such that*

$$F(x',x'') \leq -\delta_x < 0,\ \forall x' \in B(x,\rho_x),\ \forall x'' \in A(x'). \tag{3}$$

Suppose that Algorithm Model 2.3.1 constructs an infinite sequence $\{x_i\}_{i=0}^{\infty}$.

(a) If there exists an $i_0 \in \mathbb{N}$ such that $\psi(x_{i_0}) \leq 0$, then $\psi(x_i) \leq 0$ for all $i \geq i_0$.

(b) Every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ is quasi-stationary, i.e., $\hat{x} \in \mathcal{QS}$. Furthermore, if $\mathcal{QS} = S$, then $\hat{x} \in S$.

Proof. *(a)* We give a proof by induction. Suppose that $\psi(x_i) \leq 0$. Since the sequence $\{x_i\}_{i=0}^{\infty}$ is infinite, we must have that $x_i \notin \mathcal{QS}$ for all $i \in \mathbb{N}$ and hence, by (3), $F(x_i,x_{i+1}) < 0$ for all $i \in \mathbb{N}$. Now, if for any $i \in \mathbb{N}$, $\psi(x_i)_+ = 0$, then it follows from (2) that $\psi(x_{i+1}) \leq 0$. Hence, if there is an i_0 such that

$\psi(x_{i_0}) \leq 0$, then we must have $\psi(x_i) \leq 0$ for all $i \geq i_0$.

(b) Now suppose that $x_i \to^K \hat{x}$, as $i \to \infty$, and that $\hat{x} \notin \boldsymbol{QS}$. We must consider two mutually exclusive cases.

Case 1. Suppose that $\psi(x_i) > 0$ for all $i \in \mathbb{N}$. Then $\psi(x_i)_+ = \psi(x_i)$ for all $i \in \mathbb{N}$, and hence, since $F(x_i, x_{i+1}) < 0$ for all $i \in \mathbb{N}$, it follows that $\{\psi(x_i)\}_{i=0}^{\infty}$ is a monotone decreasing sequence. Since $\psi(\cdot)$ is continuous, it follows that $\psi(x_i) \to^K \psi(\hat{x})$ as $i \to \infty$, and hence it follows from Proposition 5.1.16 that $\psi(x_i) \to \psi(\hat{x})$, as $i \to \infty$. However, since $\hat{x} \notin \boldsymbol{QS}$, it follows from (3) that there exist an i_1 and a $\delta_{\hat{x}} > 0$ such that, for all $i \in K$, $i \geq i_1$,

$$\psi(x_{i+1}) - \psi(x_i) \leq F(x_i, x_{i+1}) \leq -\delta_{\hat{x}} < 0, \tag{4a}$$

which contradicts the convergence of the sequence $\{\psi(x_i)\}_{i=0}^{\infty}$.

Case 2. Suppose that there exists an i_0 such that $\psi(x_{i_0}) \leq 0$. Then, by *(a)*, $\psi(x_i)_+ = 0$ for all $i \geq i_0$. Hence for all $i \geq i_0$,

$$f^0(x_{i+1}) - f^0(x_i) \leq F(x_i, x_{i+1}) < 0, \tag{4b}$$

so that the sequence $\{f^0(x_i)\}_{i=i_0}^{\infty}$ is monotone decreasing. Since $f^0(\cdot)$ is continuous, it follows that $f^0(x_i) \to^K f^0(\hat{x})$ as $i \to \infty$, and hence it follows from Proposition 5.1.16 that $f^0(x_i) \to f^0(\hat{x})$, as $i \to \infty$. However, since $\hat{x} \notin \boldsymbol{QS}$, it follows from (3) that there exists an $i_1 \geq i_0$ and a $\delta_{\hat{x}} > 0$ such that, for all $i \in K$, $i \geq i_1$,

$$f^0(x_{i+1}) - f^0(x_i) \leq F(x_i, x_{i+1}) \leq -\delta_{\hat{x}}, \tag{4c}$$

which contradicts the convergence of the sequence $\{f^0(x_i)\}_{i=0}^{\infty}$. Thus our proof is complete. □

Algorithm Model 2.3.1 can also be used in the analysis of a class of *two-cost* phase I - phase II methods of centers for finding points in $\boldsymbol{QS}$, such as the one described in [PTM.79]. As we have already mentioned, these methods use the function $f^0(\cdot)$ as a descent function at points x_i such that $\psi(x_i) \leq 0$, and use the function $\psi(\cdot)$ as a descent function at points x_i such that $\psi(x_i) > 0$.

Exercise 2.3.3. Mimic the proof of Theorem 2.3.2 to establish the following general convergence theorem for two-cost phase I - phase II methods of centers.

Theorem 2.3.4. *Let the feasible set X be defined by (1b). Consider Algorithm Model 2.3.1, and suppose that*

(i) $A(F) \subset X$,

(ii) *For every $x \notin \boldsymbol{QS}$, there exist a $\rho_x > 0$ and a $\delta_x > 0$ such that*

if $\psi(x) > 0$, then

$$\psi(x'') - \psi(x') \leq -\delta_x < 0, \ \forall x' \in B(x, \rho_x) \cap X^c, \ \forall x'' \in A(x'), \tag{5a}$$

if $\psi(x) \leq 0$, then

$$f^0(x'') - f^0(x') \leq -\delta_x < 0, \ \forall x' \in B(x, \rho_x) \cap X, \ \forall x'' \in A(x'). \tag{5b}$$

Suppose that Algorithm Model 2.3.1 constructs an infinite sequence $\{x_i\}_{i=0}^{\infty}$.

(a) *If there exists an $i_0 \in \mathbb{N}$ such that $\psi(x_{i_0}) \leq 0$, then $\psi(x_i) \leq 0$ for all $i \geq i_0$.*

(b) *Every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ is quasi-stationary, i.e., $\hat{x} \in \boldsymbol{QS}$. Furthermore, if $\boldsymbol{QS} = \boldsymbol{S}$, then $\hat{x} \in \boldsymbol{S}$.* □

2.3.2 Algorithm Models for IECP

Let f^0, $\psi : \mathbb{R}^n \to \mathbb{R}$, and $g : \mathbb{R}^n \to \mathbb{R}^r$ be continuous functions. We will consider the problem

IECP $$\min \{f^0(x) \mid \psi(x) \leq 0, \ g(x) = 0\}. \tag{6a}$$

We assume that we have a continuous optimality function $\theta : \mathbb{R}^n \to \mathbb{R}$ (e.g., as in (2.2.25b)), and we define the set of *quasi-stationary points* $\boldsymbol{QS}$ by

$$\boldsymbol{QS} \triangleq \{x \in \mathbb{R}^n \mid \theta(x) = 0\}. \tag{6b}$$

Again, since it is possible for an optimality function for problems of the form (6a) to have zeros outside the feasible set

$$X \triangleq \{x \in \mathbb{R}^n \mid \psi(x) \leq 0, g(x) = 0\} \tag{6c}$$

and since the algorithms under consideration are not restricted to the set X, we need to introduce the additional concept of the set of *stationary points* S defined by

$$\boldsymbol{S} \triangleq \{x \in \mathbb{R}^n \mid \psi(x) \leq 0, \ g(x) = 0, \ \theta(x) = 0\}. \tag{6d}$$

Quite often, one solves the problem **IECP** by converting it into either an unconstrained min-max problem or into an inequality constrained problem by means of a variety of *exact penalty function* techniques that depend upon the correct selection of a penalty parameter π. Each value of the penalty parameter creates a problem $\mathbf{P}_\pi$ and there is usually a value $\hat{\pi}$, such that, for all $\pi \geq \hat{\pi}$, the stationary points of $\mathbf{P}_\pi$ coincide with the stationary points of **IECP**. However, it is usually impossible to determine $\hat{\pi}$ a priori. Therefore, we will present an algorithm model for determining $\hat{\pi}$ by a feedback process. We will assume that we are given a strictly monotone increasing sequence of positive penalty values

$\{\pi_j\}_{j=0}^{\infty}$ such that $\pi_j \to \infty$, as $j \to \infty$, which result in an infinite sequence of problems $\{\mathbf{P}_j\}_{j=0}^{\infty}$, with $\mathbf{P}_j \triangleq \mathbf{P}_{\pi_j}$, either of the form

$$\mathbf{P}_j: \qquad \min_{x \in \mathbb{R}^n} f_j^0(x), \tag{7a}$$

where the functions $f_j^0 : \mathbb{R}^n \to \mathbb{R}$ are continuous or of the form

$$\mathbf{P}_j: \qquad \min\{f_j^0(x) \mid \psi_j(x) \le 0\}, \tag{7b}$$

where functions $f_j^0, \psi_j : \mathbb{R}^n \to \mathbb{R}$ are both continuous.

With each problem $\mathbf{P}_j$, we associate an optimality function $\theta_j(\cdot)$, a set of *stationary points* S_j, an *algorithm function* $A_j : \mathbb{R}^n \to 2^{\mathbb{R}^n}$, and a continuous *test function* $t_j : \mathbb{R}^n \to \mathbb{R}$, which is to be used in determining when the penalty parameter is to be increased. For problems of the form (7a), we define

$$S \triangleq \{x \in \mathbb{R}^n \mid \theta_j(x) = 0\}, \tag{7c}$$

while for problems of the form (7b), we define

$$S \triangleq \{x \in \mathbb{R}^n \mid \theta_j(x) = 0, \psi_j(x) \le 0\}. \tag{7d}$$

Obviously, the optimality functions in (7c) and (7d) are not the same.

Assumption 2.3.5. *We will assume that for all $j \in \mathbb{N}$, the test functions $t_j : \mathbb{R}^n \to \mathbb{R}$ have the following properties:*

(i) *the test functions $t_j(\cdot)$ are continuous,*

(ii) *$\{x \in S_j \mid t_j(x) \le 0\} \subset S$,*

(iii) *for every $x^* \in \mathbb{R}^n$, there exist a $j^* \in \mathbb{N}$ and a $\rho^* > 0$ such that, for all $j \ge j^*$ and $x \in B(x^*, \rho^*)$, $t_j(x) \le 0$.* □

We can consider the two possibilities defined by (7a), (7b) within the same framework of the following algorithm model:

Algorithm Model 2.3.6.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$, $j = 0$.

Step 1. If $t_j(x_i) \le 0$, go to Step 2. Else, go to Step 4.

Step 2. Compute an $x_{i+1} \in A_j(x_i)$.

Step 3. If $x_{i+1} \in S_j$ and $t_j(x_{i+1}) \le 0$, stop. Else, replace i by $i+1$, and go to Step 1.

Step 4. Set $x_j^* = x_i$, replace j by $j+1$, and go to Step 1.

Theorem 2.3.7. *Suppose that*

(i) *for all $j \in \mathbb{N}$, the problems $\mathbf{P}_j$ are defined as in (7a) or (7c),*

(ii) *for all $j \in \mathbb{N}$, the stationary sets S_j are defined as in (7c) for (7a) and as in (7d) for (7b),*

(iii) *for all $j \in \mathbb{N}$, in the case of problem (7b), $\theta_j(x) < 0$ for all $x \in \mathbb{R}^n$ such that $\psi_j(x) > 0$,*

(iv) *for all $j \in \mathbb{N}$, the algorithm maps $A_j(\cdot)$ have the property that, given any $x_0 \in \mathbb{R}^n$, if the sequence $\{x_i\}_{i=0}^{\infty}$ in $\mathbb{R}^n$ is constructed according to the recursion $x_{i+1} \in A_j(x_i)$ and $\hat{x}$ is an accumulation point of the sequence $\{x_i\}_{i=0}^{\infty}$, then $\hat{x} \in S$ †, and*

(v) *Assumption 2.3.5 is satisfied.*

Under these assumptions,

(a) *if Algorithm Model 2.3.6 constructs a finite sequence $\{x_j^*\}_{j=0}^{j^*}$ and the sequence $\{x_i\}$ is infinite, then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ is in S,*

(b) *if Algorithm Model 2.3.6 constructs a finite sequence $\{x_j^*\}_{j=0}^{j^*}$ and the sequence $\{x_i\}$ is also finite, then its last element, say, x_{i^*}, is in S, and*

(c) *if Algorithm Model 2.3.6 constructs an infinite sequence $\{x_j^*\}_{j=0}^{\infty}$, then $\{x_j^*\}_{j=0}^{\infty}$ has no accumulation points.*

Proof. *(a)* In this case, because of assumption *(iii)*, every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ is in S_{j^*}. Since j has not been incremented beyond j^*, there must exist an i^* such that $t_{j^*}(x_i) \le 0$ for all $i \ge i^*$, and hence, because $t_{j^*}(\cdot)$ is continuous and $x_i \to^K \hat{x}$, as $i \to \infty$, for some $K \subset \mathbb{N}$, it follows that $t_{j^*}(\hat{x}) \le 0$. The desired conclusion now follows from Assumption 2.3.5*(ii)*.

(b) Again the desired conclusion follows directly from Assumption 2.3.5*(ii)*, since in this case, $x_{i^*} \in \{x \in S_{j^*} \mid t_{j^*}(x) \le 0\}$.

(c) For the sake of contradiction, suppose that $\{x_j^*\}_{j=0}^{\infty}$ is infinite and that it has an accumulation point x^{**}. Then there must exist an infinite subset $K \subset \mathbb{N}$ such that $x_j^* \to^K x^{**}$, as $j \to \infty$. Now, by assumption, there exists a $\rho^{**} > 0$ and a $j^{**} \in \mathbb{N}$ such that, for all $x_j^* \in B(x^{**}, \rho^{**})$ and $j \ge j^{**}$, $t_j(x^*_j) \le 0$. Clearly, there exists an $i^{**} \in \mathbb{N}$ such that in the test in Step 1 of Algorithm Model 2.3.6, $j \ge j^{**}$ for all $i \ge i^{**}$, and hence $t_j(x_i) \le 0$ for all $i \ge i^{**}$ and

† Obviously, under a suitable constraint qualification, this assumption will hold if the algorithm maps $A_j(\cdot)$ satisfy the hypotheses of Theorem 1.2.12 for (7a) and the hypotheses of Theorem 2.3.2 or Theorem 2.3.4 for (7b).

$x_i \in B(x^{**}, \rho^{**})$. Hence for all $i \geq i^{**}$, no $x_i \in B(x^{**}, \rho^{**})$ is converted into a x_j^* in Step 4 and therefore x^{**} cannot be an accumulation point of $\{x_j^*\}_{j=0}^{\infty}$. This completes our proof. □

2.3.3 Notes

The original Huard methods of centers [BuH.66., Hua.67] were *phase II* methods, i.e., they had to be initialized with a *feasible point* (a point x_0 *such that* $\psi(x_0) \leq 0$). The first extension to a "two-cost" phase I - phase II version was presented in [PTM.79], while the "unified" version was presented in [PoH.91a]. Theorems 2.3.2 and 2.3.4 summarize the ideas used in proving convergence of these algorithms, in a way that makes clear their relationship to Theorem 1.2.12. Results of the form of Algorithm Model 2.3.6 and Theorem 2.3.7 were first presented in [Pol.76].

2.4 First-Order Min-Max Algorithms

Finally, we are ready to introduce algorithms for solving the unconstrained min-max problem

$$\textbf{MMP} \qquad \min_{x \in \mathbb{R}^n} \psi(x), \tag{1}$$

where

$$\psi(x) = \max_{j \in \mathbf{q}} f^j(x), \tag{2}$$

with the functions $f^j : \mathbb{R}^n \to \mathbb{R}$ continuously differentiable.

2.4.1 The PPP Min-Max Algorithm

We begin with an algorithm that can be viewed as a natural extension of the Armijo Gradient Algorithm 1.3.3. Different versions of this algorithm were proposed by Pshenichnyi around 1970 (for an accessible reference see [PsD.75]) and later, independently, by Pironneau and Polak [PiP.72]. Although the version we present differs from the earlier versions in the step-size rule, we will refer to it as the Pshenichnyi-Pironneau-Polak min-max algorithm, or PPP min-max algorithm, for short.

Our version of the PPP algorithm is based on the the optimality function $\theta(\cdot)$ defined in (2.1.10c) and the search direction function $h(\cdot)$ defined in (2.1.10d). In addition, it uses an Armijo type step-size rule. Since it is a natural extension of the Armijo gradient Algorithm 1.3.3, hence the analysis of its convergence and rate of convergence is quite similar to that for the Armijo Gradient Algorithm 1.3.3.

Pshenichnyi-Pironneau-Polak Algorithm 2.4.1.

Parameters. $\alpha \in (0, 1]$, $\beta \in (0, 1)$, $\delta > 0$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute the *optimality function value* $\theta_i \triangleq \theta(x_i)$ and the *search direction* $h_i \triangleq h(x_i)$ according to the dual form formulas (2.1.10c,d), i.e.,

$$\theta_i = -\min_{\mu \in \Sigma_q} \{ \sum_{j=1}^{q} \mu^j [\psi(x_i) - f^j(x_i)] + \frac{1}{2\delta} \| \sum_{j=1}^{q} \mu^j \nabla f^j(x_i) \|^2 \}, \tag{3a}$$

and

$$h_i = -\frac{1}{\delta} \sum_{j=1}^{q} \mu_x^j \nabla f^j(x_i), \tag{3b}$$

where μ_x is any solution of (3a).

Step 2. If $\theta_i = 0$, stop. Else, compute the *step-size*

$$\lambda_i = \lambda(x_i) \triangleq \arg\max_{k \in \mathbb{N}} \{ \beta^k \mid \psi(x_i + \beta^k h_i) - \psi(x_i) - \beta^k \alpha \theta_i \leq 0 \}. \tag{3c}$$

Step 3. Set

$$x_{i+1} = x_i + \lambda_i h_i, \tag{3d}$$

replace i by $i + 1$, and go to Step 1.

Theorem 2.4.2. *Consider the problem* **MMP**, *defined in (1) and (2), with the assumptions stated. Suppose that Algorithm 2.4.1 constructs a sequence* $\{x_i\}_{i=0}^{\infty}$ *in solving* **MMP**. *Then every accumulation point* $\hat{x}$ *of* $\{x_i\}_{i=0}^{\infty}$ *satisfies the first-order optimality condition* $\theta(\hat{x}) = 0$.

Proof. Clearly, Algorithm 2.4.1 has the form of Algorithm Model 1.2.7. We will show that it satisfies the assumptions of Corollary 1.2.9 (and hence of Theorem 1.2.8), provided we define the set of stationary points by $S \triangleq \{x \in \mathbb{R}^n \mid \theta(x) = 0\}$ and the algorithm function by $a(x) \triangleq x + \lambda(x) h(x)$, with $h(\cdot)$ defined by (2.1.10d) and $\lambda(\cdot)$ defined by (3c).

Thus, suppose that $x^* \in \mathbb{R}^n$ is such that $\theta(x^*) < 0$ (recall that $\theta(x) \leq 0$ for all $x \in \mathbb{R}^n$). Then, by Theorem 2.1.6*(b)*,

$$d\psi(x^*, h(x^*)) \leq \theta(x^*) - \tfrac{1}{2}\delta \|h(x^*)\|^2 < 0. \tag{4a}$$

Let $\varepsilon > 0$ be such that $0 < \alpha - \varepsilon < 1$. Then it follows from the definition of the

directional derivative of $\psi(\cdot)$ that there exists a $k_\varepsilon \in \mathbb{N}$ such that

$$\psi(x^* + \beta^{k_\varepsilon} h(x^*)) - \psi(x^*) \le \beta^{k_\varepsilon}(\alpha - \varepsilon) d\psi(x^*; h(x^*))$$

$$\le \beta^{k_\varepsilon}(\alpha - \varepsilon)[\theta(x^*) - \tfrac{1}{2}\delta \|h(x^*)\|^2]. \quad (4b)$$

Hence

$$\psi(x^* + \beta^{k_\varepsilon} h(x^*)) - \psi(x^*) - \beta^{k_\varepsilon}\alpha\theta(x^*) \le -\beta^{k_\varepsilon}[\varepsilon\theta(x^*) + \tfrac{1}{2}\delta(\alpha - \varepsilon)\|h(x^*)\|^2]. \quad (4c)$$

Now,

$$\varepsilon\theta(x^*) + \tfrac{1}{2}\delta(\alpha - \varepsilon)\|h(x^*)\|^2 > 0, \quad (4d)$$

for all $\varepsilon > 0$ such that $2\varepsilon/\delta(\alpha - \varepsilon) < -\|h(x^*)\|^2/\theta(x^*) \triangleq \omega^*$, i.e., for all $\varepsilon > 0$ such that $\varepsilon < \varepsilon' \triangleq \delta\omega^*\alpha/2(2 + \delta\omega^*)$. Let $\varepsilon^* \triangleq \tfrac{1}{2}\varepsilon'$. Then, since $\psi(\cdot)$, $h(\cdot)$ and $\theta(\cdot)$ are all continuous, there exists a $\rho > 0$ such that, for all $x \in B(x^*, \rho)$,

$$\psi(x + \beta^{k_{\varepsilon^*}} h(x)) - \psi(x) - \beta^{k_{\varepsilon^*}}\alpha\theta(x) \le 0, \quad (4e)$$

which shows that for all $x \in B(x^*, \rho)$, $\lambda(x) \ge \beta^{k_{\varepsilon^*}}$. Next, since $\theta(\cdot)$ is continuous, there exists a $\rho^* \in (0, \rho)$ such that, for all $x \in B(x^*, \rho^*)$, $\theta(x) \le \theta(x^*)/2$. Therefore, it follows from the step-size rule (3c) that for all $x \in B(x^*, \rho^*)$,

$$\psi(a(x)) - \psi(x) \le \beta^{k_{\varepsilon^*}}\alpha\theta(x) \le \tfrac{1}{2}\beta^{k_{\varepsilon^*}}\alpha\theta(x^*), \quad (4f)$$

which shows that (1.2.7) (as well as (1.2.9)) holds, and hence our proof is complete. □

Note that the above proof is invalid for $\alpha > 1$.

2.4.2 Rate of Convergence of the PPP Algorithm

We will now show that the rate of convergence of the PPP Algorithm 2.4.1 is similar to that of the Armijo Gradient Algorithm 1.3.3. We will need an assumption that generalizes Assumption 1.3.5, which was used in establishing the rate of convergence of the Armijo Gradient Algorithm 1.3.3, i.e.,

Assumption 2.4.3. *In (2), the functions $f^j(\cdot)$, $j \in \mathbf{q}$, are twice continuously differentiable, and there exist $0 < m \le M < \infty$ such that*

$$m\|y\|^2 \le \langle y, f^j_{xx}(x)y \rangle \le M\|y\|^2, \quad \forall j \in \mathbf{q}, x, y \in \mathbb{R}^n. \quad (5)$$

□

When Assumption 2.4.3 is satisfied, one can obtain a $\delta \in [m, M]$ for the PPP algorithm by using the relation (1.3.19), as explained at the end of Section 1.3. Furthermore, under Assumption 2.4.3, the function $\psi(x) \triangleq \max_{j \in \mathbf{q}} f^j(x)$ is *strictly convex*, and hence it has a unique minimizer $\hat{x}$. For any $x \in \mathbb{R}^n$, Assumption 2.4.3 enables us to get a useful estimate of the quantity $\psi(\hat{x}) - \psi(x)$, as we will now see.

Lemma 2.4.4. *Suppose that Assumption 2.4.3 is satisfied and that $\delta \in [m, M]$. Let $\hat{x}$ be the unique minimizer of (1). Then, for any $x \in \mathbb{R}^n$,*

$$\psi(\hat{x}) - \psi(x) \ge \frac{\delta}{m}\theta(x). \quad (6)$$

Proof. From the second-order expansion formula (5.1.17d), Assumption 2.4.3, and the fact that $f^j(x) - \psi(x) \le 0$, it follows that

$$\psi(x') - \psi(x) \ge \max_{j \in \mathbf{q}} \{f^j(x') - \psi(x) + \langle \nabla f^j(x), x' - x \rangle + \tfrac{1}{2}m\|x' - x\|^2\}$$

$$\ge \frac{\delta}{m} \max_{j \in \mathbf{q}} \{f^j(x) - \psi(x) + \langle \nabla f^j(x), \frac{m}{\delta}(x' - x)\rangle$$

$$+ \tfrac{1}{2}\delta\|\frac{m}{\delta}(x' - x)\|^2\}. \quad (7)$$

Minimizing first the right side of (7) with respect to x' and then the left side of (7) with respect to x', we obtain (6). □

Theorem 2.4.5. *Suppose that Assumption 2.4.3 is satisfied and that $\delta \in [m, M]$. If the PPP Algorithm 2.4.1 constructs a sequence $\{x_i\}_{i=0}^{\infty}$. Then,*

(a) $x_i \to \hat{x}$ *as* $i \to \infty$,

(b)
$$[\psi(x_{i+1}) - \psi(\hat{x})] \le c[\psi(x_i) - \psi(\hat{x})], \quad \forall i \in \mathbf{N}, \quad (8a)$$

where

$$c \triangleq 1 - \frac{\beta\alpha m}{M} < 1, \quad (8b)$$

and

(c) there exists a constant $K < \infty$ such that

$$\|x_i - \hat{x}\| \le K(c^{1/2})^i \quad \forall i \in \mathbf{N}. \quad (8c)$$

Proof. *(a)* Because each $f^j(\cdot)$ is strictly convex under Assumption 2.4.3, it follows, from Proposition 5.2.15, that their level sets are convex and compact. Since the level sets of $\psi(\cdot)$ are the intersection of the corresponding level sets of the functions $f^j(\cdot)$, $j \in \mathbf{q}$, it follows that the level set $L \triangleq \{x \in \mathbb{R}^n \mid \psi(x) \le \psi(x_0)\}$ is convex and compact. Hence the sequence $\{x_i\}_{i=0}^{\infty}$ must have accumulation points $\hat{x}$, all of which satisfy $0 \in \partial\psi(\hat{x})$, by Theorem 2.4.2. Since $\psi(\cdot)$ is strictly convex, it follows that $\hat{x} = \arg\min_{x \in \mathbb{R}^n} \psi(x)$ is unique. Therefore $x_i \to \hat{x}$, as $i \to \infty$.

(b) We begin by obtaining a bound on the decrease in $\psi(x)$ at iteration i. From the second-order expansion formula (5.1.17d) and Assumption 2.4.3, it follows that for all $\lambda \in [0,1]$,

$$\begin{aligned}\psi(x_i+\lambda h_i)-\psi(x_i) &= \max_{j\in\mathbf{q}} f^j(x_i+\lambda h_i)-\psi(x_i)\\ &\le \max_{j\in\mathbf{q}} \{f^j(x_i)-\psi(x_i)+\langle \nabla f^j(x_i), \lambda h_i\rangle + \tfrac{1}{2}M\lambda^2\|h_i\|^2\}\\ &\le \lambda \max_{j\in\mathbf{q}} \{f^j(x_i)-\psi(x_i)+\langle \nabla f^j(x_i), h_i\rangle + \tfrac{1}{2}M\lambda\|h_i\|^2\}, \end{aligned} \tag{9a}$$

because $\lambda \in [0,1]$ and $f^j(x_i) \le \psi(x_i)$. Therefore, if $\lambda \le \delta/M \le 1$, then

$$\begin{aligned}\psi(x_i+\lambda h_i)-\psi(x_i) &\le \lambda \max_{j\in\mathbf{q}} \{f^j(x_i)-\psi(x_i)+\langle \nabla f^j(x_i), h_i\rangle + \tfrac{1}{2}\delta\|h_i\|^2\}\\ &= \lambda\theta(x_i). \end{aligned} \tag{9b}$$

Consequently, for $\lambda \le \delta/M$,

$$\psi(x_i+\lambda h_i)-\psi(x_i)-\lambda\alpha\theta(x_i) \le \lambda\theta(x_i)(1-\alpha),$$

which implies that $\lambda_i \ge \beta\delta/M$. Thus

$$\psi(x_{i+1})-\psi(x_i) \le \frac{\beta\delta\alpha}{M}\theta(x_i). \tag{9c}$$

Next, from (6), it follows that

$$\theta(x_i) \le \frac{m}{\delta}[\psi(\hat{x})-\psi(x_i)]. \tag{9d}$$

Combining (9d) with (9c) yields

$$\psi(x_{i+1})-\psi(x_i) \le \frac{\beta\alpha m}{M}[\psi(\hat{x})-\psi(x_i)]. \tag{9e}$$

Subtracting $\psi(\hat{x})-\psi(x_i)$ from both sides of (9e) and rearranging terms, we obtain (8a).

(c) First, from (8a), it follows that

$$\psi(x_i)-\psi(\hat{x}) \le [\psi(x_0)-\psi(\hat{x})]\,c^i. \tag{9f}$$

Next, from the second-order expansion formula (5.1.17d) and Assumption 2.4.3, it follows that for all $i \in \mathbb{N}$ and $j \in \mathbf{q}$,

$$f^j(x_i)-\psi(\hat{x}) \ge f^j(\hat{x})-\psi(\hat{x})+\langle \nabla f^j(\hat{x}), x_i-\hat{x}\rangle + \tfrac{1}{2}m\|x_i-\hat{x}\|^2. \tag{9g}$$

Let $\hat{\mu} \in \Sigma_q$ be a Danskin-Demyanov multiplier for $\hat{x}$, i.e., it satisfies (2.1.7b,c). Then, from (2.1.7b,c) and (9g), it follows that

$$\psi(x_i)-\psi(\hat{x}) \ge \sum_{j\in\mathbf{q}} \hat{\mu}^j[f^j(\hat{x})-\psi(\hat{x})] \ge \tfrac{1}{2}m\|x_i-\hat{x}\|^2. \tag{9h}$$

The relation (8c) now follows directly from (8a) and (9g). □

Note that (8b) indicates that the best choice for α is $\alpha = 1$. Also note that, when Assumption 2.4.3 holds and $\delta \notin [m, M]$ is chosen, then the effect on the rate of convergence of Algorithm 2.4.1 is that in (8b), m must be replaced by $\min\{\delta, m\}$ and M must be replaced by $\max\{\delta, M\}$. Hence the rate constant c will increase, a rather undesirable effect.

2.4.3 Algorithms for Search Direction Computation

Before we can go on much further, we must discuss methods for computing the search direction $h(x)$ and the optimality function $\theta(x)$. First, observe that the dual form for $\theta(x)$ in (2.1.10c) is a standard quadratic program (QP) and hence our first instinct is to try to solve it by commercially available QP code and then to obtain the search direction $h(x)$ using (2.1.10d). Now, in (2.1.10c), the term $\|\sum_{j=1}^q \mu^j \nabla f^j(x)\|^2$ can be expressed in the more compact form $\langle \mu, Q\mu\rangle$, where the matrix Q is defined by

$$Q \triangleq \sum_{j=1}^{q} \nabla f^j(x)\,\nabla f^j(x)^T. \tag{10}$$

Since this matrix is often only positive-semidefinite, standard quadratic programming codes occasionally fail to solve (2.1.10c). Because of this, it is preferable to use an enhanced version of the Frank-Wolfe algorithm [FrW.56], such as the one described in [HiP.90]. The Frank-Wolfe algorithm, the algorithm in [HiP.90], and its precursors (see notes) all fall into the category of *generalized support function algorithms*. The advantages of generalized support function algorithms in solving (2.1.10c) will become apparent shortly. Since the Frank-Wolfe algorithm is much simpler than the enhanced versions, we will present it now to illustrate the generic behavior of this class of algorithms and to exhibit the reasons for the enhancements. The reader will also recognize the Frank-Wolfe algorithm as a precursor of the Armijo Projected Gradient Algorithm 1.3.16.

However, before we proceed any further, we must explain what we mean by a generalized support function.

Definition 2.4.6. *Let $S \subset \mathbb{R}^n$ be a convex, compact set, and let $g: \mathbb{R}^n \to \mathbb{R}^n$ be a continuous function. We will call the function $\sigma: \mathbb{R}^n \to \mathbb{R}$, defined by*

$$\sigma(x) \triangleq \max\{\langle -g(x), y\rangle \mid y \in S\}$$

$$= -\min\{\langle g(x), y\rangle \mid y \in S\}, \tag{11}$$

a generalized support function. □

Note that if $\sigma(\cdot)$ is defined as in (11) and $y_x \in S$ is such that $\sigma(x) = \langle g(x), y_x\rangle$, then $\{y \in \mathbb{R}^n \mid \langle g(x), y - y_x\rangle = 0\}$ is a support hyperplane for the set S, with $g(x)$ an inward pointing normal (relative to S).

In applying generalized support function algorithms to the evaluation of the optimality function $\theta(\cdot)$ and the search direction function $h(\cdot)$, we must use the equivalent forms of their definitions given in (2.1.10e) and (2.1.10f), respectively. For the sake of simplicity, suppose that $\delta = 1$, in which case the minimization problem (2.1.10e) has the form

$$\min\{\gamma(\bar{\xi}) \mid \bar{\xi} \in \bar{G}\}, \tag{12}$$

where $\bar{\xi} \triangleq (\xi^0, \xi) \in \mathbb{R}^{n+1}$ and

$$\gamma(\bar{\xi}) \triangleq \xi^0 + \tfrac{1}{2}|\xi|^2, \tag{13}$$

and $\bar{G} \triangleq \bar{G}(x)$ is a convex, compact subset of $\mathbb{R}^{n+1}$. Recall that it follows from Theorem 2.1.6*(f)* that, although the function $\gamma(\cdot)$ is not strictly convex, problem (12) has a unique solution $\bar{\xi}_*$.

Exercise 2.4.7. Let $\Theta : \bar{G} \to \mathbb{R}$ be defined by

$$\Theta(\bar{\xi}) \triangleq \min\{\langle \nabla\gamma(\bar{\xi}), \bar{\zeta} - \bar{\xi}\rangle \mid \bar{\zeta} \in \bar{G}\}. \tag{14}$$

(a) Show that $-\Theta(\cdot)$ is a generalized support function for the set $\bar{G} - \{\bar{\xi}\}$.

(b) Show that $\bar{\xi}_*$ is a solution to problem (12) if and only if $\Theta(\bar{\xi}_*) = 0$, i.e., that $\Theta(\cdot)$ is an optimality function for problem (12). □

The Frank-Wolfe Algorithm 2.4.8.

Data. $\bar{\xi}_0 \in \bar{G}$.

Step 0. Set $i = 0$.

Step 1. Compute the support point $\bar{\zeta}_i$ in $\bar{G}$ according to

$$\bar{\zeta}_i \in \hat{G}(\bar{\xi}_i) \triangleq \arg\min\{\langle \nabla\gamma(\bar{\xi}_i), \bar{\zeta} - \bar{\zeta}_i\rangle \mid \bar{\zeta} \in \bar{G}\}, \tag{15a}$$

and set the *search direction* to be $\bar{\eta}_i = \bar{\zeta}_i - \bar{\xi}_i$. Stop if $\bar{\eta}_i = 0$.

Step 2. Compute the *step-length*

$$\lambda_i = \lambda(\bar{\xi}_i) \triangleq \arg\min\{\gamma(\bar{\xi}_i + \lambda\bar{\eta}_i) \mid \lambda \in [0,1]\}. \tag{15b}$$

Step 3. Update: Set

$$\bar{\xi}_{i+1} = \bar{\xi}_i + \lambda_i\bar{\eta}_i, \tag{15c}$$

replace i by $i+1$, and go to Step 1.

It turns out that we can establish both the convergence and the rate of convergence of the Frank-Wolfe algorithm within a single theorem whose proof is quite similar to that of Theorem 2.4.5.

Theorem 2.4.9. *Suppose that $\{\bar{\xi}_i\}_{i=0}^{\infty}$ is an infinite sequence constructed by the Frank-Wolfe algorithm in solving (12). Let $\bar{\xi}_*$ denote the solution to this problem. Then, (a) $\bar{\xi}_i \to \bar{\xi}_*$, as $i \to \infty$, and (b) $\gamma(\bar{\xi}_i) - \gamma(\bar{\xi}_*) \to 0$, as $i \to \infty$, at least as fast as an arithmetic progression.*

Proof. Since the largest eigenvalue of the second-derivative matrix $\gamma_{\bar{\xi}\bar{\xi}}(\bar{\xi})$ is 1,

$$\gamma(\bar{\xi}_{i+1}) - \gamma(\bar{\xi}_i) = \min\{\gamma(\bar{\xi}_i + \lambda\bar{\eta}_i) - \gamma(\bar{\xi}_i) \mid \lambda \in [0,1]\}$$

$$\leq \min\{\lambda\langle \nabla\gamma(\bar{\xi}_i), \bar{\eta}_i\rangle + \frac{\lambda^2}{2}|\bar{\eta}_i|^2 \mid \lambda \in [0,1]\}. \tag{16a}$$

Let λ_i^* be the solution of the right-hand side of (16a). Then

$$\lambda_i^* = -\frac{\langle \nabla\gamma(\bar{\xi}_i), \bar{\eta}_i\rangle}{|\bar{\eta}_i|^2} = -\frac{\Theta(\bar{\xi}_i)}{|\bar{\eta}_i|^2}, \tag{16b}$$

where $\Theta(\bar{\xi}_i)$ is defined as in (14). Hence,

$$\gamma(\bar{\xi}_{i+1}) - \gamma(\bar{\xi}_i) \leq \lambda_i^*\langle \nabla\gamma(\bar{\xi}_i), \bar{\eta}_i\rangle + \frac{(\lambda_i^*)^2}{2}|\bar{\eta}_i|^2$$

$$= -\frac{\Theta(\bar{\xi}_i)^2}{2|\bar{\eta}_i|^2} \leq -\frac{1}{2b}\Theta(\bar{\xi}_i)^2, \tag{16c}$$

where $b \triangleq \max_{\bar{\xi}, \bar{\xi}' \in \bar{G}} |\bar{\xi} - \bar{\xi}'|^2$.

Since $\gamma(\cdot)$ is convex, it follows, from Theorem 5.2.12, that

$$\gamma(\bar{\xi}_*) - \gamma(\bar{\xi}_i) \geq \langle \nabla\gamma(\bar{\xi}_i), \bar{\xi}_* - \bar{\xi}_i\rangle \geq \Theta(\bar{\xi}_i). \tag{16d}$$

Next, it follows from (16c) that

$$\gamma(\bar{\xi}_{i+1}) - \gamma(\bar{\xi}_*) \leq -\frac{1}{2b}\Theta(\bar{\xi}_i)^2 + \gamma(\bar{\xi}_i) - \gamma(\bar{\xi}_*). \tag{16e}$$

Since $\gamma(\bar{\xi}_*) - \gamma(\bar{\xi}_i) \leq 0$, it follows from (16d) that

$$-\Theta(\bar{\xi}_i)^2 \le -[\gamma(\bar{\xi}_*) - \gamma(\bar{\xi}_i)]^2 . \tag{16f}$$

Combining (16e) and (16f), we obtain

$$\gamma(\bar{\xi}_{i+1}) - \gamma(\bar{\xi}_*) \le \gamma(\bar{\xi}_i) - \gamma(\bar{\xi}_*) - \frac{1}{2b}[\gamma(\bar{\xi}_i) - \gamma(\bar{\xi}_*)]^2$$

$$= [\gamma(\bar{\xi}_i) - \gamma(\bar{\xi}_*)]\{1 - \frac{1}{2b}[\gamma(\bar{\xi}_i) - \gamma(\bar{\xi}_*)]\} . \tag{16g}$$

Let $e_i \triangleq \gamma(\bar{\xi}_i) - \gamma(\bar{\xi}_*)$, then we can rewrite (16g) as

$$e_{i+1} \le e_i(1 - \frac{1}{2b}e_i) . \tag{16h}$$

Next we normalize (16h). Defining $\mu_i = e_i / 2b$, we obtain that

$$\mu_{i+1} \le \mu_i(1 - \mu_i) , \ \forall\, i \in \mathbf{N} . \tag{16i}$$

Since $\mu_i \ge 0$ for all $i \in \mathbf{N}$, it follows from (16i) that $0 \le \mu_i \le 1$. Furthermore, since the sequence constructed by the Frank-Wolfe algorithm is infinite by assumption, it follows that $\mu_i > 0$ for all $i \in \mathbf{N}$. Next, it follows from (16i) that $\mu_i < 1$ for all $i \in \mathbf{N}$, for, if $\mu_i = 1$, it would follow from (16i) that $\mu_{i+1} = 0$, a contradiction. Hence, using the binomial expansion, we find that

$$\frac{1}{\mu_{i+1}} \ge \frac{1}{\mu_i(1-\mu_i)} \ge \frac{1}{\mu_i}(1+\mu_i) = 1 + \frac{1}{\mu_i} , \ \forall\, i \in \mathbf{N} . \tag{16j}$$

By induction, we conclude that

$$\frac{1}{\mu_i} \ge \frac{1}{\mu_0} + i \ge i , \ \forall\, i \in \mathbf{N} . \tag{16k}$$

Equivalently,

$$e_i \le \frac{e_0}{1 + (e_0/2b)i} , \ \forall\, i \in \mathbf{N} , \tag{16l}$$

which shows that the cost error converges to zero at least as fast as an arithmetic progression. In fact, in the worst case, it converges no faster than an arithmetic progression (see [CaC.68]).

Since the set $\bar{G}$ is compact, the sequence $\{\bar{\xi}_i\}_{i=0}^{\infty}$ must have accumulation points. Since at every accumulation point $\bar{\xi}_{**}$ we must have that $\gamma(\bar{\xi}_{**}) = \gamma(\bar{\xi}_*)$, and since problem (12) has a unique solution $\bar{\xi}_*$, we conclude that $\bar{\xi}_i \to \bar{\xi}_*$, as $i \to \infty$. □

Algorithm 2.4.8 can be used only when the computation of $\bar{\zeta}_i \in \hat{G}(\bar{\xi}_i)$ in (15a) can be carried out in a reasonable manner. In the case of search direction finding problems, such as (2.1.10e), this is certainly true because of the following fact.

Exercise 2.4.10. Let $\{\bar{\xi}_j\}_{j=1}^{q}$ be a collection of vectors in $\mathbb{R}^n$, let $\bar{g} \in \mathbb{R}^n$ be given, and let $\bar{G} \triangleq \text{co}\{\bar{\xi}_j\}_{j=1}^{q}$. Show that

$$\min_{\bar{\xi} \in \bar{G}} \langle \bar{g}, \bar{\xi} \rangle = \min_{j \in \mathbf{q}} \langle \bar{g}, \bar{\xi}_j \rangle . \tag{17}$$

□

Since, in the case of (2.1.10g), $\bar{G} = \bar{G}(x)$ is defined by

$$\bar{G}(x) = \underset{j \in \mathbf{q}}{\text{co}} \left\{ \begin{bmatrix} \psi(x) - f^j(x) \\ \nabla f^j(x) \end{bmatrix} \right\} , \tag{18a}$$

it is clear from Exercise 2.4.10 that if for all $j \in \mathbf{q}$, we define

$$\bar{\xi}_j^* \triangleq \begin{bmatrix} (\psi(x) - f^j(x)) \\ \nabla f^j(x) \end{bmatrix} \in \mathbb{R}^{n+1} , \tag{18b}$$

then for any $\bar{\xi} = (\xi^0, \xi) \in \bar{G}$ and $\bar{g} = (g^0, g) \triangleq \nabla\gamma(\bar{\xi})$,

$$\Theta(\bar{\xi}) = \min_{j \in \mathbf{q}} \langle \bar{g}, \bar{\xi}_j^* - \bar{\xi} \rangle$$

$$= \min_{j \in \mathbf{q}} \{ g^0(\psi(x) - f^j(x) - \xi^0) + \langle g, \nabla f^j(x) - \xi \rangle \} . \tag{18c}$$

Next, let

$$\hat{\mathbf{q}}_\Theta(\bar{\xi}) \triangleq \{ j \in \mathbf{q} \mid \Theta(\bar{\xi}) = g^0(\psi(x) - f^j(x) - \xi^0) + \langle g, \nabla f^j(x) - \xi \rangle \} . \tag{18d}$$

Then it should be clear that, for any $j \in \hat{\mathbf{q}}_\Theta(\bar{\xi})$, the vector

$$\bar{\zeta}(\bar{\xi}) = (\psi(x) - f^j(x), \nabla f^j(x)) \tag{18e}$$

is an element of the set $\hat{G}(\bar{\xi}_i)$. Thus, the computation of a $\bar{\zeta}_i \in \hat{G}(\bar{\xi}_i)$ requires only q inner product computations and the identification of the the $j \in \mathbf{q}$ which minimizes $\gamma(\bar{\xi}_j^*)$.

It should be obvious that there may be indices $j \in \mathbf{q}$ which never appear in the sets $\hat{\mathbf{q}}_\Theta(\bar{\xi}_i)$, used by the Frank-Wolfe Algorithm 2.4.8 in computing the search direction $\bar{\eta}_i$. Hence the computation of the corresponding vectors $\nabla f^j(x)$ is wasted. This waste can be avoided if we note that $\langle \nabla f^j(x), g \rangle = df^j(x; g) \simeq \Delta_j \triangleq (f^j(x + tg) - f^j(x))/t$, with $t > 0$ small. Hence we can identify the set $\hat{\mathbf{q}}_\Theta(\bar{\xi})$ by finding the smallest of the numbers $g^0[\psi(x) - f^j(x) - \xi^0] + \Delta_j - \langle g, \xi \rangle$, $j \in \mathbf{q}$. Once the set $\hat{\mathbf{q}}_\Theta(\bar{\xi})$ is identified, only one gradient $\nabla f^j(x)$ needs to be computed to complete the construction of $\bar{\zeta}_i \in \hat{G}(\bar{\xi}_i)$. Of course, for any j such that $\nabla f^j(x)$ was computed, the exact formula for Δ_j should be used.

To complete this section, we will present the enhanced version of the Frank-Wolfe algorithm described in [HiP.90] (which differs from the algorithm in [vHo.3] only by the addition of a guard step) for solving problem (12) under the additional assumption that the set $\bar{G}$ is a *polytope*, i.e., we are given a set of

vectors $\bar{X} = \{\bar{\xi}_j^*\}_{j=1}^q$ in $\mathbb{R}^{n+1}$ and

$$\bar{G} \triangleq \operatorname*{co}_{j \in \mathbf{q}} \{\bar{\xi}_j^*\}. \tag{19}$$

The original Frank-Wolfe algorithm is a *two-point algorithm* in the sense that it uses only the convex hull of two points $\bar{\xi}_i$, $\bar{\zeta}_i \in \bar{G}$ to compute $\bar{\xi}_{i+1}$, while the algorithm below is a *multi-point algorithm* in the sense that, in the construction of $\bar{\xi}_{i+1}$, it uses not only the points used by the Frank-Wolfe algorithm, but also a subset of previously constructed points. To be more precise, to construct $\bar{\xi}_{i+1}$, it uses a set of the form $\bar{B}_{i+1} = \{\bar{b}_{i+1,1}, \ldots, \bar{b}_{i+1,p_{i+1}}\} \subset \{\bar{\xi}_0, \bar{\xi}_1, \ldots, \bar{\xi}_i, \bar{\zeta}_i\}$ (with $\bar{\zeta}_i$ defined by (15a)) of $p_{i+1} \leq n+1$ affinely independent vectors.

We will say that a point $\bar{\xi}$ is in the *relative interior* of $\operatorname{co}\bar{X}$ (which we write as $\bar{\xi} \in \operatorname{ri}\operatorname{co}\bar{X}$), if $\bar{\xi} = \sum_{j=1}^p \mu^j \bar{\xi}_j$, with $\sum_{j=1}^p \mu^j = 1$, and all the $\mu^j > 0$.

Higgins-Polak Algorithm 2.4.11.

Data. $\bar{G} = \operatorname{co}\{\bar{\xi}_j^*\}_{j=1}^q \subset \mathbb{R}^{n+1}$, $\bar{\xi}_0 \in \bar{G}$.

Step 0. Set $i = 0$, set $\bar{b}_{0,1} = \bar{\xi}_0$, and set $\bar{B}_0 = \{\bar{b}_{0,1}\}$.

Step 1. If $\Theta(\bar{\xi}_i) = 0$, stop ($\bar{\xi}_i$ solves (12)).

Else, compute a $\bar{\zeta}_i \in \hat{G}(\bar{\xi}_i)$ according to (15a).

Step 2. Compute a $\bar{\xi}_{i+1} \in \bar{G}$ and a set $\bar{B}_{i+1} = \{\bar{b}_{i+1,j}\}_{j=1}^{p_{i+1}} \subset \bar{B}_i \cup \{\bar{\zeta}_i\}$ satisfying the following conditions:

(i) $\gamma(\bar{\xi}_{i+1}) \leq \min\{\gamma(\lambda\bar{\xi}_i + (1-\lambda)\bar{\zeta}_i) \mid \lambda \in [0,1]\}$,

(ii) the set of vectors $\bar{B}_{i+1}$ is affinely independent,

(iii) $\bar{\xi}_{i+1} \in \operatorname{ri}\operatorname{co}\bar{B}_{i+1}$,

(iv) $\bar{\xi}_{i+1}$ minimizes $\gamma(\cdot)$ on $\operatorname{co}\bar{B}_{i+1}$.

Step 3. Replace i by $i+1$, and go to Step 1.

Step 2 in the Algorithm 2.4.11 is rather complex and requires a special subprocedure which will be presented shortly. The statement, as well as the explanation, of the subprocedure for implementing Step 2 of Algorithm 2.4.11 are simplified if we use the following terminology. Given an affinely independent set of vectors $\bar{X} = \{\bar{\xi}_j\}_{j=1}^p$ in $\mathbb{R}^{n+1}$ and a vector $\bar{\xi} \in \operatorname{aff}\bar{X}$, we know by Exercise 2.1.9 that there exist unique multipliers $\{\alpha^j\}_{j \in \mathbf{p}}$ such that $\sum_{j=1}^p \alpha^j = 1$ and $\bar{\xi} = \sum_{j=1}^p \alpha^j \bar{\xi}_j$. We will refer to these multiplers as the *barycentric coordinates* of $\bar{\xi}$, with respect to $\bar{X}$, and we will say that a vector $\bar{\xi}j \in \bar{X}$ *carries* $\bar{\xi}$ if $\alpha^j \neq 0$.

The construction in Step 2 of Algorithm 2.4.11 is well defined. For example, one may choose

$$\bar{\xi}_{i+1} \in \arg\min_{\bar{\xi} \in \bar{B}} \gamma(\bar{\xi}),$$

with $\bar{B} = \bar{B}_i \cup \{\bar{\zeta}_i\}$, and let $\bar{B}_{i+1}$ be the set of points in $\bar{B}$ that carry $\bar{\xi}_{i+1}$. Obviously, this procedure is not practical. Later, we will give an efficient procedure for carrying out the construction in Step 2.

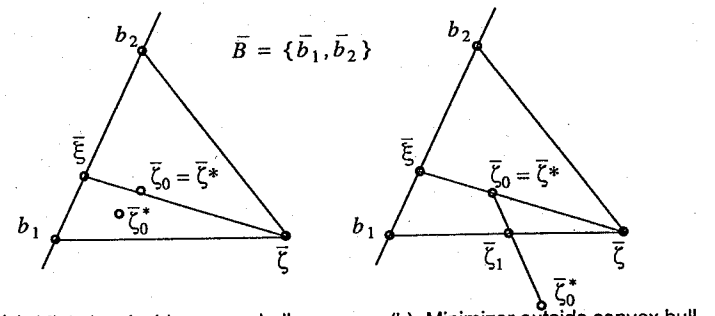
Theorem 2.4.12. *For $\bar{G}$ defined as in (19), Algorithm 2.4.11 solves problem (12) in a finite number of iterations.*

Proof. Obviously, if the algorithm terminates, it is at a solution. Suppose that, for some i, $\Theta(\bar{\xi}_i) < 0$. Since $\Theta(\bar{\xi}_i) < 0$ and

$$\gamma(\bar{\xi}_{i+1}) \leq \min\{\gamma(\lambda\bar{\xi}_i + (1-\lambda)\bar{\zeta}_i) \mid \lambda \in [0,1]\},$$

it follows that $\gamma(\bar{\xi}_{i+1}) < \gamma(\bar{\xi}_i)$. Hence, since $\bar{\xi}_i$ minimizes $\gamma(\cdot)$ on $\operatorname{co}\bar{B}_i$ and $\bar{\xi}_{i+1}$ minimizes $\gamma(\cdot)$ on $\operatorname{co}\bar{B}_{i+1}$, it follows that $\bar{B}_{i+1} \neq \bar{B}_i$, and therefore, by induction, that $\bar{B}_{i+1} \neq \bar{B}_j$, for all $j \leq i$. Since, by assumption, $\bar{G}$ is finite, it contains only a finite number of distinct, affinely independent subsets. Because, by construction, each $\bar{B}_i \subset \bar{G}$ and the sets $\bar{B}_i$ are affinely independent, we conclude that the algorithm must terminate in a finite number of iterations. □

The idea underlying the subprocedure which implements Step 2 of Algorithm 2.4.11 is as follows. Suppose the subprocedure is passed a triple $(\bar{\xi}, \bar{B}, \bar{\zeta})$, where $\bar{B} = \{\bar{b}_1, \bar{b}_2, \ldots \bar{b}_p\}$ is a set of affinely independent vectors in $\mathbb{R}^{n+1}$, $\bar{\xi} \in \operatorname{ri}\operatorname{co}\bar{B}$, $\bar{\zeta} \in \mathbb{R}^{n+1}$ is a point such that $\bar{\zeta} \notin \operatorname{aff}\bar{B}$, and $\langle \nabla\gamma(\bar{\xi}), \bar{\zeta} - \bar{\xi}\rangle < 0$. To compute a pair $(\bar{\xi}^*, \bar{B}^*)$ satisfying conditions *(i)-(iv)* of Step 2, we proceed as follows. First, we obtain a point $\bar{\zeta}^* = \lambda^*\bar{\xi} + (1-\lambda^*)\bar{\zeta}$, where $\lambda^* \in \arg\min_{\lambda \in [0,1]} \gamma(\lambda\bar{\xi} + (1-\lambda)\bar{\zeta})$. If $\bar{\zeta}^* = \bar{\zeta}$, we return the pair $(\bar{\xi}^*, \bar{B}^*)$, where $\bar{\xi}^* = \bar{\zeta}^*$ and $\bar{B}^* = \{\bar{\zeta}^*\}$, which, clearly, satisfy conditions



(a) Minimizer inside convex hull (b) Minimizer outside convex hull

Fig. 2.4.1. Construction for Step 2 of Algorithm 2.4.11.

(i)-(iv) of Step 2 in Algorithm 2.4.11. If not, we set $\bar{\zeta}_0 \triangleq \bar{\zeta}^*$ and $\bar{B}_0^* \triangleq \bar{B} \cup \{\bar{\zeta}\}$. It follows from Exercise 2.1.8 that $\bar{B}_0^*$ is affinely independent and hence that any subset of $\bar{B}_0^*$ is also affinely independent. If a minimizer $\bar{\zeta}_0^*$ of $\gamma(\cdot)$ on $\operatorname{aff}\bar{B}_0^*$ exists and $\bar{\zeta}_0^* \in \operatorname{co}\bar{B}_0^*$ (as in Fig. 2.4.1a), then the subprocedure returns the pair $(\bar{\xi}^*, \bar{B}^*)$, where $\bar{\xi}^* = \bar{\zeta}_0^*$ and $\bar{B}^* \subset \bar{B}_0^*$ is the set of points which carry $\bar{\zeta}_0^*$. If no such minimizer exists or if the minimizer $\bar{\zeta}_0^*$ lies outside of $\operatorname{co}\bar{B}_0^*$, a point $\bar{\zeta}_0^* \in \operatorname{aff}\bar{B}_0^* \setminus \operatorname{co}\bar{B}_0^*$ is obtained such that $\gamma(\bar{\zeta}^*_0) \leq \gamma(\bar{\zeta}_0)$. Note that, because $\gamma(\cdot)$ is convex,

$$\gamma(\lambda\bar{\zeta}_0 + (1-\lambda)\bar{\zeta}^*_0) \leq \gamma(\bar{\zeta}_0)$$

for all $\lambda \in [0,1]$. Then the subprocedure defines $\bar{\zeta}(\lambda) \triangleq \bar{\zeta}^* + \lambda(\bar{\zeta}_0 - \bar{\zeta}_0^*)$ and (see Fig. 2.4.1b) computes the point $\bar{\zeta}_1 = \bar{\zeta}(\lambda^*)$, where

$$\lambda^* = \min\{\lambda \in [0,1] \mid \bar{\zeta}(\lambda) \in \operatorname{co}\bar{B}_0^*\},$$

on the relative boundary of $\operatorname{co}\bar{B}_0^*$. The set $\bar{B}_1^*$ is now formed by dropping all points in $\bar{B}_0^*$ which do not carry $\bar{\zeta}_1$. Since $\gamma(\bar{\zeta}_1) \leq \gamma(\bar{\zeta}_0)$, the pair $(\bar{\zeta}_1, \bar{B}_1^*)$ satisfies conditions *(i)-(iii)* of Step 2 in Algorithm 2.4.11. To satisfy condition *(iv)*, the above procedure is repeated, if necessary, starting with $(\bar{\zeta}_1, \bar{B}_1^*)$, until a suitable point is obtained. We will now state formally the subprocedure which implements Step 2 of Algorithm 2.4.11.

Subprocedure 2.4.13 (Implementation of Step 2 of Algorithm 2.4.11).

Data. $\bar{B}$, a finite, affinely independent set in $\mathbb{R}^{n+1}$, $\bar{\xi} \in \operatorname{ri}\operatorname{co}\bar{B}$, $\bar{\zeta} \in \mathbb{R}^{n+1}$ such that $\bar{\zeta} \notin \operatorname{aff}\bar{B}$ and $\langle \nabla\gamma(\bar{\xi}), \bar{\zeta} - \bar{\xi}\rangle < 0$.

Step 0. (*Guard step*) Compute $\lambda^* = \arg\min_{\lambda \in [0,1]} \gamma(\lambda\bar{\xi} + (1-\lambda)\bar{\zeta})$, and set $\bar{\zeta}^* = \lambda^*\bar{\xi} + (1-\lambda^*)\bar{\zeta}$.

Step 1. If $\bar{\zeta}^* = \bar{\zeta}$, set $\bar{\xi}^* \triangleq \bar{\zeta}^*$, $\bar{B}^* \triangleq \{\bar{\zeta}^*\}$, return the pair $(\bar{\xi}^*, \bar{B}^*)$, and stop.

Step 2. Set $i = 0$, $\bar{\zeta}_0 \triangleq \bar{\zeta}^*$, $\bar{B}_0^* \triangleq \bar{B} \cup \{\bar{\zeta}\}$.

Step 3. If $\bar{\zeta}_i^* \triangleq \arg\min\{\gamma(\bar{\zeta}) \mid \bar{\zeta} \in \operatorname{aff}\bar{B}_i^*\}$ exists, go to Step 4.
Else, go to Step 5.

Step 4. If $\bar{\zeta}_i^* \in \operatorname{co}\bar{B}_i^*$, $\bar{B}^* \subset \bar{B}_i$ to be the set of points which carry $\bar{\zeta}_i^*$, set $\bar{\xi}^* \triangleq \bar{\zeta}_i^*$, return the pair $(\bar{\xi}^*, \bar{B}^*)$, and stop.
Else, go to Step 6.

Step 5. (*No minimizer exists*) Obtain a point $\bar{\zeta}_i^*$, such that
- *(a)* $\bar{\zeta}_i^* \in \operatorname{aff}\bar{B}_i^*$,
- *(b)* $\bar{\zeta}_i^* \notin \operatorname{co}\bar{B}_i^*$, and
- *(c)* $\gamma(\bar{\zeta}_i^*) \leq \gamma(\bar{\zeta}_i)$.

Step 6. Define $\bar{\zeta}(\lambda) \triangleq \lambda\bar{\zeta}_i + (1-\lambda)\bar{\zeta}_i^*$, then
- *(a)* compute
$$\lambda^* = \min\{\lambda \in [0,1] \mid \bar{\zeta}(\lambda) \in \operatorname{co}\bar{B}_i^*\},$$
- *(b)* set $\bar{\zeta}_{i+1} = \bar{\zeta}(\lambda^*)$, and
- *(c)* set $\bar{B}_{i+1}^* \subset \bar{B}_i$ to be the set of points in $\bar{B}_i^*$ which carry $\bar{\zeta}_{i+1}$.

Step 7. Replace i by $i+1$, and go to Step 3.

Theorem 2.4.14. *Suppose that (i) $\bar{B}$ is a finite, affinely independent subset of $\mathbb{R}^{n+1}$, (ii) $\bar{\xi} \in \operatorname{ri}\operatorname{co}\bar{B}$, and (iii) $\bar{\zeta} \in \mathbb{R}^{n+1}$ is such that $\bar{\zeta} \notin \operatorname{aff}\bar{B}$ and $\langle \nabla\gamma(\bar{\xi}), \bar{\zeta} - \bar{\xi}\rangle < 0$. Then Subprocedure 2.4.13 will terminate in a finite number of iterations returning a pair $(\bar{\xi}^*, \bar{B}^*)$ satisfying the following conditions:*

(a) $\gamma(\bar{\xi}^*) \leq \min_{\lambda \in [0,1]} \gamma(\lambda\bar{\xi} + (1-\lambda)\bar{\zeta})$.

(b) $\bar{B}^*$ *is affinely independent.*

(c) $\bar{\xi}^* \in \operatorname{ri}\operatorname{co}\bar{B}^*$.

(d) $\bar{\xi}^*$ *minimizes* $\gamma(\cdot)$ *on* $\operatorname{co}\bar{B}^*$.

(e) $\bar{B}^* \subset \bar{G} \cup \{\bar{\zeta}\}$.

Proof. Since $\bar{B}$ is finite and Step 6 always removes at least one point from $\bar{B}_i$, it is clear that the algorithm terminates in a finite number of steps. In addition, Subprocedure 2.4.13 stops only in Steps 1 and 4, and, consequently, conditions *(c)* and *(d)* are satisfied. As noted earlier, the set $\bar{B} \cup \{\bar{\zeta}\}$ is affinely independent, and hence so is $\bar{B}^*$. By construction $\gamma(\bar{\zeta}_{i+1}) \leq \gamma(\bar{\zeta}_i)$, for all i, and $\gamma(\bar{\zeta}_0) = \min_{\lambda \in [0,1]} \gamma(\lambda\bar{\xi} + (1-\lambda\bar{\zeta})$, and so, condition *(a)* is satisfied. Condition *(e)* is trivially true. □

To conclude the discussion of generalized support function algorithms, we will present an efficient method for carrying out the operations required in Steps 3 and 6 of Subprocedure 2.4.13. We will not address the situation in Step 5, since it is virtually never encountered in practice. We begin with two facts relevant to Step 3.

Proposition 2.4.15. *Suppose that $\bar{B}$ is a finite set in $\mathbb{R}^n$. Then, either the problem $\min_{\bar{\xi} \in \operatorname{aff}\bar{B}} \gamma(\bar{\xi})$ has a unique solution, $\bar{\xi}^*$, or $\inf_{\bar{\xi} \in \operatorname{aff}\bar{B}} \gamma(\bar{\xi}) = -\infty$.*

Proof. Clearly, we need to show only that if a minimizer exists in $\operatorname{aff}\bar{B}$, then

it is unique. Thus, suppose that $\bar{\xi}_* = (\xi_*^0, \xi_*)$, and $\bar{\xi}_{**} = (\xi_{**}^0, \xi_{**})$ are two minimizers of $\gamma(\cdot)$ over $\operatorname{aff}\bar{B}$. First suppose that $\xi_* \neq \xi_{**}$. Then, for all $\lambda \in (0,1)$, $\lambda\bar{\xi}_* + (1-\lambda)\bar{\xi}_* \in \operatorname{aff}\bar{B}$, and

$$\gamma(\lambda\bar{\xi}_* + (1-\lambda)\bar{\xi}_{**}) = \lambda\xi_*^0 + (1-\lambda)\xi_{**}^0 + \tfrac{1}{2}\mathbf{l}\lambda\xi_* + (1-\lambda)\xi_{**}\mathbf{l}^2$$

$$< \lambda\gamma(\bar{\xi}_*) + (1-\lambda)\gamma(\bar{\xi}_{**}) = \gamma(\bar{\xi}_*), \tag{20}$$

which contradicts the optimality of $\bar{\xi}_*$. Hence we must have that $\xi_* = \xi_{**}$. Since $\gamma(\bar{\xi}_*) = \gamma(\bar{\xi}_{**})$ it follows that $\xi_*^0 = \xi_{**}^0$, which implies that $\bar{\xi}_* = \bar{\xi}_{**}$. Hence we conclude that, if a minimizer of $\gamma(\cdot)$ over $\operatorname{aff}\bar{B}$ exists, then it must be unique. □

Corollary 2.4.16. *Suppose that $\bar{B} = \{\bar{b}_j\}_{j=1}^p$ is a finite, affinely independent set in $\mathbb{R}^{n+1}$, with $\bar{b}_j = (b_j^0, b_j)$, $j \in \mathbf{p}$. Then the problem*

$$\min_{\bar{\xi} \in \operatorname{aff}\bar{B}} \gamma(\bar{\xi}) \tag{21a}$$

has a unique minimizer $\bar{\xi}^* = \bar{\mathbf{B}}\alpha^*$, where $\bar{\mathbf{B}}$ is an $n+1 \times p$ matrix whose jth column is $\bar{b}_j$ and $\alpha^* \in \mathbb{R}^p$. Furthermore, $\bar{\mathbf{B}}\alpha^*$ is the unique minimizer of (21a) if and only if (α^*, β^*), with $\beta^* \in \mathbb{R}$, is the unique solution of the equation

$$\begin{bmatrix} \mathbf{B}^T\mathbf{B} & e^T \\ e & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \mathbf{B}^0 \\ 1 \end{bmatrix}, \tag{21b}$$

where $\mathbf{B}$ is an $n \times p$ matrix whose jth column is b_j, $\mathbf{B}^0 = (b_1^0, b_2^0, \ldots, b_p^0)^T$ and $e = (1,1,1,\ldots,1)$ is a $1 \times p$ matrix.

Proof. First we rewrite (21a) in the equivalent form

$$\min_{\alpha \in \mathbb{R}^p} \{\gamma(\bar{\mathbf{B}}\alpha) \mid e\,\alpha = 1\}. \tag{22}$$

It should be obvious that the optimality conditions stated in Corollary 2.2.20 are both necessary and sufficient for this case. For problem (22), the condition (2.2.23b) assumes the form (21b), with β replacing ζ as the multiplier, to avoid notational confusion. Since it follows from Proposition 2.4.15 that the solution $\bar{\xi}^*$ to (21a) must be unique and since its barycentric coordinates α^{*j} with respect to the affinely independent set $\bar{B}$ must be unique, we conclude that (22) has a unique solution (α^*, β^*), if and only if (21a) has a unique solution $\bar{\xi}^*$ and $\bar{\xi}^* = \bar{\mathbf{B}}\alpha^*$. □

Thus we see from the above corollary that the computation of a minimizer of $\gamma(\cdot)$ on an affinely independent set $\bar{B}$ requires only the solution of the system of equations (21b).

Exercise 2.4.17. Show that (21b) has a unique solution (α^*, β^*) if and only if the vectors $\{b_j\}_{j=1}^p$ are affinely independent. □

Next we turn to the computations in Step 6 of Subprocedure 2.4.13. Let $\bar{\mathbf{B}}_i$ be a matrix whose columns are the p_i vectors in $\bar{B}_i$, let $e = (1,1,\ldots,1)$ be a $1 \times p_i$ row matrix, and suppose that $\bar{\xi}_i = \bar{\mathbf{B}}_i\,\mu_i$, with $\mu_i^j > 0$ for all j, and $\bar{\xi}_i^* \notin \operatorname{co}\bar{B}$ is such that $\bar{\xi}_i^* = \bar{\mathbf{B}}_i\,\nu_i$, where $e\,\mu_i = e\,,\nu_i = 1$, with $\nu_i^j < 0$ for some j (because $\xi_i^* \notin \operatorname{co}B_i$). Then a simple calculation shows that

$$\bar{\xi}_{i+1} = \bar{\mathbf{B}}_i(\mu_i + \lambda_i(\nu_i - \mu_i)), \tag{23a}$$

where

$$\lambda_i \triangleq \min_{j \in \mathbf{p}} \{\frac{\mu_i^j}{\mu_i^j - \nu_i^j} \mid \mu_i^j - \nu_i^j > 0\}. \tag{23b}$$

The set $\bar{B}_{i+1}$ is computed by dropping all points in $\bar{B}_i$ for which the associated multiplier $\mu_i^j + \lambda_i(\nu_i^j - \mu_i^j)$ is zero (at least one such point exists).

2.4.4 Quadratic Convergence to a Haar Point

To conclude the discussion of the PPP Algorithm 2.4.1, we will examine the special case where the PPP Algorithm 2.4.1 converges to a Haar point. We will see that the PPP Algorithm converges quadratically in this case. To simplify matters slightly, we will assume that $\delta = 1$ in the definitions of $\theta(\cdot)$ and $h(\cdot)$ in (2.1.10c) and (2.1.10d), respectively.

Definition 2.4.18. *We will say that a point $\hat{x} \in \mathbb{R}^n$ is a* Haar point *for the min-max problem* **MMP** in (1) and (2), if

(a) $0 \in \partial\psi(\hat{x})$,

(b) *the set $\hat{\mathbf{q}}(\hat{x})$ contains exactly $n+1$ indices,*

(c) *vectors $\{\nabla f^j(\hat{x})\}_{j \in \hat{\mathbf{q}}(\hat{x})}$ are affinely independent, and*

(d) *if $\hat{\mu} \in \Sigma_q$ is such that (2.1.7a,b) hold, then $\hat{\mu}^j > 0$ for all $j \in \hat{\mathbf{q}}(\hat{x})$.* □

We note that if $\hat{x} \in \mathbb{R}^n$ is a Haar point for the min-max problem (1), then there exists a unique multiplier $\hat{\mu} \in \Sigma_q$ such that (2.1.7b,c) are satisfied.

Lemma 2.4.19. *If $\hat{x}$ is a Haar point for the min-max problem* **MMP** *in (1) and (2), then $\hat{x}$ is an isolated local minimizer for* **MMP**.

Proof. We begin by showing that there exists a $\kappa > 0$ such that

$$d\psi(\hat{x}\,;h) \geq \kappa\mathbf{l}h\mathbf{l}, \tag{24a}$$

for all $h \in \mathbb{R}^n$. For suppose that this is not so. Then, for any sequence $\kappa_i \to 0$, as $i \to \infty$, with $\kappa_i > 0$, there exists a sequence of unit vectors h_i such that $d\psi(\hat{x}\,;h_i) < \kappa_i$. Since the unit sphere is compact, without loss of generality, we

can assume that $h_i \to \hat{h}$ as $i \to \infty$, with $\|\hat{h}\| = 1$ and hence, since $d\psi(\hat{x}\,;\cdot)$ is continuous, $d\psi(\hat{x}\,,\hat{h}) \le 0$. Now, by assumption, $d\psi(\hat{x}\,;h) \ge 0$ for all $h \in \mathbb{R}^n$, and therefore we must have that $d\psi(\hat{x}\,;\hat{h}) = 0$. Let $\hat{\mu} \in \Sigma_q$ be such that (2.1.7b,c) are satisfied, i.e., $\hat{\mu}^j = 0$ for all $j \notin \hat{\mathbf{q}}(\hat{x})$, and $\sum_{j=1}^q \hat{\mu}^j \nabla f^j(\hat{x}) = 0$. Hence

$$\sum_{j \in \hat{\mathbf{q}}(\hat{x})} \hat{\mu}^j \langle \nabla f^j(\hat{x}), \hat{h} \rangle = 0, \tag{24b}$$

and, in addition, because $d\psi(\hat{x}\,;\hat{h}) = 0$, $\langle \nabla f^{sp} - 0.2\nu j(\hat{x}), \hat{h} \rangle \le 0$ for all $j \in \hat{\mathbf{q}}(\hat{x})$. Since by definition of a Haar point, $\hat{\mu}^j > 0$ for all $j \in \hat{\mathbf{q}}(\hat{x})$, we conclude that

$$\langle \nabla f^j(\hat{x}), \hat{h} \rangle = 0 \tag{24c}$$

for all $j \in \hat{\mathbf{q}}(\hat{x})$. Now, $\hat{\mathbf{q}}(\hat{x}) = \{j_1, j_2, \ldots, j_{n+1}\}$ by assumption, i.e., it contains exactly $n+1$ indices, and the vectors $\{\nabla f^{j_k}(\hat{x})\}_{k=1}^{n+1}$ are affinely independent. Hence, by Proposition 2.1.10, the vectors $\{(\nabla f^{j_k}(\hat{x}) - \nabla f^{j_{n+1}}(\hat{x}))\}_{k=1}^{n}$ are linearly independent. However, (c) implies that

$$\langle [\nabla f^{j_k}(\hat{x}) - \nabla f^{j_{n+1}}(\hat{x})], \hat{h} \rangle = 0 \tag{24d}$$

for $k = 1, 2, \ldots, n$, which contradicts the linear independence of the vectors $\{(\nabla f^{j_k}(\hat{x}) - \nabla f^{j_{n+1}}(\hat{x}))\}_{k=1}^{n}$. Hence there is no $\hat{h} \neq 0$ such that $d\psi(\hat{x}\,;\hat{h}) = 0$, i.e., (24a) must hold with $\kappa > 0$.

Next, by continuity of the gradients, there exists a $\rho > 0$ such that for all $x \in B(\hat{x}\,,\rho)$,

$$\max_{j \in \hat{\mathbf{q}}(\hat{x})} \|\nabla f^j(x) - \nabla f^j(\hat{x})\| \le \tfrac{1}{2}\kappa. \tag{24e}$$

Hence, for any $x \in B(\hat{x}\,,\rho)$ and any $h \in \mathbb{R}^n$,

$$\begin{aligned} \max_{j \in \hat{\mathbf{q}}(\hat{x})} \langle \nabla f^j(x), h \rangle &\ge \max_{j \in \hat{\mathbf{q}}(\hat{x})} \{\langle \nabla f^j(\hat{x}), h \rangle - \|\nabla f^j(x) - \nabla f^j(\hat{x})\|\|h\|\} \\ &\ge \tfrac{1}{2}\kappa\|h\|. \end{aligned} \tag{24f}$$

Therefore, for any $x \in B(\hat{x}\,,\rho)$, making use of the Mean-Value Theorem 5.1.28*(a)*, we find that

$$\begin{aligned} \psi(x) - \psi(\hat{x}) &\ge \max_{j \in \hat{\mathbf{q}}(\hat{x})} [f^j(x) - \psi(\hat{x})] \\ &= \max_{j \in \hat{\mathbf{q}}(\hat{x})} [f^j(x) - f^j(\hat{x})] \\ &= \max_{j \in \hat{\mathbf{q}}(\hat{x})} \langle \nabla f^j(\hat{x} + s^j(x - \hat{x})), x - \hat{x} \rangle \\ &\ge \tfrac{1}{2}\kappa\|x - \hat{x}\|, \end{aligned} \tag{24g}$$

with $s^j \in [0,1]$. Since (24g) implies that $\hat{x}$ is an isolated local minimizer of $\psi(\cdot)$, our proof is complete. □

Next, we associate the following two functions with the optimality function $\theta(\cdot)$ and search direction function $h(\cdot)$, defined in (2.1.9a,b) in primal form and in (2.1.10c,d) in dual form.. For any $x \in \mathbb{R}^n$, let

$$\hat{\mathbf{q}}_\theta(x) \triangleq \{j \in \mathbf{q} \mid \theta(x) = f^j(x) - \psi(x) + \langle \nabla f^j(x), h(x) \rangle + \tfrac{1}{2}\|h(x)\|^2\}, \tag{25a}$$

and

$$\hat{\Sigma}_q(x) \triangleq \{\mu \in \Sigma_q \mid \theta(x) = -\sum_{j=1}^q \mu^j(\psi(x) - f^j(x)) - \tfrac{1}{2}\|\sum_{j=1}^q \mu^j \nabla f^j(x)\|^2\}. \tag{25b}$$

Since by (2.1.10d), $h(x) = -\sum_{j=1}^q \mu_x^j \nabla f^j(x)$, with $\mu_x \in \hat{\Sigma}_q(x)$ arbitrary, it follows from (25b) that for any $\mu_x \in \hat{\Sigma}_q(x)$,

$$\begin{aligned} 0 &= \sum_{j=1}^q \mu_x^j \{\theta(x) - [f^j(x) - \psi(x) - \tfrac{1}{2}\|h(x)\|^2]\} \\ &= \sum_{j=1}^q \mu_x^j \{\theta(x) - [f^j(x) - \psi(x) + \langle \nabla f^j(x), h(x) \rangle + \tfrac{1}{2}\|h(x)\|^2]\}, \end{aligned} \tag{25c}$$

where we have used the fact that $h(x) = -\sum_{j=1}^q \mu_x^j \nabla f^j(x)$. Since all the $\mu_x^j \ge 0$ and the terms they multiply in (25c) are all nonnegative, it follows that for all $\mu_x \in \hat{\Sigma}_q(x)$ and $j \in \mathbf{q}$,

$$\mu_x^j \{\theta(x) - [f^j(x) - \psi(x) + \langle \nabla f^j(x), h(x) \rangle + \tfrac{1}{2}\|h(x)\|^2]\} = 0, \tag{25d}$$

which shows that $\mu_x^j = 0$ for all $j \notin \hat{\mathbf{q}}_\theta(x)$.

Lemma 2.4.20. *Suppose that $\hat{x}$ is a Haar point for the problem* **MMP**, *defined in (1) and (2). Then there exists a $\rho_1 > 0$ such that, for all $x \in B(\hat{x}\,,\rho_1)$, $\hat{\mathbf{q}}_\theta(x) = \hat{\mathbf{q}}(\hat{x})$ (with $\hat{\mathbf{q}}(\hat{x})$ as in (2.1.2b)), and, for all $\mu_x \in \hat{\Sigma}_q(x)$, $\mu_x^j > 0$ for all $j \in \hat{\mathbf{q}}(\hat{x})$.*

Proof. First, we note that because $h(\hat{x}) = 0$ at any local minimizer $\hat{x}$, $\hat{\mathbf{q}}_\theta(\hat{x}) = \hat{\mathbf{q}}(\hat{x})$. Furthermore, by Corollary 5.4.4, $\hat{\mathbf{q}}_\theta(\cdot)$ is outer semicontinuous, and hence there exists a $\rho_0 > 0$ such that $\hat{\mathbf{q}}_\theta(x) \subset \hat{\mathbf{q}}(\hat{x})$ for all $x \in B(\hat{x}\,,\rho_0)$. Since $\hat{x}$ is a Haar point, by assumption, $\hat{\Sigma}_q(\hat{x})$ is a singleton $\hat{\mu}$, with $\hat{\mu}^j > 0$, for

all $j \in \hat{\mathbf{q}}(\hat{x})$. Since $\hat{\Sigma}_q(\cdot)$ is outer semicontinuous by Theorem 5.4.3, there must exist a $\rho_1 \in (0, \rho_0]$ such that, for all $x \in B(\hat{x}, \rho_1)$ and $\mu_x \in \hat{\Sigma}_q(x)$, $\mu_x^j > 0$ for all $j \in \hat{\mathbf{q}}(\hat{x})$. Now, it follows from (25d) that for any $\mu_x \in \hat{\Sigma}_q(x)$ and $j \in \mathbf{q}$ such that $\mu_x^j > 0$, we must that $j \in \hat{\mathbf{q}}_\theta(x)$. Since, for all $x \in B(\hat{x}, \rho_1)$, $\hat{\mathbf{q}}_\theta(x) \subset \hat{\mathbf{q}}(\hat{x})$ and $\mu_x^j(x) = 0$ for all $j \notin \hat{\mathbf{q}}(x)$, we conclude that $\hat{\mathbf{q}}_\theta(x) = \hat{\mathbf{q}}_\theta(\hat{x})$ for all $x \in B(\hat{x}, \rho_1)$, completing our proof. □

Corollary 2.4.21. *Suppose that $\hat{x}$ is a Haar point for the problem* **MMP** *in (1) and (2). For any $x \in \mathbb{R}^n$, let*

$$y(x) \triangleq \theta(x) - \tfrac{1}{2}\|h(x)\|^2 + \psi(x). \tag{26a}$$

Then there exists a $\rho_2 \in (0, \rho_1]$, where $\rho_1 > 0$ is as defined in Lemma 2.4.20, such that, for all $x \in B(\hat{x}, \rho_2)$, the pair $(y(x), h(x))$ is the unique solution of the system of equations

$$\langle \nabla f^j(x), h \rangle - y = -f^j(x), \quad j \in \hat{\mathbf{q}}(\hat{x}). \tag{26b}$$

Proof. Since $\hat{\mathbf{q}}_\theta(x) = \hat{\mathbf{q}}(\hat{x})$ for all $x \in B(\hat{x}, \rho_1)$, it follows that, for all $j \in \hat{\mathbf{q}}(\hat{x})$,

$$f^j(x) + \langle \nabla f^j(x), h(x) \rangle = \theta(x) - \tfrac{1}{2}\|h(x)\|^2 + \psi(x) = y(x), \tag{27}$$

which shows that $(y(x), h(x))$ is a solution of (26b). If we rewrite (26b) in the matrix form:

$$\begin{bmatrix} -1 & \dots & -1 \\ \nabla f^{j_1}(x) & \cdots & \nabla f^{j_{n+1}}(x) \end{bmatrix}^T \begin{pmatrix} y \\ h \end{pmatrix} = -F(x), \tag{28}$$

where $F(x) = (f^{j_1}(x), \dots, f^{j_{n+1}}(x))$, we see that, because the vectors $\{\nabla f^{j_k}(x)\}_{k=1}^{n+1}$ are affinely independent near $\hat{x}$, there exists a $\rho_2 \in (0, \rho_1]$ such that for all $x \in B(\hat{x}, \rho_2)$, the solution $(y(x), h(x))$ of the system of equations (26b) is unique. □

Theorem 2.4.22. *Consider the min-max problem* **MMP** *in (1) and (2), and suppose that the gradients $\nabla f^j(\cdot)$, $j \in \mathbf{q}$, are Lipschitz continuous on bounded sets. If the PPP Algorithm 2.4.1 constructs a sequence $\{x_i\}_{i=0}^{\infty}$ with an accumulation point $\hat{x}$, which is a Haar point, then $x_i \to \hat{x}$, as $i \to \infty$ Q-quadratically.*

Proof. First, because $\hat{x}$ is an isolated local minimizer, it follows from Proposition 1.2.38 that $x_i \to \hat{x}$, as $i \to \infty$. Next there must exist an i_0 such that $x_i \in B(\hat{x}, \rho_2)$, for all $i \geq i_0$, where ρ_2 is as defined in Corollary 2.4.21. Hence

we see that for all $i \geq i_0$, $(y(x_i), h(x_i))$ solves (26b), with $x = x_i$. Now, (26b) happens to be the system of equations that determines the local Newton search direction for solving the system of $n+1$ equations

$$f^j(x) + y = 0, \ j \in \hat{\mathbf{q}}(\hat{x}), \tag{29a}$$

or equivalently, with y eliminated, for solving the system of n equations

$$f^{j_k}(x) - f^{j_{n+1}}(x) = 0, \ k = 1, 2, .., n, \tag{29b}$$

assuming that $\hat{\mathbf{q}}(\hat{x}) = \{j_1, \dots, j_{n+1}\}$. Therefore, since the assumptions for the quadratic convergence of the local Newton method on (29b) are satisfied near a Haar point, to show that the PPP method converges quadratically, we need to show only that there exists an $i_1 \geq i_0$ such that $\lambda_i = 1$ for all $i \geq i_1$.

First, for all $i \geq i_0$, by definition of $y(x_i)$, and for all $j \in \hat{\mathbf{q}}(\hat{x})$,

$$f^j(x_i) - y(x_i) = -\langle \nabla f^j(x_i), h(x_i) \rangle. \tag{29c}$$

Hence, since (24a) holds, it follows from (29c) that

$$\psi(x_i) - y(x_i) = \max_{j \in \hat{\mathbf{q}}(\hat{x})} \langle \nabla f^j(x_i), -h(x_i) \rangle \geq \kappa \|h(x_i)\|. \tag{29d}$$

Since by definition of $y(x_i)$,

$$\theta(x_i) = y(x_i) - \psi(x_i) + \tfrac{1}{2}\|h(x_i)\|^2, \tag{29e}$$

it follows that

$$\begin{aligned} -\theta(x_i) &= \psi(x_i) - y(x_i) - \tfrac{1}{2}\|h(x_i)\|^2 \\ &\geq \kappa \|h(x_i)\| - \tfrac{1}{2}\|h(x_i)\|^2 \\ &= \kappa \|h(x_i)\| \left[1 - \frac{1}{2\kappa}\|h(x_i)\| \right]. \end{aligned} \tag{29f}$$

Consequently, since $h(x_i) \to 0$, as $i \to \infty$, there exists an $i' \geq i_1$ such that for all $i \geq i'$,

$$-\theta(x_i) \geq \tfrac{1}{2}\kappa \|h(x_i)\|. \tag{29g}$$

Hence for all $i \geq i'$, using the expansion formula (5.1.18a), we find that

$$\begin{aligned} \psi(x_i + h(x_i)) - \psi(x_i) = \max_{j \in \mathbf{q}} \Big\{ & f^j(x_i) - \psi(x_i) + \langle \nabla f^j(x_i), h(x_i) \rangle \\ & + \int_0^1 \langle \nabla f^j(x_i + s h(x_i)) - \nabla f^j(x_i), h(x_i) \rangle \, ds \Big\} \\ \leq \max_{j \in \mathbf{q}} \{ & f^j(x_i) - \psi(x_i) + \langle \nabla f^j(x_i), h(x_i) \rangle + \tfrac{1}{2} L \|h(x_i)\|^2 \} \end{aligned}$$

$$\leq \theta(x_i) + \tfrac{1}{2}(L-1)\|h(x_i)\|^2, \tag{29h}$$

where $L \in [1,\infty)$ is an appropriate local Lipschitz constant for the gradients $\nabla f^j(\cdot)$. Since by (29g), $\|h(x_i)\|^2 \leq (2/\kappa)^2\theta(x_i)^2$ and since $\theta(x_i) \to 0$, as $i \to \infty$, (29h) implies that there exists an $i_2 \geq i'$ such that, for all $i \geq i_2$,

$$\psi(x_i + h(x_i)) - \psi(x_i) \leq \left[1 + \frac{2(L-1)}{\kappa^2}\theta(x_i)\right]\theta(x_i) \leq \alpha\theta(x_i), \tag{29i}$$

with $\alpha \in (0,1)$, as in (3c), i.e., that the step-size computed according to (3c) will be unity. Hence our proof is complete. □

2.4.5 Box-Constrained Min-Max Algorithm

Next we will consider the following simple variation of the problem **MMP** defined in (1) and (2) that occurs routinely in discrete optimal control:

$$\mathbf{MMP}_X \qquad \min_{x \in X} \psi(x), \tag{30a}$$

where $\psi(\cdot)$ is defined as in (2), with the functions $f^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, continuously differentiable, and X is a compact, convex subset of $\mathbb{R}^n$. In the case of optimal control, X is usually a box of the form

$$X = \{x \in \mathbb{R}^n \mid |x^i| \leq b,\ i = 1,2,\ldots,n\}. \tag{30b}$$

However, other simple descriptions can also appear, such as

$$X = \{x \in \mathbb{R}^n \mid \|x\| \leq b\} \tag{30c}$$

or

$$X = \{x \in \mathbb{R}^n \mid \underline{b}_i \leq x^i \leq \bar{b}_i,\ i = 1,2,\ldots,n\}. \tag{30d}$$

Theorem 2.4.23. *Suppose that $\hat{x}$ is optimal for $\mathbf{MMP}_X$. Then*

(i)
$$d\psi(\hat{x}; x - \hat{x}) \geq 0, \ \forall x \in X, \tag{31a}$$

(ii) relation (31a) holds at a vector $\hat{x} \in X$ if and only if there exists a $\hat{\mu} \in \Sigma_q$ such that

$$\Big\langle \sum_{j \in \mathbf{q}} \hat{\mu}^j \nabla f^j(\hat{x}), x - \hat{x} \Big\rangle \geq 0, \ \forall x \in X, \tag{31b}$$

and

$$\sum_{j \in \mathbf{q}} \hat{\mu}^j [\psi(\hat{x}) - f^j(\hat{x})] = 0, \tag{31c}$$

and

(iii) relation (31a) holds at a vector $\hat{x} \in X$ if and only if $\theta_X(\hat{x}) = 0$, where, for any $x \in X$,

$$\theta_X(x) \triangleq \min_{x' \in X} \max_{j \in \mathbf{q}} \{f^j(x) - \psi(x) + \langle \nabla f^j(x), x' - x \rangle + \tfrac{1}{2}\delta\|x' - x\|^2\}, \tag{31d}$$

with $\delta > 0$. □

Exercise 2.4.24. *(a)* Make use of Corollary 5.5.3 to prove Theorem 2.4.23. *(b)* Show that *(i)* $\theta_X(\cdot)$ is continuous on X and *(ii)* for all $x, x' \in X$, $d\psi(x; x' - x) \leq \theta_X(x)$. *(c)* Show that $\theta_X(x)$ can be computed by solving the following quadratic programming problem in the $n+1$ variables $(x', x'^{n+1}) \in \mathbb{R}^{n+1}$:

$$\theta_X(x) = \min\{x'^{n+1} + \tfrac{1}{2}\delta\|x' - x\|^2 \mid f^j(x) - \psi(x) + \langle \nabla f^j(x), x' - x \rangle - x'^{n+1} \leq 0,\ j \in \mathbf{q},\ x \in X\}. \tag{32a}$$

(d) Show that if (x'_x, x'^{n+1}_x) is the solution of (32a), then x'_x is the solution of (31d), i.e.,

$$x'_x = \arg\min_{x' \in X} \max_{j \in \mathbf{q}} \{f^j(x) - \psi(x) + \langle \nabla f^j(x), x' - x \rangle + \tfrac{1}{2}\|x' - x\|^2\}. \tag{32b}$$

□

The Box PPP Algorithm 2.4.25.

Parameters. $\alpha, \beta \in (0,1), \delta > 0$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute the *optimality function value* θ_i and the *search direction* h_i according to

$$\theta_i = \min_{h \in X - \{x_i\}} \max_{j \in \mathbf{q}} \{f^j(x_i) - \psi(x_i) + \langle \nabla f^j(x_i), h \rangle + \tfrac{1}{2}\delta\|h\|^2\}, \tag{33a}$$

$$h_i = \arg\min_{h \in X - \{x_i\}} \max_{j \in \mathbf{q}} \{f^j(x_i) - \psi(x_i) + \langle \nabla f^j(x_i), h \rangle + \tfrac{1}{2}\delta\|h\|^2\}. \tag{33b}$$

Step 2. If $\theta_i = 0$, stop.

Else, compute the *step-size*

$$\lambda_i = \lambda(x_i) \triangleq \arg\max_{k \in \mathbb{N}} \{\beta^k \mid \psi(x_i + \beta^k h_i) - \psi(x_i) - \beta^k \alpha\theta_i \leq 0\}. \tag{33c}$$

Step 3. Set

$$x_{i+1} = x_i + \lambda_i h_i , \tag{33d}$$

replace i by $i+1$, and go to Step 1.

Note that because for any $x', x \in X$, $h \triangleq (x' - x) \in X - \{x\}$, it follows that the sequence constructed by Algorithm 2.4.25 is automatically contained in X.

Theorem 2.4.26. *Consider problem (30a) with the assumptions stated. Suppose that Algorithm 2.4.25 constructs a sequence* $\{x_i\}_{i=0}^{\infty}$ *in solving (30a). Then every accumulation point* $\hat{x}$ *of* $\{x_i\}_{i=0}^{\infty}$ *satisfies the first-order optimality condition* $\theta_X(\hat{x}) = 0$. □

Exercise 2.4.27.

(a) Prove Theorem 2.4.26.

(b) Suppose that Assumption 2.4.3 is satisfied and that $\delta \in [m, M]$. Show that *(i)* problem (30a) has a unique solution $\hat{x}$ and *(ii)* if a sequence $\{x_i\}_{i=0}^{\infty}$ is constructed by Algorithm 2.4.25 in solving (30a), then $x_i \to \hat{x}$ as $i \to \infty$, and (8a,b,c) hold. □

2.4.6 A Barrier Function Method

The Algorithms 2.4.1 and 2.4.25 were constructed by replacing the functions $f^j(\cdot)$ in (2) by quadratic approximations. An alternative approach is to use a barrier function instead. This approach becomes attractive when q in (2) is very large, which makes computation of search directions for Algorithms 2.4.1 and 2.4.25 problematic, and only an approximate solution to the problem **MMP** defined in (1) and (2) needs to be found. This is the case when we solve approximations to semi-infinite optimization problems and optimal control problems with state constraints, as we will see in Chapters 3 and 4.

Consider problem **MMP**, defined in (1), (2), with the assumptions stated. Let

$$\hat{\alpha} \triangleq \min_{x \in \mathbb{R}^n} \psi(x), \tag{34a}$$

and, for every $\alpha \in (\hat{\alpha}, \infty)$, let

$$C(\alpha) \triangleq \{x \in \mathbb{R}^n \mid \psi(x) < \alpha\}. \tag{34b}$$

Let $b : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ be defined by

$$b(x, \alpha) \triangleq \frac{1}{q} \sum_{j=1}^{q} \frac{1}{\alpha - f^j(x)} . \tag{34c}$$

Then, for every $\alpha > \hat{\alpha}$, $b(\cdot, \alpha)$ is a *barrier function* for $C(\alpha)$, i.e., for every $x \in C(\alpha)$, $b(x, \alpha) > 0$ and $b(x, \alpha) \to \infty$, as $\psi(x) \to \alpha$. The barrier function $b(\cdot, \alpha)$ is continuously differentiable on $C(\alpha)$ and, for $x \in C(\alpha)$, its gradient is given by

$$\nabla_x b(x, \alpha) = \frac{1}{q} \sum_{j=1}^{q} \frac{1}{(\alpha - f^j(x))^2} \nabla f^j(x) . \tag{34d}$$

Hence $b(\cdot, \alpha)$ can be minimized by means of various unconstrained optimization algorithms encountered in Chapter 1. These observations suggests that, given an $x_i \in \mathbb{R}^n$, a practical alternative to the construction of a successor point x_{i+1}, according to the rules in Algorithm 2.4.1, might be the construction of any $x_{i+1} \in C(\psi(x_i))$ which is an "approximate" minimizer of $b(\cdot, \psi(x_i))$ over $C(\psi(x_i))$. These considerations lead to the following algorithm.

Basic Barrier Function Algorithm 2.4.28.

Parameters. $\gamma \in [0, 2)$, $\kappa \in (0, \infty)$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute an $x_{i+1} \in C(\psi(x_i))$ such that

$$\|\nabla_x b(x_{i+1}, \psi(x_i))\| \leq \frac{\kappa}{[\psi(x_i) - \psi(x_{i+1})]^{\gamma}} . \tag{35}$$

Step 2. Replace i by $i+1$, and go to Step 1.

As we will soon see, the values $[\psi(x_i) - \psi(x_{i+1})]$ converge to zero, which tends to make both $b(x, \psi(x_i))$ and its gradient large for $x \in C(\psi(x_i))$. The term $[\psi(x_i) - \psi(x_{i+1})]^{\gamma}$ in (35) is used to slow the increase in precision with which the problems $\min_{x \in C(\psi(x_i))} b(x, \psi(x_i))$ must be solved at each iteration.

Theorem 2.4.29. *Consider the problem* **MMP** *defined in (1) and (2) with the assumptions stated. Suppose that Algorithm 2.4.28 constructs a sequence* $\{x_i\}_{i=0}^{\infty}$ *in solving the problem* **MMP**. *Then every accumulation point* $\hat{x}$ *of* $\{x_i\}_{i=0}^{\infty}$ *satisfies the first-order optimality condition* $\theta(\hat{x}) = 0$, *with* $\theta(\cdot)$ *defined in (2.1.10c).*

Proof. First, for every $i \in \mathbb{N}$, since $x_{i+1} \in C(\psi(x_i))$, we must have that $\psi(x_{i+1}) < \psi(x_i)$. Hence the sequence $\{\psi(x_i)\}_{i=0}^{\infty}$ is monotone decreasing, and,

since, by continuity of $\psi(\cdot)$, $\psi(\hat{x})$ is an accumulation point of $\{\psi(x_i)\}_{i=0}^{\infty}$, it follows from Proposition 5.1.16 that $\psi(x_i) \to \psi(\hat{x})$, as $i \to \infty$. Next, it follows from (35) that

$$[\psi(x_{i-1}) - \psi(x_i)]^2 \|\nabla_x b(x_i, \psi(x_{i-1}))\| \leq \kappa[\psi(x_{i-1}) - \psi(x_i)]^{(2-\gamma)}. \tag{36a}$$

For $i \in \mathbb{N}$, $i \geq 1$, and $j \in \mathbf{q}$, let

$$\nu_i^j \triangleq \frac{1}{q}\left\{\frac{\psi(x_{i-1}) - \psi(x_i)}{\psi(x_{i-1}) - f^j(x_i)}\right\}^2. \tag{36b}$$

Then it follows from (36a) and (34d) that

$$\|\sum_{j=1}^{q} \nu_i^j \nabla f^j(x_i)\| \leq \kappa[\psi(x_{i-1}) - \psi(x_i)]^{(2-\gamma)}. \tag{36c}$$

Now let $\omega_i \triangleq \sum_{j=1}^{q} \nu_i^j$, and let $\mu_i^j = \nu_i^j / \omega_i$. Then $\mu_i^j \geq 0$ for all $j \in \mathbf{q}$ and $\sum_{j=1}^{q} \mu_i^j = 1$. Since each $\nu_i^j \in [0, 1/q]$ and at least one $\nu_i^j = 1/q$, it follows that $1/q \leq \omega_i \leq 1$. Hence, dividing (36c) by ω_i, we find that

$$\|\sum_{j=1}^{q} \mu_i^j \nabla f^j(x_i)\| \leq q\kappa[\psi(x_{i-1}) - \psi(x_i)]^{(2-\gamma)}. \tag{36d}$$

Now suppose that $K \subset \mathbb{N}$ is such that $x_i \xrightarrow{K} \hat{x}$, as $i \to \infty$. Since $\psi(x_i) \to \psi(\hat{x})$, as $i \to \infty$, we conclude from (36d) that

$$\lim_{i \in K} \|\sum_{j=1}^{q} \mu_i^j \nabla f^j(x_i)\| = 0, \tag{36e}$$

and from (36b) that

$$\lim_{i \in K} \sum_{j=1}^{q} \mu_i^j [\psi(x_{i-1}) - f^j(x_i)] = 0. \tag{36f}$$

Now, it follows from (2.1.10c) that

$$\theta(x_i) \geq -\left[\sum_{j=1}^{q} \mu_i^j [\psi(x_i) - f^j(x_i)] + \frac{1}{2\delta}\|\sum_{j=1}^{q} \mu_i^j \nabla f^j(x_i)\|^2\right]. \tag{36g}$$

Since $\psi(x_i) \to \psi(\hat{x})$, as $i \to \infty$, it follows that $\psi(x_i) - \psi(x_{i-1}) \to 0$, as $i \to \infty$, and hence we conclude from (36g) that $\lim_{i \in K} \theta(x_i) = 0$. Since $\theta(\cdot)$ is continuous, we conclude that $\theta(\hat{x}) = 0$, which completes our proof. □

The only problem with Algorithm 2.4.28 is that because $b(x_i, \psi(x_i)) = \infty$, x_i cannot be used as a "hot" starting point in minimizing $b(\cdot, \psi(x_i))$. This problem can be eliminated by slightly enlarging the sets $C(\psi(x_i))$, as in the algorithm below.

Polak-Higgins-Mayne Barrier Function Algorithm 2.4.30.

Parameters. $\gamma \in [0, 2)$, $\kappa \in (0, \infty)$, $\{\eta_i\}_{i=0}^{\infty}$, with $\eta_i > 0$, $\sum_{i=0}^{\infty} \eta_i < \infty$.

Data. $x_{-1}, x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Set

$$\alpha_i \triangleq \begin{cases} \frac{1}{2}\{\psi(x_{i-1}) + \psi(x_i)\} & \text{if } \psi(x_{i-1}) \neq \psi(x_i) \\ \frac{1}{2}\{\psi(x_{i-1}) + \psi(x_i)\} + \eta_i & \text{if } \psi(x_{i-1}) = \psi(x_i) \end{cases} \tag{37a}$$

and

$$y_i \triangleq \begin{cases} x_i & \text{if } \psi(x_{i-1}) \geq \psi(x_i) \\ x_{i-1} & \text{if } \psi(x_{i-1}) < \psi(x_i) \end{cases}. \tag{37b}$$

Step 2. Use y_i as a "hot" starting point to compute an $x_{i+1} \in C(\alpha_i)$ such that

$$\|\nabla_x b(x_{i+1}, \alpha_i)\| \leq \frac{\kappa}{[\alpha_i - \psi(x_{i+1})]^{\gamma}}. \tag{37c}$$

Step 3. Replace i by $i + 1$, and go to Step 1.

It is not difficult to construct max functions for which the sets $C(\alpha)$ are long and narrow, implying that the barrier function $b(\cdot, \alpha)$ can be ill-conditioned. Hence it is advisable to use either the BFGS variable metric method or conjugate gradient methods with restart for computing the x_{i+1} in Algorithm 2.4.30. Not surprisingly, Algorithm 2.4.30 retains the convergence properties of Algorithm 2.4.28.

Theorem 2.4.31. *Consider the problem* **MMP** *defined in (1) and (2) with the assumptions stated. Suppose that Algorithm 2.4.30 constructs a sequence* $\{x_i\}_{i=0}^{\infty}$ *in solving* **MMP**. *Then every accumulation point* $\hat{x}$ *of* $\{x_i\}_{i=0}^{\infty}$ *satisfies the first-order optimality condition* $\theta(\hat{x}) = 0$, *with* $\theta(\cdot)$ *defined in (2.1.9a).* □

Exercise 2.4.32. Prove Theorem 2.4.31. *Hint*: Make use of Lemmas 3.1 and 3.2 in [PHM.92]. □

Exercise 2.4.33. Define a version of Algorithm 2.4.30 for min-max problems with box constraints, and show that the corresponding version of Theorem 2.4.31 holds for it. □

2.4.7 Notes

The first version of the PPP algorithm was proposed by Pshenichnyi around 1970 (see [PsD.75]), who called it the method of linearizations and used a preassigned sequence of step lengths $\{\lambda_i\}_{i=0}^{\infty}$. A different version, using the same search direction function, but with exact line searches, was introduced independently by Pironneau and Polak [PiP.72], who used it as a subprocedure in an implementable method of centers. The version in this paper retains the Pshenichnyi search direction function, but uses an Armijo type step-size rule.

The Frank-Wolfe algorithm was described in [FrW.56]. Algorithm 2.4.11 was first presented in [HiP.90] as the last link in a chain which began with the Frank-Wolfe algorithm, followed by an algorithm for finding the nearest point to a convex polytope [Wol.76] and a generalization in [vHo.77] for the minimization of a pseudo convex function on a convex polytope.

The linear rate of convergence of the PPP Algorithm 2.4.1 was first presented in [Pol.89]. Basic elements of our proof of quadratic convergence to a Haar point can be found in [DaM.81].

In addition to the two min-max algorithms that we have discussed, there are a number of others in the literature, see e.g., [Ber.82, ChC.78, MuO.80, Han.77, Han.81, Fle.82, MaS.78, HaM.81, LiX.92, PMH.91, PMH.92, Pol.88, WoF.86, WiP.91, Yua.85, ZhT.93, ZhT.96]. The algorithms in [MaS.78, HaM.81, Fle.82, Yua.85] use trust regions. While some of the cited algorithms converge quadratically near a Haar point or when the Maratos effect [Mar.78] does not manifest itself, only the algorithms in [PMH.91, PMH.92, ZhT.93, ZhT.96] are demonstrably superlinearly converging under fairly general assumptions.

The barrier function algorithm that we have presented first appeared in [PHM.92]. A more efficient, but somewhat more complex version can be found in [HiP.91]. Both of these algorithms were inspired by barrier function methods for nonlinear programming (see, e.g., [Jar.88a, Jar.88b, SoS.88, Son.87, YeY.87]). An interesting alternative method based on modified barrier functions is given in [Pol.88].

Algorithm 2.4.30 makes use of the parameter-free Fiacco-McCormick penalty function [FiM.67], the Mifflin truncation rule [Mif.76], and the Tremolières penalty adjustment rule [Tre.68]. It is inherently a low-precision method. On well-conditioned min-max problems, the numerical performance of the barrier function algorithm is comparable to that of other first-order min-max algorithms, including the PPP algorithm. However, it offers advantages on ill-conditioned problems, because it computes solutions when excellent competing algorithms, such as [ChC.78, MuO.80, Han.78, HaM.81, Wom.82] and the PPP Algorithm 2.4.1, fail.

Algorithm 2.4.30 reduces the solution of the min-max problem (1) and (2) to the solution of a sequence of smooth, unconstrained optimization problems. Alternative methods for reducing the min-max problem to the minimization of a smooth function are presented in [Ber.82], page 313, [Zan.80], [Pol.88], and [LiX.92]. For example, [Ber.82] and [LiX.92] use the parametrized, smooth function

$$F_p(x) \triangleq (1/p)\ln\left[\sum_{j\in\mathbf{q}} \exp[pf^j(x)]\right], \tag{38a}$$

where $p > 0$ is a parameter. Note that an alternative expression for $F_p(x)$ is

$$F_p(x) = \psi(x) + (1/p)\ln\left[\sum_{j\in\mathbf{q}} \exp\big(p[f^j(x)-\psi(x)]\big)\right]. \tag{38b}$$

It follows from (38b) that

$$0 \le F_p(x) - \psi(x) \le (1/p)\ln q, \tag{38c}$$

and hence that $F_p(x) \to \psi(x)$, as $p \to \infty$. In practice, it turns out that when p is set in the range of 10^{-4}–10^{-6}, the minimization of $F_p(x)$ yields an excellent approximation to the minimum value of $\psi(x)$. However, note that the use of exponentials can lead to numerical problems, and in addition, the Hessian of $F_p(\cdot)$ can be highly ill-conditioned, even when the Hessians of all the $f^j(\cdot)$ are very well conditioned.

A better method for transforming a min-max problem into a smooth optimization problem can be found in [DGL.93].

When computer code for Algorithm 2.4.11 is not available and one has to evaluate $\theta(\cdot)$ using quadratic programming code, one can use the following version of the PPP Algorithm 2.4.1, inferred from the phase I - phase II method of feasible directions in [PTM.79]. This version requires fewer gradient evaluations and hence also constructs a smaller quadratic programming problem for search direction computations. The price one pays for this economy is that the search direction is no longer continuous in the iterates, and hence a larger amount of "zigzagging" may be observed than with Algorithm 2.4.1.

Note that for any fixed $\varepsilon > 0$, $\theta_\varepsilon(\cdot)$ defined by (39b) is an optimality function which, unlike the function $\theta(\cdot)$, defined by (2.1.9a), is only u.s.c. because the set-valued function $\hat{\mathbf{q}}_\varepsilon(\cdot)$ is only outer semi-continuous. Also note that the proof of Theorem 2.4.35 becomes easier under the assumption that the functions $f^j(\cdot)$, $j \in \mathbf{q}$, are Lipschitz continuously differentiable on bounded sets.

ε-Active PPP Algorithm 2.4.34.

Parameters. $\alpha \in (0,1]$, $\beta \in (0,1)$, ε_0, $\delta > 0$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Set $\varepsilon = \varepsilon_0$.

Step 2. Compute the set

$$\hat{\mathbf{q}}_\varepsilon(x_i) \triangleq \{j \in \mathbf{q} \mid \psi(x_i) - f^j(x_i) \le \varepsilon\}. \tag{39a}$$

Step 3. Compute the *ε–optimality function value* $\theta_\varepsilon(x_i)$ and the *search direction* $h_\varepsilon(x_i)$ according to the dual form formulas

$$\theta_\varepsilon(x_i) \triangleq -\min_{\mu \in \Sigma_q} \Big\{ \sum_{j=1}^{q} \mu^j [\psi(x_i) - f^j(x_i)] + \frac{1}{2\delta} \| \sum_{j=1}^{q} \mu^j \nabla f^j(x_i) \|^2 \mid \mu^j = 0, \forall\, j \notin \hat{\mathbf{q}}_\varepsilon(x_i) \Big\}, \tag{39b}$$

and

$$h_\varepsilon(x_i) \triangleq -\frac{1}{\delta} \sum_{j=1}^{q} \mu_x^j \nabla f^j(x_i), \tag{39c}$$

where μ_x is any solution of (39b).

Step 4. If $\theta_\varepsilon(x_i) \geq -\varepsilon$, replace ε by $\beta\varepsilon$, and go to Step 1.
Else, set $\theta_i = \theta_\varepsilon(x_i)$, $h_i = h_\varepsilon(x_i)$, and go to Step 5.

Step 5. If $\theta_i = 0$, stop.
Else, compute the *step-size*

$$\lambda_i = \lambda(x_i) \triangleq \arg\max_{k \in \mathbf{N}} \{\beta^k \mid \psi(x_i + \beta^k h_i) - \psi(x_i) - \beta^k \alpha \theta_i \leq 0\}. \tag{39d}$$

Step 6. Set

$$x_{i+1} = x_i + \lambda_i h_i, \tag{39e}$$

replace i by $i+1$, and go to Step 1.

To prove the following theorem one must use the fact that if $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 2.4.34 with an accumulation point $\hat{x}$ such that $\theta(\hat{x}) < 0$, then there exists an $\hat{\varepsilon} > 0$ such that, for all x_i sufficiently near $\hat{x}$, Algorithm 2.4.34 accepts a value of $\varepsilon \geq \hat{\varepsilon}$.

Theorem 2.4.35. *Consider the problem* **MMP**, *defined in (1) and (2), with the assumptions stated. Suppose that Algorithm 2.4.34 constructs a sequence* $\{x_i\}_{i=0}^{\infty}$ *in solving* **MMP**. *Then every accumulation point* $\hat{x}$ *of* $\{x_i\}_{i=0}^{\infty}$ *satisfies the first-order optimality condition* $\theta(\hat{x}) = 0$. □

2.5 Newton's Method for Min-Max Problems

We will now present local and global versions of Newton's method for solving the problem

$$\min_{x \in \mathbb{R}^n} \psi(x), \tag{1a}$$

where

$$\psi(x) = \max_{j \in \mathbf{q}} f^j(x), \tag{1b}$$

where the functions $f^j : \mathbb{R}^n \to \mathbb{R}$ are convex. We will see that these versions of Newton's method are natural extensions of those presented in Section 1.4.

The versions of Newton's method in this section were first presented by Polak, Mayne and Higgins in [PMH.92]. To distinguish them from other versions, such as those discussed in the Notes subsection, we will refer to them as Polak-Mayne-Higgins algorithms or PMH algorithms, for short.

As in Section 1.4, we begin with the local version.

2.5.1 The Local Newton Method

Following the pattern introduced in Section 1.4, we make the following hypotheses.

Assumption 2.5.1.

(i) For all $j \in \mathbf{q}$, *the functions* $f^j : \mathbb{R}^n \to \mathbb{R}$ *in (1b) are twice Lipschitz continuously differentiable, i.e., there exists an* $L_S < \infty$ *such that*

$$\| f_{xx}^j(x') - f_{xx}^j(x) \| \leq L_S \| x' - x \|, \quad \forall\, x, x' \in \mathbb{R}^n, \; j \in \mathbf{q}. \tag{2a}$$

(ii) There exist constants $0 < m \leq M$ *such that for all* $x \in \mathbb{R}^n$,

$$m\|h\|^2 \leq \langle h, f_{xx}^j(x) h \rangle \leq M\|h\|^2, \quad \forall\, h \in \mathbb{R}^n, \; \forall\, t \in [0,1], \; \forall\, j \in \mathbf{q}. \tag{2b}$$

□

Since by Theorem 5.2.14, the functions $f^j(\cdot)$, $j \in \mathbf{q}$ are strictly convex under Assumption 2.5.1*(ii)*, it follows that $\psi(\cdot)$ is also strictly convex, and hence that it has a unique minimizer $\hat{x}$.

Proposition 2.5.2. *Suppose that Assumption 2.5.1 holds and that* $\hat{x}$ *is the unique minimizer of* $\psi(\cdot)$. *Then, for all* $x \in \mathbb{R}^n$,

$$\psi(x) - \psi(\hat{x}) \geq \frac{m}{2} \|x - \hat{x}\|^2. \tag{3}$$

Proof. Making use of the second-order expansion formula (5.1.17d) and of (2b), we find that

$$\begin{aligned} \psi(x) - \psi(\hat{x}) &\geq \max_{j \in \mathbf{q}} \{f^j(\hat{x}) - \psi(\hat{x}) + \langle \nabla f^j(\hat{x}), x - \hat{x} \rangle + \frac{m}{2}\|x - \hat{x}\|^2\} \\ &\geq \max_{j \in \hat{\mathbf{q}}(\hat{x})} \{f^j(\hat{x}) - \psi(\hat{x}) + \langle \nabla f^j(\hat{x}), x - \hat{x} \rangle + \frac{m}{2}\|x - \hat{x}\|^2\} \\ &= d\psi(\hat{x}, x - \hat{x}) + \frac{m}{2}\|x - \hat{x}\|^2, \end{aligned} \tag{4}$$

where $\hat{\mathbf{q}}(\hat{x}) \triangleq \{j \in \mathbf{q} \mid f^j(\hat{x}) = \psi(\hat{x})\}$, as before. Since $\hat{x}$ is the minimizer of

$\psi(\cdot), d\psi(\hat{x}, x - \hat{x}) \geq 0$, and hence (3) follows. □

By analogy with Newton's method for differentiable functions, we define a quadratic approximation $\tilde{\psi}(y, \cdot)$ to $\psi(\cdot)$ around the point $y \in \mathbb{R}^n$ by

$$\tilde{\psi}(y, x) \triangleq \max_{j \in \mathbf{q}} \{f^j(y) + \langle \nabla f^j(y), x - y \rangle + \tfrac{1}{2} \langle (x - y), f^j_{xx}(y)(x - y) \rangle\}. \quad (5)$$

Local Polak-Mayne-Higgins Algorithm 2.5.3 (Local Newton Method).

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute

$$x_{i+1} = \arg \min_{x \in \mathbb{R}^n} \tilde{\psi}(x_i, x). \quad (6)$$

Step 2. Replace i by $i+1$, and go to Step 1.

Since the functions

$$f^j(y) + \langle \nabla f^j(y), x - y \rangle + \tfrac{1}{2} \langle (x - y), f^j_{xx}(y)(x - y) \rangle,$$

$j \in \mathbf{q}$, in (5) are strictly convex, $\tilde{\psi}(x_i, \cdot)$ is strictly convex and therefore x_{i+1} is uniquely defined by (6).

To establish the local convergence and rate of convergence of the above algorithm, we will need the following lemmas.

Lemma 2.5.4. *Suppose that Assumption 2.5.1 holds. Then there exists a* $\kappa < \infty$ *such that for any* $x, y \in \mathbb{R}^n$,

$$|\psi(x) - \tilde{\psi}(y, x)| \leq \kappa \|x - y\|^3. \quad (7)$$

Proof. Let $L < \infty$ be a common Lipschitz constant for the Hessians $f^j_{xx}(\cdot)$. Then, making use of the second-order expansion formula (5.1.17d), we obtain that

$$\psi(x) = \max_{j \in \mathbf{q}} \Big\{ f^j(y) + \langle \nabla f^j(y), x - y \rangle + \tfrac{1}{2} \langle (x - y), f^j_{xx}(y)(x - y) \rangle + \int_0^1 (1 - s) \langle (x - y), [f^j_{xx}(y + s(x - y)) - f^j_{xx}(y)](x - y) \rangle ds \Big\}$$

$$\leq \tilde{\psi}(y, x) + \frac{L}{6} \|x - y\|^3. \quad (8)$$

The other half of the inequality in (7) follows similarly (with $\kappa = L/6$). □

The following result is a simple modification of Theorem 2.1.6.

Lemma 2.5.5. *Suppose that Assumption 2.5.1 is satisfied. Let* $\theta : \mathbb{R}^n \to \mathbb{R}$ *and* $h : \mathbb{R}^n \to \mathbb{R}^n$ *be defined by*

$$\theta(x) \triangleq \min_{h \in \mathbb{R}^n} \tilde{\psi}(x, x + h) - \psi(x), \quad (9a)$$

and

$$h(x) \triangleq \arg \min_{h \in \mathbb{R}^n} \tilde{\psi}(x, x + h). \quad (9b)$$

Then,

(a) both $\theta(\cdot)$ *and* $h(\cdot)$ *are continuous,*

(b) for all $x \in \mathbb{R}^n$, $d\psi(x, h(x)) \leq \theta(x)$,

(c) if $\hat{x}$ *is a solution of (1a,b), then both* $\theta(\hat{x}) = 0$ *and* $h(\hat{x}) = 0$, *and*

(d) for all $x \in \mathbb{R}^n$ *such that* $x \neq \hat{x}$, *the unique minimizer of (1a,b),* $\theta(x) < 0$.

Exercise 2.5.6. Prove Lemma 2.5.5. *Hint*: Mimic the proof of Theorem 2.1.6. □

Our proof of superlinear convergence for Algorithm 2.5.3 will require the following technical results.

Lemma 2.5.7. *Suppose that* $K \in (0, \infty)$ *and that* $t, s \geq 0$ *are such that*

$$t^2 \leq K[(s + t)^3 + s^3], \quad (10a)$$

and

$$0 \leq t \leq \frac{1}{9K}, \quad 0 \leq s \leq \frac{1}{9K}. \quad (10b)$$

Then $t \leq s$ *and*

$$t \leq 3\sqrt{K}\, s^{3/2}. \quad (10c)$$

Furthermore, if $s \leq \gamma/9K$, *with* $\gamma \in (0, 1)$, *then*

$$t \leq \sqrt{\gamma}\, s. \quad (10d)$$

Proof. Let $\lambda \triangleq 1/9K$. Then, from (10a, b),

$$t^2 \leq K(2s^3 + 3s^2 t + 3st^2 + t^3)$$

$$\leq K(2\lambda s^2 + 3\lambda s^2 + 3\lambda t^2 + \lambda t^2). \quad (11a)$$

Hence,

$$(1 - 4\lambda K)t^2 \leq 5K\lambda s^2. \quad (11b)$$

Since $(1 - 4\lambda K) = 5K\lambda = 5/9$, it follows that $t \leq s$. Hence, replacing t by s in

(10a), we obtain (10c).

Now, if $s \le \gamma/9K$, then $\sqrt{s} \le \sqrt{\gamma}/3\sqrt{K}$. Substituting for $\sqrt{s}$ in (10c), we obtain (10d), which completes our proof. □

Corollary 2.5.8. *Suppose that $K \in (0,\infty)$, $\gamma \in (0,1)$, and that $\{\alpha_i\}_{i=0}^{\infty}$ is a sequence of reals such that*

$$\alpha_{i+1}^2 \le K[(\alpha_i + \alpha_{i+1})^3 + \alpha_i^3] \tag{12a}$$

and

$$0 \le \alpha_i \le \frac{\gamma}{9K}, \quad \forall i \in \mathbb{N}. \tag{12b}$$

Then $\alpha_i \to 0$, as $i \to \infty$ superlinearly, with Q-rate 3/2.

Proof. It follows from Lemma 2.5.7 (with $t = \alpha_{i+1}$ and $s = \alpha_i$) (see (10d)) that $\alpha_{i+1} \le \sqrt{\gamma}\,\alpha_i$, for all $i \in \mathbb{N}$. Hence $\alpha_i \to 0$ as $i \to \infty$. The 3/2 Q-rate follows from (10c). □

Finally, we are ready to establish the convergence properties of Algorithm 2.5.3.

Theorem 2.5.9. *Suppose that Assumption 2.5.1 is satisfied. Let $\hat{x}$ be the solution of (1a,b). Then there exists a $\rho > 0$ such that if $\|x_0 - \hat{x}\| \le \rho$ and $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 2.5.3, then $x_i \to \hat{x}$, as $i \to \infty$, Q-superlinearly, with rate at least 3/2.*

Proof. Let $\alpha = m/2$, where $m > 0$ is as in (2). Then, making use of (3) and (6), we find that, for $i = 0, 1, 2, \ldots$,

$$\begin{aligned}\alpha\|x_{i+1} - \hat{x}\|^2 &\le \psi(x_{i+1}) - \psi(\hat{x}) \\ &= \psi(x_{i+1}) - \tilde{\psi}(x_i, x_{i+1}) + \tilde{\psi}(x_i, x_{i+1}) - \psi(\hat{x}) \\ &\le \psi(x_{i+1}) - \tilde{\psi}(x_i, x_{i+1}) + \tilde{\psi}(x_i, \hat{x}) - \psi(\hat{x}),\end{aligned} \tag{13a}$$

because $\tilde{\psi}(x_i, x_{i+1}) \le \tilde{\psi}(x_i, \hat{x})$, by construction of x_{i+1}. It now follows from (12a) and Lemma 2.5.4 that

$$\begin{aligned}\|x_{i+1} - \hat{x}\|^2 &\le K[\|x_{i+1} - x_i\|^3 + \|x_i - \hat{x}\|^3] \\ &\le K[\|(x_{i+1} - \hat{x}) - (x_i - \hat{x})\|^3 + \|x_i - \hat{x}\|^3] \\ &\le K[(\|x_{i+1} - \hat{x}\| + \|x_i - \hat{x}\|)^3 + \|x_i - \hat{x}\|^3],\end{aligned} \tag{13b}$$

where $K = \kappa/\alpha$ and κ is as in Lemma 2.5.4.

Next, since, by Lemma 2.5.5, $h(\cdot)$ is continuous and $h(\hat{x}) = 0$, it follows that given any $\gamma^* \in (0,1)$, there exists a $\bar{\rho} > 0$ such that if $\|x_i - \hat{x}\| \le \bar{\rho}$, then $\|h(x_i)\| = \|x_{i+1} - x_i\| \le \gamma^*/18K$. Let $\rho^* = \min\{\bar{\rho}, \gamma^*/18K\}$. Then, for all x_i such that $\|x_i - \hat{x}\| \le \rho^*$, we must have that

$$\|x_{i+1} - \hat{x}\| \le \|x_{i+1} - x_i\| + \|x_i - \hat{x}\| \le \frac{\gamma^*}{18K} + \rho^* \le \frac{\gamma^*}{9K}. \tag{13c}$$

Letting $t \triangleq \|x_{i+1} - \hat{x}\|$ and $s \triangleq \|x_i - \hat{x}\|$, and using Lemma 2.5.7 (see (10d)), we find that

$$\|x_{i+1} - \hat{x}\| \le \sqrt{\gamma^*}\,\|x_i - \hat{x}\|. \tag{13d}$$

Hence, if $\|x_0 - \hat{x}\| \le \rho^*$, then $\|x_i - \hat{x}\| \le \rho^*$ for all $i = 1, 2, 3, \ldots$, and therefore, by (13d), $\|x_i - \hat{x}\| \to 0$, as $i \to \infty$. It now follows from (13b) and Corollary 2.5.8 (via (10c)) that

$$\|x_{i+1} - \hat{x}\| \le 3\sqrt{K}\,\|x_i - \hat{x}\|^{3/2}, \quad \forall i \in \mathbb{N}, \tag{13e}$$

which completes our proof. □

2.5.2 The Global Newton Method

We will now present an extension of the Newton-Armijo Algorithm 1.4.8, for the minimization of the maximum of a set of twice Lipschitz continuously differentiable, convex functions, i.e., for solving (1a,b) under Assumption 2.5.1. As was also the case with Algorithm 1.4.8, stabilization is achieved by adding an Armijo-type step-size rule to the Local Newton Method, Algorithm 2.5.3. The rate of convergence of the local Newton method is preserved, because, as we will show, near the solution of (1a,b), under Assumption 2.5.1, the step size becomes unity, i.e., the Global Newton Method reverts to the Local Newton Method.

Global Polak-Mayne-Higgins Algorithm 2.5.10
(Global Newton Method).

Data. $x_0 \in \mathbb{R}^n$, $\alpha, \beta \in (0,1)$.

Step 0. Set i = 0.

Step 1. Compute $\theta(x_i)$ and $h_i = h(x_i)$, according to (9a) and (9b). If $\theta(x_i) = 0$, stop.

Step 2. Compute the step size

$$\lambda_i \triangleq \max_{k \in \mathbb{N}} \{\beta^k \mid \psi(x_i + \beta^k h_i) - \psi(x_i) \leq \beta^k \alpha\theta(x_i)\}. \tag{14}$$

Step 3. Set $x_{i+1} = x_i + \lambda_i h_i$. Replace i by $i+1$, and go to Step 1.

First we show that Algorithm 2.5.10 is globally convergent.

Theorem 2.5.11. *Suppose that Assumption 2.5.1 holds and that $\hat{x}$ is the solution of (1a,b). Then any sequence $\{x_i\}_{i=0}^{\infty}$, constructed by Algorithm 2.5.10 converges to $\hat{x}$.*

Proof. First, because of Assumption 2.5.1*(ii)*, it follows from Proposition 5.2.15 that the level sets of $\psi(\cdot)$ are bounded and, by construction in Step 2, $\psi(x_{i+1}) < \psi(x_i)$. Hence any sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 2.5.10 must have accumulation points. For the sake of contradiction, suppose that the sequence $\{x_i\}_{i=0}^{\infty}$ does not converge to $\hat{x}$. Then it must have an accumulation point $x^* \neq \hat{x}$. By Lemma 2.5.5, we then have that $\theta(x^*) < 0$ and $h(x^*) \neq 0$. Since, by Lemma 2.5.5, the directional derivative $d\psi(x^*, h(x^*)) \leq \theta(x^*) < 0$, it follows that, for any $\varepsilon > 0$ such that $\alpha + \varepsilon < 1$, there is a $\lambda^* \in \{1, \beta, \beta^2, \ldots\}$ such that

$$\psi(x^* + \lambda^* h(x^*)) - \psi(x^*) \leq \lambda^*(\alpha + \varepsilon) d\psi(x^*, h(x^*))$$

$$\leq \lambda^*(\alpha + \varepsilon)\theta(x^*), \tag{15a}$$

so that

$$\psi(x^* + \lambda^* h(x^*)) - \psi(x^*) - \lambda^*\alpha\theta(x^*) \leq \lambda^*\varepsilon\theta(x^*). \tag{15b}$$

Hence, making use of the continuity of $\theta(\cdot)$ and $h(\cdot)$, we conclude that, for all x_i sufficiently near x^*, the step-size $\lambda_i \geq \lambda^*$ and $\theta(x_i) \leq \theta(x^*)/2$. Therefore, for all such x_i,

$$\psi(x_i + \lambda_i h(x_i)) - \psi(x_i) \leq \lambda_i \alpha\theta(x_i) \leq \lambda^*\alpha\theta(x^*)/2. \tag{15c}$$

Since the sequence $\{\psi(x_i)\}_{i=0}^{\infty}$ is monotone decreasing, (15b) implies that $\psi(x_i) \to -\infty$, as $i \to \infty$, which is a contradiction, because the level sets of $\psi(\cdot)$ are bounded. Hence the theorem must be true. □

Next we establish superlinear convergence.

Theorem 2.5.12. *Suppose that Assumption 2.5.1 holds and that $\hat{x}$ is the solution of problem (1a,b). Then any sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 2.5.10 converges to $\hat{x}$ superlinearly, with Q-rate at least 3/2.*

Proof. Since $\{x_i\}_{i=0}^{\infty}$ converges to $\hat{x}$ by Theorem 2.5.11, we need to show only that there exists an i_0 such that $\lambda_i = 1$ for all $i \geq i_0$, so that Algorithm 2.5.10 reduces to Algorithm 2.5.3, and invoke Theorem 2.5.8.

Now, it follows from (7) that

$$\theta(x_i) = \tilde{\psi}(x_i, x_i + h(x_i)) - \psi(x_i + h(x_i)) + \psi(x_i + h(x_i)) - \psi(x_i)$$

$$\geq \psi(x_i + h(x_i)) - \psi(x_i) - \kappa\|h(x_i)\|^3. \tag{16a}$$

Hence

$$\psi(x_i + h(x_i)) - \psi(x_i) \leq \alpha\theta(x_i) + [(1-\alpha)\theta(x_i) + \kappa\|h(x_i)\|^3]. \tag{16b}$$

Next we establish a relationship between $\theta(x)$ and $\|h(x)\|$. Since $x + h(x)$ is the minimizer of $\tilde{\psi}(x, \cdot)$, it follows that it satisfies the first-order optimality condition

$$0 \in \partial_2 \tilde{\psi}(x, x + h(x)). \tag{16c}$$

Then it follows from (16c), the definition of the subgradient $\partial_2\tilde{\psi}(x, x + h(x))$ (see (5.4.7b)) and the Caratheodory Theorem 5.2.5, that there exists a multiplier $\mu \in \Sigma_q$ such that

$$0 = \sum_{j=1}^{q} \mu^j [\nabla f^j(x) + f^j_{xx}(x) h(x)]. \tag{16d}$$

Since the $\mu^j \geq 0$ in (16d), it follows from (2b) that the matrix $\sum_{j=1}^{q} \mu^j f^j_{xx}(x)$ is invertible and hence that

$$h(x) = -\Big[\sum_{j=1}^{q} \mu^j f^j_{xx}(x)\Big]^{-1} \sum_{j=1}^{q} \mu^j \nabla f^j(x). \tag{16e}$$

Furthermore (since $\sum_{j=1}^{q} \hat{\mu}^j \psi(\hat{x}) = \psi(\hat{x})$), the following complementary slackness condition (see (2.1.7c)) is satisfied:

$$\theta(x) = \sum_{j=1}^{q} \mu^j \{[f^j(x) - \psi(x)] + \langle \nabla f^j(x), h(x)\rangle + \tfrac{1}{2}\langle h(x), f^j_{xx}(x) h(x)\rangle\}. \tag{16f}$$

Substituting for $h(x)$ from (16e) into (16f), we find, in view of Assumption 2.5.1*(ii)*, that

$$\theta(x) = \sum_{j=1}^{q} \mu^j [f^j(x) - \psi(x)] - \tfrac{1}{2}\langle h(x), \Big[\sum_{j=1}^{q} \mu^j f^j_{xx}(x)\Big] h(x)\rangle$$

$$\leq -\tfrac{1}{2} m \|h(x)\|^2, \tag{16g}$$

with the last line following from the fact that $f^j(x) - \psi(x) \leq 0$ for all $j \in \mathbf{q}$.

Substituting for $\theta(x)$ from (16g) into (16b), we find that

$$\psi(x_i + h(x_i)) - \psi(x_i) \leq \alpha\theta(x_i) - [m(1-\alpha)/2 - \kappa\|h(x_i)\|]\|h(x_i)\|^2 . \quad (16h)$$

Since $h(x_i) \to 0$, as $i \to \infty$, it follows that there exists an i_0 such that, for all $i \geq i_0$,

$$\psi(x_i + h(x_i)) - \psi(x_i) \leq \alpha\theta(x_i) \quad (16i)$$

i.e., that $\lambda_i = 1$. This completes our proof. □

It should be obvious that the "dual" expression for $\theta(x)$ is

$$\theta(x) = \max_{\mu \in \Sigma_q} \sum_{j=1}^{q} \mu^j [f^j(x) - \psi(x)]$$

$$- \tfrac{1}{2}\langle \Big[\sum_{j=1}^{q} \mu^j f^j_{xx}(x)\Big]^{-1} \sum_{j=1}^{q} \mu^j \nabla f^j(x), \sum_{j=1}^{q} \mu^j \nabla f^j(x)\rangle . \quad (16j)$$

We see that this is no ordinary quadratic programming problem. However, as was shown in [PMH92], it can be solved reasonably efficiently using the Levitin-Polyak projected Newton method, described in [LeP.66].

2.5.3 Notes

The material for this section was drawn from [PMH.92], and uses a novel technique for establishing superlinear convergence. Earlier versions of Newton's method for min-max problems, e.g., [Han.78, PMH.91], used the results in [Rob.74] to establish quadratic convergence. Since the results in [Rob.74] depend on the Implicit Function Theorem, earlier proofs of superlinear convergence required an assumption of linear independence of the active gradients at all local minimizers, which is rather restrictive for large problems.

The algorithms in [PMH.91] also use the search direction finding subprocedure (9b).

The algorithm in [Han.78] is only a local algorithm. It updates not only the variable x, but also a multiplier μ. Thus, given $x_i \in \mathbb{R}^n$, $\mu_i \in \mathbb{R}^q$, it computes their update by solving the direction finding problem

$$\min_{x \in \mathbb{R}^n} \max_{\mu \in \Sigma_q} \Big\{ \sum_{j \in \mathbf{q}} \mu^j [f^j(x_i) - \psi(x_i)$$

$$+ \langle \nabla f^j(x_i), x - x_i \rangle] + \tfrac{1}{2}\langle x - x_i, \sum_{j \in \mathbf{q}} \mu_i^j f^j_{xx}(x_i)(x - x_i)\rangle \Big\} . \quad (17)$$

Note that (17) is easier to solve than (9b). A globalized version of the algorithm in [Han.78], using a nonmonotone line search, was presented in [ZhT.93].

2.6 Phase I - Phase II Methods of Centers

We recall that algorithm models for methods of centers were introduced in Section 2.3. In this section we will be concerned with phase I - phase II methods of centers for solving the inequality constrained problem **ICP** that we have already encountered in Section 2.2, i.e., the problem

$$\textbf{ICP} \qquad \min \{ f^0(x) \mid f^j(x) \leq 0,\ j \in \mathbf{q} \}, \quad (1a)$$

where we will assume that

$$f^0(x) \triangleq \max_{k \in \mathbf{p}} c^k(x), \quad (1b)$$

with the functions $f^j, c^k : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, $k \in \mathbf{p}$, at least once continuously differentiable.

We will continue to use the definitions of $\psi(x)$ and $\psi(x)_+$ given in (2.2.1a) and (2.2.8b), respectively, i.e.,

$$\psi(x) \triangleq \max_{j \in \mathbf{q}} f^j(x), \quad (1c)$$

$$\psi(x)_+ \triangleq \max \{0, \psi(x)\}. \quad (1d)$$

We will refer to the function $\psi(\cdot)$ as the *constraint violation function*.

It turns out that many phase I - phase II methods of centers can be presented in two closely related versions: a *novel* "single-cost" version satisfying the conditions in Theorem 2.3.2 and a *classical* "two-cost" version satisfying the conditions in Theorem 2.3.4. The "single-cost" versions evaluate both the cost function $f^0(\cdot)$ and the constraint violation function $\psi(\cdot)$ at both feasible and infeasible points while the "two-cost" versions evaluate both of these functions at feasible points, but evaluate only the constraint violation function at infeasible points. Although, at first glance, the "single-cost" versions seem to be less efficient, they make up for the extra function evaluations by being more favorably influenced by the cost function at infeasible points, and hence their overall performance turns out to be, if anything, slightly superior to that of "two-cost" versions. The most serious advantage of "single-cost" versions lies in the fact that they satisfy the hypotheses of the theory of consistent approximations theory presented in Chapters 3 and 4 and hence can be used in the solution of semi-infinite optimization and optimal control problems, while it is not clear that the same is true for "two-cost" versions.

We will present two types of phase I - phase II method of centers. The first is derived from a min-max definition of a center, while the second one is derived from the definition of a center as a minimizer of barrier function.

2.6.1 Min-Max-Type Phase I - Phase II Methods

We begin with a method of the form depicted in Fig. 2.3.1, which, given x_i, computes x_{i+1} by applying one iteration of the PPP Algorithm 2.4.1 to the problem $\min_{x \in \mathbb{R}^n} F(x_i, x)$, where the parametrized function $F(\cdot,\cdot)$ is defined as in (2.2.8a), i.e.,

$$F(z,x) \triangleq \max\{f^0(x) - f^0(z) - \gamma\psi(z)_+, \psi(x) - \psi(z)_+\}$$

$$= \max\left\{\max_{k \in \mathbf{p}}\{c^k(x) - f^0(z) - \gamma\psi(z)_+\}, \max_{j \in \mathbf{q}}\{f^j(x) - \psi(z)_+\}\right\}, \quad (1e)$$

where $\gamma > 0$. As we will see in the subsection dealing with its rate of convergence, when the parameter γ is sufficiently large, under a convexity assumption, a feasible point will be computed in a finite number of iterations, and this number decreases as γ increases. The algorithm uses the optimality function $\theta(\cdot)$ and search direction function $h(\cdot)$ defined by (2.2.8f) (or, equivalently, (2.2.9e)) and (2.2.8g) (or, equivalently, (2.2.9f)), respectively. Recall that the definition of $\theta(\cdot)$ and $h(\cdot)$ requires two parameters $\gamma, \delta > 0$, whose actual values have no impact on the truth of the first-order optimality condition stated in Theorem 2.2.8*(f)*, but which, as we will see, do affect the qualitative behavior of the following algorithms. The algorithm also uses the function $F(\cdot,\cdot)$ which was defined in (2.2.8a).

Polak-He Algorithm 2.6.1.

Parameters. $\gamma, \delta > 0$, $\alpha \in (0,1]$, $\beta \in (0,1)$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute the the *optimality function* value $\theta_i = \theta(x_i)$ and the corresponding *search direction* $h_i = h(x_i)$, according to the dual form formulas (2.2.9e) and (2.2.9f), respectively, i.e.,

$$\theta_i = -\min_{\substack{\mu \in \Sigma_q^0 \\ \nu \in \Sigma_p}} \left\{\mu^0 \sum_{k=1}^{p} \nu^k [f^0(x_i) - c^k(x_i) + \gamma\psi(x_i)_+] + \sum_{j=1}^{q} \mu^j [\psi(x_i)_+ - f^j(x_i)] + \frac{1}{2\delta}\|\mu^0 \sum_{k=1}^{p} \nu^k \nabla c^k(x_i) + \sum_{j=1}^{q} \mu^j \nabla f^j(x_i)\|^2 \right\}, \quad (2a)$$

and

$$h(x_i) = -\frac{1}{\delta}\left\{\mu_x^0 \sum_{k=1}^{p} \nu_x^k \nabla c^k(x_i) + \sum_{j=1}^{q} \mu_x^j \nabla f^j(x_i)\right\}, \quad (2b)$$

where (μ_x, ν_x) is any solution of (2).

Step 2. Compute the *step-size* λ_i:

$$\lambda_i = \max_{k \in \mathbb{N}}\{\beta^k \mid F(x_i, x_i + \beta^k h_i) \le \beta^k \alpha\theta_i\}. \quad (2c)$$

Step 3. Set $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i+1$, and go to Step 1.

Theorem 2.6.2. *Consider the problem (1a,b), and suppose that the functions $f^j, c^k : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, $k \in \mathbf{p}$, are at least once continuously differentiable. If $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 2.6.1 in solving (1a,b), then any accumulation point $\hat{x}$ of the sequence $\{x_i\}_{i=0}^{\infty}$ satisfies $\theta(\hat{x}) = 0$.*

Proof. First we note that, by the construction of the sequence $\{x_i\}_{i=0}^{\infty}$, $F(x_i, x_{i+1}) \le F(x_i, x_i) = 0$. Consequently, we conclude from the definition of $F(x_i,\cdot)$, that, for all $i \in \mathbb{N}$,

$$f^0(x_{i+1}) \le f^0(x_i) + \gamma\psi(x_i)_+ \quad (3a)$$

and

$$\psi(x_{i+1}) \le \psi(x_i)_+. \quad (3b)$$

Now suppose that, for some $K \subset \mathbb{N}$, $x_i \to^K x^*$, as $i \to \infty$, and that $\theta(x^*) < 0$ (we recall that by part *(a)* of Theorem 2.2.8, $\theta(x) \le 0$ for all $x \in \mathbb{R}^n$). To complete our proof, we repeat the steps used in the Proof of Theorem 2.4.2. Thus, by part *(b)* of Theorem 2.2.8,

$$d_2F(x^*, x^*; h(x^*)) \le \theta(x^*) - \tfrac{1}{2}\delta\|h(x^*)\|^2 < 0. \quad (3c)$$

It now follows from the definition of the directional derivative $d_2F(\cdot,\cdot;\cdot)$ that for any $\varepsilon > 0$, such that $0 < \alpha - \varepsilon < 1$, that there exists a $k_\varepsilon \in \mathbb{N}$ such that

$$F(x^*, x^* + \beta^{k_\varepsilon} h(x^*)) - F(x^*, x^*) \le \beta^{k_\varepsilon}(\alpha - \varepsilon) d_2F(x^*, x^*; h(x^*))$$

$$\le \beta^{k_\varepsilon}(\alpha - \varepsilon)[\theta(x^*) - \tfrac{1}{2}\delta\|h(x^*)\|^2]. \quad (3d)$$

Hence, taking into account that $F(x^*, x^*) = 0$,

$$F(x^*, x^* + \beta^{k_\varepsilon} h(x^*)) - \beta^{k_\varepsilon}\alpha\theta(x^*) \le -\beta^{k_\varepsilon}[\varepsilon\theta(x^*) + \tfrac{1}{2}\delta(\alpha - \varepsilon)\|h(x^*)\|^2]. \quad (3e)$$

Now

$$\varepsilon\theta(x^*) + \tfrac{1}{2}\delta(\alpha - \varepsilon)\|h(x^*)\|^2 > 0, \quad (3f)$$

for all $\varepsilon > 0$ such that

$$\frac{2\varepsilon}{\delta(\alpha-\varepsilon)} < -\frac{\|h(x^*)\|^2}{\theta(x^*)} \triangleq \omega^*, \tag{3g}$$

i.e., for all $\varepsilon > 0$ such that $\varepsilon < \varepsilon' \triangleq \delta\omega^*\alpha/(2+\delta\omega^*)$. Let $\varepsilon^* \triangleq \varepsilon'/2$. Since $F(z,x)$ is continuous both in its argument x and its parameter z, and both $h(\cdot)$ and $\theta(\cdot)$ are continuous, and since $F(x_i,x_i) = 0$, it now follows that there exists a $\rho > 0$ such, that for all $x_i \in B(x^*,\rho)$,

$$F(x_i, x_i + \beta^{k_{\varepsilon^*}} h(x_i)) - \beta^{k_{\varepsilon^*}}\alpha\theta(x_i) \le 0, \tag{3h}$$

which shows that, for all $x_i \in B(x^*,\rho)$, $\lambda(x_i) \ge \beta^{k_{\varepsilon^*}}$. Next, since $\theta(\cdot)$ is continuous, there exists a $\rho^* \in (0,\rho)$ such that for all $x_i \in B(x^*,\rho^*)$, $\theta(x_i) \le \theta(x^*)/2$. Therefore it follows from the step-size rule (2c) and (3f) that for all $x_i \in B(x^*,\rho^*)$,

$$F(x_i, x_{i+1}) \le \beta^{k_{\varepsilon^*}}\alpha\theta(x_i) \le \tfrac{1}{2}\beta^{k_{\varepsilon^*}}\alpha\theta(x^*). \tag{3i}$$

Since $x_i \to^K x^*$, as $i \to \infty$, there exists an $i_0 \in \mathbb{N}$ such that $x_i \in B(x^*,\rho^*)$ for all $i \in K$, $i > i_0$, and hence (3i) is satisfied for all $i \in K$, $i > i_0$.

Now we must consider two cases.

Case (i). There exists $i_1 \in \mathbb{N}$, $i_1 \ge i_0$, such that $\psi(x_{i_1}) \le 0$. Then it follows from (3a) and (3b) that $\psi(x_i) \le 0$ and $f^0(x_{i+1}) \le f^0(x_i)$ for all $i > i_1$. In addition, it follows from (3i) that, for all $i \in K$, $i \ge i_1$,

$$f^0(x_{i+1}) - f^0(x_i) \le F(x_i, x_{i+1}) \le \alpha\beta^{k_{\varepsilon^*}}\theta(x^*)/2. \tag{3j}$$

Since the sequence $\{f^0(x_i)\}_{i=i_1}^{\infty}$ is monotonically decreasing, we conclude from (3j) and Proposition 5.1.16 that $f^0(x_i) \to -\infty$, as $i \to \infty$, which contradicts the fact that, by continuity of $f^0(\cdot)$, $f^0(x_i) \to^K f^0(x^*)$ as $i \to \infty$.

Case *(ii).* $\psi(x_i) > 0$ for all i. Then it follows from (3b) that $\psi(x_{i+1}) \le \psi(x_i)$ for all i. It now follows from (3j), that for all $i \in K$, $i > i_0$,

$$\psi(x_{i+1}) - \psi(x_i) \le F(x_i, x_{i+1}) \le \alpha\beta^{k_{\varepsilon^*}}\theta(x^*)/2. \tag{3k}$$

Since the sequence $\{\psi(x_i)\}_{i=0}^{\infty}$ is monotonically decreasing, we conclude from (3j) and Proposition 5.1.16 that $\psi(x_i) \to -\infty$, as $i \to \infty$, which contradicts the fact that $\psi(x_i) > 0$ for all $i \in \mathbb{N}$.

Hence we conclude that no matter which one of the above two cases holds, $\theta(x^*) = 0$ must hold, which completes our proof. □

Note that Theorem 2.6.2 does not claim that any accumulation point $\hat{x}$ of a sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 2.6.1 satisfies not only $\theta(\hat{x}) = 0$, but also $\psi(\hat{x}) \le 0$, i.e., it does not claim that the accumulation points will be stationary (i.e., quasi-stationary and feasible (see (2.3.1c,d))). The reason for this can be found by examining the expression (2.2.9e) for $\theta(x)$. We see that, when $\psi(x) > 0$ and $0 \in \partial\psi(x)$, i.e., when there exists a multiplier $\mu \in \Sigma_q^0$ such that

$\mu^0 = 0$ and *(i)* $\sum_{j=1}^q \mu^j[\psi(x) - f^j(x)] = 0$, and *(ii)* $\sum_{j=1}^q \mu^j \nabla f^j(x) = 0$, then $\theta(x) = 0$. Thus we see that Algorithm 2.6.1 can be trapped by an infeasible local minimizer of $\psi(\cdot)$. This cannot occur when the problem **ICP** (1a,b) satisfies the following commonly made assumption:

Assumption 2.6.3. *Consider problem* **ICP** *(1a,b). We will assume that for all* $x \in \mathbb{R}^n$ *such that* $\psi(x) \ge 0$, $0 \notin \partial\psi(x)$. □

The "two-cost" version of Algorithm 2.6.1 also uses the optimality function $\theta(\cdot)$ and search direction function $h(\cdot)$, defined by (2.2.9e) and (2.2.9f), respectively. Hence it also uses the two parameters $\gamma, \delta > 0$. It has the following form:

Polak-Trahan-Mayne Algorithm 2.6.4.

Parameters. $\gamma, \delta > 0$, $\alpha, \beta \in (0,1)$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute the the *optimality function* value $\theta_i = \theta(x_i)$ and the corresponding *search direction* $h_i = h(x_i)$ according to (2) and (2b), respectively.

Step 2. If $\psi(x_i) > 0$, set the step-size

$$\lambda_i = \max\{\beta^k \mid k \in \mathbb{N},\ \psi(x_i + \beta^k h_i) - \psi(x_i) \le \beta^k\alpha\theta_i\}. \tag{4a}$$

Else, set the step-size

$$\lambda_i = \max\{\beta^k \mid k \in \mathbb{N}, f^0(x_i + \beta^k h_i) - f^0(x_i) \le \beta^k\alpha\theta_i, \psi(x_i + \beta^k h_i) \le 0\} \tag{4b}$$

Step 3. Set $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i+1$, and go to Step 1.

Remark 2.6.5. The behavioral difference between Algorithms 2.6.1 and 2.6.4 can be explained as follows. The infeasible points x_i generated by Algorithm 2.6.4 are attracted to the set $S_1 \triangleq \{x \in \mathbb{R}^n \mid 0 \in \partial\psi(x)\}$. The feasible points x_i generated by Algorithm 2.6.4 and *all* the points x_i generated by Algorithm 2.6.1 are attracted to the set $S_2 \triangleq \{x \in \mathbb{R}^n \mid \theta(x) = 0\}$. Hence points generated by Algorithm 2.6.1 are likely to enter the feasible set at a more favorable location than those generated by Algorithm 2.6.4. □

Exercise 2.6.6. Mimic the proof of Theorem 2.6.2 to establish the following result.

Theorem 2.6.7. *Consider problem (1a,b) and suppose that the functions*

$f^j, c^k : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, $k \in \mathbf{p}$, are at least once continuously differentiable. If $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 2.6.4 in solving (1a,b), then any accumulation point $\hat{x}$ of the sequence $\{x_i\}_{i=0}^{\infty}$ satisfies $\theta(\hat{x}) = 0$. □

2.6.2 Rate of Convergence

Next we turn to an analysis of the rate of convergence of Algorithm 2.6.1 and of the effects of the parameters γ and δ.

Assumption 2.6.8. *We will assume that, in problem (1a,b),*

(i) the functions $f^j(\cdot)$, $j \in \mathbf{q}$, and $c^k(\cdot)$, $k \in \mathbf{p}$, are twice continuously differentiable,

(ii) the set $\{x \mid \psi(x) < 0\}$ is not empty, and

(iii) there exist constants $0 < m \le \delta \le M < \infty$ such that, for all $x, y \in \mathbb{R}^n$, $j \in \mathbf{q}$, and $k \in \mathbf{p}$,

$$m\|y\|^2 \le \langle y, f^j_{xx}(x)y \rangle \le M\|y\|^2 \tag{5a}$$

and

$$m\|y\|^2 \le \langle y, c^k_{xx}(x)y \rangle \le M\|y\|^2 . \tag{5b}$$

□

Referring to Remark 1.3.19*(b)*, we see that it is quite easy to estimate a value of $\delta \in [m, M]$ for twice continuously differentiable convex functions.

The following result is a direct consequence of the second-order expansion formula (5.1.17d) and Assumption 2.6.8:

Lemma 2.6.9. *Suppose that parts (i) and (iii) of Assumption 2.6.8 hold. Then, for any $x, h \in \mathbb{R}^n$, $j \in \mathbf{q}$, and $k \in \mathbf{p}$,*

$$f^j(x) + \langle \nabla f^j(x), h \rangle + \frac{m}{2}\|h\|^2 \le f^j(x+h) \le f^j(x) + \langle \nabla f^j(x), h \rangle + \frac{M}{2}\|h\|^2, \tag{6a}$$

$$c^k(x) + \langle \nabla c^k(x), h \rangle + \frac{m}{2}\|h\|^2 \le c^k(x+h) \le c^k(x) + \langle \nabla c^k(x), h \rangle + \frac{M}{2}\|h\|^2. \tag{6b}$$

□

Lemma 2.6.10. *Suppose that Assumption 2.6.8 is satisfied. Then,*

(a) for all x such that $\psi(x) \ge 0$, $0 \notin \partial\psi(x)$,

(b) problem (1a,b) has a unique solution, and

(c) the unique solution of (1a,b) is the unique zero of $\theta(\cdot)$.

Proof. *(a)* Because of Assumption 2.6.8*(i)* and *(iii)*, it follows from Proposition 5.2.14 that the functions $f^j(\cdot)$, $j \in \mathbf{q}$, are strictly convex and have convex, compact level sets. Hence the function $\psi(\cdot)$ is also strictly convex and has convex, compact level sets that are the intersection of the level sets of the $f^j(\cdot)$. Therefore $\psi(\cdot)$ has a unique global minimizer $\bar{x}$, which is also the only point satisfying the relation $0 \in \partial\psi(x)$. It now follows from Assumption 2.6.8*(ii)* that $\psi(\bar{x}) < 0$.

(b) Because the functions $c^k(\cdot)$, $k \in \mathbf{p}$, are strictly convex, by Proposition 5.2.15, the cost function $f^0(\cdot)$ is strictly convex. Since both $f^0(\cdot)$ and the feasible set $\{x \in \mathbb{R}^n \mid \psi(x) \le 0\}$ are strictly convex, it follows that problem (1a,b) has a unique solution.

(c) Suppose that x^* is such that $\theta(x^*) = 0$. Then it follows from (2.2.9g) and (2.2.9i) that $0 \in \overline{GF}(x^*)$. Now suppose that $\psi(x^*) > 0$. Then it follows from the form of (2.2.9i) and the fact that $0 \in \overline{GF}(x^*)$, that $0 \in \partial\psi(x^*)$, contradicting *(a)* above. Hence we must have that $\psi(x^*) \le 0$. It now follows from Theorem 2.2.8*(f)* and Corollary 2.2.6 that x^* is a global minimizer of problem (1a,b). Since the global minimizer of (1a,b) is unique and there are no other stationary points, the zero of $\theta(\cdot)$ is unique. □

Next, referring to (2.2.9e), for any $x \in \mathbb{R}^n$, let the set-valued maps $\hat{\Sigma}_q^0(x) \subset \Sigma_q^0$ and $\hat{\Sigma}_p(x) \subset \Sigma_p$ be defined by

$$(\hat{\Sigma}_q^0(x), \hat{\Sigma}_p(x)) \triangleq \arg \min_{\substack{\mu \in \Sigma_q^0 \\ \nu \in \Sigma_p}} \Big\{ \mu^0 \sum_{k=1}^{p} \nu^k [f^0(x) - c^k(x) + \gamma\psi(x)_+]$$

$$+ \sum_{j=1}^{q} \mu^j [\psi(x)_+ - f^j(x)]$$

$$+ \tfrac{1}{2}\delta\|\mu^0 \sum_{k=1}^{p} \nu^k \nabla c^k(x) + \sum_{j=1}^{q} \mu^j \nabla f^j(x)\|^2 \Big\}. \tag{7a}$$

Let $\hat{x}$ denote the unique solution of (1a,b). Then, because $\theta(\hat{x}) = 0$, $\psi(\hat{x})_+ = 0$, and $\xi^0 \ge 0$ for all $\bar{\xi} = (\xi^0, \xi) \in \overline{GF}(\hat{x})$, it follows directly from (7a) that

$$\hat{\Sigma}_q^0(\hat{x}) = \{\mu \in \Sigma_q^0 \mid 0 \in \sum_{j=0}^{q} \mu^j \partial f^j(\hat{x}),\ \sum_{j=1}^{q} \mu^j f^j(\hat{x}) = 0\}. \tag{7b}$$

Since the subgradient $\partial f^0(\hat{x})$ is compact and convex and the subgradients $\partial f^j(x)$, $j \in \mathbf{q}$, are singletons, the set $\hat{\Sigma}_q^0(\hat{x})$ is compact and convex. Let

$$\underline{\mu}(\hat{x})^0 \triangleq \min\{\mu^0 \mid \mu \in \hat{\Sigma}_q^0(\hat{x})\}, \tag{7c}$$

and let

$$\bar{\mu}^0(\hat{x}) \triangleq \max\{\mu^0 \mid \mu \in \hat{\Sigma}_q^0(\hat{x})\}. \tag{7d}$$

Then, we have following result:

Lemma 2.6.11. *Suppose that Assumptions 2.6.3 and 2.6.8 hold, and that $\hat{x}$ is the unique solution of (1a,b). Then,*

(a) $0 < \underline{\mu}^0(\hat{x}) \le \bar{\mu}^0(\hat{x}) \le 1$,

(b) for any $\varepsilon \in (0,1)$, there exists a $\hat{\rho} > 0$ such that for any $x \in B(\hat{x}, \hat{\rho})$ and $\mu = (\mu^0, \mu^1, \ldots, \mu^q) \in \hat{\Sigma}_q^0(x)$,

$$\underline{\mu}^0(\hat{x})(1-\varepsilon) \le \mu^0 \le \bar{\mu}^0(\hat{x})(1+\varepsilon), \tag{8a}$$

and

(c) for all $x \in \mathbb{R}^n$,

$$\bar{\mu}^0(\hat{x})[f^0(\hat{x}) - f^0(x)] \le [1 - \bar{\mu}^0(\hat{x})]\psi_+(x) \tag{8b}$$

and

$$\tfrac{1}{2}m\|x-\hat{x}\|^2 \le \underline{\mu}^0(\hat{x})[f^0(x) - f^0(\hat{x})] + [1 - \underline{\mu}^0(\hat{x})]\psi_+(x). \tag{8c}$$

Proof. *(a)* If $\psi(\hat{x}) < 0$, then it follows from (7b) that $\hat{\mu} = (1, 0, \ldots, 0)$ is the only element in $\hat{\Sigma}_q^0(\hat{x})$. If $\psi(\hat{x}) = 0$, then, by Assumption 2.6.3, $0 \notin \partial\psi(\hat{x})$ and therefore, from (7b), for any $\mu \in \hat{\Sigma}_q^0(\hat{x})$ we must have that $\mu^0 > 0$. It now follows from the compactness of $\hat{\Sigma}_q^0(\hat{x})$ that $\underline{\mu}^0 > 0$.

(b) Since by Theorem 5.4.3 the set-valued map $\hat{\Sigma}_q^0(\cdot)$ is outer semicontinuous, and it is compact valued and uniformly bounded, by inspection, the desired result follows directly from part *(b)* of Theorem 5.3.7.

(c) For all $j \in \mathbf{q}$, let $\hat{\xi}_j \triangleq \nabla f^j(\hat{x})$. Then, by (2.2.5c,d), for any $\mu = (\mu^0, \mu^1, \ldots, \mu^q) \in \hat{\Sigma}_q^0(\hat{x})$, there exists a $\hat{\xi}_0 \in \partial f^0(\hat{x})$ such that

$$\sum_{j=0}^{q} \mu^j \hat{\xi}_j = 0, \tag{9a}$$

and, in addition,

$$0 = \sum_{j=1}^{q} \mu^j f^j(\hat{x}). \tag{9b}$$

Next, it follows from (6b) and (5.4.7c) (with x replaced by $\hat{x}$ and h replaced by $x - \hat{x}$) that for any $x \in \mathbb{R}^n$,

$$\begin{aligned} f^0(x) &= \max_{k \in \mathbf{p}} c^k(x) \ge \max_{k \in \hat{\mathbf{p}}(\hat{x})} c^k(x) \\ &\ge \max_{k \in \hat{\mathbf{p}}(\hat{x})} c^k(\hat{x}) + \langle \nabla c^k(\hat{x}), x - \hat{x} \rangle + \tfrac{1}{2}m\|x-\hat{x}\|^2 \\ &= f^0(\hat{x}) + df^0(\hat{x}; x - \hat{x}) + \tfrac{1}{2}m\|x-\hat{x}\|^2 \\ &\ge f^0(\hat{x}) + \langle \hat{\xi}_0, x - \hat{x} \rangle + \frac{m}{2}\|x-\hat{x}\|^2. \end{aligned} \tag{9c}$$

Making use of the fact that $\psi(x)_+ \ge f^j(x)$ for all $j \in \mathbf{q}$, of (6a) (with x replaced by $\hat{x}$ and h replaced by $x - \hat{x}$), and of (9a,b,c), we obtain that for any $x \in \mathbb{R}^n$,

$$\begin{aligned} \mu^0 f^0(x) + (1-\mu^0)\psi(x)_+ &\ge \sum_{j=0}^{q} \mu^j f^j(x) \\ &\ge \sum_{j=0}^{q} \mu^j f^j(\hat{x}) + \tfrac{1}{2}m\|x-\hat{x}\|^2 \\ &= \mu^0 f^0(\hat{x}) + \tfrac{1}{2}m\|x-\hat{x}\|^2. \end{aligned} \tag{9d}$$

Replacing μ^0 by $\bar{\mu}^0(\hat{x})$ and $\underline{\mu}^0(\hat{x})$ in (9d), we obtain, respectively, (8b,c). □

Lemma 2.6.12. *Suppose that Assumption 2.6.8 is satisfied and that $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 2.6.1. Then, for all $i \in \mathbb{N}$ and $\mu_i = (\mu_i^0, \ldots, \mu_i^q) \in \hat{\Sigma}_q^0(x_i)$,*

$$F(x_i, x_{i+1}) \le \frac{\alpha\beta m\,\mu_i^0}{M}[f^0(\hat{x}) - f^0(x_i)] - \frac{\alpha\beta m}{M}[1 + (\gamma-1)\mu_i^0]\psi(x_i)_+, \tag{10a}$$

$$\begin{aligned} f^0(x_{i+1}) - f^0(\hat{x}) &\le (1 - \frac{\alpha\beta m\,\mu_i^0}{M})[f^0(x_i) - f^0(\hat{x})] \\ &\quad + [\gamma - \frac{\alpha\beta m\,(1 + (\gamma-1)\mu_i^0)}{M}]\psi(x_i)_+, \end{aligned} \tag{10b}$$

and

$$\psi(x_{i+1}) \le \frac{\alpha\beta m\,\mu_i^0}{M}[f^0(\hat{x}) - f^0(x_i)] + [1 - \alpha\beta\frac{m}{M}(1 + (\gamma-1)\mu_i^0)]\psi(x_i)_+. \tag{10c}$$

Proof. We begin by obtaining a bound on the decrease in $F(\cdot, x_i)$ at the ith iteration. In view of (6a,b), the definition of $F(\cdot,\cdot)$ in (1e), and the fact that $c^k(x_i) - f^0(x_i) - \gamma\psi(x_i)_+ \le 0$ for all $k \in \mathbf{p}$ and $f^j(x_i) - \psi(x_i)_+ \le 0$ for all

$j \in \mathbf{q}$, we find that, for all $\lambda \in [0, \delta/M]$,

$$F(x_i, x_i + \lambda h(x_i)) = \max \Big\{ f^0(x_i + \lambda h(x_i)) - f^0(x_i) - \gamma\psi(x_i)_+, \max_{j \in \mathbf{q}} \{ f^j(x_i + \lambda h(x_i)) - \psi(x_i)_+ \} \Big\}$$

$$\leq \lambda \max \Big\{ \max_{k \in \mathbf{p}} \{ c^k(x_i) - f^0(x_i) - \gamma\psi(x_i)_+ + \langle \nabla c^k(x_i), h(x_i) \rangle + \tfrac{1}{2}\lambda M \|h(x_i)\|^2 \}, \max_{j \in \mathbf{q}} \{ f^j(x_i) - \psi(x_i)_+ + \langle \nabla f^j(x_i), h(x_i) \rangle + \tfrac{1}{2}\lambda M \|h(x_i)\|^2 \} \Big\}$$

$$\leq \lambda\theta(x_i) \leq \lambda\alpha\theta(x_i). \tag{11a}$$

Therefore (2c) is satisfied with $\lambda_i \geq \beta\delta/M$, and thus

$$F(x_i, x_{i+1}) \leq \alpha\lambda_i\theta(x_i) \leq \alpha\beta\delta\theta(x_i)/M. \tag{11b}$$

Next we will establish relationships between $\theta(x_i)$ and $f^0(\hat{x}) - f^0(x_i)$ and between $\theta(x_i)$ and $\psi(x_i)_+$. Let $\mu_i \in \hat{\Sigma}_q^0(x_i)$ and $\nu_i \in \hat{\Sigma}_p(x_i)$ be arbitrary, and let

$$\begin{bmatrix} \xi_{0,i}^0 \\ \xi_{0,i} \end{bmatrix} = \sum_{k=1}^{p} \nu_i^k \begin{bmatrix} c^k(x_i) - f^0(x_i) - \gamma\psi(x_i)_+ \\ \nabla c^k(x_i) \end{bmatrix}, \tag{11c}$$

and

$$\begin{bmatrix} \xi_{j,i}^0 \\ \xi_{j,i} \end{bmatrix} = \begin{bmatrix} f^j(x_i) - \psi(x_i)_+ \\ \nabla f^j(x_i) \end{bmatrix}, \ j \in \mathbf{q}. \tag{11d}$$

Then, making use of the definitions (1e) and (7a), the Discrete Minimax Theorem stated in Corollary 5.5.3, (5b), (6a,b), the fact that $m/\delta \leq 1$, the fact that $\xi_{j,i}^0 \leq 0$, for $j = 0, \ldots, q$, and the fact that for any $x \in \mathbb{R}^n$, $f^0(x) \geq \sum_{k=1}^{p} \nu_i^k c^k(x)$, we find, with $m' \triangleq m/\delta$, that

$$\theta(x_i) = \min_{h \in \mathbb{R}^n} \sum_{j=0}^{q} \mu_i^j [\xi_{j,i}^0 + \langle \xi_{j,i}, h \rangle + \tfrac{1}{2}\delta\|h\|^2]$$

$$\leq m' \min_{h \in \mathbb{R}^n} \sum_{j=0}^{q} \mu_i^j [\xi_{j,i}^0 + \langle \xi_{j,i}, h/m' \rangle + \tfrac{1}{2}m\|h/m'\|^2]$$

$$\leq m' \min_{h \in \mathbb{R}^n} \Big\{ \mu_i^0 [f^0(x_i + h/m') - f^0(x_i) - \gamma\psi(x_i)_+] + \sum_{j=1}^{q} \mu_i^j [f^j(x_i + h/m') - \psi(x_i)_+] \Big\}. \tag{11e}$$

Replacing h by $m'(\hat{x} - x_i)$ in (11e) and using the fact that $\sum_{j \in \mathbf{q}} \mu_i^j f^j(\hat{x}) \leq \psi(\hat{x}) \leq 0$ and the fact that $\sum_{j \in \mathbf{q}} \mu_i^j = (1 - \mu_i^0)$, we deduce from (11e) that

$$\theta(x_i) \leq m' \{ \mu_i^0 [f^0(\hat{x}) - f^0(x_i) - \gamma\psi(x_i)_+] + \sum_{j=1}^{q} \mu_i^j [f^j(\hat{x}) - \psi(x_i)_+] \}$$

$$\leq \frac{m}{\delta} \{ \mu_i^0 [f^0(\hat{x}) - f^0(x_i)] - [1 + (\gamma - 1)\mu_i^0] \psi(x_i)_+ \}. \tag{11f}$$

Combining (11b) with (11f), we obtain (10a).

Finally, (10b) and (10c) follow from (10a) and the following inequalities

$$\left. \begin{aligned} f^0(x_{i+1}) - f^0(x_i) - \gamma\psi_+(x_i) &\leq F(x_i, x_{i+1}), \\ \psi(x_{i+1}) - \psi(x_i)_+ &\leq F(x_i, x_{i+1}). \end{aligned} \right\} \tag{11g}$$

□

Lemma 2.6.13. *Suppose that Assumption 2.6.8 is satisfied and that $\hat{x}$ is the unique solution of (1a,b). Then, for any $\varepsilon \in (0,1)$, there exists a $\rho > 0$ such that, for any sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 2.6.1, if $x_i \in B(\hat{x}, \rho)$, then*

$$\psi_+(x_{i+1}) \leq \delta_1(\varepsilon)\psi(x_i)_+, \tag{12a}$$

and

$$[f^0(x_{i+1}) - f^0(\hat{x})]_+ \leq \delta_2(\varepsilon)\big[[f^0(x_i) - f^0(\hat{x})]_+ + \gamma\psi(x_i)_+\big], \tag{12b}$$

where we have used our notation $a_+ \triangleq \max\{0, a\}$, for any $a \in \mathbb{R}$, and

$$\delta_1(\varepsilon) = \max\left\{ 0, 1 - \frac{\varepsilon\gamma\underline{\mu}^0\alpha\beta m}{M} \right\} \in [0,1) \tag{12c}$$

and

$$\delta_2(\varepsilon) = \left[1 - \frac{\varepsilon\underline{\mu}^0\alpha\beta m}{M} \right] \in (0,1). \tag{12d}$$

Proof. We begin with (12a). For any fixed $\varepsilon \in (0,1)$, we can choose $\varepsilon_1 > 0$, small enough, so that

$$1-\varepsilon_1-\frac{\varepsilon_1}{\gamma\underline{\mu}^0}\geq\varepsilon\,. \tag{13a}$$

By part *(c)* of Lemma 2.6.11, there exists a $\rho>0$ such that for any $x_i\in B(\hat{x},\rho)$ and any $\mu_i\in\hat{\Sigma}_q^0(x_i)$,

$$\underline{\mu}^0(\hat{x})(1-\varepsilon_1)\leq\mu_i^0\leq\bar{\mu}^0(\hat{x})(1+\varepsilon_1)\,. \tag{13b}$$

Using (10c), (8b), and (13a,b), we find that, for any $x_i\in B(\hat{x},\rho)$,

$$\begin{aligned}\psi(x_{i+1})&\leq\frac{\alpha\beta m\,\mu_i^0[1-\bar{\mu}^0(\hat{x})]}{M\,\bar{\mu}^0(\hat{x})}\psi(x_i)_+ +\Big[1-\frac{\alpha\beta m\,[1+(\gamma-1)\mu_i^0]}{M}\Big]\psi(x_i)_+\\ &=[1+\frac{\alpha\beta m}{M}(-1+\frac{\mu_i^0}{\bar{\mu}^0(\hat{x})}-\gamma\mu_i^0)]\psi(x_i)_+\\ &\leq\Big[1+\frac{\alpha\beta m}{M}[-1+\bar{\mu}^0(\hat{x})\frac{(1+\varepsilon_1)}{\bar{\mu}^0(\hat{x})}-\gamma\underline{\mu}^0(\hat{x})(1-\varepsilon_1)]\Big]\psi(x_i)_+\\ &=\Big[1-\frac{\gamma\underline{\mu}^0(\hat{x})\alpha\beta m}{M}[1-\varepsilon_1-\frac{\varepsilon_1}{\gamma\underline{\mu}^0(\hat{x})}]\Big]\psi(x_i)_+\\ &\leq[1-\frac{\varepsilon\gamma\underline{\mu}^0(\hat{x})\alpha\beta m}{M}]\psi(x_i)_+\,.\end{aligned} \tag{13c}$$

Therefore (12a) must hold.

Next, using (10b) and the fact that $1+(\gamma-1)\mu_i^0\geq\gamma\mu_i^0$, we find that for $x_i\in B(\hat{x},\rho)$,

$$\begin{aligned}f^0(x_{i+1})-f^0(\hat{x})&\leq(1-\frac{\alpha\beta m\,\mu_i^0}{M})[f^0(x_i)-f^0(\hat{x})]\\ &\quad+[\gamma-\frac{\alpha\beta m\,(1+(\gamma-1)\mu_i^0)}{M}]\psi(x_i)_+\\ &\leq(1-\frac{\alpha\beta m\,\mu_i^0}{M})[f^0(x_i)-f^0(\hat{x})]_+ +\gamma(1-\frac{\alpha\beta m\,\mu_i^0}{M})\psi(x_i)_+\\ &\leq(1-\frac{\alpha\beta m\,\mu_i^0}{M})\big[[f^0(x_i)-f^0(\hat{x})]_+ +\gamma\psi(x_i)_+\big]\,.\end{aligned} \tag{13d}$$

Since $1-\varepsilon_1\geq\varepsilon$ and $\mu_i^0\geq\underline{\mu}^0(\hat{x})(1-\varepsilon_1)$, $\mu_i^0\geq\underline{\mu}^0\varepsilon$. Thus, (12b) must hold. □

Now we are ready to establish the linear convergence of the Polak-He Algorithm 2.6.1.

Theorem 2.6.14. *Suppose that Assumption 2.6.8 is satisfied, that $\hat{x}$ is the unique solution of (1a,b), and that $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 2.6.1 in solving problem (1a,b). Then,*

(a) the sequence $\{x_i\}_{i=0}^{\infty}$ converges to $\hat{x}$,

(b) with $\delta_1(\varepsilon)$, $\delta_2(\varepsilon)$ defined as in (12c), (12d), respectively, for any $\varepsilon\in(0,1)$, there exists a $\rho>0$ such that, for all $x_i\in B(\hat{x},\rho)$,

(i) if $\psi(x_i)>0$, then

$$\psi(x_{i+1})\leq\delta_1(\varepsilon)\,\psi(x_i)\,, \tag{14a}$$

(ii) if $\psi(x_i)\leq 0$, then

$$f^0(x_{i+1})-f^0(\hat{x})\leq\delta_2(\varepsilon)[f^0(x_i)-f^0(\hat{x})] \tag{14b}$$

and

(iii) if $\gamma>M/(\underline{\mu}^0\alpha\beta m)$, then $\psi(x_{i+1})\leq 0$.

Proof. *(a)* Since the functions $f^j(\cdot)$, $j=1,\ldots,m$, satisfy parts *(i)* and *(iii)* of Assumption 2.6.8, $\psi(\cdot)$ has bounded level sets. Since by (3b), $\psi(x_i)_+\leq\psi(x_0)_+$ for all $i\in\mathbb{N}$, it follows that the sequence $\{x_i\}_{i=0}^{\infty}$ is bounded. Because $\hat{x}$ is the unique zero of $\theta(\cdot)$, it follows from Theorem 2.6.2 that $\{x_i\}_{i=0}^{\infty}$ has only one accumulation point $\hat{x}$. Therefore $\{x_i\}_{i=0}^{\infty}$ converges to $\hat{x}$.

(b) Since $x_i\to\hat{x}$, as $i\to\infty$, it follows that there exists an $i_0\in\mathbb{N}$ such that, for all $i\geq i_0$, $x_i\in B(\hat{x},\rho)$ with $\rho>0$ as specified in Lemmas 2.6.12 and 2.6.13. Hence *(i)* follows directly from Lemma 2.6.13 and the fact that $\psi(x)_+=\psi(x)$ when $\psi(x)>0$, and *(ii)* follows directly from Lemma 2.6.13 and the fact that $\psi(x)_+=0$ and $f^0(x)-f^0(\hat{x})>0$ when $\psi(x)\leq 0$.

(iii) Since $1-\gamma\underline{\mu}^0(\hat{x})\alpha\beta m/M<0$ for all $\gamma>M/(\underline{\mu}^0(\hat{x})\alpha\beta m)$, we can pick an $\varepsilon\in(0,1)$ such that $\delta_1(\varepsilon)=0$. Then the desired result follows from Lemma 2.6.13 for this particular ε. □

The following corollary is a consequence of the above theorem, the fact that $\delta_1(1)=0$ for all $\gamma>0$ such that $\gamma\geq M/\underline{\mu}^0(\hat{x})\alpha\beta m$, and the fact that $\delta_2(1)=1-\underline{\mu}^0(\hat{x})\alpha\beta m/M$.

Corollary 2.6.15. *Suppose that Assumption 2.6.8 is satisfied, that $\hat{x}$ is the unique solution of (1a,b), and that $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by*

Algorithm 2.6.1. Then,

(a) if $\gamma > M/\underline{\mu}^0(\hat{x})\alpha\beta m$, *then there exist an* i_0 *such that* $\psi(x_i) \le 0$ *for all* $i \ge i_0$,

(b) if $\psi(x_i) > 0$ *for all* $i \ge 0$, *then*

$$\overline{\lim_{i\to\infty}} \frac{\psi(x_{i+1})}{\psi(x_i)} \le 1 - \frac{\gamma\underline{\mu}^0\alpha\beta m}{M}, \tag{15a}$$

and

(c) if there exists an $i_1 \in \mathbb{N}$ *such that* $\psi(x_{i_1}) \le 0$, *then* $\psi(x_i) \le 0$ *for all* $i \ge i_1$, *and*

$$\overline{\lim_{i\to\infty}} \frac{f^0(x_{i+1}) - f^0(\hat{x})}{f^0(x_i) - f^0(\hat{x})} \le 1 - \frac{\underline{\mu}^0\alpha\beta m}{M}. \tag{15b}$$

□

In practice, it has been observed that starting Algorithm 2.6.1 from a particular point x_0, increasing the value of γ results in a feasible point being found faster. Thus γ can be thought of as a "steering" parameter, i.e., it can be used to adjust the speed with which a sequence constructed by Algorithm 2.6.1 approaches the feasible set. Also, note again, that should the value of δ used in (2a,b) not satisfy the condition $\delta \in [m, M]$, then, in (15a,b), m must be replaced by $m' \triangleq \min\{\delta, m\}$ and M must be replaced by $M' \triangleq \max\{\delta, M\}$.

The following theorem establishes the R-linear convergence of the iterates constructed by Algorithm 2.6.1.

Theorem 2.6.16. *Suppose that Assumption 2.6.8 is satisfied and that* $\hat{x}$ *is the unique solution of (1a,b). Then, for any* $\varepsilon \in (0,1)$ *and any sequence* $\{x_i\}_{i=0}^{\infty}$ *constructed by Algorithm 2.6.1, there exists an* $i_0 \in \mathbb{N}$ *and a* $\rho_1 > 0$, *such that, for all* $i \ge i_0$,

$$\|x_i - \hat{x}\| \le \rho_1[\delta_3(\varepsilon)^{1/2}]^{i-i_0}, \tag{16a}$$

where

$$\delta_3(\varepsilon) \triangleq \max\{\delta_1(\varepsilon), \delta_2(\varepsilon)\} = 1 - \min\{1,\gamma\}\varepsilon\underline{\mu}^0\frac{\alpha\beta m}{M}, \tag{16b}$$

with $\delta_1(\varepsilon)$, $\delta_2(\varepsilon)$ *defined by (12c) and (12d), respectively.*

Proof. For any fixed $\varepsilon \in (0,1)$, let $\varepsilon_1 \in (\varepsilon, 1)$ be arbitrary. Then there exists $\rho_1 > 0$ such that Lemma 2.6.13 holds for $\varepsilon = \varepsilon_1$ and $\rho = \rho_1$. Since $\delta_3(\varepsilon_1) < \delta_3(\varepsilon)$, there exists a $K > 0$ such that, for all $i \ge 0$,

$$i[\delta_3(\varepsilon_1)/\delta_3(\varepsilon)]^i < K. \tag{17a}$$

Since $f^0(\cdot)$ and $\psi(\cdot)$ are continuous, we can find a $\rho_0 \in (0, \rho_1)$ such that

$$\max_{x \in B(\hat{x}, \rho_0)} \{\underline{\mu}^0(\hat{x})[f^0(x) - f^0(\hat{x})]_+ + [\underline{\mu}^0(\hat{x})\gamma K + 1 - \underline{\mu}^0(\hat{x})]\psi(x)_+\} \le \tfrac{1}{2}m\rho_1^2. \tag{17b}$$

We will prove by induction that (16a) holds for any sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 2.6.1. Let x_{i_0} be the first element of this sequence in the ball $B(\hat{x}, \rho_0)$.

Since $\rho_0 \le \rho_1$, (16a) holds for $i = i_0$. Suppose that (16a) holds for $i = i_0, \ldots, k+i_0$. Then $x_i \in B(\hat{x}, \rho_1)$ for $i = i_0, \ldots, k+i_0$. Thus, due to the selection of ρ_1, for $i = i_0, \ldots, k+i_0$, x_i satisfies inequalities (12a,b), where ε is replaced by ε_1, i.e.,

$$\psi_+(x_{i+1}) \le \delta_1(\varepsilon_1)\,\psi_+(x_i), \tag{17c}$$

and

$$[f^0(x_{i+1}) - f^0(\hat{x})]_+ \le \delta_2(\varepsilon_1)\big[[f^0(x_i) - f^0(\hat{x})]_+ + \gamma\psi(x_i)_+\big]. \tag{17d}$$

Therefore, for $i = i_0, \ldots, k+i_0$, we can recursively deduce that

$$\psi_+(x_{i+1}) \le [\delta_3(\varepsilon_1)]^{i+1-i_0}\psi(x_{i_0})_+, \tag{17e}$$

$$[f^0(x_{i+1}) - f^0(\hat{x})]_+ \le \delta_3(\varepsilon_1)^{i+1-i_0}\big([f^0(x_{i_0}) - f^0(\hat{x})]_+ + \gamma(i+1-i_0)\psi(x_{i_0})_+\big), \tag{17f}$$

where $\gamma(i+1-i_0) \ge \gamma$ is used for technical reasons, soon to become clear. Using (8c), (17e,f), and the fact that $\delta_3(\varepsilon) > \delta_3(\varepsilon_1)$, we conclude that

$$\begin{aligned}
\tfrac{1}{2}m\|x_{k+1+i_0} - \hat{x}\|^2 &\le \underline{\mu}^0(\hat{x})[f^0(x_{k+1+i_0}) - f^0(\hat{x})] + [1 - \underline{\mu}^0(\hat{x})]\psi(x_{k+1+i_0})_+ \\
&\le \delta_3(\varepsilon_1)^{k+1}\big[\underline{\mu}^0(\hat{x})[f^0(x_{i_0}) - f^0(\hat{x})]_+ \\
&\quad + (\underline{\mu}^0(\hat{x})\,\gamma(k+1) + 1 - \underline{\mu}^0(\hat{x}))\,\psi(x_{i_0})_+\big] \\
&\le \delta_3(\varepsilon)^{k+1}\big[\underline{\mu}^0(\hat{x})[f^0(x_{i_0}) - f^0(\hat{x})]_+ \\
&\quad + \big(\underline{\mu}^0(\hat{x})\gamma(k+1)[\delta_3(\varepsilon_1)/\delta_3(\varepsilon)]^{k+1} + 1 - \underline{\mu}^0(\hat{x})\big)\psi(x_{i_0})_+\big].
\end{aligned} \tag{17g}$$

Since $x_{i_0} \in B(\hat{x}, \rho_0)$, we conclude from (17a,b) and (17g) that

$$\tfrac{1}{2}m\|x_{k+1+i_0} - \hat{x}\|^2 \le [\delta_3(\varepsilon)]^{k+1}\big[\underline{\mu}^0(\hat{x})[f^0(x_{i_0}) - f^0(\hat{x})]_+$$

$$+[\underline{\mu}^0(\hat{x})\gamma K + 1 - \underline{\mu}^0(\hat{x})]\psi(x_{i_0})_+]$$

$$\leq \tfrac{1}{2} m \delta_3(\varepsilon)^{k+1}\rho_1^2. \tag{17h}$$

Consequently, (16a) holds for $i = k+1+i_0$. Therefore the proof by induction of (16a) is completed. □

Corollary 2.6.17. *Suppose that Assumption 2.6.8 is satisfied and that $\hat{x}$ is the unique solution of (1a,b). Let $\delta_3(\varepsilon)$ be defined as in (16b). If $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 2.6.1, then*

$$\overline{\lim}_{i\to\infty} (\|x_i - \hat{x}\|)^{(1/i)} \leq [\delta_3(1)]^{1/2}. \tag{18}$$

□

2.6.3 A Barrier Function Method

Just like Algorithm 2.4.1, Algorithm 2.6.1 was constructed by replacing the functions $c^k(\cdot)$ and $f^j(\cdot)$ in (1a,b) by quadratic approximations. As was also the case with min-max problems, an alternative approach, also envisaged in phase II form by Huard, in [BuH.66, Hua.67, Hua.68], is to use a barrier function instead. This approach becomes attractive when q, p in (1a,b) are very large, which makes computing search directions for Algorithms 2.6.1 problematic, and only an approximate solution to (1a,b) needs to be found. This is the case when we solve approximations to semi-infinite optimization problems and optimal control problems with state constraints, as we will see in Chapters 3 and 4.

For the sake of simplicity, consider problem (1a,b) under the assumption that, for $j = 0, 1, \ldots, q$, *all* the functions $f^j : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable. Let $F(\cdot,\cdot)$ be defined as in (1e) for some $\gamma > 0$. Then, for every $z \in \mathbb{R}^n$, we define the set

$$C(z) \triangleq \{x \in \mathbb{R}^n \mid F(z,x) < 0\}, \tag{19a}$$

and we associate, with the function $F(\cdot,\cdot)$, the barrier function $b : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, defined by

$$b(z,x) \triangleq \frac{1}{f^0(z) + \gamma\psi(z)_+ - f^0(x)} + \frac{1}{q}\sum_{j=1}^{q} \frac{1}{\psi(z)_+ - f^j(x)}. \tag{19b}$$

Then, for every $z \in \mathbb{R}^n$, $b(z,\cdot)$ is a *barrier function* for $C(z)$, i.e., for all $x \in C(z)$, $b(z,x) > 0$ and $b(z,x) \to \infty$, as $\psi(x) \to \psi(z)_+$. The barrier function $b(z,\cdot)$ is continuously differentiable on $C(z)$ and, for $x \in C(z)$, its gradient is given by

$$\nabla_x b(z,x) = \frac{1}{[f^0(z) + \gamma\psi(z)_+ - f^0(x)]^2}\nabla f^0(x)$$

$$+ \frac{1}{q}\sum_{j=1}^{q} \frac{1}{[\psi(z)_+ - f^j(x)]^2}\nabla f^j(x). \tag{19c}$$

Hence $b(\cdot,z)$ can be minimized using various unconstrained optimization algorithms encountered in Chapter 1. These observations suggest that, given an $x_i \in \mathbb{R}^n$, a practical alternative to the construction of a successor point x_{i+1} according to the rules in Algorithm 2.6.1 might be the construction of any $x_{i+1} \in C(x_i)$ which is an "approximate" minimizer of $b(x_i,\cdot)$. These considerations lead to the following algorithm.

Basic Barrier Function Algorithm 2.6.18.

Parameters. $\zeta \in [0,2)$, $\kappa \in (0,\infty)$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute an $x_{i+1} \in C(x_i)$ such that

$$\|\nabla_x b(x_i, x_{i+1})\| \leq \frac{\kappa}{|F(x_i, x_{i+1})|^{\zeta}}. \tag{20}$$

Step 2. Replace i by $i+1$, and go to Step 1.

As we will soon see, the values $F(x_i, x_{i+1})$ converge to zero, which tends to make both $b(x_i,x)$ and its gradient large, for $x \in C(x_i)$. The term $|F(x_i,x_{i+1})|^{\zeta}$ in (20) is used to slow the increase in precision with which the problems $\min_{x \in C(x_i)} b(x_i,x)$ must be solved at each iteration.

Theorem 2.6.19. *Consider problem (1a). Suppose that, for $j = 0, 1, \ldots, q$, the functions $f^j : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable, that Assumption 2.6.3 is satisfied, and that Algorithm 2.6.18 has constructed a bounded sequence $\{x_i\}_{i=0}^{\infty}$ in solving (1a). Then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies the first-order optimality condition $\psi(\hat{x}) \leq 0$ and $\theta(\hat{x}) = 0$, with $\theta(\cdot)$ defined in (2.2.8f), with $\gamma > 0$, as in the definition of $F(\cdot,\cdot)$ (see (1e)) and $\delta > 0$ arbitrary.*

Proof. Suppose that Algorithm 2.6.18 has constructed a bounded sequence $\{x_i\}_{i=0}^{\infty}$ in solving (1a,b). Since this sequence is bounded, it must have accumulation points. Thus, suppose that $\hat{x}$ is an accumulation point and that $K \subset \mathbb{N}$ is such that $x_i \to^K \hat{x}$ as $i \to \infty$. Now, for every $i \in \mathbb{N}$, since

$x_{i+1} \in C(x_i)$, we must have that $\psi(x_{i+1}) < \psi(x_i)$. Hence the sequence $\{\psi(x_i)\}_{i=0}^{\infty}$ is monotone decreasing, and, since, by continuity, $\psi(\hat{x})$ is an accumulation point of this sequence, by Proposition 5.1.16, we must have that $\psi(x_i) \to \psi(\hat{x})$, as $i \to \infty$. First, it follows from (20) that

$$F(x_{i-1}, x_i)^2 \|\nabla_x b(x_{i-1}, x_i)\| \leq \kappa |F(x_{i-1}, x_i)|^{(2-\zeta)}. \tag{21a}$$

For $i \geq 1$, let

$$\nu_i^0 \triangleq \frac{F(x_{i-1}, x_i)^2}{[f^0(x_{i-1}) + \gamma\psi(x_{i-1})_+ - f^0(x_i)]^2}, \tag{21b}$$

and, for $i \geq 1$ and $j \in \mathbf{q}$, let

$$\nu_i^j \triangleq \frac{1}{q} \frac{F(x_{i-1}, x_i)^2}{[\psi(x_{i-1})_+ - f^j(x_i)]^2}. \tag{21c}$$

Then we find that for all $i \in \mathbb{N}$, $\nu_i^0 \in [0, 1]$ and $\nu_i^j \in [0, 1/q]$ for $j \in \mathbf{q}$. Now let $\omega_i \triangleq \sum_{j=0}^{q} \nu_i^j$, and for $j = 0, 1, \ldots, q$, let $\mu_i^j = \nu_i^j/\omega_i$. Then $\mu_i^j \geq 0$ for all $j = 0, 1, \ldots, q$ and $\sum_{j=0}^{q} \mu_i^j = 1$, i.e., $\mu_i \triangleq (\mu_i^0, \ldots, \mu_i^q) \in \Sigma_q^0$. Since each $\nu_i^j \in [0, 1/q]$, $j \in \mathbf{q}$, and either at least one $\nu_i^j = 1/q$, $j \in \mathbf{q}$, or $\nu_i^0 = 1$, or both, must hold, we conclude that $1/q \leq \omega_i \leq 2$. Hence, using (19c) and dividing (21a) by ω_i, we find that

$$\|\sum_{j=0}^{q} \mu_i^j \nabla f^j(x_i)\| \leq q\kappa |F(x_{i-1}, x_i)|^{(2-\zeta)}. \tag{21d}$$

Now we must consider two cases.

Case 1. There exists an i_0 such that $\psi(x_{i_0}) \leq 0$, so that $\psi(x_{i_0})_+ = 0$. Clearly, since $x_{i_0+1} \in C(x_{i_0})$ and

$$\psi(x_{i_0+1}) - \psi(x_{i_0})_+ \leq F(x_{i_0}, x_{i_0+1}) < 0, \tag{21e}$$

it follows that $\psi(x_{i_0+1} < 0$ and hence, by induction, that $\psi(x_i) < 0$ for all $i \geq i_0$. Hence, for all $i \geq i_0$,

$$f^0(x_{i+1}) - f^0(x_i) \leq F(x_i, x_{i+1}) < 0, \tag{21f}$$

i.e., the sequence $\{f^0(x_i)\}_{i=i_0}^{\infty}$, is monotone decreasing, and since $f^0(x_i) \to^K f^0(\hat{x})$, as $i \to \infty$, by continuity of $f^0(\cdot)$, it follows from Proposition 5.1.16 that $f^0(x_i) \to f^0(\hat{x})$, as $i \to \infty$. It now follows from the definition of $F(x_{i-1}, x_i)$ that $F(x_{i-1}, x_i) \to 0$, as $i \to \infty$. Since, for $i \geq i_0$ and $j \in \mathbf{q}$,

$$\nu_i^j = \frac{1}{q} \frac{F(x_{i-1}, x_i)^2}{[f^j(x_i)]^2}, \tag{21g}$$

it follows that $\nu_i^j \to^K 0$, as $i \to \infty$, for all $j \notin \mathbf{q}_A(\hat{x})$ and hence that

$$\lim_{i \in K} \sum_{j=1}^{q} \mu_i^j f^j(x_i) = 0. \tag{21h}$$

Next, it follows from (21d) that

$$\lim_{i \in K} \|\sum_{j=0}^{q} \mu_i^j \nabla f^j(x_i)\| = 0. \tag{21i}$$

Since, by (2.2.9e),

$$\theta(x_i) \geq -\left[-\sum_{j=1}^{q} \mu_i^j f^j(x_i) + \tfrac{1}{2}\delta \|\sum_{j=0}^{q} \mu_i^j \nabla f^j(x_i)\|^2\right], \tag{21j}$$

it follows that $\lim_{i \in K} \theta(x_i) = 0$. Since $\theta(\cdot)$ is continuous, we conclude that $\theta(\hat{x}) = 0$, which completes our proof for Case 1.

Case 2. Suppose that $\psi(x_i) > 0$ for all $i \in \mathbb{N}$. Then we must have that

$$\psi(x_{i+1}) - \psi(x_i) \leq F(x_i, x_{i+1}) < 0 \tag{21k}$$

for all $i \in \mathbb{N}$. Therefore, because, by continuity, $\psi(x_i) \to^K \psi(\hat{x})$, as $i \to \infty$, it follows from Proposition 5.1.6, that $\psi(x_i) \to \psi(\hat{x})$, as $i \to \infty$, which implies that $F(x_{i-1}, x_i) \to 0$, as $i \to \infty$. Hence it follows from (21d) that (21i) must hold again. First suppose that $\psi(\hat{x}) = 0$. Then, reasoning as above, we conclude that (21f) must hold, and hence, using (21j), that $\theta(\hat{x}) = 0$. Consequently, it only remains to be shown that $\psi(\hat{x}) = 0$. For the sake of contradiction, suppose that $\psi(\hat{x}) > 0$. Now we have two possibilities. The first is that $f^0(x_{i-1}) + \gamma\psi(x_{i-1})_+ - f^0(x_i) \to 0$, as $i \to \infty$. In this case, there must exist an i_1 such that, for all $i \geq i_1$,

$$f^0(x_i) - f^0(x_{i-1}) \geq \tfrac{1}{2}\gamma\psi(\hat{x})_+, \tag{21l}$$

which implies that $f^0(x_i) \to +\infty$, as $i \to \infty$. Since, by assumption, $\{x_i\}_{i=0}^{\infty}$ is bounded and $f^0(\cdot)$ is continuous, this is impossible. Hence we must assume that the alternative is true, i.e., that there exists an infinite subset $K' \subset \mathbb{N}$, and an $\varepsilon > 0$ such that, for all $i \in K'$,

$$f^0(x_i) - f^0(x_{i-1}) - \gamma\psi(x_{i-1})_+ \geq \varepsilon. \tag{21m}$$

Since $\nu_i^0 \leq F(x_{i-1}, x_i)^2/\varepsilon^2$, for all $i \in K'$, it follows that $\mu_i^0 \to^{K'} 0$, as $i \to \infty$. Since $\{x_i\}_{i=0}^{\infty}$ is bounded, there must exist an infinite subset $K'' \subset K'$ and an $x^* \in \mathbb{R}^n$ such that $x_i \to^{K''} x^*$, as $i \to \infty$. Clearly, we must have that $\psi(x^*) = \psi(\hat{x}) > 0$ and, since $\mu_i^0 \to^{K''} 0$, as $i \to \infty$, that

$$\lim_{i \in K''} \|\sum_{j=1}^{q} \mu_i^j \nabla f^j(x_i)\| = 0. \tag{21n}$$

Since (21h) together with (21n) imply that $0 \in \partial\psi(x^*)$, we find that we have a violation of Assumption 2.6.3, and hence our proof is complete. □

The only problem with Algorithm 2.6.18 is that because $b(x_i, x_i) = \infty$, x_i cannot be used to as "hot" starting point in minimizing $b(x_i, \cdot)$. This problem can be eliminated by slightly "relaxing" both the sets $C(z)$ and the barrier function, as follows. For any $\alpha \geq 0$ and $z \in \mathbb{R}^n$, we define

$$C_\alpha(z) \triangleq \{x \in \mathbb{R}^n \mid F(z, x) < \alpha\}, \tag{22a}$$

and for any $\alpha \geq 0$, we define the relaxed barrier function $b_\alpha : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ by

$$b_\alpha(z, x) \triangleq \frac{1}{\alpha + f^0(z) + \gamma\psi(z)_+ - f^0(x)} + \frac{1}{q}\sum_{j=1}^{q} \frac{1}{\alpha + \psi(z)_+ - f^j(x)}. \tag{22b}$$

This leads to the following simple modification of Algorithm 2.6.18.

Polak-Yang-Mayne Barrier Function Algorithm 2.6.20.

Parameters. $\zeta \in [0, 2)$, $\kappa \in (0, \infty)$, $\{\alpha_i\}_{i=0}^{\infty}$, with $\alpha_i > 0$, $\sum_{i=0}^{\infty} \alpha_i < \infty$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Using x_i as a "hot" starting point for an unconstrained minimization algorithm on the problem $\min_{x \in C_{\alpha_i}(x_i)} b_{\alpha_i}(x_i, x)$, compute an $x_{i+1} \in C_{\alpha_i}(x_i)$ such that

$$\|\nabla_x b_{\alpha_i}(x_i, x_{i+1})\| \leq \frac{\kappa}{[\alpha_i - F(x_i, x_{i+1})]^{\zeta}}. \tag{23}$$

Step 2. Replace i by $i + 1$, and go to Step 1.

Since the barrier function $b_{\alpha_i}(x_i, \cdot)$ can be ill-conditioned, it is advisable to use either the BFGS variable metric method or conjugate gradient methods with restart for computing the x_{i+1} in Algorithm 2.6.20. Not surprisingly, Algorithm 2.6.20 retains the convergence properties of Algorithm 2.6.18:

Theorem 2.6.21. *Consider problem (1a), and suppose that, for $j = 0, 1, \ldots, q$, the functions $f^j : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable, that Assumption 2.6.3 is satisfied, and that Algorithm 2.6.20 has constructed a bounded sequence $\{x_i\}_{i=0}^{\infty}$ in solving (1a). Then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies the first-order optimality condition $\psi(\hat{x}) \leq 0$ and $\theta(\hat{x}) = 0$, with $\theta(\cdot)$ defined in (2.2.9e).* □

Exercise 2.6.22. Prove Theorem 2.6.21. *Hint*: Make use of Lemma 3.1 in [PHM.92]. □

2.6.4 Notes

All methods of centers result from a fairly lengthy evolutionary process which began with the phase II conceptual method of Huard [BuH.66, Hua.67, Hua.68, Tre.68].

Methods based on the notion of a center as a minimizer of a max function were evolved into an implementable phase II method by Pironneau and Polak [PiP.72, PiP.73], then into an implementable "two-cost" phase I - phase II method by Polak, Trahan and Mayne [PTM.79] (Algorithm 2.6.4), and finally into an implementable "single-cost" phase I - phase II method by Polak and He [PoH.91a] (Algorithm 2.6.1). When initialized with a feasible point x_0, Algorithms 2.6.1 and 2.6.4 construct identical sequences; in fact, they revert to the Pironneau-Polak Algorithm in [PiP.72, PiP.73]. In general, there appears to be no benefit in using better than first-order approximations to the function $F(z, x)$ in (1e). Thus, if we let all the functions in (1a) be affine and define a center as the *actual* minimizer of $F(z, \cdot)$, we still get an algorithm that converges only linearly, and the bound is exact.

Methods based on the notion of a center as the minimizer of a barrier function were evolved into an implementable phase II method by Mifflin [Mif.76], and then into a "single-cost" phase I - phase II method by Polak, Yang and Mayne [PYM.93] (Algorithm 2.6.20). Algorithm 2.6.20 makes use of the parameter-free Fiacco-McCormick penalty function [FiM.67], the Mifflin truncation rule [Mif.76], and the Tremolières penalty adjustment rule [Tre.68].

There are some common features in Algorithm 2.6.20 and other barrier function methods for nonlinear programming (see, e.g., [Jar.88a, Jar.88b, SoS.88, Son.87, YeY.87]).

When code for Algorithm 2.4.11 is not available and one has to evaluate $\theta(\cdot)$ and $h(\cdot)$, for Algorithm 2.6.1, using quadratic programming code, one can use the following ε–*active* version of Algorithm 2.6.1, derived from the phase II methods in [Pol.71] and the phase I - phase II methods of feasible directions in [PTM.77]. This version requires fewer gradient evaluations and hence also constructs a smaller quadratic programming problem for search direction computations. The price one pays for this economy is that the search direction is no longer continuous in the iterates, and hence a larger amount of "zigzagging" may be observed than with Algorithm 2.6.1.

ε-Active Polak-He Algorithm 2.6.23.

Parameters. $\gamma, \delta > 0$, $\alpha \in (0, 1]$, $\beta \in (0, 1)$, $\varepsilon_0 > 0$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. set $\varepsilon = \varepsilon_0$.

Step 2. Compute the set

$$\mathbf{q}_{A,\varepsilon}(x_i) \triangleq \{j \in \mathbf{q} \mid f^j(x_i) \geq \psi(x_i)_+ - \varepsilon\}. \tag{24a}$$

Step 3. Compute the the ε-*optimality function* value $\theta_\varepsilon(x_i)$ and the corresponding *search direction* $h_\varepsilon(x_i)$, according to the dual form formulas

$$\theta_\varepsilon(x_i) = -\min_{\substack{\mu \in \Sigma_q^0 \\ \nu \in \Sigma_p}} \Big\{ \mu^0 \sum_{k=1}^{p} \nu^k [f^0(x_i) - c^k(x_i) + \gamma\psi(x_i)_+] + \sum_{j=1}^{q} \mu^j [\psi(x_i)_+ - f^j(x_i)]$$

and

$$+ \frac{1}{2\delta} \| \mu^0 \sum_{k=1}^{p} \nu^k \nabla c^k(x_i) + \sum_{j=1}^{q} \mu^j \nabla f^j(x_i) \|^2 \mid \mu^j = 0,\ \forall\, j \in \mathbf{q}_{A,\varepsilon}(x_i) \Big\}, \tag{24b}$$

$$h_\varepsilon(x_i) = -\frac{1}{\delta} \Big\{ \mu_x^0 \sum_{k=1}^{p} \nu_x^k \nabla c^k(x_i) + \sum_{j=1}^{q} \mu_x^j \nabla f^j(x_i) \Big\}, \tag{24c}$$

where (μ_x, ν_x) is any solution of (24b).

Step 4. If $\theta_\varepsilon(x_i) > -\varepsilon$, replace ε by $\beta\varepsilon$, and go to Step 1.
Else, set $\theta_i = \theta_\varepsilon(x_i)$, $h_i = h_\varepsilon(x_i)$, and go to Step 5.

Step 5. Compute the *step-size* λ_i:

$$\lambda_i = \max_{k \in \mathbf{N}} \{ \beta^k \mid F(x_i, x_i + \beta^k h_i) \le \beta^k \alpha \theta_i \}. \tag{24d}$$

Step 6. Set $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i+1$, and go to Step 1.

Note that, because μ^j is required to be zero, for all $j \in \mathbf{q}_{A,\varepsilon}(x_i)$, in (24b), there is no need to compute the gradients $\nabla f^j(x_i)$, for all $j \in \mathbf{q}_{A,\varepsilon}(x_i)$, in order to evaluate $\theta_\varepsilon(x_i)$ and $h_\varepsilon(x_i)$.

It can be inferred from the results in [PTM.77] that the following theorem is true.

Theorem 2.6.24. *Consider the problem (1a,b), and suppose that the functions $f^j, c^k : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, $k \in \mathbf{p}$, are at least once continuously differentiable. If $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 2.6.23 in solving (1a,b), then any accumulation point $\hat{x}$ of the sequence $\{x_i\}_{i=0}^{\infty}$ satisfies $\theta(\hat{x}) = 0$.* □

2.7 Penalty Function Algorithms

Penalty functions are used to expand inequality constrained optimization problems, such as **ICP** (2.1.0b), into a sequence of unconstrained, approximating optimization problems of the form of **MMP** (2.1.0a), and inequality-equality constrained such as **IECP** (2.1.0c), either into a sequence of unconstrained, approximating optimization problems of the form of **MMP**, or into a sequence of inequality constrained, approximating optimization problems of the form of **ICP**. There are three reasons for using penalty functions. The first is that they enable us to solve problems such as **IECP**, when the only computer code available is for solving min-max or differentiable unconstrained optimization problems. The second, as we will see in Section 2.8, is that penalty functions can be used for constructing special algorithms for solving problems such as **IECP**, and the third, as we will see in Section 2.9, is that they can be used as a tool for globally stabilizing locally converging Newton-like methods for solving **IECP**.

2.7.1 Basic Theory of Penalty Functions

At first, it is simpler to consider a constrained optimization problem in the somewhat abstract form

$$\mathbf{P} \qquad \min \{ f^0(x) \mid x \in C \}, \tag{1a}$$

where $f^0 : \mathbb{R}^n \to \mathbb{R}$ is continuous and C is a subset of $\mathbb{R}^n$. To make the formulation (1a) relevant to the solution, by penalty function methods, of problems such as **IECP** it is useful to assume that C is of the form

$$C = C' \cap C'' \cap C''', \tag{1b}$$

which allows us to associate the set C' with the set defined by the equality constraints and some of the inequality constraints, to be removed using *exterior* penalty functions, the set C'' with a set defined by other inequality constraints, to be removed using *interior* penalty functions, and the set C''' with $\mathbb{R}^n$ or with a set defined by the remaining inequality constraints, to be left as is. Hence, since there may not be any equality or any inequality constraints, we must assume that one of the two sets C', C'' can be $\mathbb{R}^n$.

Assumption 2.7.1. *We will assume that*

(i) *the sets C', C'', and C''' are all closed,*

(ii) *there exists a point $x^* \in C$ such that the level set $\{x \in \mathbb{R}^n \mid f^0(x) \le f^0(x^*)\}$ is compact, and*

(iii) *there exists an optimal solution $\hat{x} \in C$ for problem (1a) such that, for any $\rho > 0$, the set $B(\hat{x}, \rho) \cap C' \cap \operatorname{int} C'' \cap C'''$ is not empty.* □

It should be obvious that Assumptions 2.7.1*(i, ii)* ensure that problem (1a) has a solution. Assumption 2.7.1*(iii)* is an abstract form of Assumption 2.6.3, and will be needed later to rule out pathological cases, such as

$$C'' \triangleq \{ x \in \mathbb{R}^2 \mid (\|x - c\|^2 - \|r\|^2)(r, x - (c + r))^2 \le 0 \}, \tag{1c}$$

which consists of a ball, centered at c, of radius $\|r\|$, and of a tangent line passing through $c + r$, on which the solution to **P** lies, but not at $c + r$. Finally, Assumption 2.7.1*(iii)* ensures that $C' \cap \operatorname{int} C'' \cap C'''$ is not empty and that the closure of this set contains an optimal solution for (1a).

To obtain an approximate solution to problem (1a) using penalty functions, one transforms the constraints in (1a) into *penalty functions* that are added to the

cost function. The role of the penalty functions is to impose a penalty for failure to satisfy the constraints. There are two kinds of penalty functions: *exterior* and *interior*. Exterior penalty functions can be used for arbitrary closed sets, and hence for sets defined by either equality or inequality constraints, while interior penalty functions can be used only for closed sets with interiors and hence, in practice, only for sets defined by inequality constraints.

Definition 2.7.2. *Let X be a closed subset of $\mathbb{R}^n$. A sequence of continuous functions $p_k^e : \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbb{N}$, is a sequence of* exterior penalty functions *for the set X if*

$$p_k^e(x) = 0, \quad \forall x \in X, \; k \in \mathbb{N}, \tag{2a}$$

$$0 < p_k^e(x) < p_{k+1}^e(x), \quad \forall x \notin X, \; k \in \mathbb{N}, \tag{2b}$$

and

$$p_k^e(x) \to \infty, \quad \text{as } k \to \infty, \quad \forall x \notin X. \tag{2c}$$

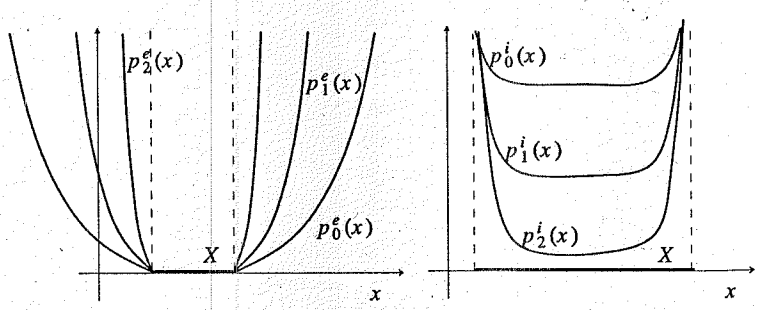
□

A typical set of exterior penalty functions is shown in Fig. 2.7.1a.

Exercise 2.7.3. *(a)* Suppose that X', X'' are closed subsets of $\mathbb{R}^n$ and that $\{p'^e_k(\cdot)\}_{k=0}^\infty$ and $\{p''^e_k(\cdot)\}_{k=0}^\infty$ are sequences of exterior penalty functions for the sets X', X'', respectively. Show that if, for $k \in \mathbb{N}$, $p_k^e : \mathbb{R}^n \to \mathbb{R}$ is defined by

$$p_k^e(x) \triangleq \max\{p'^e_k(x), p''^e_k(x)\}, \tag{3a}$$

or by



(a) Exterior penalty functions (b) Interior penalty functions

Fig. 2.7.1. Typical graphs of penalty functions.

$$p_k^e(x) \triangleq p'^e_k(x) + p''^e_k(x), \tag{3b}$$

then $\{p_k^e(\cdot)\}_{k=0}^\infty$ is a sequence of exterior penalty functions for the set $X' \cap X''$.

(b) Suppose that $f : \mathbb{R}^n \to \mathbb{R}^q$ is continuous and that

$$X' \triangleq \{x \in \mathbb{R}^n \mid f^j(x) \leq 0, \; j \in \mathbf{q}\}. \tag{3c}$$

For $k \in \mathbb{N}$, let π_k be such that $\pi_{k+1} > \pi_k > 0$ and $\pi_k \to \infty$, as $k \to \infty$, and let $\beta \geq 1$. Next, for $j \in \mathbf{q}$, let $f^j(x)_+ \triangleq \max\{0, f^j(x)\}$, and let

$$f(x)_+ \triangleq (f^1(x)_+, \ldots, f^q(x)_+). \tag{3d}$$

Show that if for $k \in \mathbb{N}$, $p_k^e : \mathbb{R}^n \to \mathbb{R}$ is defined by

$$p_k^e(x) \triangleq \pi_k \|f(x)_+\|_p^\beta, \tag{3e}$$

where $p = 1, 2, \infty$, then the $p_k^e(\cdot)$ form a sequence of exterior penalty functions for the set X'.

(c) Suppose that $g : \mathbb{R}^n \to \mathbb{R}^r$ is continuous, and that

$$X'' \triangleq \{x \in \mathbb{R}^n \mid g(x) = 0\}. \tag{3f}$$

Let π_k and β be as in *(b)*. Show that if we define $p_k^e : \mathbb{R}^n \to \mathbb{R}$ by

$$p_k^e(x) \triangleq \pi_k \|g(x)\|_p^\beta, \tag{3g}$$

where $p = 1, 2, \infty$, then the $p_k^e(\cdot)$ form a sequence of exterior penalty functions for the set X''. □

Next we will define interior penalty functions.

Definition 2.7.4. *Let X be a closed subset of $\mathbb{R}^n$, which is equal to the closure of its interior. Then a sequence of continuous functions $p_k^i : \operatorname{int} X \to \mathbb{R}$, $k \in \mathbb{N}$, is a sequence of* interior penalty functions *for the set X if*

$$0 < p_{k+1}^i(x) < p_k^i(x), \quad \forall x \in \operatorname{int} X, \; k \in \mathbb{N}, \tag{4a}$$

$$p_k^i(x) \to 0, \quad \text{as } k \to \infty, \quad \forall x \in \operatorname{int} X, \tag{4b}$$

and

$$p_k^i(x) \to \infty, \quad \text{as } x \to \partial X, \text{ with } x \in \operatorname{int} X, \quad \forall k \in \mathbb{N}, \tag{4c}$$

where ∂X denotes the boundary of X. □

A typical set of interior penalty functions is shown in Fig. 2.7.1b.

Exercise 2.7.5. *(a)* Suppose that $f : \mathbb{R}^n \to \mathbb{R}^q$ is continuous, that the set

$$X \triangleq \{x \in \mathbb{R}^n \mid f^j(x) \leq 0, \; j \in \mathbf{q}\}, \tag{5a}$$

is equal to the closure of its interior, and that, for all $j \in \mathbf{q}$, $f^j(x) < 0$ for all $x \in \operatorname{int} X$. For $k \in \mathbb{N}$, let π_k be such that $0 < \pi_{k+1} < \pi_k$ and $\pi_k \to 0$, as $k \to \infty$. Show that if, for $k \in \mathbb{N}$, the functions $p_k^i : \operatorname{int} X \to \mathbb{R}$ are defined by

$$p_k^i(x) \triangleq -\pi_k \sum_{j=1}^{q} \frac{1}{f^j(x)} \tag{5b}$$

or by

$$p_k^i(x) \triangleq -\pi_k \sum_{j=1}^{q} \left(\log[-f^j(x)] - \log[-\alpha^j]\right), \tag{5c}$$

where $\alpha^j \triangleq \inf\{f^j(x) \mid x \in X\}$, then $\{p_k^i(\cdot)\}_{k=0}^{\infty}$ is a sequence of interior penalty functions† for X, provided that $\min_{j \in \mathbf{q}} \alpha^j > -\infty$.

(b) Suppose that the functions $f^j(\cdot)$, $j \in \mathbf{q}$, defining the set X in (5a) are continuously differentiable. Let $\psi(x) \triangleq \max_{j \in \mathbf{q}} f^j(x)$. Show that if, for all $x \in X$ such that $\psi(x) = 0$, $0 \notin \partial\psi(x)$, then X is equal to the closure of its interior. □

Now consider the sequence of "simpler" (unconstrained) minimization problems that approximate problem **P** in (1a):

$$\mathbf{P}_k \qquad \min\{f^0(x) + p_k^e(x) + p_k^i(x) \mid x \in \operatorname{int} C'' \cap C'''\}, \quad k \in \mathbb{N}, \tag{6}$$

where $f^0(\cdot)$ and C'', C''' are as in (1a,b), the $p_k^e(\cdot)$ are exterior penalty functions for C' in (1b), and the $p_k^i(\cdot)$ are interior penalty functions for C'' in (1b). When $C' = \mathbb{R}^n$, we assume that $p_k^e(x) \equiv 0$ for all $k \in \mathbb{N}$ and, when $C'' = \mathbb{R}^n$, we assume that $p_k^i(x) \equiv 0$ for all $k \in \mathbb{N}$, since, in these two cases, we need not impose the restrictions stated in Definitions 2.7.2 and 2.7.4.

Note that the fact that the problem **P** has a solution does not ensure that the problems $\mathbf{P}_k$ have solutions. For example, consider the problem **P**

$$\min_{x \in \mathbb{R}} \{-x^3 \mid x = 0\},$$

whose solution is $\hat{x} = 0$. If we use an exterior penalty function of the form (3g), we get problems $\mathbf{P}_k$ of the form

$$\min_{x \in \mathbb{R}} -x^3 + \tfrac{1}{2}\pi_k x^2$$

that have no solution.

Definition 2.7.6. *Consider the problems* **P** *in (1a) and* $\mathbf{P}_k$ *in (6). By abuse of notation, we will call the cost function in (6)* $f_k^0 : \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbb{N}$, *defined by*

$$f_k^0(x) \triangleq f^0(x) + p_k^e(x) + p_k^i(x), \tag{7}$$

a penalty function *for the problem* **P** *in (1a). If there exists a finite integer* k_0 *such that, for all* $k \geq k_0$, *the solutions of* $\mathbf{P}_k$ *are also solutions of* **P**, *then we say that* $f_k^0(\cdot)$ *is an* exact penalty function *for* **P**. □

† Assuming that $p_k^i(\cdot)$ in (6) is defined by (5c), we see that the minimizer of $f_k^0(\cdot)$ does not depend on the terms $\log[-\alpha^j]$. Hence it is usual to remove this term from (5b). Although this results in a violation of the axioms for an interior penalty function, it does not invalidate the conclusions of Theorem 2.7.8.

Exercise 2.7.7. Consider the problem **P** and suppose that Assumptions 2.7.1*(i, ii)* are satisfied.

(a) Show that both **P** in (1a) and all $\mathbf{P}_k$ in (6) have solutions. *Hint*: First show that, for any $k \in \mathbb{N}$, the level set $L_{f(x^*)}(f_k^0)$ is a closed subset of the level set $L_{f(x^*)}(f^0)$, where

$$L_{f^0(x^*)}(f_k^0) \triangleq \{x \in \operatorname{int} C'' \cap C''' \mid f_k^0(x) \leq f^0(x^*)\} \tag{8a}$$

and

$$L_{f^0(x^*)}(f^0) \triangleq \{x \in \mathbb{R}^n \mid f^0(x) \leq f^0(x^*)\}. \tag{8b}$$

(b) Suppose that $C'' = C''' = \mathbb{R}^n$ holds, so that problem (1a) assumes the form $\min\{f^0(x) \mid x \in C'\}$, with minimum value $\underline{\alpha}$, and the problems (6) assume the form $\min_{x \in \mathbb{R}^n} f^0(x) + p_k^e(x)$, with minimum values $\underline{\alpha}_k$, $k \in \mathbb{N}$. Show that $\underline{\alpha}_0 \leq \underline{\alpha}_1 \leq \underline{\alpha}_2 \leq \ldots \leq \underline{\alpha}$ must hold.

(c) Suppose that $C' = C''' = \mathbb{R}^n$ holds, so that problem (1a) assumes the form $\min\{f^0(x) \mid x \in C''\}$ and the problems $\mathbf{P}_k$ in (6) assume the form $\min_{x \in \operatorname{int} C''} f^0(x) + p_k^i(x)$, with minimum values $\underline{\alpha}$, and $\underline{\alpha}_k$, respectively, $k \in \mathbb{N}$. Show that $\underline{\alpha}_0 \geq \underline{\alpha}_1 \geq \underline{\alpha}_2 \geq \ldots \geq \underline{\alpha}$ must hold. □

We are now ready to state the major result of this subsection:

Theorem 2.7.8. *Suppose that Assumption 2.7.1 is satisfied, that, for all* $k \in \mathbb{N}$, $p_k^e(x) \equiv 0$ *if* $C' = \mathbb{R}^n$, *and that, for all* $k \in \mathbb{N}$, $p_k^i(x) \equiv 0$ *if* $C'' = \mathbb{R}^n$. *For every* $k \in \mathbb{N}$, *let* x_k *be an optimal solution for problem* $\mathbf{P}_k$ *in (6). Then any accumulation point* $\hat{x}$ *of* $\{x_k\}_{k=0}^{\infty}$ *is an optimal solution for problem* **P** *in (1a).*

Proof. Without loss of generality, we may assume that $x_k \to \hat{x}$, as $k \to \infty$. First we will show that $\hat{x} \in C$. Since C'', C''' are closed by assumption, we must have that $\hat{x} \in C'' \cap C'''$. If $C' = \mathbb{R}^n$, then, clearly, $\hat{x} \in C$. Hence, suppose that $C' \neq \mathbb{R}^n$ and, for the sake of contradiction, that $\hat{x} \notin C'$. By Assumption 2.7.1*(iii)*, there exists an $x^* \in C' \cap \operatorname{int} C'' \cap C'''$. Since $p_k^e(x^*) = 0$ for all $k \in \mathbb{N}$, in view of (4b), we must have that

$$f^0(x^*) + p_k^e(x^*) + p_k^i(x^*) \to f^0(x^*), \tag{9a}$$

as $k \to \infty$. Let $\delta > 0$ be arbitrary. Then there exists a $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$,

$$f^0(x^*) + p_k^e(x^*) + p_k^i(x^*) \leq f^0(x^*) + \delta . \tag{9b}$$

Next, in view of (2b), since $\hat{x} \notin C'$, there exists a $k_1 \geq k_0$ such that, for all $k \geq k_1$, $f^0(\hat{x}) + p_k^e(\hat{x}) \geq f^0(x^*) + 4\delta$. Since $p_{k_1}^e(\cdot)$ is continuous, it follows that there exists a $\rho > 0$ such that, for all $\in B(\hat{x}, \rho)$, $k \geq k_1$,

$$f^0(x) + p_k^e(x) \geq f^0(x) + p_{k_1}^e(x) \geq f^0(x^*) + 2\delta . \tag{9c}$$

Since $x_k \to \hat{x}$, as $k \to \infty$, there exists a $k_2 \geq k_1$ such that $x_k \in B(\hat{x}, \rho)$, for all $k \geq k_2$ and hence for all $k \geq k_2$,

$$f^0(x_k) + p_k^e(x_k) + p_k^i(x_k) \geq f^0(x^*) + 2\delta > f^0(x^*) + p_k^e(x^*) + p_k^i(x^*), \tag{9d}$$

which contradicts the optimality of x_k, $k \geq k_2$. Thus we must have that $\hat{x} \in C$.

Now, again for the sake of contradiction, suppose that $\hat{x}$ is not optimal for (1a). By Assumption 2.7.1*(iii)*, there exists an $\tilde{x} \in C$ that is optimal for (1a) such that for any $\rho > 0$, $B(\tilde{x}, \rho) \cap C' \cap \operatorname{int} C'' \cap C'''$ is not empty. Since we must have that $f^0(\hat{x}) > f^0(\tilde{x})$ and $f^0(\cdot)$ is continuous, there exists a $\tilde{\rho} > 0$ such that

$$f^0(x) < f^0(\hat{x}), \quad \forall\, x \in B(\tilde{x}, \tilde{\rho}) . \tag{9e}$$

Let x' be any point in $B(\tilde{x}, \tilde{\rho}) \cap C' \cap \operatorname{int} C'' \cap C'''$. Then

$$f^0(x') + p_k^e(x') + p_k^i(x') = f^0(x') + p_k^i(x') \to f^0(x') < f^0(\hat{x}), \tag{9f}$$

as $k \to \infty$. Now, for all $k \in \mathbb{N}$, by definition of x_k, because of (2b) and (4a), we must have that

$$f^0(x_k) \leq f^0(x_k) + p_k^e(x_k) + p_k^i(x_k) \leq f^0(x') + p_k^i(x') . \tag{9g}$$

Hence, since $f^0(x_k) \to f^0(\hat{x})$ and $p_k^i(x') \to 0$, as $k \to \infty$, (9f,g) imply that

$$f^0(x') < f^0(\hat{x}) \leq \lim_{k \to \infty} \{ f^0(x_k) + p_k^e(x_k) + p_k^i(x_k) \} \leq f^0(x'), \tag{9h}$$

which is clearly impossible. Hence we have a contradiction, and our proof is complete. □

Exercise 2.7.9. Suppose that Assumption 2.7.1 is satisfied, that for all $k \in \mathbb{N}$, $p_k^e(x) \equiv 0$ if $C' = \mathbb{R}^n$ and that, for all $k \in \mathbb{N}$, $p_k^i(x) \equiv 0$ if $C'' = \mathbb{R}^n$. For every $k \in \mathbb{N}$, let x_k be a strict local minimizer for problem $\mathbf{P}_k$ in (6), so that, for some $\rho_k > 0$, $f_k^0(x_k) < f_k^0(x)$ for all $x \in B(x_k, \rho_k)$. Show that, if $\hat{x}$ is an accumulation point of $\{x_k\}_{k=0}^{\infty}$ and there exists a $\rho > 0$ such that $\rho_k \geq \rho$ for all $k \in \mathbb{N}$, then $\hat{x}$ is a local minimizer for problem $\mathbf{P}$ in (1a). □

Now consider the problem

$$\textbf{IECP} \qquad \min \{ f^0(x) \mid f(x) \leq 0,\ g(x) = 0 \}, \tag{10}$$

first discussed in Section 2.2, where we will assume that $f^0 : \mathbb{R}^n \to \mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}^q$, and $g : \mathbb{R}^n \to \mathbb{R}^r$ are at least once continuously differentiable, and $r < n$. By the vector inequality $f(x) \leq 0$, we mean that $f^j(x) \leq 0$ for all $j \in \mathbf{q}$.

If we elect to use the exterior penalty functions in (3e) and (3g) for the inequality and equality constraints, with $p = \beta = 2$, we obtain the penalized problems

$$\mathbf{P}_k \qquad \min_{x \in \mathbb{R}^n} f^0(x) + \tfrac{1}{2}\pi_k [\|f(x)_+\|^2 + \|g(x)\|^2], \quad k \in \mathbb{N}, \tag{11}$$

where $\{\pi_k\}_{k=0}^{\infty}$ is such that $0 < \pi_k < \pi_{k+1}$ for all k and $\pi_k \to \infty$, as $k \to \infty$. By inspection, the cost function $f_k^0 : \mathbb{R}^n \to \mathbb{R}$ for problem (11), defined by

$$f_k^0(x) \triangleq f^0(x) + \tfrac{1}{2}\pi_k [\|f(x)_+\|^2 + \|g(x)\|^2], \tag{12}$$

is continuously differentiable, and hence we can use an unconstrained optimization algorithm to compute, in a finite number of iterations, a vector x_k such that $\nabla f_k^0(x_k) \approx 0$. Here is what we can say about such vectors x_k.

Theorem 2.7.10. *Consider the problems (10) and (11). For any $x \in \mathbb{R}^n$, let $G(x) \triangleq g_x(x)$, let*

$$\mathbf{q}_A(x) \triangleq \{ j \in \mathbf{q} \mid f^j(x) \geq 0 \}, \tag{13}$$

as before, and let $F_A(x)$ be a matrix with with rows $\nabla f^j(x)^T$, $j \in \mathbf{q}_A(x)$. Suppose that, for every $x \in \mathbb{R}^n$, the matrices $F_A(x)$ and $G(x)$ satisfy the PLI condition in Definition 1.8.1(b), and that the sequence $\{x_k\}_{k=0}^{\infty}$, in $\mathbb{R}^n$ is such that $\|\nabla f_k^0(x_k)\| \leq \varepsilon_k$ with $\varepsilon_k > 0$ for all $k \in \mathbb{N}$ and $\varepsilon_k \to 0$, as $k \to \infty$. Then any accumulation point $\hat{x}$ of $\{x_k\}_{k=0}^{\infty}$ satisfies $f(\hat{x}) \leq 0$, $g(\hat{x}) = 0$, and, together with some $\hat{\mu} \in \Sigma_q^0$ such that $\hat{\mu}^0 > 0$ and $\hat{\zeta} \in \mathbb{R}^r$, the first-order optimality condition (2.2.23d,e).

Proof. Without loss of generality, we may assume that $x_k \to \hat{x}$, as $k \to \infty$. Now, for every $k \in \mathbb{N}$,

$$\nabla f_k^0(x_k) = \nabla f^0(x_k) + \pi_k [f_x(x_k)^T f(x)_+ + g_x(x_k)^T g(x_k)], \tag{14a}$$

where, as before, $f(x)_+ \triangleq (f^1(x)_+, \ldots, f^q(x)_+)$. For $k \in \mathbb{N}$, let

$$\left.\begin{aligned} \eta_k &\triangleq \pi_k f(x_k)_+ , \\ \xi_k &\triangleq \pi_k g(x_k) . \end{aligned}\right\} \tag{14b}$$

We begin by showing that the sequences $\{\eta_k\}_{k=0}^{\infty}$ and $\{\xi_k\}_{i=0}^{\infty}$ are bounded.

For suppose that at least one of these sequences is not bounded. For any $k \in \mathbb{N}$, let $\alpha_k = \sum_{j=1}^{q} \eta_k^j + \sum_{l=1}^{r} |\xi_k^l|$. Then we must have that $\alpha_k \to \infty$, as $k \to \infty$. Let $\eta'_k = \eta_k / \alpha_k$ and $\xi'_k = \xi_k / \alpha_k$, $k \in \mathbb{N}$. Then the sequences $\{\eta'_k\}_{k=0}^{\infty}$ and $\{\xi'_k\}_{i=0}^{\infty}$ are bounded and hence must have accumulation points. Without loss of generality we may assume that $\eta'_k \to \eta'$ and $\xi'_k \to \xi'$, as $k \to \infty$. Clearly, $\eta' \geq 0$, and $\eta'^j = 0$ for all $j \notin \mathbf{q}_A(\hat{x})$. It now follows from (14a) and the fact that, by assumption, $\varepsilon_k \to 0$, as $k \to \infty$, that

$$0 = \lim_{k \to \infty} \| \frac{1}{\alpha_k} \nabla f^0(x_k) + f_x(\hat{x})^T \eta' + g_x(\hat{x})^T \xi' \|$$

$$= \| f_x(\hat{x})^T \eta' + g_x(\hat{x})^T \xi' \| . \tag{14c}$$

Since this contradicts our assumption that $F_A(\hat{x})$ and $G(\hat{x})$ satisfy the PLI condition, we have a contradiction.

Because the sequences $\{\eta_k\}_{k=0}^{\infty}$ and $\{\xi_k\}_{i=0}^{\infty}$ are bounded, they must have accumulation points. Hence, because $f(\cdot)$ and $g(\cdot)$ are continuous, and $\pi_k \to \infty$, as $k \to \infty$, it follows from the definition of the η_k and the ξ_k that $f(\hat{x})_+ = 0$ and $g(\hat{x}) = 0$. Let $\eta_* \in \mathbb{R}^m$ and $\xi_* \in \mathbb{R}^r$ be accumulation points of $\{\eta_k\}_{k=0}^{\infty}$ and $\{\xi_k\}_{i=0}^{\infty}$, respectively. Then, because $f^0(\cdot)$, $f(\cdot)$ and $g(\cdot)$ are all continuously differentiable, we must have that

$$\nabla f^0(\hat{x}) + f_x(\hat{x})^T \eta_* + g_x(\hat{x})^T \xi_* = 0 . \tag{14d}$$

Furthermore, we must have that $\eta_* \geq 0$ and $\eta_*^j = 0$ for all $j \in \mathbf{q}$ such that $f^j(\hat{x}) < 0$. If we now divide all terms in (14d) by $\beta \triangleq (1 + \sum_{j=1}^{q} \eta_*^j)$, set $\hat{\eta}^0 = 1/\beta$, $\hat{\mu}^j = \eta_*^j / \beta$, $j \in \mathbf{q}$, and $\hat{\zeta} = \xi_* / \beta$, we see that the first-order optimality condition (2.2.23d,e) is satisfied, which completes our proof. □

Exercise 2.7.11. Consider the problem

$$\min \{ f^0(x) \mid g(x) = 0 \} , \tag{15a}$$

where the functions $f^0 : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^r$ are twice continuously differentiable and the Jacobian matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$. For all $x \in \mathbb{R}^n$ and $\pi > 0$, let

$$f_\pi^0(x) \triangleq f^0(x) + \tfrac{1}{2}\pi \| g(x) \|^2 . \tag{15b}$$

Suppose that $\{\pi_k\}_{k=0}^{\infty}$ is such that for all $k \in \mathbb{N}$, $\pi_k > 0$ and $\pi_k \to \infty$, as $k \to \infty$, and that $\{x_k\}_{k=0}^{\infty}$ is such that $x_k \to \hat{x}$, as $k \to \infty$. Then

$$\nabla f_{\pi_k}^0(x_k) = 0 , \quad \forall k \in \mathbb{N}, \tag{15c}$$

and there exists an $m > 0$ such that

$$\langle h , f_{\pi_k, xx}^0(x_k) h \rangle \geq m \| h \|^2 , \quad \forall h \in \mathbb{R}^n , \forall k \in \mathbf{N} . \tag{15d}$$

Show that $g(\hat{x}) = 0$, that there exists a multiplier $\hat{\zeta} \in \mathbb{R}^r$ such that

$$\nabla f^0(\hat{x}) + g_x(\hat{x})^T \hat{\zeta} = 0 , \tag{15e}$$

and that, with the Lagrangian defined by $L(x, \zeta) \triangleq f^0(x) + \langle \hat{\zeta} , g(x) \rangle$,

$$\langle h , L_{xx}(\hat{x} , \hat{\zeta}) h \rangle \geq m \| h \|^2 , \tag{15f}$$

for all $h \in \mathbb{R}^n$ such that $g_x(\hat{x}) h = 0$. Thus, show that local minimizers of $f_{\pi_k}^0(\cdot)$, satisfying a uniform, second-order sufficient condition, can converge only to strict local minimizers of (15a). □

Now consider again problem (15a). For $k \in \mathbb{N}$, let π_k be such that $0 < \pi_k < \pi_{k+1}$ and $\pi_k \to \infty$, as $k \to \infty$, and let x_k be such that (15c) holds. Suppose that $x_k \to \hat{x}$, as $k \to \infty$, so that, for some $\hat{\zeta} \in \mathbb{R}^r$, (15e) holds. Furthermore, suppose that $(\hat{x} , \hat{\zeta})$ satisfy the second-order sufficient condition stated in Corollary 2.2.30, i.e., (15e) holds and there exists an $m > 0$, such that, with the Lagrangian defined by $L(x , \hat{\zeta}) = f^0(x) + \langle \hat{\zeta} , g(x) \rangle$,

$$\langle h , L_{xx}(\hat{x} , \hat{\zeta}) h \rangle \geq m \| h \|^2 , \tag{16a}$$

for all $h \in \mathbb{R}^n$ such that $g_x(\hat{x}) h = 0$. Let $\zeta_k \triangleq \pi_k g(x_k)$, $k \in \mathbb{N}$, so that $\zeta_k \to \hat{\zeta}$, as $k \to \infty$, since the $\hat{\zeta}$ satisfying (15e) is unique because $g_x(\hat{x})$ has maximum row rank. Then we see that

$$f_{\pi_k, xx}^0(\hat{x}) = L_{xx}(\hat{x} , \hat{\zeta}) + \pi_k g_x(\hat{x})^T g_x(\hat{x}) . \tag{16b}$$

Referring to Proposition 2.8.1, we see that there exists a $\hat{\pi} > 0$ such that for all $\pi \geq \hat{\pi}$, $f_{\pi_k, xx}^0(\hat{x})$ is positive-definite. Clearly, this implies that there exists a $\hat{k}$ such that $f_{\pi_k, xx}^0(x_k)$ is positive-definite for all $k \geq \hat{k}$, so that $f_{\pi_k}^0(x)$ can be minimized using Newton's method. However, the conditioning of the matrices $f_{\pi_k, xx}^0(x_k)$ becomes progressively worse, which requires a better and better starting point for Newton's method. Hence, rather than choosing a very large penalty π_0 and minimizing $f_\pi^0(x)$, it makes better sense to proceed recursively, as follows. Start with a relatively small penalty π_0, and, after a global version of Newton's method has computed an x_0 such that $\nabla f_{\pi_0}^0(x_0) \approx 0$, increase it to $\pi_1 > \pi_0$, and repeat the entire process, over and over again, generating a sequence $\{x_i\}_{k=0}^{\infty}$. The increase of the penalty from k to $k+1$ should be kept sufficiently small to ensure that Newton's method continues to converge superlinearly. Clearly, a considerable amount of heuristics is needed to set up such a scheme.

Although, for the penalty function $f^0_{\pi_k}(\cdot)$ defined by (12), the matrix $f^0_{\pi_k,xx}(\cdot)$ is not always defined and not continuous, even when it is defined, again, one can reason as above to conclude that it is much preferable to increase the penalty slowly, rather than choose a very large penalty from the start.

One way to eliminate the need for heuristics in deciding when and how much to augment the penalty, in using (exterior or interior) penalty functions to solve a constrained optimization problem, is to use homotopies, which embed a difficult-to-solve equation into a parametrized equation for which the solution is known for one value of the parameter. The simplest situation arises in the case of problem (15a), under the assumption that the functions $f^0 : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^r$ are twice continuously differentiable, that the matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$, and that all the stationary points of (15a) satisfy second-order sufficient conditions. Let $\{\pi_k\}_{k=0}^{\infty}$ and $\{x_k\}_{k=0}^{\infty}$ be as in Exercise 2.7.11. For every $k \in \mathbb{N}$, let $\varepsilon_k \triangleq 1/\pi_k$. Then (15c) can be replaced by the following two equations:

$$\nabla_x L(x_k, \zeta_k) = 0, \tag{17a}$$

and

$$g(x_k) - \varepsilon_k \zeta_k = 0. \tag{17b}$$

Let $G : \mathbb{R}^n \times \mathbb{R}^r \times \mathbb{R} \to \mathbb{R}^n \times \mathbb{R}^r$ be defined by

$$G(x, \zeta, \varepsilon) \triangleq \begin{bmatrix} \nabla_x L(x, \zeta) \\ g(x) - \varepsilon\zeta \end{bmatrix}. \tag{17c}$$

It should be obvious that the limit point $(\hat{x}, \hat{\zeta})$, in Exercise 2.7.11 satisfies $G(\hat{x}, \hat{\zeta}, 0) = 0$. Now, the Jacobian of $G(\cdot, \cdot, \cdot)$ is given by

$$G_{(x,\zeta)}(\hat{x}, \hat{\zeta}, 0) = \begin{bmatrix} L_{xx}(\hat{x}, \hat{\zeta}) & g_x(\hat{x})^T \\ g_x(\hat{x}) & 0 \end{bmatrix} \tag{17d}$$

and, since $(\hat{x}, \hat{\zeta})$ satisfy second-order sufficient conditions, $G_{(x,\zeta)}(\hat{x}, \hat{\zeta}, 0)$ can be shown to be nonsingular (see Proposition 2.9.10). Hence, since $G(\cdot, \cdot, \cdot)$ is continuously differentiable, we can invoke the Implicit Function Theorem 5.1.33 to conclude that there exist continuously differentiable functions $x(\varepsilon)$ and $\zeta(\varepsilon)$ such that $x(0) = \hat{x}$, $\zeta(0) = \hat{\zeta}$, and, for all $\varepsilon > 0$ sufficiently small, $G(x(\varepsilon), \zeta(\varepsilon), \varepsilon) = 0$.

An important difference between minimizing the penalty function $f^0_{\pi}(x)$ and solving the equation $G(x, \zeta, \varepsilon) = 0$ with $\varepsilon = 1/\pi$ is that while the Hessian matrix $f^0_{\pi xx}(x)$ becomes progressively more ill-conditioned, as $\pi \to \infty$, and makes the minimization of $f^0_{\pi}(x)$ a progressively more ill-conditioned problem,

the Jacobian $G_{(x,\zeta)}(x, \zeta, \varepsilon)$ remains well-conditioned, as $\varepsilon \to 0$, which makes solving the equation $G(x, \zeta, \varepsilon) = 0$ a well-conditioned problem.

The final step is to set up a differential equation that will take us from x_0, computed by minimizing $f^0_{\pi_0}(x)$, to $\hat{x}$. Let $z = (x, \zeta)$, then, introducing an additional scalar variable $s \in [0, 1]$, we find that

$$G_z(z(s), \varepsilon(s))\frac{dz(s)}{ds} + G_\varepsilon(z(s), \varepsilon(s))\frac{d\varepsilon(s)}{ds} = 0, \tag{17e}$$

and hence that

$$\frac{dz(s)}{ds} = -G_z(z(s), \varepsilon(s))^{-1} G_\varepsilon(z(s), \varepsilon(s))\frac{d\varepsilon(s)}{ds}, \; s \in [0, 1], \; z(0) = z_0. \tag{17f}$$

To complete the specification of the differential equation, we can set

$$\frac{d\varepsilon(s)}{ds} = -\varepsilon_0, \; s \in [0, 1], \; \varepsilon(0) = \varepsilon_0. \tag{17g}$$

Assuming that $\varepsilon_0 = 1/\pi_0$ is sufficiently small for the above results to be valid, we can integrate (17f,g) numerically to compute $\hat{z} = (\hat{x}, \hat{\zeta})$. Referring to [AlP.88, LPY.90], we see that this approach applies to a large class of penalty functions, that other normalizations of the variable s are possible, and that special numerical methods for integrating equations such as (17f,g) have been developed. In fact, there is a considerable literature dealing with homotopy methods.

2.7.2 Exact Penalty Functions

The differentiable penalty functions used in the problems $\mathbf{P}_k$ defined in (11) and (15b), respectively, have the positive consequence that the cost functions for these problems are differentiable. They also have a negative consequence: as suggested by Fig. 2.7.2(a) (where $F(\cdot)$ is as defined in (18f)), there is no finite penalty π_k such that the solutions of $\mathbf{P}_k$ are also the solutions of $\mathbf{P}$, and as π_k increases, so does the ill-conditioning of the $\mathbf{P}_k$. Hence one must use either homotopy methods or extrapolation techniques, partial conjugate gradient methods, or Newton-like methods for solving the $\mathbf{P}_k$, as described in [FiM.68].

We will now establish a class of exact penalty functions for the most general of the problems considered in Section 2.2, i.e., **IECP**. Thus, consider the problem

$$\mathbf{P} \qquad \min\{f^0(x) \mid f^j(x) \le 0, \; j \in \mathbf{q}, \; g(x) = 0\}, \tag{18a}$$

where

$$f^0(x) \triangleq \max_{k \in \mathbf{p}} c^k(x), \tag{18b}$$

under the assumption that the functions $c^k : \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbf{p}$, $f^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, and $g : \mathbb{R}^n \to \mathbb{R}^r$ are all continuously differentiable, and $r < n$.

Now suppose that we chose to use the exterior penalty functions $\pi_k \max\{\|f(x)_+\|_\infty, \|g(x)\|_\infty\}$, with $\{\pi_k\}_{k=0}^\infty$ such that $\pi_{k+1} > \pi_k > 0$ for all $k \in \mathbb{N}$, and $\pi_k \to \infty$, as $k \to \infty$, that we have already encountered in (3d,f). Then we get the family of approximating problems

$$\mathbf{P}_{\pi_k} \qquad \min_{x \in \mathbb{R}^n} f_{\pi_k}^0(x), \quad k \in \mathbb{N}, \tag{18c}$$

where, for any $\pi > 0$, the function $f_\pi^0 : \mathbb{R}^n \to \mathbb{R}$ is defined by

$$f_\pi^0(x) \triangleq f^0(x) + \pi \max\{\|f(x)_+\|_\infty, \|g(x)\|_\infty\}$$

$$= f^0(x) + \pi \max\{0, \max_{j \in \mathbf{q}} f^j(x), \max_{l \in 2\mathbf{r}} g^l(x)\}, \tag{18d}$$

where for all $l \in \mathbf{r}$, $g^{l+r}(x) \triangleq -g^l(x)$.

Soon we will see that, under fairly unrestrictive assumptions, the functions $f_\pi^0(\cdot)$ are exact penalty functions for the problem (18a). We begin with a heuristic exploration. To this end, consider the special case of (18a)

$$\min\{f^0(x) \mid g(x) = 0\}, \tag{18e}$$

with $p = r = 1$ (so that $f^0(\cdot)$ is differentiable), and suppose that $\hat{x}$ is an optimal solution of (18e). Then, letting

$$F(x) \triangleq \begin{bmatrix} f^0(x) \\ g(x) \end{bmatrix}, \tag{18f}$$

we can draw $F(\mathbb{R}^n)$, the image of $\mathbb{R}^n$ under $F(\cdot)$ in $\mathbb{R}^2$, as shown in Fig. 2.7.2. First, referring to Fig. 2.7.2(a), we see that when the differentiable exterior penalty function $\pi g(x)^2$ is used, the values of the approximating problems are lower than the optimal value $f^0(\hat{x})$ for all $\pi > 0$ and hence that they cannot lead to exact penalty functions.

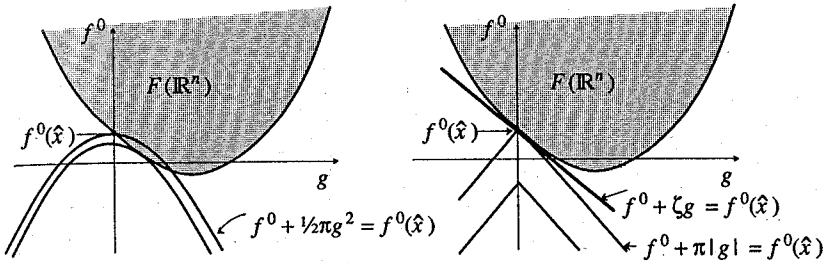
Suppose that the boundary of $F(\mathbb{R}^n)$ is smooth and that its slope is finite at $(f^0(\hat{x}), 0)$. Then the line tangent to the boundary at this point has the equation

$$f^0 + \hat{\zeta} g = f^0(\hat{x}), \tag{18g}$$

where $-\hat{\zeta}$ is the slope of the line (see Fig. 2.7.2*(b)*). Since, *locally*, all of $F(\mathbb{R}^n)$ lies to one side of this line, expanding $f^0(x) + \hat{\zeta} g(x)$ around $\hat{x}$ to first-order terms, we find that for some $\rho > 0$,

$$f^0(\hat{x}) + \langle \nabla f^0(\hat{x}) + \hat{\zeta} \nabla g(\hat{x}), (x - \hat{x}) \rangle \geq f^0(\hat{x}), \quad \forall x \in \mathbf{B}(\hat{x}, \rho). \tag{18h}$$

This leads to the conclusion that



(a) Differentiable penalty function (b) Exact penalty function

Fig. 2.7.2. Level sets of penalty functions.

$$\nabla f^0(\hat{x}) + \hat{\zeta} \nabla g(\hat{x}) = 0 \tag{18i}$$

must hold, i.e., $\hat{\zeta}$ is the Lagrange multiplier at $\hat{x}$. Next, if $\pi > |\hat{\zeta}|$, then it follows from Fig. 2.7.2b that

$$\min_{x \in \mathbb{R}^n} \{f^0(x) + \pi |g(x)|\} = f^0(\hat{x}), \tag{18j}$$

i.e., there is an unconstrained, but nondifferentiable optimization problem with the same solution point $\hat{x}$ and optimal value $f^0(\hat{x})$ as (18e).

The following straightforward generalizations of two theorems due to Han and Mangasarian [HaM.79] demonstrate the extent to which the above heuristic observations are true for the general case of problem (18a). For the sake of notational compactness we will use the definitions $c(x) \triangleq (c^1(x), \ldots, c^p(x))$ and $f(x) \triangleq (f^1(x), \ldots, f^q(x))$, and we will use the following two facts whose proof we leave as an exercise for the reader:

(i) Given scalars α^j, $j = 1, 2, \ldots, a$, and β^k, $k = 1, 2, \ldots, b$,

$$\max_{j \in \mathbf{a}} \alpha^j + \max_{k \in \mathbf{b}} \beta^k = \max_{j \in \mathbf{a}} \max_{k \in \mathbf{b}} (\alpha^j + \beta^k). \tag{18k}$$

(ii) Given $C_1, \ldots, C_k$, two convex subsets of $\mathbb{R}^n$,

$$\operatorname{co}_{j \in \mathbf{k}} C_j = \{x \in \mathbb{R}^n \mid x = \textstyle\sum_{j=1}^k \mu^j x_j, \ \mu \in \Sigma_k, \ x_j \in C_j, \ j \in \mathbf{k}\}. \tag{18l}$$

Theorem 2.7.12. *Consider problem* $\mathbf{P}$ *defined in (18a,b), and suppose that the functions* $c^k : \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbf{p}$, $f^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, *and* $g : \mathbb{R}^n \to \mathbb{R}^r$ *are all continuously differentiable and that the Jacobian matrix* $g_x(x)$ *has maximum row rank for all* $x \in \mathbb{R}^n$.

(a) Suppose that $\hat{x} \in \mathbb{R}^n$ *is such that* $\psi(\hat{x}) \leq 0$, $g(\hat{x}) = 0$, *and that there exist a* $\hat{\mu} \in \Sigma_q^0$, *with* $\hat{\mu}^0 > 0$, *a* $\hat{\nu} \in \Sigma_p$, *and a* $\hat{\zeta} \in \mathbb{R}^r$ *satisfying the first-order*

optimality condition (2.2.21c,d), i.e.,

$$\hat{\mu}^0 c_x(\hat{x})^T\hat{\nu} + f_x(\hat{x})^T\hat{\mu} + g_x(\hat{x})^T\hat{\zeta} = 0, \tag{19a}$$

$$\left.\begin{aligned} \sum_{k\in\mathbf{p}} \hat{\nu}^k[c^k(\hat{x}) - f^0(\hat{x})] = 0, \\ \sum_{j\in\mathbf{q}} \hat{\mu}^j f^j(\hat{x}) = 0. \end{aligned}\right\} \tag{19b}$$

Let

$$\hat{\pi} \triangleq \Big(\sum_{j\in\mathbf{q}} \hat{\mu}^j + \sum_{l\in\mathbf{r}} |\hat{\zeta}^l|\Big)/\hat{\mu}^0. \tag{19c}$$

Then, for all $\pi > \hat{\pi}$, $\hat{x}$ satisfies the first-order optimality condition (2.1.5a) (with $\psi(\cdot) \triangleq f_\pi^0(\cdot)$) for $\hat{x}$ to be a local minimizer of $f_\pi^0(\cdot)$, namely, $0 \in \partial f_\pi^0(\hat{x})$ (see Theorem 2.1.3).

(b) Suppose that the functions $c^k(\cdot)$, $k = 1,2,\ldots,p$, $f^j(\cdot)$, $j = 1,2,\ldots,q$, and $g(\cdot)$ are twice continuously differentiable and that $\hat{x}$ is a strict local minimizer of (18a) satisfying the second-order sufficient conditions in Theorem 2.2.29, i.e., (i) there exist a $\hat{\mu} \in \Sigma_q^0$, with $\hat{\mu}^0 > 0$, a $\hat{\nu} \in \Sigma_p$ and a $\hat{\zeta} \in \mathbb{R}^r$ satisfying (19a,b), and (ii) there exists an $m > 0$ such that, for the Lagrangian defined by

$$L(x,\hat{\mu},\hat{\nu},\hat{\zeta}) \triangleq \hat{\mu}^0\langle\hat{\nu}, c(x)\rangle + \langle\hat{\mu}, f(x)\rangle + \langle\hat{\zeta}, g(x)\rangle, \tag{19d}$$

$$\langle h, L_{xx}(\hat{x},\hat{\mu},\hat{\nu},\hat{\zeta})h\rangle \geq m\|h\|^2, \tag{19e}$$

for all $h \in H_{IE}'(\hat{x})$, where $L(\cdot,\cdot,\cdot,\cdot,\cdot)$ was defined in (2.2.27b) and $H_{IE}'(\hat{x})$ was defined in (2.2.33c). Let $\hat{\pi}$ be defined as in (19c). Then, for all $\pi > \hat{\pi}$, $\hat{x}$ is also a strict local minimizer of $f_\pi^0(\cdot)$.

Proof. (a) Suppose that (19a,b) hold with $\hat{\mu}^0 > 0$, and let $\pi > \hat{\pi}$. We begin by establishing a formula for the subgradient $\partial f_\pi^0(\hat{x})$. Let $\psi(x) \triangleq \max_{j\in\mathbf{q}} f^j(x)$, as before, and $\gamma(x) \triangleq \|g(x)\|_\infty$. Since $\gamma(\hat{x}) = 0$, it is easy to see that

$$\partial\gamma(\hat{x}) = \{g_x(\hat{x})^T\beta \mid \beta \in S\}, \tag{20a}$$

where $S \triangleq \{\beta \in \mathbb{R}^r \mid \sum_{l=1}^r |\beta^l| \leq 1\}$. With this new notation, it follows from (18d) that $f_\pi^0(x) = f^0(x) + \pi\max\{0, \psi(x), \gamma(x)\}$. Since $\psi(\hat{x}) \leq 0$ and $\gamma(\hat{x}) = 0$, by assumption, it follows from Theorems 5.4.8 and 5.4.9 that

$$\partial f_\pi^0(\hat{x}) = \partial f^0(\hat{x}) + \pi\,\mathrm{co}\{0, \delta\partial\psi(\hat{x}), \partial\gamma(\hat{x})\}, \tag{20b}$$

where $\delta = 1$ if $\psi(\hat{x}) = 0$ and $\delta = 0$ otherwise. Hence it follows from (18l) that[†]

$$\partial f_\pi^0(\hat{x}) = \partial f^0(\hat{x}) + \pi\{\lambda^2\partial\psi(\hat{x}) + \lambda^3\partial\gamma(\hat{x}) \mid \lambda \in \Sigma_3,\ \lambda^2\psi(\hat{x}) = 0\}. \tag{20b'}$$

Therefore, $y \in \partial f_\pi^0(\hat{x})$ if and only if there exist a $\hat{\nu} \in \Sigma_p$, an $\alpha \in \Sigma_q$, a $\beta \in S$, and a $\lambda \in \Sigma_3$, such that

$$y = c_x(\hat{x})^T\hat{\nu} + \pi\{\lambda^2 f_x(\hat{x})^T\alpha + \lambda^3 g_x(\hat{x})^T\beta\}, \tag{20c}$$

$$\left.\begin{aligned} \sum_{k\in\mathbf{p}} \hat{\nu}^k[c^k(\hat{x}) - f^0(\hat{x})] = 0, \\ \lambda^2\psi(\hat{x}) = 0, \\ \sum_{j=1}^q \alpha^j f^j(\hat{x}) = 0. \end{aligned}\right\} \tag{20d}$$

We now return to (19a). Let $\hat{\eta} \in \mathbb{R}^q$ be defined by $\hat{\eta}^j \triangleq \hat{\mu}^j/\hat{\mu}^0$, $j \in \mathbf{q}$, and let $\hat{\xi} \triangleq \hat{\zeta}/\hat{\mu}^0$. Then, for any $\pi > 0$, (19a) can be rewritten as

$$0 = c_x(\hat{x})^T\hat{\nu} + \pi\{f_x(\hat{x})^T\hat{\eta}/\pi + g_x(\hat{x})^T\hat{\xi}/\pi\}. \tag{20e}$$

Now, suppose that $\pi > \hat{\pi}$. Let $\lambda^2 \triangleq (\sum_{j\in\mathbf{q}} \hat{\eta}^j)/\pi$ and let $\lambda^3 \triangleq (\sum_{l\in\mathbf{r}} |\hat{\xi}^l|)/\pi$. If $\lambda^2 > 0$, let $\alpha \triangleq \hat{\eta}/\pi\lambda^2 = \hat{\eta}/\sum_{j\in\mathbf{q}}\hat{\eta}^j$, otherwise take any $\alpha \in \Sigma_q$. If $\lambda^3 > 0$, let $\beta \triangleq \hat{\xi}/\pi\lambda^3 = \hat{\xi}/\sum_{l\in\mathbf{r}}|\hat{\xi}^l|$, otherwise take any $\beta \in S$. Then we see that $\lambda^2 + \nu^3 \leq 1$, $\alpha \in \Sigma_q$, $\beta \in S$, and that (20e) reduces to (20c), with $y = 0$. If we now divide the second term in (19b) by $\pi\hat{\mu}^0$, we obtain (20d), which completes our demonstration that $0 \in \partial f_\pi^0(\hat{x})$ for all $\pi \geq \hat{\pi}$.

(b) Let $\pi > \hat{\pi}$ with $\hat{\pi}$ as in (19c), and suppose that $\hat{x}$ is not a strict local minimizer of $f_\pi^0(\cdot)$. Then there must exist a sequence $\{x_i\}_{i=0}^\infty$ such that $x_i \to \hat{x}$, as $i \to \infty$, and $f_\pi^0(x_i) \leq f_\pi^0(\hat{x})$ for all $i \in \mathbb{N}$. Let $x_i - \hat{x} = t_i h_i$ with $\|h_i\| = 1$. Then $t_i \to 0$, as $i \to \infty$. Without loss of generality, we can assume that $h_i \to \hat{h}$, as $i \to \infty$, and, obviously, $\|\hat{h}\| = 1$. Next, since the functions $c^k(\cdot)$, $k \in \mathbf{p}$, $f^j(\cdot)$, $j \in \mathbf{q}$, and $r^l(\cdot)$, $l \in \mathbf{r}$, are twice continuously differentiable, there must exist a

[†] Note that λ^1 does not appear in any of the expressions below because λ^1 multiplies the zero vector, $\partial 0(\hat{x})$, where $0(x)\equiv 0$ is the zero function.

local Lipschitz constant $L < \infty$ for their gradients, as used in obtaining the inequalities below:

$$\begin{aligned} f^0(x_i) - f^0(\hat{x}) &= \max_{k \in \mathbf{p}} c^k(x_i) - f^0(\hat{x}) \\ &\geq \max_{k \in \hat{\mathbf{p}}(\hat{x})} \{c^k(x_i) - c^k(\hat{x})\} \\ &= \max_{k \in \hat{\mathbf{p}}(\hat{x})} \{t_i \langle \nabla c^k(\hat{x}), h_i \rangle + t_i \langle [\nabla c^k(\hat{x} + s_i^k t_i h_i) - \nabla c^k(\hat{x})], h_i \rangle\} \\ &\geq t_i \max_{k \in \hat{\mathbf{p}}(\hat{x})} \langle \nabla c^k(\hat{x}), h_i \rangle - t_i^2 L, \end{aligned} \tag{20f}$$

where we have used the Mean-Value Theorem 5.1.28*(a)*, so that $s_i^k \in [0,1]$. Proceeding in a similar fashion, in turn we obtain

$$\begin{aligned} \max_{j \in \mathbf{q}} f^j(x_i)_+ &\geq \max_{j \in \mathbf{q}_A(\hat{x})} \max\{0, f^j(x_i)\} \\ &\geq \max_{j \in \mathbf{q}_A(\hat{x})} \max\{0, t_i \langle \nabla f^j(\hat{x}), h_i \rangle - t_i^2 L\} \\ &\geq t_i \max_{j \in \mathbf{q}_A(\hat{x})} \langle \nabla f^j(\hat{x}), h_i \rangle_+ - t_i^2 L, \end{aligned} \tag{20g}$$

where

$$\langle \nabla f^j(\hat{x}), h_i \rangle_+ \triangleq \max\{0, \langle \nabla f^j(x_i), h_i \rangle\}. \tag{20g'}$$

Similarly, for some $s_i^l \in [0,1]$,

$$\begin{aligned} \max_{l \in \mathbf{r}} |g^l(x_i)| &= \max_{l \in \mathbf{r}} |t_i(\langle \nabla g^l(\hat{x}), h_i \rangle + \langle \nabla g^l(\hat{x} + s_i^l t_i h_i) - \nabla g^l(\hat{x}), h_i \rangle)| \\ &\geq t_i \max_{j \in \mathbf{r}} (|\langle \nabla g^l(\hat{x}), h_i \rangle| - |\langle \nabla g^l(\hat{x} + s_i^l t_i h_i) - \nabla g^l(\hat{x}), h_i \rangle|) \\ &\geq t_i \max_{j \in \mathbf{r}} |\langle \nabla g^l(\hat{x}), h_i \rangle| - t_i^2 L. \end{aligned} \tag{20h}$$

Hence, for all $i \in \mathbb{N}$,

$$\begin{aligned} 0 &\geq f_\pi^0(x_i) - f_\pi^0(\hat{x}) \\ &= f^0(x_i) - f^0(\hat{x}) + \pi \max_{\substack{j \in \mathbf{q} \\ l \in \mathbf{r}}} \{f^j(x_i)_+, |g^l(x_i)|\} \\ &\geq t_i \max_{k \in \hat{\mathbf{p}}(\hat{x})} \langle \nabla c_k(\hat{x}), h_i \rangle \\ &\quad + t_i \pi \max_{\substack{j \in \mathbf{q}_A(\hat{x}) \\ l \in \mathbf{r}}} \{\langle \nabla f^j(\hat{x}), h_i \rangle_+, |\langle \nabla g^l(x_i), h_i \rangle|\} - t_i^2 L(\pi + 1). \end{aligned} \tag{20i}$$

Hence, dividing by t_i and letting $i \to \infty$, we conclude, from (20i), that

$$0 \geq \max_{k \in \hat{\mathbf{p}}(\hat{x})} \langle \nabla c^k(\hat{x}), \hat{h} \rangle + \pi \max_{\substack{j \in \mathbf{q}_A(\hat{x}) \\ l \in \mathbf{r}}} \{\langle \nabla f^j(\hat{x}), \hat{h} \rangle_+, |\langle \nabla g^l(\hat{x}), \hat{h} \rangle|\}. \tag{20j}$$

Now, making use of (19a,b), after dividing by $\hat{\mu}^0$, we find that

$$\sum_{k \in \hat{\mathbf{p}}(\hat{x})} \hat{\nu}^k \nabla c^k(\hat{x}) + \sum_{j \in \mathbf{q}_A(\hat{x})} \hat{\eta}^j \nabla f^j(\hat{x}) + \sum_{l \in \mathbf{r}} \hat{\xi}^l \nabla g^l(\hat{x}) = 0. \tag{20k}$$

Therefore, taking the inner product of (20k) with $\hat{h}$ and subtracting from (20j), we obtain

$$\begin{aligned} 0 \geq &\Big[\max_{k \in \hat{\mathbf{p}}(\hat{x})} \langle \nabla c_k(\hat{x}), \hat{h} \rangle - \sum_{k \in \hat{\mathbf{p}}(\hat{x})} \hat{\nu}^k \langle \nabla c^k(\hat{x}), \hat{h} \rangle\Big] \\ &+ \Big[\pi \max_{\substack{j \in \mathbf{q}_A(\hat{x}) \\ l \in \mathbf{r}}} \{\langle \nabla f^j(\hat{x}), \hat{h} \rangle_+, |\langle \nabla g^l(\hat{x}), \hat{h} \rangle|\} \\ &- \sum_{j \in \mathbf{q}_A(\hat{x})} \hat{\eta}^j \langle \nabla f^j(\hat{x}), \hat{h} \rangle - \sum_{l \in \mathbf{r}} \hat{\xi}^l \langle \nabla g^l(\hat{x}), \hat{h} \rangle\Big]. \end{aligned} \tag{20l}$$

Next,

$$\begin{aligned} &\sum_{j \in \mathbf{q}_A(\hat{x})} \hat{\eta}^j \langle \nabla f^j(\hat{x}), \hat{h} \rangle + \sum_{l \in \mathbf{r}} \hat{\xi}^l \langle \nabla g^l(\hat{x}), \hat{h} \rangle \\ &\leq \sum_{j \in \mathbf{q}_A(\hat{x})} \hat{\eta}^j \langle \nabla f^j(\hat{x}), \hat{h} \rangle_+ + \sum_{l \in \mathbf{r}} |\hat{\xi}^l| |\langle \nabla g^l(\hat{x}), \hat{h} \rangle| \\ &\leq \Big(\sum_{j \in \mathbf{q}_A(\hat{x})} \hat{\eta}^j + \sum_{l \in \mathbf{r}} |\hat{\xi}^l|\Big) \max_{\substack{j \in \mathbf{q}_A(\hat{x}) \\ l \in \mathbf{r}}} \{\langle \nabla f^j(\hat{x}), \hat{h} \rangle_+, |\langle \nabla g^l(\hat{x}), \hat{h} \rangle|\}: \end{aligned} \tag{20m}$$

Combining (20m) with (20l), we find that

$$0 \geq \Big[\max_{k \in \hat{\mathbf{p}}(\hat{x})} \langle \nabla c_k(\hat{x}), \hat{h} \rangle - \sum_{k \in \hat{\mathbf{p}}(\hat{x})} \hat{\nu}^k \langle \nabla c^k(\hat{x}), \hat{h} \rangle \Big]$$

$$+ (\pi - \sum_{j \in \mathbf{q}_A(\hat{x})} \hat{\eta}^j + \sum_{l \in \mathbf{r}} |\hat{\xi}^l|) \max_{\substack{j \in \mathbf{q}_A(\hat{x}) \\ l \in \mathbf{r}}} \{ \langle \nabla f^j(\hat{x}), \hat{h} \rangle_+, |\langle \nabla g^l(\hat{x}), \hat{h} \rangle| \}, \tag{20n}$$

By inspection, the first term in square brackets in (20n) is nonnegative, and because, by choice, $\pi > \sum_{j \in \mathbf{q}_A(\hat{x})} \hat{\eta}^j + \sum_{l \in \mathbf{r}} |\hat{\xi}^l|$, the second term in (20n) is also nonnegative. Since the entire expression (20n) is nonpositive, it follows that each term in square brackets is zero. Therefore,

$$\max_{k \in \hat{\mathbf{p}}(\hat{x})} \langle \nabla c^k(\hat{x}), \hat{h} \rangle = \sum_{k \in \hat{\mathbf{p}}(\hat{x})} \hat{\nu}^k \langle \nabla c^k(\hat{x}), \hat{h} \rangle, \tag{20o}$$

and

$$\max_{\substack{j \in \mathbf{q}_A(\hat{x}) \\ l \in \mathbf{r}}} \{ \langle \nabla f^j(\hat{x}), \hat{h} \rangle_+, |\langle \nabla g^l(\hat{x}), \hat{h} \rangle| \} = 0. \tag{20p}$$

Hence

$$\langle \nabla f^j(\hat{x}), \hat{h} \rangle_+ = 0, \quad \forall j \in \mathbf{q}_A(\hat{x}), \tag{20q}$$

and

$$\langle \nabla g^l(\hat{x}), \hat{h} \rangle = 0, \quad \forall l \in \mathbf{r}. \tag{20r}$$

Next, it follows from (20j) that

$$\max_{k \in \hat{\mathbf{p}}(\hat{x})} \langle \nabla c^k(\hat{x}), \hat{h} \rangle \leq 0. \tag{20s}$$

Finally, since by (20k)

$$\sum_{k \in \hat{\mathbf{p}}(\hat{x})} \hat{\nu}^k \langle \nabla c^k(\hat{x}), \hat{h} \rangle + \sum_{j \in \mathbf{q}_A(\hat{x})} \hat{\mu}^j \langle \nabla f^j(\hat{x}), \hat{h} \rangle + \sum_{l \in \mathbf{r}} \hat{\zeta}^l \langle \nabla g^l(\hat{x}), \hat{h} \rangle = 0, \tag{20t}$$

we conclude from (20q), (20r), and (20s) that

$$\langle \nabla c^k(\hat{x}), \hat{h} \rangle = 0, \quad \forall k \in \mathbf{p}_+(\hat{x}, \hat{\mu}, \hat{\nu}, \hat{\zeta}), \tag{20u}$$

and

$$\langle \nabla f^j(\hat{x}), \hat{h} \rangle = 0, \quad \forall j \in \mathbf{q}_{A+}(\hat{x}, \hat{\mu}, \hat{\nu}, \hat{\zeta}) \tag{20v}$$

(see (2.2.31a,b) for definitions of $\mathbf{p}_+(\hat{x}, \hat{\mu}, \hat{\nu}, \hat{\zeta})$ and $\mathbf{q}_{A+}(\hat{x}, \hat{\mu}, \hat{\nu}, \hat{\zeta})$). Hence

(20v), together with (20q,r,s), implies that $\hat{h} \in H_{IE}'(\hat{x})$, where $H_{IE}'(\hat{x})$ was defined in (2.2.31c). Now, by assumption, $\langle \hat{h}, L_{xx}(\hat{x}, \hat{\mu}, \hat{\nu}, \hat{\zeta}) \hat{h} \rangle \geq m$ (since $|\hat{h}| = 1$), and therefore, if we define $\hat{L}(x) \triangleq L(x, \hat{\mu}, \hat{\nu}, \hat{\zeta})$, then

$$\langle \hat{h}, \hat{L}_{xx}(\hat{x}) \hat{h} \rangle \geq m / \hat{\mu}^0. \tag{20w}$$

Next, because $\hat{L}(\cdot)$ is twice continuously differentiable,

$$\hat{L}(x_i) - \hat{L}(\hat{x}) = t_i \langle \nabla \hat{L}(\hat{x}), h_i \rangle + t_i^2 \int_0^1 (1-s) \langle h_i, \hat{L}_{xx}(\hat{x} + s t_i h_i) h_i \rangle \, ds. \tag{20x}$$

Now, by (19a), $\nabla \hat{L}(\hat{x}) = 0$. Hence, because $h_i \to \hat{h}$ and $t_i \to 0$, as $i \to \infty$, in view of (20w,x), there must exist an integer i_0 such that for all $i \geq i_0$,

$$\hat{L}(x_i) - \hat{L}(\hat{x}) \geq t_i^2 m / 4\hat{\mu}^0 > 0. \tag{20y}$$

Because $\pi > \hat{\pi}$, we must have that for all $i \geq i_0$,

$$\begin{aligned} f_\pi^0(\hat{x}) = f^0(\hat{x}) &= \hat{L}(\hat{x}) \\ &< \hat{L}(x_i) = f^0(x_i) + \sum_{j \in \mathbf{q}} \hat{\eta}^j f^j(x_i) + \sum_{l \in \mathbf{r}} \hat{\xi}^l g^l(x_i) \\ &\leq f^0(x_i) + \sum_{j \in \mathbf{q}} \hat{\eta}^j f^j(x_i)_+ + \sum_{l \in \mathbf{r}} |\hat{\xi}^l| |g^l(x_i)| \\ &\leq f^0(x_i) + s \Big[\sum_{j \in \mathbf{q}} \alpha^j f^j(x_i)_+ + \sum_{l \in \mathbf{r}} \beta^l |g^l(x_i)| \Big] \\ &\leq f_\pi^0(x_i), \end{aligned} \tag{20z}$$

where $s \triangleq \sum_{j \in \mathbf{q}} \hat{\eta}^j + \sum_{l \in \mathbf{r}} |\hat{\xi}^l|$, $\alpha^j \triangleq \hat{\eta}^j / s$, $\beta^l \triangleq \hat{\xi}^l / s$, and we have used the facts that $\pi > s$, that $\sum_{j=1}^q \alpha^j + \sum_{l=1}^r \beta^l = 1$, and that the maximum over a set of points in $\mathbb{R}$ is equal to the maximum over their convex hull. Hence we conclude that, for all $i \geq i_0$, $f_\pi(x_i) > f_\pi(\hat{x})$, which contradicts our hypothesis that $f_\pi(x_i) \leq f_\pi(\hat{x})$ for all $i \in \mathbb{N}$. Hence our proof is complete. □

Theorem 2.7.12 is obviously also valid for the simpler cases of problem (18a,b) that have only inequality constraints or only equality constraints, provided the absent terms are removed from all the relevant expressions. For example, for the problem

$$\min \{ f^0(x) \mid f^j(x) \leq 0, \ j \in \mathbf{q} \}, \tag{21a}$$

with $f^0(\cdot), f^j(\cdot), j \in \mathbf{q}$, as in (18a), the approximating problems have the form

$$\min_{x \in \mathbb{R}^n} \{ f^0(x) + \pi \| f(x)_+ \|_\infty \}, \tag{21b}$$

where $f : \mathbb{R}^n \to \mathbb{R}^q$ is defined by $f(x) \triangleq (f^1(x), \ldots, f^q(x))$. In this case, the sufficient condition Theorem 2.2.26 is replaced by its special form given in Theorem 2.2.13, so that the set $H_{IE}'(\hat{x})$ is replaced by the set $H_f'(\hat{x})$.

The next theorem supplies the remaining results needed to establish the equivalence of a problem of the form $\min_{x \in \mathbb{R}^n} f_\pi^0(x)$ to the original problem (18a,b).

Theorem 2.7.13. *Consider the problem (18a,b), with the associated exact penalty function $f_\pi^0(x)$, with $\pi > 0$ defined in (18d), and suppose that the functions $c^k : \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbf{p}$, $f^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, and $g : \mathbb{R}^n \to \mathbb{R}^r$ are all continuously differentiable and that the Jacobian matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$.*

(a) Suppose that $\hat{x} \in \mathbb{R}^n$ satisfies $\psi(\hat{x}) \leq 0$, $g(\hat{x}) = 0$. If $\hat{x}$ is a local minimizer of $f_\pi^0(\cdot)$, then $\hat{x}$ is a local minimizer for (18a).

(b) Suppose that $\hat{x} \in \mathbb{R}^n$ satisfies $\psi(\hat{x}) \leq 0$, $g(\hat{x}) = 0$. If $\hat{x}$ satisfies the first-order condition $0 \in \partial f_\pi^0(\hat{x})$ for a local minimizer of $f_\pi^0(\cdot)$, then there exist a $\hat{\mu} \in \Sigma_q^0$, with $\hat{\mu}^0 > 0$, a $\hat{\nu} \in \Sigma_p$, and a $\hat{\zeta} \in \mathbb{R}^r$ satisfying the optimality condition (2.2.21c,d), i.e. (19a,b).

(c) Let

$$\Psi(x) \triangleq \max \{ \| f(x)_+ \|_\infty, \| g(x) \|_\infty \} \tag{22a}$$

denote the constraint violation function for (18a,b), and let

$$\theta_\Psi(x) \triangleq \min_{h \in \mathbb{R}^n} \Big\{ \tfrac{1}{2}\|h\|^2 + \max \{ \| (f(x) + f_x(x)h)_+ \|_\infty, \| g(x) + g_x(x)h \|_\infty \} \Big\} - \Psi(x) \tag{22b}$$

be its associated optimality function (c.f. (2.1.9m)). Then, for any $x \in \mathbb{R}^n$ such that $\theta_\Psi(x) < 0$, there exists a $\pi_x < \infty$ such that $\theta_\pi(x) < 0$ for all $\pi > \pi_x$, where $\theta_\pi(\cdot)$ is the optimality function for $f_\pi^0(\cdot)$ defined, according to (2.1.9m), by

$$\theta_\pi(x) \triangleq \min_{h \in \mathbb{R}^n} \Big\{ \tfrac{1}{2}\|h\|^2 + \max_{k \in \mathbf{p}} [c^k(x) - f^0(x)] + \langle \nabla c^k(x), h \rangle + \pi \max \{ \| [f(x) + f_x(x)h]_+ \|_\infty, \| g(x) + g_x(x)h \|_\infty \} \Big\} - \pi\Psi(x). \tag{22c}$$

(d) If the matrices $f_x(x)$ and $g_x(x)$ satisfy the PLI condition (see Definition 1.8.1) for all $x \in \mathbb{R}^n$, then, given any compact set $X \subset \mathbb{R}^n$, there exists a $\pi_X < \infty$ such that $\theta_\pi(x) < 0$ for all $x \in X$ such that $\Psi(x) > 0$.

Proof. *(a)* Since $f_\pi^0(x) = f^0(x)$ for all $x \in \mathbb{R}^n$ such that $f^j(x) \leq 0$, $j \in \mathbf{q}$, and $g(x) = 0$, this part is obvious.

(b) We will use the representation of $\partial f_\pi^0(x)$ established in (20a-d). Since by assumption $0 \in \partial f_\pi^0(\hat{x})$, we may set $\xi = 0$ in (20c). Having done so, let $\hat{\mu}^0 \triangleq 1/(1 + \pi\lambda^2)$, $\hat{\mu}^j = \hat{\mu}^0 \pi\lambda^2 \alpha^j$, $j \in \mathbf{q}$, and $\hat{\zeta} = \hat{\mu}^0 \pi\lambda^3 \beta$. Then, multiplying (20c) by $\hat{\mu}^0$, we obtain (19a), and multiplying the second term in (20d) by $\hat{\mu}^0\pi$, we obtain (19b), which completes our proof of part *(b)*.

(c) Suppose that $x \in \mathbb{R}^n$ is such that $\theta_\Psi(x) < 0$. Let $h_\Psi(x) \in \mathbb{R}^n$ be the solution of (22b). Then, for any $\pi \geq 1$, we must have that

$$\theta_\pi(x) \leq \max_{k \in \mathbf{p}} \langle \nabla c^k(x), h_\Psi(x) \rangle + \pi\theta_\Psi(x). \tag{23a}$$

It follows, by inspection, that there is a finite $\pi_x > 0$ such that for all $\pi > \pi_x$, $\theta_\pi(x) < 0$.

(d) For any $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^{q+r}$ with $d = (d', d'')$, $d' \in \mathbb{R}^q$, $d'' \in \mathbb{R}^r$, let the set-valued map $\Omega(\cdot, \cdot)$ be defined by

$$\Omega(x, d) \triangleq \{ h \in \mathbb{R}^n \mid f_x(x)h \leq d', \; g_x(x)h = d'' \}. \tag{23b}$$

Then it follows from Theorem 5.3.10 that $\Omega(x, d)$ is not empty for all $x \in \mathbb{R}^n$, $d \in \mathbb{R}^{q+r}$, and is continuous. Hence, by Corollary 5.4.2, the function

$$\omega(x, d) \triangleq \min_{h \in \Omega(x,d)} \|h\| = -\max_{h \in \Omega(x,d)} -\|h\| \tag{23c}$$

is also continuous, and so is the function

$$w(x) \triangleq \max_{\|d\|_\infty = 1} \omega(x, d). \tag{23d}$$

Consequently, given any compact set $X \subset \mathbb{R}^n$, there exists a $\hat{w} < \infty$ such that $w(x) \leq \hat{w}$ for all $x \in X$. It now follows from Corollary 1.8.4 that, for any $x \in X$, the system of equations and inequalities

$$\left.\begin{aligned} f_x(x)h &\leq -f(x)_+, \\ g_x(x)h &= -g(x) \end{aligned}\right\} \tag{23e}$$

has a solution $h(x)$ such that $\|h(x)\| \leq \hat{w}\Psi(x)$ and, in addition, $f^j(x) + \langle \nabla f^j(x), h(x) \rangle \leq 0$ for all $j \in \mathbf{q}$. Placing this $h(x)$ in (22c), we conclude that

$$\theta_\pi(x) \leq \tfrac{1}{2}\|h(x)\|^2 + \max_{k \in \mathbf{p}} \langle \nabla c^k(x), h(x) \rangle - \pi\Psi(x)$$

$$\leq (\tfrac{1}{2}\hat{w}^2\Psi(x) + \hat{w}\max_{k\in\mathbf{p}}\|\nabla c^k(x)\| - \pi)\,\Psi(x)\,. \tag{23f}$$

Since X is compact and $\Psi(\cdot)$ and the $\nabla c^k(\cdot)$ are continuous, it now follows from (23f) that there exists a $\pi_X < \infty$ such that $\theta_\pi(x) < 0$ for all $\pi > \pi_X$, and all $x \in X$ such that $\Psi(x) > 0$, which completes our proof. □

Remark 2.7.14. Note that the PLI assumption, in Theorem 2.7.13*(d)*, can be replaced by the following somewhat less restrictive assumption in the form of the conclusion that we derived from the PLI assumption:

(a) for any $x \in \mathbb{R}^n$, the system of equations and inequalities

$$\left.\begin{aligned} f(x) + f_x(x)h &\leq 0 \\ g(x) + g_x(x)h &= 0 \end{aligned}\right\} \tag{23g}$$

has a solution h.

(b) For any compact set $X \subset \mathbb{R}^n$, there exists a $K_X < \infty$ such that, for every $x \in X$, (23g) has a solution $h(x)$, satisfying

$$\|h(x)\| \leq K_X \max\{\|f(x)_+\|_\infty, \|g(x)\|_\infty\}. \tag{23h}$$

□

To summarize, we find that, under the conditions invoked in Theorems 2.7.12 and 2.7.13, if we restrict ourselves to a sufficiently large, compact subset of $\mathbb{R}^n$ that intersects (or contains) the feasible set for the problem **P** in (18a,b), then there exists a $\hat{\pi} > 0$ such that for all $\pi > \hat{\pi}$, the problem $\mathbf{P}_\pi$ in (18c) is equivalent to the problem **P** in (18a,b) in the sense that they have common local and global minimizers, as well as common stationary points which satisfy appropriate first-order optimality conditions.

Since we have already encountered algorithms for solving optimization problems with inequality constraints, we may choose to remove only the equality constraints from the problem (18a,b) by exact penalty function techniques.

Exercise 2.7.15. Consider problem (18a,b), and suppose that the functions $c^k : \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbf{p}$, $f^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, and $g : \mathbb{R}^n \to \mathbb{R}^r$ are all continuously differentiable and the matrix $g_x(x)$ has maximum row rank for all $s \in \mathbb{R}^n$.

For any $\pi > 0$, let the penalized problem $\mathbf{P}_\pi^g$ be defined by

$$\mathbf{P}_\pi^g \qquad \min\{f^0(x) + \pi\|g(x)\|_\infty \mid f^j(x) \leq 0,\ j \in \mathbf{q}\}. \tag{24a}$$

(a) Suppose that $\hat{x} \in \mathbb{R}^n$ is such that $\psi(\hat{x}) \leq 0$, $g(\hat{x}) = 0$ and that there exist a $\hat{\mu} \in \Sigma_q^0$, with $\hat{\mu}^0 > 0$, a $\hat{\nu} \in \Sigma_p$, and a $\hat{\zeta} \in \mathbb{R}^r$ satisfying the first-order optimality condition (2.2.21c,d), i.e., (19a,b). Let $\hat{\pi} \triangleq \sum_{j=1}^r |\hat{\zeta}^j| / \hat{\mu}^0$. Show that for all

$\pi \geq \hat{\pi}$, $\hat{x}$ is a stationary point for (24a).

(b) Suppose that the functions $c^k(\cdot)$, $k = 1,2,\ldots,p$, $f^j(\cdot)$, $j = 1,2,\ldots,q$, and $g(\cdot)$ are twice continuously differentiable and that $\hat{x}$ is a strict local minimizer of (18a,b) satisfying the second-order sufficient conditions in Theorem 2.2.29, i.e., *(i)* there exist a $\hat{\mu} \in \Sigma_q^0$, with $\hat{\mu}^0 > 0$, a $\hat{\nu} \in \Sigma_p$ and a $\hat{\zeta} \in \mathbb{R}^r$ satisfying (19a,b), and *(ii)* there exists an $m > 0$ such that, for $L(x, \hat{\mu}, \hat{\nu}, \hat{\zeta})$ defined by (19d), (19e) holds. Show that for all $\pi > \hat{\pi} \triangleq \sum_{j=1}^r |\hat{\zeta}^j| / \hat{\mu}^0$, $\hat{x}$ is also a strict local minimizer for $\mathbf{P}_\pi^g$.

(c) Suppose that $\hat{x}$ satisfies $g(\hat{x}) = 0$, $f^j(\hat{x}) \leq 0$, for all $j \in \mathbf{q}$, and $\pi > 0$. Show that if $\hat{x}$ is a local minimizer (stationary point) for $\mathbf{P}_{\pi_s}$, then it is also a local minimizer (stationary point) for (18a,b).

(d) Suppose that, for all $x \in \mathbb{R}^n$ such that $\psi(x) \triangleq \max_{j\in\mathbf{q}} f^j(x) \geq 0$, $0 \notin \partial\psi(x)$ and that, for all $x \in \mathbb{R}^n$ such that $\psi(x) = 0$, the matrices $[\nabla f^j(x)]_{j\in\mathbf{q}_A(x)}$ and $g_x(x)^T$ satisfy the PLI condition in Definition 1.8.1 (where $[\nabla f^j(x)]_{j\in\mathbf{q}_A(x)}$ denotes a matrix with the indicated columns). Make use of Corollary 1.8.4 to show that, for every $x \in \mathbb{R}^n$ such that $\Psi(x) > 0$ (defined in (22a)), there exists a π_x such that, for all $\pi \geq \pi_x$, $\theta_\pi^g(x) < 0$, where $\theta_\pi^g(\cdot)$ is the optimality function for $\mathbf{P}_\pi^g$, defined by

$$\theta_\pi(x) \triangleq \min_{h\in\mathbb{R}^n}\Big\{\tfrac{1}{2}\|h\|^2 + \max\{\max_{k\in\mathbf{p}}\{c^k(x) - f^0(x) + \langle\nabla c^k(x), h\rangle\}$$
$$+ \pi[\|g(x) + g_x(x)h\|_\infty - \|g(x)\|_\infty], f^j(x)\}\Big\} - \psi(x)_+\,. \tag{24b}$$

□

2.7.3 Exact Penalty Function Algorithms

From an algorithmic point of view, our results on exact penalty functions are incomplete, because they do not include a rule for chosing the penalty π. As we will now show, this problem can be resolved by using Algorithm Model 2.3.6, which enables us to construct algorithms for solving constrained optimization problems with equality constraints of the form (18a,b). We begin with a simple, but important special case of (18a). Consider the problem

$$\min\{f^0(x) \mid g(x) = 0\}, \tag{25a}$$

where the functions $f^0 : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^r$ $(r < n)$ are both continuously differentiable and the Jacobian matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$. For this problem, the functions $f_\pi^0(\cdot)$ in (18d) assume the following form:

$$f_\pi^0(x) \triangleq f^0(x) + \pi \|g(x)\|_\infty = \max_{l \in 2\mathbf{r}} \{f^0(x) + \pi g^l(x)\}, \tag{25b}$$

where, for $l \in \mathbf{r}$, $g^{l+r}(\cdot) \triangleq -g^l(\cdot)$.

Under the above assumptions, for any $x \in \mathbb{R}^n$, the problem

$$\min_{\zeta \in \mathbb{R}^l} \|\nabla f^0(x) + g_x(x)^T \zeta\|^2 \tag{25c}$$

has a unique solution $\zeta(x)$, given by

$$\zeta(x) = -[g_x(x) g_x(x)^T]^{-1} g_x(x) \nabla f^0(x), \tag{25d}$$

and, for an algorithm of the form of Algorithm Model 2.3.6, we can use the test functions $t_j(\cdot)$, defined, in terms of parameters $\pi_0 > 0$ and $\sigma > 1$, $\tau > 1$, by

$$t_j(x) \triangleq -[\tau^j \pi_0 - \sigma \sum_{l=1}^r |\zeta^l(x)|]. \tag{25e}$$

We will now show that the test functions $t_j(\cdot)$ above satisfy Assumption 2.3.5, i.e., it has the properties *(a)*, *(b)*, and *(c)*, in the lemma below.

Lemma 2.7.16. *Consider the problem (25a) and suppose that the functions $f^0 : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^r$ are continuously differentiable and that the Jacobian matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$. Let the Lagrangian be defined by $L(x, \zeta) \triangleq f^0(x) + \langle \zeta, g(x) \rangle$. Then,*

(a) the test functions $t_j(\cdot)$, defined by (25e) are continuous,

(b) for every $x^* \in \mathbb{R}^n$ there exist a $\rho^* > 0$ and a $j^* \in \mathbb{N}$, such that, for all $x \in B(x^*, \rho^*)$ and $j \geq j^*$, $t_j(x) \leq 0$, and

(c) for any $j \in \mathbb{N}$, if $\hat{x}$ is such that $t_j(\hat{x}) \leq 0$ and $0 \in \partial f_\pi^0(\hat{x})$, with $\pi = \tau^j \pi_0$, then $\nabla_x L(\hat{x}, \zeta(\hat{x})) = 0$, and $g(\hat{x}) = 0$.

Proof. *(a)* The fact that the test functions $t_j(\cdot)$ are continuous follows directly from the continuous differentiability of $f^0(\cdot)$ and $g(\cdot)$, (25d), and our assumption that the Jacobian matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$.

(b) Given any $x^* \in \mathbb{R}^n$, there exists a j^* such that $\tau^j \pi_0 \geq 2\sigma \sum_{l=1}^r |\zeta^l(x^*)|$ for all $j \geq j^*$. It now follows by continuity that there exist a $\rho^* > 0$ such that, for all $x \in B(x^*, \rho^*)$ and $j \geq j^*$, $t_j(x) \leq 0$.

(c) First, since $0 \in \partial f_\pi^0(\hat{x})$, we must have that, for some $\hat{\mu} \in \Sigma_{2r}$,

$$\sum_{l=1}^{2r} \hat{\mu}^l [\nabla f^0(\hat{x}) + \pi \nabla g^l(\hat{x})] = \nabla f^0(\hat{x}) + \pi \sum_{l=1}^{2r} \hat{\mu}^l \nabla g^l(\hat{x}) = 0, \tag{26a}$$

and

$$\sum_{l=1}^{2r} \hat{\mu}^l \{[f^0(\hat{x}) + \pi g^l(\hat{x})] - [f^0(\hat{x}) + \pi \|g(\hat{x})\|_\infty\} = 0, \tag{26b}$$

which implies that

$$\sum_{l=1}^{2r} \hat{\mu}^l [g^l(\hat{x})] - \|g(\hat{x})\|_\infty = 0. \tag{26c}$$

For all $j \in \mathbf{r}$, let $\xi^l \triangleq \hat{\mu}^l - \hat{\mu}^{l+r}$. Then (26a) becomes

$$\nabla f^0(\hat{x}) + \pi \sum_{l=1}^r \xi^l \nabla g^l(\hat{x}) = 0, \tag{26d}$$

and since the vectors $\nabla g^l(\hat{x})$, $j \in \mathbf{r}$, are linearly independent by assumption, we must have that $\pi \xi^l = \zeta^l(\hat{x})$, $j \in \mathbf{r}$, which proves that $\nabla_x L(\hat{x}, \zeta(\hat{x})) = 0$. Next, suppose that $g(\hat{x}) \neq 0$. Then we must have that $\hat{\mu}^l \hat{\mu}^{l+r} = 0$ for all $l \in \mathbf{r}$, and hence that $\sum_{l=1}^r |\zeta^l| = 1$, which leads to the conclusion that

$$\sum_{l=1}^r |\zeta^l(\hat{x})| = \pi \sum_{l=1}^r |\xi^l| = \pi. \tag{26e}$$

But, because $t_j(\hat{x}) \leq 0$, $\pi \geq \sigma \sum_{l=1}^r |\zeta^l(\hat{x})|$, with $\sigma > 1$. Therefore we have a contradiction, which shows that $g(\hat{x}) = 0$ must hold. □

The following algorithm is of the form of Algorithm Model 2.3.6 and solves problem (25a). It uses the test functions $t_j(\cdot)$, defined by (25e), and the algorithm functions $A_j(\cdot)$, defined by the PPP min-max Algorithm 2.4.1, applied to the minimization of $f_{\pi_j}^0(\cdot)$, with $\pi_j = \tau^j \pi_0$.

Algorithm 2.7.17.

Parameters. $\alpha \in (0, 1]$, $\beta \in (0, 1)$, $\tau, \sigma > 1$, $\pi_0 > 0$, $\delta > 0$.

Data. $x_0 \in \mathbb{R}^n$,

Step 0. Set $i = 0$, $j = 0$.

Step 1. Compute $t_j(x_i)$ according to (25d,e).

Step 2. If $t_j(x_i) \leq 0$, go to Step 3.
Else, set $x_j^* = x_i$, set $\pi_{j+1} = \tau^{j+1} \pi_0$, replace j by $j+1$, and go to Step 1.

Step 3. Compute the PPP *search direction* (note that terms that cancel have been omitted below)

$$h_i \triangleq \arg\min_{h \in \mathbb{R}^n} \{ \tfrac{1}{2}\delta\|h\|^2 + \langle \nabla f^0(x_i), h \rangle + \max_{l \in 2\mathbf{r}} \pi_j [g^l(x_i) - \|g(x_i)\|_\infty + \langle \nabla g^l(x_i), h \rangle] \}, \tag{27a}$$

and the value of the *optimality function*

$$\theta_i \triangleq \min_{h \in \mathbb{R}^n} \{ \tfrac{1}{2}\delta\|h\|^2 + \langle \nabla f^0(x_i), h \rangle + \max_{l \in 2\mathbf{r}} \pi_j [g^l(x_i) - \|g(x_i)\|_\infty + \langle \nabla g^l(x_i), h \rangle] \}. \tag{27b}$$

Step 4. Compute the *step-size*

$$\lambda_i = \arg\max_{k \in \mathbb{N}} \{ \beta^k \mid f^0_{\pi_j}(x_i + \beta^k h_i) - f^0_{\pi_j}(x_i) - \beta^k \alpha \theta_i \le 0 \}. \tag{27c}$$

Step 5. Update: Set

$$x_{i+1} = x_i + \lambda_i h_i, \tag{27d}$$

replace i by $i+1$, and go to Step 1.

The properties of Algorithm 2.7.17 can be summarized as follows:

Theorem 2.7.18. *Consider problem (25a). Suppose that the functions $f^0 : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^r$ are both continuously differentiable and that the Jacobian matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$.*

(a) If Algorithm 2.7.17 constructs a finite sequence $\{x^*_j\}_{j=0}^{j^*}$ and the sequence $\{x_i\}_{i=0}^{\infty}$ is infinite, then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies $g(\hat{x}) = 0$ and the first-order optimality condition (2.2.23b) for (25a), i.e., $\nabla f^0(\hat{x}) + g_x(\hat{x})^T \zeta(\hat{x}) = 0$.

(b) If Algorithm 2.7.17 constructs an infinite sequence $\{x^*_j\}_{j=0}^{\infty}$, then $\{x^*_j\}_{j=0}^{\infty}$ has no accumulation points. □

Exercise 2.7.19. Use Theorem 2.3.7 to prove Theorem 2.7.18. □

While devising a test function for problem (25a) for use in Algorithm 2.7.17 and proving Theorem 2.7.18 is quite easy, devising a test function for the general case of problem (18a,b) is much more difficult. We will give full proofs only for the following, notationally simpler, problem:

$$\min \{ f^0(x) \mid g(x) = 0 \}, \tag{28a}$$

where $f^0(\cdot)$ is defined by (18b), the functions $c^k : \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbf{p}$, and $g : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable, and the Jacobian matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$. For this problem, the functions $f^0_\pi(\cdot)$ in (18d) assume the following form:

$$f^0_\pi(x) \triangleq f^0(x) + \pi \|g(x)\|_\infty = \max_{\substack{k \in \mathbf{p} \\ l \in 2\mathbf{r}}} \{ c^k(x) + \pi g^l(x) \}, \tag{28b}$$

where, again, for $l \in \mathbf{r}$, $g^{l+r}(\cdot) \triangleq -g^l(\cdot)$. We will use the test functions $t_\pi(\cdot)$ defined by

$$t_\pi(x) \triangleq \theta_\pi(x) + \frac{1}{\pi}\|g(x)\|_\infty, \tag{28c}$$

where $\theta_\pi(\cdot)$ is the optimality function for $f^0_\pi(\cdot)$, defined by

$$\theta_\pi(x) \triangleq \min_{h \in \mathbb{R}^n} \{ \tfrac{1}{2}\delta\|h\|^2 + \max_{k \in \mathbf{p}} \max_{l \in 2\mathbf{r}} \{ c^k(x) - f^0(x) + \langle \nabla c^k(x), h \rangle + \pi \{ g^l(x) - \|g(x)\|_\infty + \langle \nabla g^l(x), h \rangle \} \} \}, \tag{28d}$$

where $\delta > 0$. The associated PPP search direction for $f^0_\pi(\cdot)$ at $x \in \mathbb{R}^n$ is defined by

$$h_\pi(x) \triangleq \arg\min_{h \in \mathbb{R}^n} \{ \tfrac{1}{2}\delta\|h\|^2 + \max_{k \in \mathbf{p}} \max_{l \in 2\mathbf{r}} \{ c^k(x) - f^0(x) + \langle \nabla c^k(x), h \rangle + \pi \{ g^l(x) - \|g(x)\|_\infty + \langle \nabla g^l(x), h \rangle \} \} \}. \tag{28e}$$

We are now ready to state an algorithm of the form of Algorithm Model 2.3.6 for solving (28a), with $f^0(\cdot)$ defined by (18b). We will assume that $\{\pi_j\}_{j=0}^{\infty}$ is a sequence of positive numbers such that $\pi_{j+1} > \pi_j > 0$ for all $j \in \mathbb{N}$ and that $\pi_j \to \infty$, as $j \to \infty$. We will say that such a sequence is *diverging*. The algorithm below uses the test functions $t_\pi(\cdot)$ defined in (28c,d) and the algorithm functions $A_j(\cdot)$, defined by the PPP Min-Max Algorithm 2.4.1, applied to the minimization of $f^0_{\pi_j}(\cdot)$.

Algorithm 2.7.20.

Parameters. $\alpha \in (0,1]$, $\beta \in (0,1)$, $\delta > 0$, a diverging sequence of penalties $\{\pi_j\}_{j=0}^{\infty}$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$, $j = 0$.

Step 1. If $t_{\pi_j}(x_i) \le 0$ (where $t_{\pi_j}(x_i)$ is defined by (28c)), go to Step 2.
Else, go to Step 5.

Step 2. Compute the PPP *search direction* $h_i = h_{\pi_j}(x_i)$ and the value of the *optimality function* $\theta_i = \theta_{\pi_j}(x_i)$ according to (28d,e).

Step 3. Compute the *step-size*

$$\lambda_i = \arg\max_{k \in \mathbb{N}} \{\beta^k \mid f^0_{\pi_j}(x_i + \beta^k h_i) - f^0_{\pi_j}(x_i) - \beta^k \alpha \theta_i \le 0\}. \tag{29a}$$

Step 4. Update: Set

$$x_{i+1} = x_i + \lambda_i h_i, \tag{29b}$$

replace i by $i+1$, and go to Step 1.

Step 5. Set $x^*_j = x_i$, replace j by $j+1$, and go to Step 1.

Algorithm 2.7.20 is quite obviously of the form of Algorithm Model 2.3.6. The most difficult part in showing that Algorithm 2.7.20 satisfies the assumptions in Theorem 2.3.7 is to show that the test function $t_\pi(\cdot)$ satisfies Assumption 2.3.5, since the rest follows directly from the properties of the PPP Algorithm 2.4.1.

Lemma 2.7.21. *Consider the test function $t_\pi(\cdot)$ defined in (28c) for the problem defined by (28a) and (18b), and suppose that the functions $c^k : \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbf{p}$, and $g : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable, and the Jacobian matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$. Then,*

(a) for every $\pi > 0$, $t_\pi(\cdot)$ is continuous,

(b) for any $\pi > 0$, if $\hat{x}$ is such that $\theta_\pi(\hat{x}) = 0$ and $t_\pi(\hat{x}) \le 0$, then $g(\hat{x}) = 0$, and the first-order optimality conditions (2.2.21c,d) hold for the problem defined by (28a) and (18b), viz., for some $\hat{\nu} \in \Sigma_p$ and $\hat{\zeta} \in \mathbb{R}^r$,

$$\sum_{k=1}^{p} \hat{\nu}^k \nabla c^k(\hat{x}) + g_x(\hat{x})^T \hat{\zeta} = 0, \tag{30a}$$

and

$$\sum_{k=1}^{p} \hat{\nu}^k [c^k(\hat{x}) - f^0(\hat{x})] = 0, \tag{30b}$$

and

(c) for every $x^* \in \mathbb{R}^n$, there exists a $\rho^* > 0$ and a $\pi^* \in (0, \infty)$ such that, for all $x \in B(x^*, \rho^*)$ and $\pi \ge \pi^*$, $t_\pi(x) \le 0$.

Proof. To simplify notation, let $\gamma(x) \triangleq \|g(x)\|_\infty = \max_{l \in 2\mathbf{r}} g^l(x)$.

(a) The continuity of $t_\pi(\cdot)$ follows directly from the continuity of $\theta_\pi(\cdot)$ (see Theorem 2.1.6 *(e)*) and the continuity of $\gamma(\cdot)$.

(b) Suppose that $\hat{x} \in \mathbb{R}^n$ is such that, for some $\pi > 0$, $\theta_\pi(\hat{x}) = 0$ and $t_\pi(\hat{x}) \le 0$. Then it follows from the definition of $t_\pi(\hat{x})$ that $\gamma(\hat{x}) = 0$ and hence that $g(\hat{x}) = 0$. Next, it follows from Theorem 2.1.6*(d)* that $0 \in \bar{G} f^0_\pi(\hat{x})$, i.e., there exist multipliers $\mu^{k,l}$, $k \in \mathbf{p}$, $l \in 2\mathbf{r}$, such that $\sum_{k=1}^{p} \sum_{l=1}^{2r} \mu^{k,l} = 1$, and

$$0 = \sum_{k=1}^{p} \sum_{l=1}^{2r} \mu^{k,l} [\nabla c^k(\hat{x}) + \pi \nabla g^l(\hat{x})]$$

$$= \sum_{k=1}^{p} \hat{\nu}^k \nabla c^k(\hat{x}) + \pi \sum_{l=1}^{2r} \hat{\mu}^l \nabla g^l(\hat{x}), \tag{31a}$$

and, since $g(\hat{x}) = 0$,

$$\sum_{k=1}^{p} \sum_{l=1}^{2r} \mu^{k,l} [c^k(\hat{x}) - f^0(\hat{x})] = \sum_{k=1}^{p} \hat{\nu}^k [c^k(\hat{x}) - f^0(\hat{x})] = 0, \tag{31b}$$

where $\hat{\nu}^k \triangleq \sum_{l=1}^{2r} \mu^{k,l}$ and $\hat{\mu}^l \triangleq \sum_{k=1}^{p} \mu^{k,l}$. Hence we see that (30a) holds with the $\hat{\nu}^k$ as defined above and with $\hat{\zeta}^l \triangleq \pi[\hat{\mu}^l - \hat{\mu}^{l+r}]$, $l \in \mathbf{r}$, and that (30b) holds with the $\hat{\nu}^k$ as defined above.

(c) First we note that, for any $x \in \mathbb{R}^n$ and $\pi > 0$, $\theta_\pi(x)$ can be written in the more compact form

$$\theta_\pi(x) = \min_{h \in \mathbb{R}^n} \{ \tfrac{1}{2}\delta\|h\|^2 + \max_{k \in \mathbf{p}} \{c^k(x) - f^0(x) + \langle \nabla c^k(x), h \rangle\} + \pi\{\|g(x) + g_x(x)h\|_\infty - \|g(x)\|_\infty\} \}. \tag{31c}$$

Next, since by assumption, $g_x(x)$ has maximum rank for all $x \in \mathbb{R}^n$, the matrix $G(x) = [g_x(x) g_x(x)^T]^{-1}$ is well defined and continuous. Hence, if we define $h : \mathbb{R}^n \to \mathbb{R}^n$ by

$$h(x) \triangleq -g_x(x)^T G(x) g(x), \tag{31d}$$

then we see that $h(\cdot)$ is continuous and that $g(x) + g_x(x)h(x) = 0$ always holds. Since $c^k(x) - f^0(x) \le 0$ for all $k \in \mathbf{p}$, we conclude that

$$\theta_\pi(x) \le \tfrac{1}{2}\delta\|h(x)\|^2 + \max_{k \in \mathbf{p}} \langle \nabla c^k(x), h(x) \rangle - \pi\|g(x)\|_\infty. \tag{31e}$$

Now, using the fact that, for any $x \in \mathbb{R}^n$, $\|x\| \le \sqrt{n}\,\|x\|_\infty$, we conclude that $\|h(x)\|^2 \le A(x)\|g(x)\|^2_\infty$ and that $\max_{k \in \mathbf{p}} \langle \nabla c^k(x), h(x) \rangle \le B(x)\|g(x)\|_\infty$, where

$$A(x) \triangleq n \|g_x(x)^T G(x)\|^2, \tag{31f}$$

$$B(x) \triangleq \max_{k \in \mathbf{p}} \sqrt{n}\, \|G(x) g_x(x) \nabla c^k(x)\|. \tag{31g}$$

Hence

$$\theta_\pi(x) \leq \|g(x)\|_\infty \{ \tfrac{1}{2}\delta A(x) \|g(x)\|_\infty + B(x) - \pi \}, \tag{31h}$$

and, as a consequence,

$$t_\pi(x) \leq \|g(x)\|_\infty \{ \tfrac{1}{2}\delta A(x) \|g(x)\|_\infty + B(x) - \pi + \frac{1}{\pi} \}. \tag{31i}$$

Since $A(\cdot)$, $B(\cdot)$, and $g(\cdot)$ are all continuous, it follows that given any $x^* \in \mathbb{R}^n$, there exists a $\rho^* > 0$ such that, for all $x \in B(x^*, \rho^*)$,

$$[\tfrac{1}{2}\delta A(x) \|g(x)\|_\infty + B(x)] \leq 2[\tfrac{1}{2}\delta A(x^*) \|g(x^*)\|_\infty + B(x^*)] \triangleq \kappa^*. \tag{31j}$$

Let π^* be such that $\pi^* - 1/\pi^* \geq \kappa^*$. Then we see that, for all $x \in B(x^*, \rho^*)$ and any $\pi \geq \pi^*$, $t_\pi(x) \leq 0$, which completes our proof. □

In view of Lemma 2.7.21, the following result is a direct consequence of Theorems 2.3.7 and 2.4.2.

Theorem 2.7.22. *Consider the problem defined by (28a) and (18b), and suppose that the functions $c^k : \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbf{p}$, and $g : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable and the Jacobian matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$.*

(a) If Algorithm 2.7.20 constructs a finite sequence $\{x_j^*\}_{j=0}^{j^*}$ and the sequence $\{x_i\}_{i=0}^{\infty}$ is infinite, then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies $g(\hat{x}) = 0$ and the first-order optimality conditions (30a,b) for some $\hat{\nu} \in \Sigma_p$ and $\hat{\zeta} \in \mathbb{R}^r$.

(b) If Algorithm 2.7.20 constructs an infinite sequence $\{x_j^*\}_{j=0}^{\infty}$, then $\{x_j^*\}_{j=0}^{\infty}$ has no accumulation points. □

Finally, we are ready to tackle the full problem (18a,b), for which we define the test function $t_\pi(\cdot)$ by

$$t_\pi(x) \triangleq \theta_\pi(x) + \frac{1}{\pi}\Psi(x), \tag{32}$$

with $\Psi(\cdot)$ defined by (22a) and the optimality function $\theta_\pi(\cdot)$ for $f_\pi(\cdot)$ in (18d) defined by (22c).

Exercise 2.7.23. Consider the test function $t_\pi(\cdot)$ in (32) for problem (18a,b), and suppose that the functions $c^k : \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbf{p}$, $f^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbf{q}$, and $g : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable and the Jacobian matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$. Furthermore, suppose that, for every $x \in \mathbb{R}^n$ such that $\Psi(x) > 0$, $\theta_\Psi(x) < 0$, with $\theta_\Psi(\cdot)$ defined in (22b). Show that, under these assumptions,

(a) for every $\pi > 0$, $t_\pi(\cdot)$ is continuous,

(b) for any $\pi > 0$, if $\hat{x}$ is such that $\theta_\pi(\hat{x}) = 0$ and $t_\pi(\hat{x}) \leq 0$, then $f^j(\hat{x}) \leq 0$ for all $j \in \mathbf{q}$, $g(\hat{x}) = 0$, and the first-order optimality conditions (19a,b) hold for problem (18a,b), and

(c) for every $x^* \in \mathbb{R}^n$, there exists a $\rho^* > 0$ and a $\pi^* \in (0, \infty)$ such that, for all $x \in B(x^*, \rho^*)$ and $\pi \geq \pi^*$, $t_\pi(x) \leq 0$.

Hint: Make use of Theorem 2.7.13. □

In view of the result claimed in Exercise 2.7.23, the following result is a direct consequence of Theorem 2.3.7.

Theorem 2.7.24. *Consider problem (18a,b), and suppose that the functions $c^k : \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbf{p}$, and $g : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable and the Jacobian matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$. Suppose that, for every $x \in \mathbb{R}^n$ such that $\Psi(x) > 0$, $\theta_\Psi(x) < 0$, and that, in Algorithm 2.7.20, $f_\pi(\cdot)$ is computed according to (18d), $t_\pi(\cdot)$ according to (32), $\theta_\pi(\cdot)$ according to (22c), and $h_\pi(\cdot)$ according to*

$$h_\pi(x) \triangleq \arg\min_{h \in \mathbb{R}^n} \Big\{ \tfrac{1}{2}\|h\|^2 + \max_{k \in \mathbf{p}} [c^k(x) - f^0(x)] + \langle \nabla c^k(x), h \rangle + \pi \max \{ \|(f(x) + f_x(x)h)_+\|_\infty, \|g(x) + g_x(x)h\|_\infty \} \Big\} - \pi\Psi(x). \tag{33}$$

(a) If Algorithm 2.7.20 constructs a finite sequence $\{x_j^*\}_{j=0}^{j^*}$ and the sequence $\{x_i\}_{i=0}^{\infty}$ is infinite, then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies $f^j(\hat{x}) \leq 0$ for all $j \in \mathbf{q}$, $g(\hat{x}) = 0$, and the first-order optimality conditions stated in Theorem 2.2.19 (i.e., (19a,b)), with $\hat{\mu}^0 > 0$.

(b) If Algorithm 2.7.20 constructs an infinite sequence $\{x_j^*\}_{j=0}^{\infty}$, then $\{x_j^*\}_{j=0}^{\infty}$ has no accumulation points. □

Exercise 2.7.25. Develop an exact penalty function algorithm for solving the problem (18a,b) by using an equivalent problem of the form $\mathbf{P}_\pi^g$ in (24a), a new test function $t_\pi(\cdot)$, and the phase I - phase II Algorithm 2.6.1. □

2.7.4 Notes

Sequential unconstrained minimization methods, based on differentiable penalty functions, requiring that the penalty be driven to infinity, such as those at the beginning of this section, were first proposed by Courant [Cou.43], using exterior penalty functions. A systematic development of sequential unconstrained minimization methods, based on both exterior and interior penalty functions, can be found in [FiM.67, FiM.68] as well as in the much less accessible report [Loo.68].

Exact penalty functions methods were proposed, independently, by Eremin [Ere.66] and Zangwill [Zan.67]. Since then, a sizable literature has grown up on this subject. For surveys and general treatments of nondifferentiable exact penalty function methods, see [Con. 81, Bon.90, Bur.91, Bur.91a, Dip.94]. Conditions for and definitions of exactness can be found in [Ber.75, HaM.79, BaG.82, DiG.89, Bur.91]. A sampling of exact penalty function methods can be found in [Pie.69, Con.73, PsD.75, CoP.77, Han.77, CoC.82, Fle.84].

The theory of test functions used in this section, which makes it possible to determine, adaptively, a finite satisfactory penalty value, was first presented in [Pol.76] as a generalization of a test function used in an augmented Lagrangian method described in [MuP.75].

An alternative to the method presented in Exercise 2.7.15 for the removal of equality constraints from problem (18a), when all the functions $f^j(\cdot)$, $j = 0, \ldots, q$ and $g(\cdot)$ are continuously differentiable, was presented in [MaP.76]. What is particularly interesting about [MaP.76] is that it presents an algorithm based on a *differentiable* exact penalty function to remove the equality constraints from problem (18a). This approach has recently been implemented in the sequential quadratic programming computer code CFSQP, produced at the University of Maryland, under the direction of Prof. A. L. Tits. CFSQP is based on [PaT.91, PaT.93].

We will briefly describe the results in [MaP.76], which depend on the following assumption.

Assumption 2.7.26. *We will assume that the cost function $f^0(\cdot)$, as well as the constraint functions $f^j(\cdot)$, $j \in \mathbf{q}$, and $g(\cdot)$ in (18a) are continuously differentiable and that, for every $x \in \mathbb{R}^n$, the vectors $\nabla f^j(x)$, $j \in \mathbf{q}_A(x)$, together with the vectors $\nabla g^l(x)$, $l \in \mathbf{r}$, are linearly independent ($\mathbf{q}_A(x)$ was defined in (13)).* □

One of the consequences of Assumption 2.7.26, is as follows. For any $x \in \mathbb{R}^n$, let $F(x) \triangleq \operatorname{diag}([f^1(x)]^2, \ldots, [f^q(x)]^2)$ be an $q \times q$ matrix, and let $\mu : \mathbb{R}^n \to \mathbb{R}^q$ and $\zeta : \mathbb{R}^n \to \mathbb{R}^r$ be defined by

$$(\mu(x), \zeta(x)) \triangleq \arg \min_{\substack{\mu \in \mathbb{R}^q \\ \zeta \in \mathbb{R}^r}} \{ \|\nabla f^0(x) + f_x(x)^T \mu + g_x(x)^T \zeta\|^2 + \langle \mu, F(x)\mu \rangle \}, \tag{34a}$$

where $f(x) = (f^1(x), \ldots, f^q(x))$. Clearly, if $\hat{x}$ is a local minimizer for (18a), then the value of the quadratic program (34a) is zero. Furthermore, in Proposition 2.8.16, it is shown that, under Assumption 2.7.26, $(\mu(\hat{x}), \zeta(\hat{x}))$ are the unique Karush-Kuhn-Tucker multipliers for (18a) and both $\mu(\cdot)$ and $\zeta(\cdot)$ are continuous functions.

Now, for any $\pi > 0$, consider the penalized problem

$$\mathbf{P}'^g_\pi \qquad \min \{ f^0(x) - \pi \sum_{l \in \mathbf{r}} f^{l+q}(x) \mid f^j(x) \le 0,\ j \in \mathbf{q}' \}, \tag{34b}$$

where $\pi \sum_{l \in \mathbf{r}} f^{l+q}(x)$ is the penalty term, $q' = q + r$, and, for $l \in \mathbf{r}$, $f^{l+q}(x) \triangleq -g^l(x)$. Let

$$f^0_\pi(x) \triangleq f^0(x) - \pi \textstyle\sum_{l \in \mathbf{r}} f^{l+q}(x). \tag{34c}$$

Then it follows from (2.1.8b) that an optimality function for $\mathbf{P}'^g_\pi$ is given (with $\delta = \gamma = 1$) by

$$\theta'^g_\pi(x) \triangleq \min_{h \in \mathbb{R}^n} \left\{ \tfrac{1}{2}\|h\|^2 + \max\left\{ \langle \nabla f^0_\pi(x), h \rangle, \max_{j \in \mathbf{q}'} \{ f^j(x) + \langle \nabla f^j(x), h \rangle \} \right\} \right\}$$

$$- \psi(x)_+, \tag{34d}$$

where $\psi(x) \triangleq \max_{j \in \mathbf{q}'} f^j(x)$.

We can establish the following relationships between the problems (18a) and (34b).

Theorem 2.7.27. *Suppose that Assumption 2.7.26 is satisfied.*

(a) If $(\hat{x}, \hat{\mu}, \hat{\zeta})$ is a triplet satisfying the necessary conditions (2.2.23d,e) for (18a) and $\pi \ge \hat{\pi} \triangleq \|\hat{\zeta}\|_\infty / \hat{\mu}^0$, then there exists a $\mu^* \in \Sigma^0_{q'}$, such that $(\hat{x}, \mu^*)$ satisfy the necessary conditions (2.2.7a,b) for (34b).

(b) If $f^j(\cdot)$, $j = 0, \ldots, q$ and $g(\cdot)$ are twice continuously differentiable and $(\hat{x}, \hat{\mu}, \hat{\zeta})$ is a triplet satisfying the necessary conditions (2.2.23d,e), as well as the second-order sufficient conditions for a strict local minimizer of (18a), stated in Corollary 2.2.27, and $\pi \ge \hat{\pi} \triangleq \|\hat{\zeta}\|_\infty / \hat{\mu}^0$, then there there exists a $\mu^* \in \Sigma^0_{q'}$, such that $(\hat{x}, \mu^*)$ satisfy both the necessary conditions (2.2.7a,b) and the second-order sufficient conditions for a strict local minimizer of (34b), stated in Corollary 2.2.14.

(c) For any $\pi > 0$, if $\hat{x}$ is a local/global minimizer or stationary point for (34b) and $g(\hat{x}) = 0$, then $\hat{x}$ is a local/global minimizer or stationary point for (18a).

(d) For any $x \in \mathbb{R}^n$ such that $\sum_{l \in \mathbf{r}} f^{l+q}(x) > 0$, if $\pi > \|\zeta(x)\|_\infty$, then $\theta'^g_\pi(x) < 0$. □

Theorem 2.7.27*(c,d)* has the following important corollary:

Corollary 2.7.28. *Suppose that Assumption 2.7.26 is satisfied. If $\hat{x}$ is a local/global minimizer or stationary point for $\mathbf{P}'^g_\pi$, and $\pi > \|\zeta(\hat{x})\|$, then $\hat{x}$ is a local/global minimizer or stationary point for (18a).* □

Theorem 2.7.27 and Corollary 2.7.28 suggest that, if one uses a test function based on the function $\zeta(\cdot)$, and, starting from a *feasible point* for (34b), uses a phase II method of centers to solve problem (34b), then any accumulation point, thus produced, must be a feasible and stationary point for (18a). Indeed, this is the case. Let the test functions $t_j(\cdot)$, $j \in \mathbb{N}$, be defined, in terms of parameters $\pi_0 > 0$ and $\kappa > 1$, $\tau > 1$, by

$$t_j(x) \triangleq -[\tau^j \pi_0 - \kappa \|\zeta(x)\|_\infty], \tag{35}$$

with $\zeta(\cdot)$ as in (34a). It can be shown that the $t_j(\cdot)$ satisfy the conclusions of an appropriate modification of Lemma 2.7.16 and hence can be incorporated in an exact penalty function algorithm, using a phase II method of centers, based on the optimality function $\theta'^g_\pi(\cdot)$, as a subroutine, as follows.

Algorithm 2.7.29.

Parameters. $\alpha, \beta \in (0,1)$, $\tau, \kappa > 1$, $\pi_0 > 0$.

Data. $x_0 \in \mathbb{R}^n$, such that $f^j(x_0) \leq 0$ for all $j \in \mathbf{q}'$.

Step 0. Set $i = 0$, $j = 0$.

Step 1. Compute $t_j(x_i)$ according to (34a) and (35).

Step 2. If $t_j(x_i) \leq 0$, go to Step 3.
Else, set $x_j^* = x_i$, set $\pi_{j+1} = \tau^{j+1}\pi_0$, replace j by $j+1$, and go to Step 1.

Step 3. Compute the phase II *search direction* for problem $\mathbf{P}'^g_{\pi_j}$, given by

$$h_i \triangleq \arg\min_{h \in \mathbb{R}^n} \left\{ \tfrac{1}{2}\|h\|^2 + \max\left\{ \langle \nabla f^0_{\pi_j}(x_i), h \rangle, \max_{j \in \mathbf{q}'} \{ f^j(x_i) + \langle \nabla f^j(x_i), h \rangle \} \right\} \right\} \tag{36a}$$

and the value of the *optimality function*

$$\theta_i \triangleq \min_{h \in \mathbb{R}^n} \left\{ \tfrac{1}{2}\|h\|^2 + \max\left\{ \langle \nabla f^0_{\pi_j}(x_i), h \rangle, \max_{j \in \mathbf{q}'} \{ f^j(x_i) + \langle \nabla f^j(x_i), h \rangle \} \right\} \right\}. \tag{36b}$$

Step 4. Compute the *step-size*

$$\lambda_i = \arg\max_{k \in \mathbf{N}} \left\{ \beta^k \mid f^0_{\pi_j}(x_i + \beta^k h_i) - f^0_{\pi_j}(x_i) - \beta^k \alpha \theta_i \leq 0, \; f^j(x_i + \beta^k h_i) \leq 0, \; \forall j \in \mathbf{q}' \right\}. \tag{36d}$$

Step 5. Set

$$x_{i+1} = x_i + \lambda_i h_i, \tag{36e}$$

replace i by $i+1$, and go to Step 1.

To compute an initial x_0 satisfying $f^j(x_0) \leq 0$ for all $j \in \mathbf{q}'$, apply the PPP min-max Algorithm 2.4.1 to the function $\psi(x) \triangleq \max_{j \in \mathbf{q}} f^j(x)$ for the finite number of iterations needed to compute an x_0 such that $\psi(x_0) \leq 0$. Then, for all $l \in \mathbf{r}$ such that $g^l(x_0) < 0$, replace $g^l(\cdot)$ by $-g^l(\cdot)$.

The properties of Algorithm 2.7.29 can be summarized as follows:

Theorem 2.7.30. *Consider problem (18a), and suppose that Assumption 2.7.26 is satisfied.*

(a) If Algorithm 2.7.29 constructs a finite sequence $\{x_j^*\}_{j=0}^{j^*}$ and the sequence $\{x_i\}_{i=0}^{\infty}$ is infinite, then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies $f^j(\hat{x}) \leq 0$, $j \in \mathbf{q}$, $g(\hat{x}) = 0$, and, together with $\mu(\hat{x})$, $\zeta(\hat{x})$ (defined by (34a)), the first-order optimality conditions (2.2.23d,e).

(b) If Algorithm 2.7.29 constructs an infinite sequence $\{x_j^*\}_{j=0}^{\infty}$, then $\{x_j^*\}_{j=0}^{\infty}$ has no accumulation points. □

2.8 Augmented Lagrangian Methods

Augmented Lagrangian methods are based on exact, differentiable penalty functions that are an outgrowth of a study of second-order conditions for equality constrained optimization problems. The first versions of such methods were proposed, independently, by Hestenes [Hes.69] and Powell [Pow.69]. Since then, several other versions have been proposed. We will describe two algorithms, one for optimization problems with equality constraints only and one for optimization problems with mixed constraints. Both of these algorithms are based on Algorithm Model 2.3.6.

2.8.1 Problems with Equality Constraints

Consider the equality constrained problem

$$\mathbf{P} \qquad \min\{ f^0(x) \mid g(x) = 0 \}, \tag{1a}$$

where $f^0 : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^r$, and $r < n$.

Suppose that $f^0(\cdot)$ and $g(\cdot)$ are twice continuously differentiable functions and that $g_x(x)$ is of maximum row rank for all $x \in \mathbb{R}^n$. Let $\pi > 0$, and consider the "augmented" problem

$$\mathbf{P}_\pi \qquad \min\{ f^0(x) + \tfrac{1}{2}\pi\|g(x)\|^2 \mid g(x) = 0 \}. \tag{1b}$$

Then it is obvious that $\hat{x}$ is a local minimizer for $\mathbf{P}$ if and only if it is a local minimizer for $\mathbf{P}_\pi$.

Now suppose that $\hat{x}$ is a local minimizer for $\mathbf{P}$, and that there exists a multiplier $\hat{\zeta} \in \mathbb{R}^r$ such that the pair $(\hat{x}, \hat{\zeta})$ satisfies the second-order sufficient condition stated in Corollary 2.2.31, i.e.,

$$\nabla_x L(\hat{x}, \hat{\zeta}) = \nabla f^0(\hat{x}) + g_x(\hat{x})^T \hat{\zeta} = 0, \tag{1c}$$

and there exists an $m > 0$ such that

$$\langle h, L_{xx}(\hat{x}, \hat{\zeta}) h \rangle \geq m\|h\|^2, \quad \forall h \in \mathcal{H}_E(\hat{x}), \tag{1d}$$

where

$$\mathcal{H}_E(\hat{x}) \triangleq \{ h \in \mathbb{R}^n \mid g_x(\hat{x}) h = 0 \}, \tag{1e}$$

and the Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}$ is defined by

$$L(x, \zeta) \triangleq f^0(x) + \langle \zeta, g(x) \rangle. \tag{1f}$$

For the augmented problem $\mathbf{P}_\pi$, the (augmented) Lagrangian $L_\pi : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}$ is defined by

$$L_\pi(x,\zeta) \triangleq f^0(x) + \langle \zeta, g(x)\rangle + \tfrac{1}{2}\pi\|g(x)\|^2. \tag{1g}$$

Since $g(\hat{x}) = 0$, we see that

$$\nabla_x L_\pi(\hat{x},\hat{\zeta}) = \nabla_x L(\hat{x},\hat{\zeta}) = 0, \tag{1h}$$

and, because

$$L_{\pi xx}(\hat{x},\hat{\zeta}) = L_{xx}(\hat{x},\hat{\zeta}) + \pi g_x(\hat{x})^T g_x(\hat{x}), \tag{1i}$$

$$\langle h, L_{\pi xx}(\hat{x},\hat{\zeta})h\rangle \geq m\|h\|^2, \tag{1j}$$

for all $h \in \mathcal{H}_E(\hat{x})$, i.e., the pair $(\hat{x},\hat{\zeta})$ also satisfy the second-order sufficient condition, stated in Corollary 2.2.31, for the problem $\mathbf{P}_\pi$. However, a really interesting observation is contained in the following result:

Proposition 2.8.1. *Consider the problem* **P**. *Suppose that*

(i) the functions $f^0(\cdot)$ and $g(\cdot)$ are twice continuously differentiable, and that the matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$,

(ii) $\hat{x} \in \mathbb{R}^n$, $\hat{\zeta} \in \mathbb{R}^r$ are such that $g(\hat{x}) = 0$ and the pair $(\hat{x},\hat{\zeta})$ satisfies the second-order sufficient conditions (1c,d).

Then there exists a $\hat{\pi} \geq 0$ such that, for all $\pi \geq \hat{\pi}$, the matrix $L_{\pi xx}(\hat{x},\hat{\zeta})$ is positive-definite.

Proof. For the sake of contradiction, suppose that $\{\pi_i\}_{i=0}^\infty$ is a sequence of positive penalties such that $\pi_i \to \infty$, as $i \to \infty$, and that, for each $i \in \mathbb{N}$, there exists a vector $h_i \in \mathbb{R}^n$ such that $\|h_i\| = 1$ and

$$\langle h_i, L_{\pi_i xx}(\hat{x},\hat{\zeta})h_i\rangle = \langle h_i, L_{xx}(\hat{x},\hat{\zeta})h_i\rangle + \pi_i\|g_x(\hat{x})h_i\|^2 \leq 0. \tag{2a}$$

Since the unit sphere is compact, without loss of generality, we can assume that $h_i \to \hat{h}$, as $i \to \infty$. Hence,

$$\langle \hat{h}, L_{xx}(\hat{x},\hat{\zeta})\hat{h}\rangle + \overline{\lim}\,\pi_i\|g_x(\hat{x})h_i\|^2 \leq 0. \tag{2b}$$

Because $\pi_i \to \infty$, as $i \to \infty$, it follows from (2b) that $g_x(\hat{x})\hat{h} = 0$, i.e., that $\hat{h} \in \mathcal{H}_E(\hat{x})$ and hence that $\langle \hat{h}, L_{xx}(\hat{x},\hat{\zeta})\hat{h}\rangle \leq 0$, a contradiction. □

As a corollary to Proposition 2.8.1, we note that, if $\hat{x} \in \mathbb{R}^n$, $\hat{\zeta} \in \mathbb{R}^r$ are such that $g(\hat{x}) = 0$ and the pair $(\hat{x},\hat{\zeta})$ satisfies the second-order sufficient conditions (1c,d), then there exists a $\hat{\pi} \geq 0$ such that, for all $\pi \geq \hat{\pi}$, $\hat{x}$ also satisfies the second-order sufficient conditions for a local minimizer of the function $L_\pi(\cdot,\hat{\zeta})$. Hence, if we only knew $\hat{\zeta}$ and $\hat{\pi}$, we could solve problem (1a) using a

fast unconstrained optimization algorithm, such as Newton's method. Now we will describe one way of dealing with the fact that neither $\hat{\zeta}$ nor $\hat{\pi}$ are known in advance. Specifically, we will use the formula

$$\zeta(x) \triangleq \arg\min_{\zeta \in \mathbb{R}^r} \|\nabla f^0(x) + g_x(x)^T\zeta\|^2 \tag{3a}$$

to estimate the multiplier $\hat{\zeta}$ and Algorithm Model 2.3.6 as a means of identifying $\hat{\pi}$. Thus, for every $\pi > 0$, we define the functions $f_\pi^0 : \mathbb{R}^n \to \mathbb{R}$ by

$$f_\pi^0(x) \triangleq L(x,\zeta(x)) = f^0(x) + \langle \zeta(x), g(x)\rangle + \tfrac{1}{2}\pi\|g(x)\|^2, \tag{3b}$$

and the associated unconstrained optimization problem

$$\mathbf{P}_\pi^* \qquad \min_{x \in \mathbb{R}^n} f_\pi^0(x). \tag{3c}$$

It is clear from Fig. 2.8.1 (where $g \in \mathbb{R}$) that the function

$$\hat{f}_\pi^0(x) \triangleq f^0(x) + \langle \zeta(\hat{x}), g(x)\rangle + \frac{\pi}{2}\|g(x)\|^2, \tag{3d}$$

whose definition requires knowledge of the multiplier $\zeta(\hat{x})$ at a local minimizer $\hat{x}$, is a modified exact penalty function for **P** (modified to the extent that it fails to satisfy (2.7.2b,c)) with the term $\langle \zeta(\hat{x}), g(x)\rangle$ making it possible for this penalty function to be exact. Our modified exact penalty function $f_\pi^0(\cdot)$ obviously approximates $\hat{f}_\pi^0(\cdot)$ with both functions giving the same value at $\hat{x}$.

We will now establish the relevant properties of $f_\pi^0(\cdot)$. We begin with an auxiliary result.

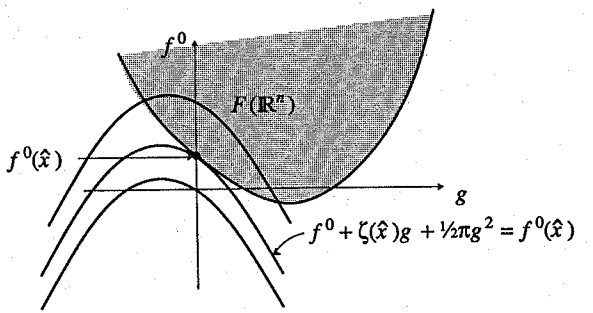


Fig. 2.8.1. A modified exact penalty function.

Lemma 2.8.2. *Suppose that the functions $f^0(\cdot)$ and $g(\cdot)$ in (1a) are continuously differentiable and that the matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$. Then, for every $x \in \mathbb{R}^n$,*

$$g_x(x)\nabla_x L(x, \zeta(x)) = 0. \tag{4}$$

Proof. Let $\phi(\zeta) \triangleq \|\nabla f^0(x) + g_x(x)^T\zeta\|^2$. Then it follows from (3a) that $\nabla\phi(\zeta(x)) = 0$. Computing $\nabla\phi(\zeta(x))$, we obtain (4). □

Lemma 2.8.3. *Suppose that the functions $f^0(\cdot)$ and $g(\cdot)$ are three times continuously differentiable and that the matrix $g_x(x)$ has maximum rank for all $x \in \mathbb{R}^n$. Then the function $\zeta(\cdot)$, defined in (3a), is twice continuously differentiable and its Jacobian is given by*

$$\zeta_x(x) = -[g_x(x)g_x(x)^T]^{-1}[g_x(x)L_{xx}(x, \zeta(x)) + \sum_{j=1}^{r} e_j \nabla_x L(x, \zeta(x))^T g_{xx}^j(x)], \tag{5}$$

where e_j is the jth column of the $r \times r$ identity matrix.

Proof. It follows from (3a) that

$$\zeta(x) = -[g_x(x)g_x(x)^T]^{-1}g_x(x)\nabla f^0(x). \tag{6a}$$

Therefore, it follows from our assumptions and (6a) that $\zeta(\cdot)$ is twice continuously differentiable.

Next, note that $g_x(x) = \sum_{j=1}^{r} e_j g_x^j(x)$. Hence, writing (4) in expanded form, we find that

$$\sum_{j=1}^{r} e_j g_x^j(x)\nabla_x L(x, \zeta(x)) = 0. \tag{6b}$$

Differentiating (6b), we conclude that

$$\begin{aligned} 0 &= \sum_{j=1}^{r} e_j \big(L_x(x, \zeta(x))g_{xx}^j(x) \\ &\quad + g_x^j(x)[L_{xx}(x, \zeta(x)) + L_{\zeta x}((x, \zeta(x))\zeta_x(x)]\big), \\ &= \sum_{j=1}^{r} e_j L_x(x, \zeta(x))g_{xx}^j(x) \\ &\quad + g_x(x)[L_{xx}(x, \zeta(x)) + L_{\zeta x}((x, \zeta(x))\zeta_x(x)], \end{aligned} \tag{6c}$$

where we have followed the notational convention $g_x^j(x) = \nabla g^j(x)^T$ and

$L_x(x, \zeta(x)) = \nabla_x L(x, \zeta(x))^T$. Since $L_{\zeta x}((x, \zeta(x)) = g_x(x)^T$, the desired result follows from (6c). □

The following result is a direct consequence of Lemma 2.8.3 and the chain rule.

Corollary 2.8.4. *Suppose that the functions $f^0(\cdot)$ and $g(\cdot)$ are three times continuously differentiable and that the matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$. Then, for any $\pi \geq 0$, the function $f_\pi^0(\cdot)$ is twice continuously differentiable.* □

The next result establishes a crucial relationship between the stationary points of **P** and those of the $\mathbf{P}_\pi^*$, defined in (1a) and (3c) respectively.

Lemma 2.8.5. *Suppose that the functions $f^0(\cdot)$ and $g(\cdot)$ are continuously differentiable and that the matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$.*

(a) If $\hat{x} \in \mathbb{R}^n$ and $\hat{\zeta} \in \mathbb{R}^r$ are such that $g(\hat{x}) = 0$ and $\nabla_x L(\hat{x}, \hat{\zeta}) = 0$, then $\hat{\zeta} = \zeta(\hat{x})$, and, for any $\pi \geq 0$, $\nabla f_\pi^0(\hat{x}) = 0$.

(b) For every compact subset S of $\mathbb{R}^n$, there exists a $\pi_S < \infty$ such that, for all $\pi \geq \pi_S$, if $\hat{x} \in S$ is such that $\nabla f_\pi^0(\hat{x}) = 0$, then $g(\hat{x}) = 0$ and $\nabla_x L(\hat{x}, \zeta(\hat{x})) = 0$.

Proof. *(a)* Because $\nabla_x L(\hat{x}, \hat{\zeta}) = 0$, we must have that $\hat{\zeta} = \zeta(\hat{x})$, and because $g(\hat{x}) = 0$, $\nabla_\zeta L(\hat{x}, \zeta(\hat{x})) = g(\hat{x}) = 0$. Hence, using the chain rule, we find that for any $\pi \geq 0$,

$$\begin{aligned} 0 &= \nabla f_\pi^0(\hat{x}) \\ &= \nabla_x L(\hat{x}, \zeta(\hat{x})) + \zeta_x(\hat{x})^T\nabla_\zeta L(\hat{x}, \zeta(\hat{x})) + \pi g_x(\hat{x})^T g(\hat{x}). \end{aligned} \tag{7a}$$

(b) Next, let S be a compact subset of $\mathbb{R}^n$. Suppose that $\hat{x} \in S$ is such that $\nabla f_\pi^0(\hat{x}) = 0$ for some $\pi \geq 0$. Then, using (4), we conclude that

$$\begin{aligned} 0 &= g_x(\hat{x})\nabla f_\pi^0(\hat{x}) \\ &= g_x(\hat{x})[\nabla_x L(\hat{x}, \zeta(\hat{x})) + \zeta_x(\hat{x})^T g(\hat{x}) + \pi g_x(\hat{x})^T g(\hat{x})] \\ &= [\pi g_x(\hat{x})g_x(\hat{x})^T + g_x(\hat{x})\zeta_x(\hat{x})^T]g(\hat{x}). \end{aligned} \tag{7b}$$

Since by assumption, $g_x(\hat{x})$ has maximum row rank, the matrix $g_x(\hat{x})g_x(\hat{x})^T$ is positive definite. Hence, because S is compact, and all the matrices in (7b) are continuous, there must exist a $\pi_S < \infty$ such that the matrix $[\pi g_x(\hat{x})g_x(\hat{x})^T + g_x(\hat{x})\zeta_x(\hat{x})^T]$ is positive-definite, and, therefore, nonsingular,

for all $\pi \geq \pi_S$ and $\hat{x} \in S$. Hence for all $\pi \geq \pi_S$, (7b) implies that $g(\hat{x}) = 0$. It now follows from the fact that $\nabla f_\pi^0(\hat{x}) = 0$ and $\nabla_\zeta L(\hat{x}, \zeta(\hat{x})) = g(\hat{x}) = 0$, together with (7a), that $\nabla_x L(\hat{x}, \hat{\zeta}) = 0$, with $\hat{\zeta} = \zeta(\hat{x})$, which completes our proof. □

Corollary 2.8.6. *Suppose that the functions $f^0(\cdot)$ and $g(\cdot)$ are three times continuously differentiable and that the matrix $g_x(x)$ has maximum rank for all $x \in \mathbb{R}^n$. Then, for every compact subset S of $\mathbb{R}^n$, there exists a $\pi_S < \infty$ such that for all $\pi \geq \pi_S$, if $\hat{x} \in S$ is a local minimizer of $\mathbf{P}_\pi^*$, then $\hat{x}$ is a local minimizer for $\mathbf{P}$. Furthermore, if for some $m > 0$, $\hat{x}$ satisfies the second-order sufficient condition $\langle h, f_{\pi xx}^0(\hat{x})h \rangle \geq m\|h\|^2$, for all $h \in \mathbb{R}^n$, then $\hat{x}$ together with $\hat{\zeta} = \zeta(\hat{x})$ satisfy the second or sufficient condition (1c,d) for $\mathbf{P}$.*

Proof. Let $S \subset \mathbb{R}^n$ be a compact subset, let $\pi_S < \infty$ be as postulated in Lemma 2.8.5, and suppose that $\pi \geq \pi_S$. If $\hat{x} \in S$ is a local minimizer of $\mathbf{P}_\pi^*$, then $\nabla f_\pi^0(\hat{x}) = 0$, and hence it follows from Lemma 2.8.5*(b)* that $g(\hat{x}) = 0$. Hence $\hat{x}$ is also a local minimizer for the problem $\min\{f_\pi^0(x) \mid g(x) = 0\}$. Since $f^0(x) = f_\pi^0(x)$ for all x such that $g(x) = 0$, the first result follows.

Next, by direct calculation, it follows that, for any $x \in \mathbb{R}^n$,

$$f_{\pi xx}^0(x) = G_\pi(x) + \sum_{j=1}^{r} g^j(x)[\zeta_{xx}^j(x) + g_{xx}^j(x)], \tag{8a}$$

where

$$G_\pi(x) \triangleq L_{xx}(x, \zeta(x)) + g_x(x)^T\zeta_x(x) + \zeta_x(x)^T g_x(x) + \pi g_x(x)^T g_x(x). \tag{8b}$$

Since $g(\hat{x}) = 0$, it follows that $f_{\pi xx}^0(\hat{x}) = G_\pi(\hat{x})$. Since, by assumption, $\langle h, f_{\pi xx}^0(\hat{x})h \rangle \geq m\|h\|^2$, for all $h \in \mathbb{R}^n$, it follows from (8b) that $\langle h, L_{xx}(x, \zeta(x))h \rangle \geq m\|h\|^2$ for all $h \in \mathbb{R}^n$ such that $g_x(\hat{x})h = 0$, which completes our proof. □

Exercise 2.8.7. Consider the problem $\mathbf{P}$. Suppose that

(i) the functions $f^0(\cdot)$ and $g(\cdot)$ are twice continuously differentiable and that the matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$, and

(ii) $\hat{x} \in \mathbb{R}^n$, and $\zeta(\hat{x})$ are such that $g(\hat{x}) = 0$ and the pair $(\hat{x}, \zeta(\hat{x}))$ satisfies the second-order sufficient conditions (1c,d).

Show that there exists a $\hat{\pi} \geq 0$ such that, for all $\pi \geq \hat{\pi}$, the matrix $G_\pi(\hat{x})$ is positive-definite. *Hint*: Mimic the proof of Proposition 2.8.1. □

Next, Algorithm Model 2.3.6 requires a family of test functions $t_j : \mathbb{R}^n \to \mathbb{R}$, $j \in \mathbb{N}$, that satisfy Assumption 2.3.5. The following theorem presents such test functions.

Theorem 2.8.8. *Suppose that the functions $f^0(\cdot)$ and $g(\cdot)$ are twice continuously differentiable and that the matrix $g_x(x)$ has maximum rank for all $x \in \mathbb{R}^n$. For any $\pi \geq 0$, let $t_\pi : \mathbb{R}^n \to \mathbb{R}$ be defined by*

$$t_\pi(x) \triangleq -\langle g_x(x)^T g(x), \nabla f_\pi^0(x) \rangle + \|g(x)\|^2. \tag{9}$$

Then,

(a) $t_\pi(\cdot)$ is continuous,

(b) if for some $\hat{x} \in \mathbb{R}^n$, $t_\pi(\hat{x}) \leq 0$ and $\nabla f_\pi^0(\hat{x}) = 0$, then $g(\hat{x}) = 0$ and $\nabla f^0(\hat{x}) + g_x(\hat{x})^T\zeta(\hat{x}) = 0$, i.e., $\hat{x}$ satisfies the first-order necessary condition for $\mathbf{P}$, and

(c) for every $\hat{x} \in \mathbb{R}^n$, there exists a $\hat{\pi} \in (0, \infty)$, and a $\hat{\rho} > 0$ such that, for all $x \in B(\hat{x}, \hat{\rho})$ and $\pi \geq \hat{\pi}$, $t_\pi(x) \leq 0$.

Proof.

(a) Since $\nabla f^0(\cdot)$, $\zeta(\cdot)$, $g(\cdot)$, and $g_x(\cdot)$ are all continuous, it follows that for any $\pi \geq 0$, $t_\pi(\cdot)$ is continuous.

(b) Suppose that $\hat{x} \in \mathbb{R}^n$ is such that $t_\pi(\hat{x}) \leq 0$ and $\nabla f_\pi^0(\hat{x}) = 0$. Then it follows, by inspection, from (9) that $g(\hat{x}) = 0$. We now conclude from the formula for $\nabla f_\pi^0(\hat{x})$ in (7a) and the fact that $\nabla_\zeta L(\hat{x}, \zeta(\hat{x})) = g(\hat{x})$, that $\nabla_x L(\hat{x}, \zeta(\hat{x})) = 0$, i.e., that $\hat{x}$ satisfies the first-order necessary condition for $\mathbf{P}$.

(c) Using (4) and the expression for $\nabla f_\pi^0(x)$ in (7a), we conclude that, for any $\pi \geq 0$ and $x \in \mathbb{R}^n$,

$$\begin{aligned} t_\pi(x) &= -\langle g(x), g_x(x)\nabla f_\pi^0(x) \rangle + \|g(x)\|^2 \\ &= -\langle g(x), [\pi g_x(x)g_x(x)^T - I + g_x(x)\zeta_x(x)^T]g(x) \rangle. \end{aligned} \tag{10}$$

Let $\hat{x} \in \mathbb{R}^n$ be given. Since $g_x(\hat{x})g_x(\hat{x})^T$ is a positive-definite matrix, there exists a $\pi^* \in [0, \infty)$ such that the matrix $[\pi g_x(\hat{x})g_x(\hat{x})^T - I + g_x(\hat{x})\zeta_x(\hat{x})^T]$ is positive-definite for all $\pi \geq \pi^*$. Let $\hat{\pi} = 2\pi^*$. Then, by continuity, there exists a $\hat{\rho} > 0$ such that, for all $x \in B(\hat{x}, \hat{\rho})$ and $\pi \geq \hat{\pi}$, the matrix $[\pi g_x(x)g_x(x)^T - I + g_x(x)\zeta_x(x)^T]$ is positive-definite, and hence, for all $x \in B(\hat{x}, \hat{\rho})$ and $\pi \geq \hat{\pi}$, $t_\pi(x) \leq 0$, which completes our proof. □

Since first-order derivatives of the modified exact penalty functions $f_\pi^0(\cdot)$ involve second-order derivatives of $f^0(\cdot)$ and of $g(\cdot)$ and second-order derivatives of $f_\pi^0(\cdot)$ involve third order derivatives of $f^0(\cdot)$ and of $g(\cdot)$, it is desirable to use a superlinearly converging algorithm that does not require second-order derivatives of $f_\pi^0(\cdot)$, in the minimization of the functions $f_\pi^0(\cdot)$. That leaves us with at least three choices: the BFGS variable metric method, Algorithm 1.6.16,

the global Newton's method, Algorithm 1.4.15 with the Hessian matrix $f^0_{\pi xx}(\cdot)$ replaced by the matrix $G_\pi(\cdot)$, and a trust region method using the matrix $G_\pi(\cdot)$ in its model for defining the iteration maps in Algorithm Model 2.3.6. We are now ready to present an algorithm of the form of Algorithm Model 2.3.6 for solving the problem **P** in (1a).

Thus, let $\{\pi_j\}_{j=0}^\infty$ be a sequence of penalties such that $\pi_{j+1} > \pi_j > 0$, for all $j \in \mathbb{N}$, and such that $\pi_j \to \infty$, as $j \to \infty$. To identify the algorithm below with Algorithm Model 2.3.6, we define the problems $\mathbf{P}_j$ in (2.3.7a), $j \in \mathbb{N}$, by setting $\mathbf{P}_j = \mathbf{P}^*_{\pi_j}$. Next, we define the test functions $t_j(\cdot)$ in Algorithm Model 2.3.6 by $t_j(x) = t_{\pi_j}(x)$. Finally, we let the algorithm functions $A_j(\cdot)$ in Algorithm Model 2.3.6 be defined either by one iteration of Algorithm 1.6.16 or by one iteration of Algorithm 1.4.15 with the Hessian matrix $f^0_{\pi xx}(\cdot)$ replaced by the matrix $G_\pi(\cdot)$, or by one iteration of a trust region algorithm of the form of Algorithm Model 1.2.26, using the matrix $G_\pi(\cdot)$ in its model applied to the problem $\mathbf{P}_j$. In particular, when Algorithm 1.4.15 is used to define the algorithm functions $A_j(\cdot)$, Algorithm Model 2.3.6 assumes the following specific form:

Mukai-Polak Algorithm 2.8.9.

Parameters. $\alpha \in (0, \frac{1}{2})$, $\beta \in (0,1)$, $\{\pi_j\}_{j=0}^\infty$, and machine precision parameters $\sigma_* \ll 1$, $\kappa \gg 1$.

Data: $x_0 \in \mathbb{R}^n$.

Step 0: Set $i = 0$, $j = 0$.

Step 1: If $t_{\pi_j}(x_i) \le 0$, go to Step 2.

Else, go to Step 4.

Step 2: Call Algorithm 1.4.15:

Newton Step 1. Compute the matrix $G_{\pi_j}(x_i)$ and its largest and smallest eigenvalues $\sigma_{\max}(x_i)$ and $\sigma_{\min}(x_i)$.

If $\sigma_{\min}(x_i) \ge \sigma_*$ and $\sigma_{\max}(x_i)/\sigma_{\min}(x_i) \le \kappa$, set

$$h_i \triangleq -G_{\pi_j}(x_i)^{-1}\nabla f_{\pi_j}(x_i)\,. \tag{11a}$$

Else, set

$$h_i \triangleq -\nabla f^0_{\pi_j}(x_i)\,. \tag{11b}$$

Newton Step 2. Compute the Armijo step size

$$\lambda_i = \max_{k \in \mathbb{N}} \{\beta^k \mid f^0_{\pi_j}(x_i + \beta^k h_i) - f^0_{\pi_j}(x_i) \le \alpha\beta^k \langle h_i, \nabla f^0_{\pi_j}(x_i)\rangle\}\,. \tag{11c}$$

Newton Step 3. Set

$$x_{i+1} = x_i + \lambda_i h_i\,. \tag{11d}$$

Step 3: If $\nabla f^0_{\pi_j}(x_{i+1}) = 0$ and $t_{\pi_j}(x_{i+1}) \le 0$, stop.

Else, replace i by $i+1$, and go to Step 1.

Step 4: Set $x_j^* = x_i$, replace j by $j+1$, and go to Step 1.

The convergence properties of Algorithm 2.8.8 are exactly as established in Theorem 2.3.7:

Theorem 2.8.10. *Suppose that the functions $f^0(\cdot)$ and $g(\cdot)$ are three times continuously differentiable and that the matrix $g_x(x)$ has maximum rank for all $x \in \mathbb{R}^n$.*

(a) If Algorithm 2.8.9 constructs a finite sequence $\{x_j^*\}_{j=0}^{j^*}$ and the sequence $\{x_i\}_{i=0}^\infty$ is infinite, then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^\infty$ satisfies $g(\hat{x}) = 0$ and the first-order necessary condition $\nabla_x L(\hat{x}, \zeta(\hat{x})) = 0$ stated in Corollary 2.2.20(a) for **P**.

(b) If Algorithm 2.8.9 constructs a finite sequence $\{x_j^*\}_{j=0}^{j^*}$ and the sequence $\{x_i\}$ is also finite, then its last element, say, x_{i^*}, satisfies $g(x_{i^*}) = 0$ and the first-order necessary condition $\nabla_x L(x_{i^*}, \zeta(x_{i^*})) = 0$ stated in Corollary 2.2.20(a) for **P**.

(c) If Algorithm 2.8.9 constructs an infinite sequence $\{x_j^*\}_{j=0}^\infty$, then $\{x_j^*\}_{j=0}^\infty$ has no accumulation points. □

Note that it follows from Exercise 2.8.7 that, if $\hat{x}$ satisfies second-order sufficient conditions for (1a), then there exists a $\hat{\pi} \in (0, \infty)$ such that, for all $\pi \ge \hat{\pi}$, $G_\pi(\hat{x})$ is positive-definite. Hence it should be clear from the analysis of rate of convergence of the local Newton method, that, if Algorithm 2.8.9 constructs a finite sequence $\{x_j^*\}_{j=0}^{j^*}$, and the sequence $\{x_i\}_{i=0}^\infty$ is infinite and has an accumulation point $\hat{x}$ which satisfies second-order sufficient conditions for (1a), and π has been increased enough to ensure that the Hessian matrix $f^0_{\pi xx}(\hat{x}) = G_\pi(\hat{x})$ is positive-definite, then Algorithm 2.8.9 converges to this point superlinearly.

2.8.2 Problems with Mixed Constraints

Now consider the problem with mixed constraints

$$\mathbf{P} \qquad \min \{ f^0(x) \mid f^j(x) \le 0,\ j \in \mathbf{q},\ g(x) = 0 \}, \tag{12a}$$

where $f^j : \mathbb{R}^n \to \mathbb{R}$, $j \in \{0, 1, \ldots, q\}$, and $g : \mathbb{R}^n \to \mathbb{R}^r$, $r < n$, are twice continuously differentiable functions. We can convert this problem into an equality-constrained problem by introducing "slack variables" $y^j \in \mathbb{R}$, $j \in \mathbf{q}$, as follows:

$$\bar{\mathbf{P}} \qquad \min_{\substack{x \in \mathbb{R}^n \\ y \in \mathbb{R}^q}} \{ f^0(x) \mid f^j(x) + (y^j)^2 = 0,\ j \in \mathbf{q},\ g(x) = 0 \}. \tag{12b}$$

Proposition 2.8.11. *A vector $\hat{x} \in \mathbb{R}^n$ is a local minimizer of $\mathbf{P}$ in (12a) if and only if $\hat{x}$ and $\hat{y} \in \mathbb{R}^q$, defined by $\hat{y}^j = -f^j(\hat{x})^{1/2}$, $j \in \mathbf{q}$, is a local minimizer of $\bar{\mathbf{P}}$ in (12b).* □

Exercise 2.8.12. Prove Proposition 2.8.11. □

Next we turn to first-order optimality conditions. Suppose that $(\hat{x}, \hat{y})$ is a local minimizer for $\bar{\mathbf{P}}$. We can apply the first-order condition in Corollary 2.2.20*(a)* to $\bar{\mathbf{P}}$, only if the vectors

$$\begin{bmatrix} \nabla g^k(\hat{x}) \\ 0 \end{bmatrix}, \begin{bmatrix} \nabla f^j(\hat{x}) \\ 2\hat{y}^j e_j \end{bmatrix}, \quad k \in \mathbf{r},\ j \in \mathbf{q}, \tag{12c}$$

are linearly independent, where e_j is the jth column of the $q \times q$ identity matrix.

Exercise 2.8.13. Show that the vectors in (12c) are linearly independent, if and only if the vectors $\nabla g^k(\hat{x})$, $k \in \mathbf{r}$, together with the vectors $\nabla f^j(\hat{x})$, $j \in \mathbf{q}_A(\hat{x})$, are linearly independent, where $\mathbf{q}_A(\hat{x})$ was defined in (2.2.12d). □

Thus, in view of Corollary 2.2.20*(a)* and Exercise 2.8.13, if $(\hat{x}, \hat{y})$ is a local minimizer for $\bar{\mathbf{P}}$ and the vectors $\nabla g^k(\hat{x})$, $k \in \mathbf{r}$, together with the vectors $\nabla f^j(\hat{x})$, $j \in \mathbf{q}_A(\hat{x})$, are linearly independent, then there exist multipliers $\hat{\eta} \in \mathbb{R}^q$ and $\hat{\xi} \in \mathbb{R}^r$ such that

$$\nabla f^0(\hat{x}) + \sum_{j=1}^{q} \hat{\eta}^j \nabla f^j(\hat{x}) + \sum_{k=1}^{r} \hat{\xi}^j \nabla g^j(\hat{x}) = 0, \tag{12d}$$

$$2\hat{\eta}^j \hat{y}^j = 0,\ j \in \mathbf{q}. \tag{12e}$$

Referring to Corollary 2.2.20, which states first-order optimality conditions for the problem $\mathbf{P}$ in (12a) we note that, upon division by $\sigma \triangleq (1 + \sum_{j=1}^{q} \hat{\eta}^j)$, (12d) has the required form (2.2.23d) (with $\hat{\mu}^0 = 1/\sigma$, $\hat{\mu}^j = \hat{\eta}^j/\sigma$, $j \in \mathbf{q}$, and $\hat{\zeta}^l = \hat{\xi}^l/\sigma$, $l \in \mathbf{r}$) and that (12e) implies that (2.2.23e) holds. However, there is nothing above to require that $\hat{\eta} \ge 0$ hold. Thus we see that the problem $\bar{\mathbf{P}}$ may have first-order stationary points that are not stationary for the problem $\mathbf{P}$ in (12a). However, this is no longer possible when second-order conditions are satisfied.

Exercise 2.8.14. Suppose that $f^j(\cdot)$, $j \in \{0, 1, \ldots, q\}$, and $g(\cdot)$ are twice continuously differentiable functions, that $(\hat{x}, \hat{y})$ is a local minimizer for $\bar{\mathbf{P}}$, and that the vectors $\nabla g^k(\hat{x})$, $k \in \mathbf{r}$, together with the vectors $\nabla f^j(\hat{x})$, $j \in \mathbf{q}_A(\hat{x})$ are linearly independent.

(a) Suppose that $(\hat{x}, \hat{y})$, together with the multipliers $\hat{\eta} \in \mathbb{R}^q$ and $\hat{\xi} \in \mathbb{R}^r$, satisfies the second-order necessary optimality conditions for $\bar{\mathbf{P}}$ in (12b) stated in Corollary 2.2.27a. Let $\sigma \triangleq 1 + \sum_{j=1}^{q} \hat{\eta}^j$, and let $\hat{\mu} \in \Sigma_q^0$ be defined by $\hat{\mu}^0 \triangleq 1/\sigma$, $\hat{\mu}^j \triangleq \hat{\eta}^j/\sigma$, $j \in \mathbf{q}$, $\hat{\zeta}^l = \hat{\xi}^l/\sigma$, $l \in \mathbf{r}$. Show that $\hat{x}$ together with the multipliers $\hat{\mu}$ and $\hat{\zeta} \in \mathbb{R}^r$ satisfies the second-order necessary optimality conditions for $\mathbf{P}$ in (12a) stated in Corollary 2.2.27.

(b) Suppose that $(\hat{x}, \hat{y})$, together with the multipliers $\hat{\eta} \in \mathbb{R}^q$ and $\hat{\xi} \in \mathbb{R}^r$, satisfies the second-order sufficient optimality conditions for $\bar{\mathbf{P}}$ in (12b), stated in Corollary 2.2.31. Let $\hat{\mu} \in \Sigma_q^0$ and $\hat{\zeta} \in \mathbb{R}^r$ be defined as above. Show that $\hat{x}$ together with the multipliers $\hat{\mu}$ and $\hat{\zeta} \in \mathbb{R}^r$ satisfies the second-order sufficient optimality conditions for $\mathbf{P}$ in (12a) stated in Corollary 2.2.30. □

In view of the above discussion and the fact that optimization algorithms find only stationary points, it appears that the application of augmented Lagrangian methods to problem $\mathbf{P}$ in (12a) makes sense primarily if the following assumption is satisfied:

Assumption 2.8.15. *We will assume that*

(i) in (12a), the functions $f^j(\cdot)$, $j = 0, 1, \ldots, q$, and $g(\cdot)$ are three times continuously differentiable,

(ii) for any $x \in \mathbb{R}^n$, the vectors $\nabla g^k(x)$, $k \in \mathbf{r}$, together with the vectors $\nabla f^j(x)$, $j \in \mathbf{q}_A(x)$ (defined in (2.2.12d)), are linearly independent, and

(iii) if $\hat{x}$ is a stationary point for $\mathbf{P}$ in (12a), so that $f^j(\hat{x}) \le 0$, for all $j \in \mathbf{q}$, $g(\hat{x}) = 0$, and there exist a $\hat{\mu} \in \Sigma_q^0$ and a $\hat{\zeta} \in \mathbb{R}^r$ such that (2.2.23d,e) are satisfied, then both the strict complementary slackness condition *$\hat{\mu}^j > 0$, for all $j \in \mathbf{q}_A(\hat{x})$, and the second-order sufficient conditions in Corollary 2.2.30 are satisfied.* □

Next we will obtain a simplified formula for the augmented Lagrangian for $\bar{\mathbf{P}}$. First, for any $\pi \geq 0$, $x \in \mathbb{R}^n$, $\eta \in \mathbb{R}^q$ and $\xi \in \mathbb{R}^r$, the augmented Lagrangian of $\bar{\mathbf{P}}$ is of the form

$$\bar{L}_\pi(x, y, \eta, \xi) \triangleq f^0(x) + \sum_{j=1}^{q} \eta^j [f^j(x) + (y^j)^2] + \langle \xi, g(x) \rangle + \frac{\pi}{2} \Big[\sum_{j=1}^{q} [f^j(x) + (y^j)^2]^2 + \|g(x)\|^2 \Big]. \tag{13a}$$

Now, assuming that we have a pair of multipliers $\hat{\eta} \in \mathbb{R}^q$ and $\hat{\xi} \in \mathbb{R}^r$, corresponding to a local minimizer $\hat{x}$ for $\bar{\mathbf{P}}$, and a sufficiently high value of π, we can compute this local minimizer of $\bar{\mathbf{P}}$ in (12b) by solving the problem

$$\min_{\substack{x \in \mathbb{R}^n \\ y \in \mathbb{R}^q}} \bar{L}_\pi(x, y, \hat{\eta}, \hat{\xi}) = \min_{x \in \mathbb{R}^n} \min_{y \in \mathbb{R}^q} \bar{L}_\pi(x, y, \hat{\eta}, \hat{\xi}). \tag{13b}$$

Now, the minimization with respect to y can be carried out explicitly. Let $u \in \mathbb{R}^q$ be defined by $u^j = (y^j)^2$, $j \in \mathbf{q}$. Then the inner minimization problem becomes

$$\min_{u \in \mathbb{R}^q,\, u \geq 0} \sum_{j=1}^{q} \hat{\eta}^j [f^j(x) + u^j] + \langle \hat{\xi}, g(x) \rangle + \frac{\pi}{2} \sum_{j=1}^{q} [f^j(x) + u^j]^2. \tag{13c}$$

Since there is no interaction in (13c) among the u^j, (13c) can be seen as consisting of q constrained minimizations of quadratic functions of a single variable:

$$\min_{u^j \in \mathbb{R},\, u^j \geq 0} \hat{\eta}^j [f^j(x) + u^j] + \frac{\pi}{2} [f^j(x) + u^j]^2, \quad j \in \mathbf{q}. \tag{13d}$$

Because the quadratics in (13d) are convex, the unconstrained global minimizer of (13d), u_*^j, is obtained by setting the derivative of the cost function in (13d) equal to zero, and is given by

$$u_*^j = -[\hat{\eta}^j / \pi + f^j(x)], \quad j \in \mathbf{q}. \tag{13e}$$

Hence $\hat{u}^j$, $j \in \mathbf{q}$, the solution to the constrained minimization problem (13d), is given by

$$\hat{u}^j = \max \{0, -[\hat{\eta}^j / \pi + f^j(x)]\}, \quad j \in \mathbf{q}. \tag{13f}$$

Hence, adding $f^j(x)$ to both sides of (13f), we find that

$$f^j(x) + \hat{u}^j = \max \{f^j(x), -\hat{\eta}^j / \pi\}, \quad j \in \mathbf{q}. \tag{13g}$$

Let

$$L_\pi(x, \hat{\eta}, \hat{\xi}) \triangleq \min_{y \in \mathbb{R}^q} \bar{L}_\pi(x, y, \hat{\eta}, \hat{\xi}). \tag{13h}$$

Then it follows from (13h) and (13a) that

$$L_\pi(x, \hat{\eta}, \hat{\xi}) = f^0(x) + \sum_{j=1}^{q} \hat{\eta}^j \max \{f^j(x), -\hat{\eta}^j / \pi\} + \langle \hat{\xi}, g(x) \rangle + \frac{\pi}{2} \Big[\sum_{j=1}^{q} [\max \{f^j(x), -\hat{\eta}^j / \pi\}]^2 + \|g(x)\|^2 \Big]. \tag{13i}$$

Let $\alpha_j \triangleq \max \{f^j(x), -\hat{\eta}^j / \pi\}$, $j \in \mathbf{q}$. Then we see that the the terms $\max \{f^j(x), -\hat{\eta}^j / \pi\}$ in (13i) combine to form the expressions

$$\tfrac{1}{2}\pi[\alpha_j^2 + 2(\hat{\eta}^j / \pi)\alpha_j] = \tfrac{1}{2}\pi[(\alpha_j + \hat{\eta}^j / \pi)^2 - (\hat{\eta}^j / \pi)^2]. \tag{13j}$$

Since $\alpha_j + \hat{\eta}^j / \pi = (f^j(x) + \hat{\eta}^j / \pi)_+$ (recall that $a_+ \triangleq \max \{0, a\}$), we now find that (13i) becomes

$$L_\pi(x, \hat{\eta}, \hat{\xi}) = f^0(x) + \langle \hat{\xi}, g(x) \rangle + \frac{\pi}{2} \|g(x)\|^2 + \frac{1}{2\pi} \sum_{j=1}^{q} \big[[\pi f^j(x) + \hat{\eta}^j]_+^2 - (\hat{\eta}^j)^2 \big], \tag{13k}$$

which is our final form for the augmented Lagrangian for problem $\mathbf{P}$ in (12a).

Next we introduce a formula for estimating the multipliers $\hat{\eta}$ and $\hat{\xi}$. For any $x \in \mathbb{R}^n$, let $F(x) \triangleq \operatorname{diag}([f^1(x)]^2, \ldots, [f^1(x)]^2)$, be a $q \times q$ matrix, and let $\eta : \mathbb{R}^n \to \mathbb{R}^q$ and $\xi : \mathbb{R}^n \to \mathbb{R}^r$ be defined by

$$(\eta(x), \xi(x)) \triangleq \arg \min_{\substack{\eta \in \mathbb{R}^q \\ \xi \in \mathbb{R}^r}} \big\{ \|\nabla f^0(x) + f_x(x)^T \eta + g_x(x)^T \xi\|^2 + \langle \eta, F(x)\eta \rangle \big\}. \tag{13l}$$

Proposition 2.8.16. *Suppose that Assumption 2.8.15(i,ii) is satisfied. Then the functions $\eta(\cdot)$ and $\xi(\cdot)$ are well defined and twice continuously differentiable. Furthermore, if a vector $\hat{x} \in \mathbb{R}^n$ such that $f^j(\hat{x}) \leq 0$, $j \in \mathbf{q}$, and $g(\hat{x}) = 0$ together with multipliers $\hat{\mu} \in \Sigma_q^0$ and $\hat{\zeta} \in \mathbb{R}^r$ satisfies the necessary first-order conditions (2.2.23d,e) for $\mathbf{P}$ in (12a), then $\hat{\mu}^j / \hat{\mu}^0 = \eta(\hat{x})^j$ for all $j \in \mathbf{q}$, and $\hat{\zeta} = \zeta(\hat{x}) / \hat{\mu}^0$.*

Proof. To prove that the functions $\eta(\cdot)$ and $\xi(\cdot)$ are well defined, we need to show only that the quadratic function being minimized in (13l) is positive-definite. Clearly, this function is positive-semidefinite. Let $z = (\eta, \xi)$. Then, for some $(q+r) \times (q+r)$ matrix Q and $d \in \mathbb{R}^{q+r}$, the quadratic function in (13l)

can be rewritten as follows:

$$\|\nabla f^0(x) + f_x(x)^T\eta + g_x(x)^T\xi\|^2 + \langle \eta, F(x)\eta\rangle = \langle z, Qz\rangle + \langle d, z\rangle. \quad (14a)$$

Since

$$\langle z, Qz\rangle = \|f_x(x)^T\eta + g_x(x)^T\xi\|^2 + \langle \eta, F(x)\eta\rangle, \quad (14b)$$

we see that the quadratic function in (13l) is positive-definite if and only if

$$\|f_x(x)^T\eta + g_x(x)^T\xi\|^2 + \langle \eta, F(x)\eta\rangle = 0 \quad (14c)$$

implies that $\xi = 0$ and $\eta = 0$. Now, when (14c) holds, we must have that $\eta^j = 0$, for all $j \in \mathbf{q}$ such that $f^j(x) \neq 0$. Hence (14c) implies that

$$\sum_{j \in \mathbf{q}_A(x)} \eta^j \nabla f^j(x) + g_x(x)^T\xi = 0. \quad (14d)$$

It now follows from the linear independence assumption, Assumption 2.8.15*(ii)*, that (14d) and hence also (14c) hold if and only if $\xi = 0$ and $\eta = 0$, which shows that $\eta(\cdot)$ and $\xi(\cdot)$ are well defined.

It follows from first-order optimality conditions that $\eta(\cdot)$ and $\xi(\cdot)$ must satisfy the following relations, obtained by differentiating the cost function in (13l):

$$\frac{\partial}{\partial \eta} \Rightarrow f_x(x)(\nabla f^0(x) + f_x(x)^T\eta + g_x(x)^T\xi) + F(x)\eta = 0, \quad (14e)$$

and

$$\frac{\partial}{\partial \xi} \Rightarrow g_x(x)(\nabla f^0(x) + f_x(x)^T\eta + g_x(x)^T\xi) = 0. \quad (14f)$$

If we set $z(\cdot) \triangleq (\eta(\cdot), \xi(\cdot))$, then (14e,f) can be expressed in compact form as $A(x)z(x) = b(x)$ with $A(x)$ a $(q+r)\times(q+r)$ matrix and $b(x)$ a $(q+r)$ vector both determined by (14e,f). Since we know that $z(x)$ is uniquely defined by (14e,f), the matrix $A(x)$ must be nonsingular for all $x \in \mathbb{R}^n$, and hence $z(x) = A(x)^{-1}b(x)$. Since both the matrix-valued function $A(\cdot)$ and the vector-valued function $b(\cdot)$ are twice continuously differentiable, it follows that $z(\cdot)$ is twice continuously differentiable.

Now suppose that $\hat{x}$ together with multipliers $\hat{\mu} \in \Sigma_q^0$ and $\hat{\zeta} \in \mathbb{R}^r$ satisfies the first order conditions (2.2.23d,e) for $\mathbf{P}$ in (12a). Let $\hat{\eta} \in \mathbb{R}^q$ be defined by $\hat{\eta}^j \triangleq \hat{\mu}^j/\hat{\mu}^0$, $j \in \mathbf{q}$, and $\hat{\xi} \triangleq \hat{\zeta}/\hat{\mu}^0$. When substituted into the minimand in (13l), $\hat{\eta}$ and $\hat{\xi}$ give the value of zero and hence must be a solution of (13l). Since this solution is unique, the desired result follows. □

We now define the modified exact penalty functions for $\mathbf{P}$ in (12a) by replacing $\hat{\eta}$ and $\hat{\xi}$ in $L_\pi(x, \hat{\eta}, \hat{\xi})$, defined by (13k), by $\eta(x)$ and $\xi(x)$, i.e., for $\pi > 0$, we define the functions $f_\pi^0 : \mathbb{R}^n \to \mathbb{R}$ by

$$f_\pi^0(x) \triangleq f^0(x) + \langle \xi(x), g(x)\rangle + \tfrac{1}{2}\pi\|g(x)\|^2 + \frac{1}{2\pi}\sum_{j=1}^{q}[(\pi f^j(x) + \eta^j(x))_+^2 - (\eta^j(x))^2]. \quad (15a)$$

If we make use of the fact that $\nabla\sum_{j=1}^q [\mu^j(x)]^2 = 2\mu_x(x)^T\mu(x)$ and that $\nabla(\sum_{j=1}^q [\pi f^j(x) + \eta^j(x), 0]_+^2) = 2[\pi f_x(x)^T + \eta_x(x)^T][\pi f(x) + \eta(x)]_+$, then we find that, for any $x \in \mathbb{R}^n$, the gradient of $f_\pi^0(\cdot)$ is given by

$$\nabla f_\pi^0(x) = \nabla f^0(x) + g_x(x)^T[\xi(x) + \pi g(x)] + \xi_x(x)^T g(x) + f_x(x)^T\eta(x) + \pi[f_x(x) + \frac{1}{\pi}\eta_x(x)]^T a(x, \pi), \quad (15b)$$

where

$$a(x, \pi) \triangleq [f(x) + \frac{1}{\pi}\eta(x)]_+ - \frac{1}{\pi}\eta(x), \quad (15c)$$

with $f(x) \triangleq (f^1(x), \ldots, f^q(x))^T$ and, for any vector $v \in \mathbb{R}^q$, v_+ denoting a vector with components v_+^j, $j \in \mathbf{q}$.

Exercise 2.8.17. Suppose that Assumption 2.8.15 is satisfied. Prove the following results:

(a) Given $x \in \mathbb{R}^n$, if there exists a $\pi \in (0, \infty)$ such that $a(x, \pi) = 0$, then $f^j(x) \leq 0$ and $\eta^j(x) \geq 0$, for all $j \in \mathbf{q}$, and $\langle \eta(x), f(x)\rangle = 0$. Furthermore, for any $x \in \mathbb{R}^n$ such that $f^j(x) \leq 0$ and $\eta^j(x) \geq 0$ for all $j \in \mathbf{q}$, and $\langle \eta(x), f(x)\rangle = 0$, $a(x, \pi) = 0$ for all $\pi > 0$.

(b) Given $x \in \mathbb{R}^n$, if there exists a $\pi \in (0, \infty)$ such that $a(x, \pi) = 0$, then $a(x, \pi') = 0$ for all $\pi' \in (0, \infty)$.

(c) A vector $\hat{x} \in \mathbb{R}^n$ such that $f^j(\hat{x}) \leq 0$, $j \in \mathbf{q}$, and $g(\hat{x}) = 0$, together with multipliers $\hat{\mu} \in \Sigma_q^0$ (with $\hat{\mu}^0 > 0$) and $\hat{\zeta} \in \mathbb{R}^r$, satisfies the first-order necessary conditions (2.2.23d,e) for $\mathbf{P}$ in (12a), if and only if $\hat{\mu}^j/\hat{\mu}^0 = \eta^j(\hat{x})$, $j \in \mathbf{q}$, $\hat{\zeta}/\hat{\mu}^0 = \xi(\hat{x})$, and, for any $\pi \in (0, \infty)$, $\nabla f_\pi^0(\hat{x}) = 0$, $a(\hat{x}, \pi) = 0$ and $g(\hat{x}) = 0$, where $f_\pi^0(\cdot)$ is defined by (15a). □

Our next result is of the same nature as Lemma 2.8.5*(b)*.

Lemma 2.8.18. *Suppose that Assumption 2.8.15 is satisfied. Then, for every compact subset $S \subset \mathbb{R}^n$ that does not contain a point $\hat{x}$ satisfying first-order necessary conditions for $\mathbf{P}$ in (12a) stated in Corollary 2.2.20(b) (i.e., $f^j(\hat{x}) \leq 0$, $j \in \mathbf{q}$, $g(\hat{x}) = 0$, and (2.2.23d,e) hold for some $\hat{\mu} \in \Sigma_q^0$ and $\hat{\zeta} \in \mathbb{R}^r$), there exists a $\pi_S \in (0, \infty)$ and a $\delta > 0$ such that*

$$\|\nabla f_\pi^0(x)\| \geq \delta, \quad \forall\, \pi \geq \pi_S, \quad \forall\, x \in S. \tag{16}$$

Proof. For the sake of contradiction, suppose that the compact $S \subset \mathbb{R}^n$ does not contain any points satisfying the first-order necessary conditions, stated in Corollary 2.2.20*(b)*, and yet there exist two sequences $\{x_i\}_{i=0}^\infty$ in S and $\{\pi_i\}_{i=0}^\infty$ in $\mathbb{R}_+$ such that $\pi_i \to \infty$ and $\nabla f_{\pi_i}^0(x_i) \to 0$, as $i \to \infty$. Since S is compact, without loss of generality, we can assume that $x_i \to \hat{x} \in S$, as $i \to \infty$. Since, by Assumption 2.8.15*(i)*, all the functions in (15b,c) are continuous, we must have that

$$g_x(x_i)^T g(x_i) + f_x(x_i)^T [f(x_i) + \frac{1}{\pi_i}\eta(x_i)]_+ \to g_x(\hat{x})^T g(\hat{x}) + f_x(\hat{x})^T f(\hat{x})_+ \tag{17a}$$

as $i \to \infty$. Hence, since $\pi_i \to \infty$ and $\nabla f_{\pi_i}^0(x_i) \to 0$, as $i \to \infty$, it follows from (15b,c) and the fact that $\nabla f_{\pi_i}^0(x_i) \to 0$, as $i \to \infty$, that

$$g_x(\hat{x})^T g(\hat{x}) + f_x(\hat{x})^T f(\hat{x})_+ = 0. \tag{17b}$$

It now follows from Assumption 2.8.15*(ii)* that $g(\hat{x}) = 0$ and $f(\hat{x})_+ = 0$, so that $f^j(\hat{x}) \leq 0$, for all $j \in \mathbf{q}$.

Next, since $\eta(\cdot)$ is continuous and therefore bounded on S, there must exist an $i_0 \in \mathbb{N}$ such that $\pi_i f^j(x_i) + \eta^j(x_i) \leq 0$, for all $i \geq i_0$ and $j \in \mathbf{q}_A(\hat{x})^c \triangleq \{j \in \mathbf{q} \mid f^j(\hat{x}) < 0\}$. Clearly, this implies that

$$[\pi_i f^j(x_i) + \eta^j(x_i)]_+ = 0 \tag{17c}$$

for all $i \geq i_0$ and $j \in \mathbf{q}_A(\hat{x})^c$. Suppose that $\mathbf{q}_A(\hat{x}) = \{j_1, j_2, \ldots, j_p\}$. For any $x \in \mathbb{R}^n$, let $\bar{f}(x) \triangleq (f^{j_1}(x), \ldots, f^{j_p}(x))$ and $\bar{\eta}(x) \triangleq (\eta^{j_1}(x), \ldots, \eta^{j_p}(x))^T$. Then, after rearranging terms in (15b), we find that for all $i \geq i_0$,

$$\begin{aligned}\nabla f_{\pi_i}^0(x_i) = {} & \nabla f^0(x_i) + g_x(x_i)^T[\xi(x_i) + \pi_i g(x_i)] + \xi_x(x_i)^T g(x_i) \\ & + \bar{f}_x(x_i)^T[\pi_i \bar{f}(x_i) + \bar{\eta}(x_i)]_+ + \frac{1}{\pi_i}\bar{\eta}_x(x_i)^T[\pi_i \bar{f}(x_i) + \bar{\eta}(x_i)]_+ \\ & - \frac{1}{\pi_i}\eta_x(x_i)^T \eta(x_i).\end{aligned} \tag{17d}$$

Now, for all $i \in \mathbb{N}$, let $\xi_i = \xi(x_i) + \pi_i g(x_i)$, and let $\bar{\eta}_i = [\pi_i \bar{f}(x_i) + \bar{\eta}(x_i)]_+$. By Assumption 2.8.15*(ii)*, the vectors $\nabla f^j(\hat{x})$, $j \in \mathbf{q}_A(\hat{x})$, together with the vectors $\nabla g^k(\hat{x})$ are linearly independent. Consequently there exists an $i_1 \geq i_0$, such that for all $i \geq i_1$, the matrix $[\bar{f}_x(x_i)^T g_x(x_i)^T]$ has maximum rank. Hence, since $\nabla f_{\pi_i}^0(x_i) \to 0$, as $i \to \infty$, we conclude that both $\bar{\eta}_i$ and ξ_i are bounded and therefore must have accumulation points $\bar{\eta}^*$ and $\bar{\xi}^*$. Furthermore, since $\bar{\eta}_i \geq 0$ for all $i \in \mathbb{N}$, we must have that $\bar{\eta}^* \geq 0$. Finally, since both

$\xi_x(x_i)^T g(x_i) \to 0$ and $\eta_x(x_i)^T \eta_x(x_i)/\pi_i \to 0$, as $i \to \infty$, we must have that

$$\lim \nabla f_{\pi_i}^0(x_i) = \nabla f^0(\hat{x}) + \bar{f}_x(\hat{x})^T \bar{\eta}^* + g_x(\hat{x})^T \xi^* = 0, \tag{17e}$$

which, together with the previously established fact that $f^j(\hat{x}) \leq 0$, for all $j \in \mathbf{q}$, and $g(\hat{x}) = 0$, implies that $\hat{x}$ is a stationary point for **P** in (12a), a contradiction. Hence our proof is complete. □

Next we must define a test function t_π. The proof of parts *(i)* and *(ii)* of the following theorem is quite straightforward. But the proof of part *(iii)* is rather long and difficult. Hence we omit the entire proof, which can be found in [GIP.79].

Theorem 2.8.19. *Suppose that Assumption 2.8.15 is satisfied. For every $\pi > 0$, let $t_\pi : \mathbb{R}^n \to \mathbb{R}$ be defined by*

$$t_\pi(x) = -\|\nabla f_\pi^0(x)\|^2 + \frac{1}{\pi^3}[\|a(x,\pi)\|^2 + \|g(x)\|^2], \tag{18}$$

Then,

(a) $t_\pi(\cdot)$ is continuous;,

*(b) if for some $\hat{x} \in \mathbb{R}^n$, $t_\pi(\hat{x}) \leq 0$ and $\nabla f_\pi^0(\hat{x}) = 0$, then $g(\hat{x}) = 0$ and $a(x,\pi) = 0$, and hence (see Exercise 2.8.17) $f^j(x) \leq 0$ and $\eta^j(x) \geq 0$, for all $j \in \mathbf{q}$, $\langle \eta(x), f(x) \rangle = 0$, and $\nabla f^0(\hat{x}) + f_x(\hat{x})^T \eta(\hat{x}) + g_x(\hat{x})^T \xi(\hat{x}) = 0$, i.e., $\hat{x}$ satisfies the first-order necessary condition for **P**, stated in Corollary 2.2.20(b), i.e., (2.2.23d,e), and*

(c) for every $\hat{x} \in \mathbb{R}^n$ there exists a $\hat{\pi} \in (0,\infty)$ and a $\hat{\rho} > 0$ such that, for all $x \in B(\hat{x}, \hat{\rho})$ and $\pi \geq \hat{\pi}$, $t_\pi(x) \leq 0$. □

Finally, as we had done for the equality constrained problem in (8a) we will state an approximation to the Hessian of $f_\pi^0(\cdot)$ that does not involve any third derivatives of the functions in (12a).

Exercise 2.8.20. For any $\pi > 0$ and $x \in \mathbb{R}^n$, let

$$\begin{aligned}G_\pi(x) \triangleq {} & f_{xx}^0(x) + \sum_{k=1}^r \xi^k(x) g_{xx}^k(x) + g_x(x)^T \xi_x(x) \\ & + \xi_x(x)^T g_x(x) + \pi g_x(x)^T g_x(x) \\ & + \frac{1}{\pi}\sum_{j=1}^q \left\{ \frac{-|f^j(x)|}{|f^j(x)| + |\eta_j(x)|} \nabla\eta^j(x) \nabla\eta^j(x)^T \right.\end{aligned}$$

$$+\frac{\pi|\eta^j(x)|}{|f^j(x)|+|\eta_j(x)|}\Big[\pi\nabla f^j(x)\nabla f^j(x)^T+f^j_{xx}(x)$$

$$+\nabla f^j(x)\nabla\eta^j(x)^T+\nabla\eta^j(x)\nabla f^j(x)^T\Big]\Big\}. \tag{19}$$

Show that if $\hat{x}$ is such that $f^j(\hat{x})\leq 0$, for all $j\in\mathbf{q}$, and $g(\hat{x})=0$, and, in addition, together with some $\hat{\mu}\in\Sigma_q^0$, $\hat{\zeta}\in\mathbb{R}^r$, $\hat{x}$ satisfies the first-order necessary conditions (2.2.23d,e) for the problem $\mathbf{P}$ in (12a) then, for any $\pi>0$, $G_\pi(\hat{x})=f^0_{\pi xx}(\hat{x})$. □

We can obviously redefine the functions $f^0_\pi(\cdot)$, $t_\pi(\cdot)$ and $G_\pi(\cdot)$ in Algorithm 2.8.9 to be as defined for the problem with mixed constraints (12a). In that case, we obtain the following result as a direct interpretation of Theorem 2.3.7:

Theorem 2.8.21. *Suppose that Assumption 2.8.15 is satisfied and that the functions $f^0_\pi(\cdot)$, $t_\pi(\cdot)$ and $G_\pi(\cdot)$ appearing in Algorithm 2.8.9 are redefined to be as in (15a), (18), and (19), respectively.*

(a) If Algorithm 2.8.9 constructs a finite sequence $\{x_j^*\}_{j=0}^{j^*}$ and the sequence $\{x_i\}$ is infinite, then every accumulation point $\hat{x}$ of $\{x_i\}$ satisfies $f^j(\hat{x})\leq 0$, for all $j\in\mathbf{q}$, $g(\hat{x})=0$, and the first-order necessary conditions (2.2.23d,e) for $\mathbf{P}$ in (12a).

(b) If Algorithm 2.8.9 constructs a finite sequence $\{x_j^*\}_{j=0}^{j^*}$ and the sequence $\{x_i\}_{i=0}^{i^*}$ is also finite, then the last element, x_{i^*}, of $\{x_i\}_{i=0}^{i^*}$ satisfies $f^j(x_{i^*})\leq 0$, for all $j\in\mathbf{q}$, $g(x_{i^*})=0$, and the first-order necessary conditions (2.2.23d,e), for $\mathbf{P}$ in (12a).

(c) If Algorithm 2.8.9 constructs an infinite sequence $\{x_j^*\}_{j=0}^{\infty}$, then $\{x_j^*\}_{j=0}^{\infty}$ has no accumulation points. □

Assumption 2.8.15*(iii)* ensures that for any feasible point $\hat{x}$, satisfying the first-order necessary conditions (2.2.23d,e) for problem $\mathbf{P}$ in (12a), there exists a $\hat{\pi}>0$ such that, for all $\pi\geq\hat{\pi}$, $f^0_{\pi xx}(\hat{x})=G_\pi(\hat{x})$ is positive-definite. Hence, assuming that Algorithm 2.8.9 increases π above such a threshold value, it should be clear again from the analysis of rate of convergence of the local Newton method that, if Algorithm 2.8.9 constructs a finite sequence $\{x_j^*\}_{j=0}^{j^*}$ and the sequence $\{x_i\}_{i=0}^{\infty}$ is infinite and has an accumulation point $\hat{x}$ such that the Hessian matrix $f^0_\pi(\hat{x})=G_\pi(\hat{x})$ is positive-definite, then Algorithm 2.8.9 converges to this point superlinearly.

2.8.3 Notes

As we have already mentioned, augmented Lagrangian methods were first proposed in [Hes.69, Pow.69], and, slightly later, in [HaB.70]. These methods are differentiable, exact penalty function methods, in which the multiplier and penalty are not updated continuously, as in the methods presented in this section. For an analysis of these early methods, see [Ber.75] and [Ber.82a].

In a series of three papers, Fletcher [Fle.70, FlL.71, Fle.72] proposed a modification of the earlier augmented Lagrangian algorithms for equality constrained problems, which consisted of using the formula (3a) for the multiplier. The result was a new kind of algorithm whose only shortcoming was that it did not have a mechanism for automatic penalty limitation. Such a mechanism, in the form of a test function, was proposed for Fletcher's method by Mukai and Polak in [MuP.75]. For an abstract form of such test functions, see [Pol.76]. An extension of the Mukai Polak method to optimization problems with both equality and inequality constraints can be found in [GlP.79]. The test function in [GlP.79] was constructed using the theory of test functions in [Pol.76]. The augmented Lagrangian used in [GlP.79] is due to Rockafellar [Roc.73, Roc.74] and was selected because it results in superior differentiability properties over those of the augmented Lagrangian used by Fletcher for problems with inequality constraints. The results in [MuP.75, Pol.76, GlP.79] were the basis for the presentation in this book.

Significant further refinements and extensions of the algorithms in [MuP.75 , GlP.79] can be found in [DiG. 85, DiG.86, DiG.89, CDL.93].

2.9 Sequential Quadratic Programming

Sequential quadratic programming methods, commonly referred to as SQP methods, are used for solving problems of the form

$$\min\{f^0(x)\mid f^j(x)\leq 0,\ j\in\mathbf{q},\ g(x)=0\} \tag{1}$$

with $f^j:\mathbb{R}^n\to\mathbb{R}$, $j=0,1,\ldots,q$, and $g:\mathbb{R}^n\to\mathbb{R}^r$. At any particular point, we will need some or all of the assumptions below:

Assumption 2.9.1. *We will assume that*

(i) the functions $f^j:\mathbb{R}^n\to\mathbb{R}$, $j=0,1,\ldots,q$, and $g:\mathbb{R}^n\to\mathbb{R}^r$ are twice Lipschitz continuously differentiable on bounded sets, and $r<n$,

(ii) the functions $f(\cdot)\triangleq(f^1(\cdot),\ldots,f^q(\cdot))$ and $g(\cdot)$ satisfy the Mangasarian-Fromowitz constraint qualification (see Definition 1.8.1) at any $\hat{x}\in\mathbb{R}^n$, which, together with some $\hat{\mu}\in\Sigma_q^0$ and $\hat{\zeta}\in\mathbb{R}^r$, satisfies the first-order necessary conditions $f^j(\hat{x})\leq 0$, $j\in\mathbf{q}$, $g(\hat{x})=0$, and (2.2.23d,e),

(iii) at any triplet $(\hat{x},\hat{\mu},\hat{\zeta})\in\mathbb{R}^n\times\Sigma_q^0\times\mathbb{R}^r$ satisfying the second-order sufficient conditions stated in Corollary 2.2.30, $\hat{\mu}^j>0$ for all $j\in\mathbf{q}_A(\hat{x})$ and the

vectors $\nabla f^j(\hat{x})$, $j \in \mathbf{q}_A(\hat{x})$ (with $\mathbf{q}_A(\hat{x})$ defined in (2.2.12d)), together with the vectors $\nabla g^l(\hat{x})$, $l \in \mathbf{r}$, are linearly independent. □

There are a number of sequential quadratic programming (SQP) algorithms described in the literature, all of which share two basic features: *(a)* they are globally stabilized perturbations of some version of Newton's method for solving the equations and inequalities of the first-order necessary conditions in Corollary 2.2.20 and hence are superlinearly convergent under appropriate conditions, and *(b)* they use a merit function, often an exact penalty function, to prevent convergence to a local maximum rather than a local minimum of (1). Particular SQP-type algorithms are distinguished by features such as *(i)* the quadratic programming problem that they solve in constructing updates, *(ii)* whether they use actual Hessians of the Lagrangian or BFGS type approximations to these Hessians, *(iii)* the type of merit function used to ensure convergence to a local minimum, *(iv)* the mechanism used for avoiding the Maratos effect, to be described later, and *(v)* the mechanism used for obtaining global convergence. We will comment more on these algorithms in the Notes at the end of this section.

2.9.1 Wilson's Method

The first sequential quadratic programming method for solving problems of the form (1) was proposed by Wilson [Wil.63]. It consisted of the sequential minimization of second-order expansions of the Lagrangian for (1), subject to a first-order expansion of the constraints in (1).

First we will consider Wilson's method for the special case of (1) in which there are no inequality constraints, i.e., for the problem

$$\min \{ f^0(x) \mid g(x) = 0 \}. \tag{2}$$

For this case, Assumption 2.9.1*(iii)* does not apply, while Assumption 2.9.1*(ii)* reduces to the requirement that the matrix $g_x(\hat{x})$ have maximum row rank, which we may as well require to be true for arbitrary $\hat{x}$. Hence, for problem (2), Assumption 2.9.1 reduces to the following form:

Assumption 2.9.1a. *We will assume that*

(i) the functions $f^0 : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^r$ are twice Lipschitz continuously differentiable on bounded sets,

(ii) $r < n$, and

(iii) the $r \times n$ Jacobian matrix $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$. □

Thus, suppose that Assumption 2.9.1a is satisfied. The Lagrangian for problem (2), $L : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}$, is defined by

$$L(x, \zeta) \triangleq f^0(x) + \langle \zeta, g(x) \rangle. \tag{3a}$$

Given an estimate x_i of an optimal solution of (2) and an estimate of its corresponding Lagrange multiplier ζ_i (see Corollary 2.2.20*(a)*), Wilson computed the update (x_{i+1}, ζ_{i+1}) by solving the quadratic program

$$\min \{ \langle \nabla_x L(x_i, \zeta_i), \delta x \rangle + \tfrac{1}{2} \langle \delta x, L_{xx}(x_i, \zeta_i) \delta x \rangle \mid g(x_i) + g_x(x_i) \delta x = 0 \} \tag{3b}$$

for its solution δx_i and the corresponding Lagrange multiplier $\delta\zeta_i$. Then he set $x_{i+1} = x_i + \delta x_i$ and $\zeta_{i+1} = \zeta_i + \delta\zeta_i$. Note that the constant term $L(x_i, \zeta_i)$ of the second-order expansion of the Lagrangian has been omitted from (3b) since it does not affect the solution of this problem.

Next, for any δx such that

$$g(x_i) + g_x(x_i) \delta x = 0, \tag{3c}$$

we have that

$$\begin{aligned} \langle \nabla_x L(x_i, \zeta_i), \delta x \rangle &= \langle \nabla f^0(x_i), \delta x \rangle + \langle \zeta_i, g_x(x_i) \delta x \rangle \\ &= \langle \nabla f^0(x_i), \delta x \rangle - \langle \zeta_i, g(x_i) \rangle. \end{aligned} \tag{3d}$$

Clearly, the constant term $\langle \zeta_i, g(x_i) \rangle$ can be dropped from (3b), and hence the following problem has the same solution δx_i as (3b):

$$\min \{ \langle \nabla f^0(x_i), \delta x \rangle + \tfrac{1}{2} \langle \delta x, L_{xx}(x_i, \zeta_i) \delta x \rangle \mid g(x_i) + g_x(x_i) \delta x = 0 \}, \tag{3e}$$

which was the final form used by Wilson. However, as we will now show, the Lagrange multiplier for (3e) is different from that for (3b). First consider (3b). Since, by assumption, $g_x(x_i)$ has maximum rank, if δx_i solves (3b), then, by Corollary 2.2.20*(a)*, there exists a unique multiplier $\delta\zeta_i \in \mathbb{R}^r$ such that

$$\nabla_x L(x_i, \zeta_i) + L_{xx}(x_i, \zeta_i) \delta x_i + g_x(x_i)^T \delta\zeta_i = 0, \tag{3f}$$

and, in addition, we must have that

$$g(x_i) + g_x(x_i) \delta x_i = 0. \tag{3g}$$

Since $\nabla_x L(x_i, \zeta_i) = \nabla f^0(x_i) + g_x(x_i)^T \zeta_i$, (3e) is equivalent to

$$\nabla f^0(x_i) + L_{xx}(x_i, \zeta_i) \delta x_i + g_x(x_i)^T (\zeta_i + \delta\zeta_i). \tag{3h}$$

Hence, by inspection, the appropriate multiplier to use with δx_i for problem (3e) is $\zeta_{i+1} = \zeta_i + \delta\zeta_i$, i.e., the relationship between the Lagrange multipliers for (3b) and (3e) is quite simple.

A quick referral to Section 1.4 (see (1.4.2b)) leads us to the conclusion that the linear system of update equations (3f,g) is exactly the same as the update

equations for Newton's method in solving the equation

$$G(z) = 0, \tag{3i}$$

where $z \triangleq (x, \zeta) \in \mathbb{R}^{n+r}$ and $G : \mathbb{R}^{n+r} \to \mathbb{R}^{n+r}$ is defined by

$$G(z) \triangleq \begin{bmatrix} \nabla_x L(x, \zeta) \\ g(x) \end{bmatrix}, \tag{3j}$$

which must be satisfied by any local minimizer $\hat{x} \in \mathbb{R}^n$ of (2), together with a $\hat{\zeta} \in \mathbb{R}^r$ (see Corollary 2.2.20*(a)*).

We will now establish a result needed to show that Newton's local method, defined by (1.4.2c), for solving (3i) is well defined near a local minimizer of (2) satisfying second-order sufficient conditions.

Lemma 2.9.2. *Consider the $(n+r) \times (n+r)$ matrix with $r \leq n$,*

$$Q = \begin{bmatrix} H & G^T \\ G & 0 \end{bmatrix},$$

where G is an $r \times n$ full-rank matrix and H is an $n \times n$, symmetric matrix. If $\langle h, Hh \rangle \neq 0$ for all nonzero $h \in \mathbb{R}^n$ such that $Gh = 0$, then Q is nonsingular.

Proof. For the sake of contradiction, suppose that Q is singular. Then there must exist a nonzero vector $v = (h, u) \in \mathbb{R}^{(n+r)\times(n+r)}$ such that $Qv = 0$. Clearly, this implies that

$$\left.\begin{aligned} Gh &= 0 \\ Hh + G^T u &= 0. \end{aligned}\right\} \tag{4}$$

Consequently, $\langle h, G^T u \rangle = \langle Gh, u \rangle = 0$, and therefore $\langle h, Hh \rangle = 0$, which implies that $h = 0$. Since G has maximum rank, it now follows from the second expression in (4) that $u = 0$, a contradiction. □

Proposition 2.9.3. *Suppose that Assumption 2.9.1a is satisfied and that $\hat{x} \in \mathbb{R}^n$ is a strict local minimizer for (2) which, together with $\hat{\zeta} \in \mathbb{R}^n$ satisfies the second-order sufficient conditions stated in Corollary 2.2.31, i.e.,*

$$\nabla_x L(\hat{x}, \hat{\zeta}) = 0, \tag{5a}$$

and

$$\min \{ \langle h, L_{xx}(\hat{x}, \hat{\zeta}) h \rangle \mid g_x(\hat{x}) h = 0, \ \|h\| = 1 \} = m > 0. \tag{5b}$$

Then,

(a) there exists a $\rho > 0$ such that, for all $x_i \in B(\hat{x}, \rho)$, $\zeta_i \in B(\hat{\zeta}, \rho)$, the system of equations (3f,g) has a unique solution $(\delta x_i, \delta\zeta_i)$, and

(b) there exists a $\hat{\rho} \in (0, \rho)$, with ρ as in (a), such that if (x_0, ζ_0) are such that $x_0 \in B(\hat{x}, \hat{\rho})$ and $\zeta_0 \in B(\hat{x}, \hat{\rho})$, then the sequence $\{(x_i, \zeta_i)\}_{i=0}^{\infty}$, constructed by Newton's local method (see (1.4.2c)) in solving (3i), defined by

$$\left.\begin{aligned} x_{i+1} &= x_i + \delta x_i, \\ \zeta_{i+1} &= \zeta_i + \delta\zeta_i, \end{aligned}\right\} \tag{5c}$$

with δx_i, $\delta\zeta_i$ determined by (3f,g), is well defined and converges to $(\hat{x}, \hat{\zeta})$ Q-quadratically.

Proof. *(a)* Since one can deduce from Corollary 5.4.2 that the function

$$w(x, \zeta) \triangleq \min \{ \langle h, L_{xx}(x, \zeta) h \rangle \mid g_x(x) h = 0, \ \|h\| = 1 \} \tag{6}$$

is continuous and $w(\hat{x}, \hat{\zeta}) = m > 0$, there exists a $\hat{\rho} > 0$ such that $w(x_i, \zeta_i) > m/2$ for all $x_i \in B(\hat{x}, \hat{\rho})$, $\zeta_i \in B(\hat{x}, \hat{\rho})$. The desired result now follows from Proposition 2.9.10 and the form of the equations (3f,g).

(b) This part follows directly from Theorem 1.4.1. □

Thus we see that Wilson's method is an acceptable *local* method for solving problem (2).

For the full problem (1), Wilson assumed that the first-order optimality conditions in Corollary 2.2.20*(b)* can be satisfied with $\hat{\mu}^0 > 0$, and hence that, for any local minimizer $\hat{x}$ of (1), there exist Karush-Kuhn-Tucker multipliers $\hat{\eta} \in \mathbb{R}^q$ and $\hat{\xi} \in \mathbb{R}^r$ such that

$$\left.\begin{aligned} \nabla f^0(\hat{x}) + f_x(\hat{x})^T \hat{\eta} + g_x(\hat{x})^T \hat{\xi} &= 0, \\ g(\hat{x}) &= 0, \\ \langle \hat{\eta}, f(\hat{x}) \rangle &= 0, \\ f(\hat{x}) &\leq 0, \\ \hat{\eta} &\geq 0, \end{aligned}\right\} \tag{7a}$$

where $f(x) \triangleq (f^1(x), f^2(x), \ldots, f^q(x))$. If for (1), we define the Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^l \to \mathbb{R}$, by

$$L(x, \eta, \xi) \triangleq f^0(x) + \langle \eta, f(x) \rangle + \langle \xi, g(x) \rangle, \tag{7b}$$

then we see that the first line in (7a) becomes $\nabla_x L(x, \eta, \xi) = 0$. With $L(\cdot, \cdot, \cdot)$ defined as in (7b), Wilson generalized (3e) to

$$\min \{ \langle \nabla f^0(x_i), \delta x \rangle + \tfrac{1}{2} \langle \delta x, L_{xx}(x_i, \eta_i, \xi_i) \delta x \rangle \mid$$
$$f(x_i) + f_x(x_i)\delta x \le 0, \quad g(x_i) + g_x(x_i)\delta x = 0 \}. \tag{7c}$$

By Corollary 2.2.20*(b)* (in (7a) form), the solution δx_i and corresponding multipliers $\eta_{i+1} = \eta_i + \delta\eta_i$, $\xi_{i+1} = \xi_i + \delta\xi_i$ satisfy the following set of equations and inequalities (since $\nabla_x L(x_i, \eta_i, \xi_i) = \nabla f^0(x_i) + f_x(x_i)^T \eta_i + g_x(x_i)^T \xi_i$):

$$\left.\begin{aligned} \nabla_x L(x_i, \eta_i, \xi_i) + L_{xx}(x_i, \eta_i, \xi_i)\delta x_i + f_x(x_i)^T \delta\eta_i + g_x(x_i)^T \delta\xi_i = 0, \\ g(x_i) + g_x(x_i)\delta x_i = 0, \\ \langle \eta_i + \delta\eta_i, f(x_i) + f_x(x_i)\delta x_i \rangle = 0, \\ f(x_i) + f_x(x_i)\delta x_i \le 0, \\ \eta_i + \delta\eta_i \ge 0. \end{aligned}\right\} \tag{7d}$$

Referring to the Local Newton Algorithm 1.8.5, we see that, when it is applied to the solution of the system of equations and inequalities (7a), it solves the following quadratic program to compute the update $(\delta x_i, \delta\eta_i, \delta\xi_i)$:

$$\min (\|\delta x\|^2 + \|\delta\eta\|^2 + \|\delta\xi\|^2) \tag{7e}$$

subject to the constraints

$$\left.\begin{aligned} \nabla_x L(x_i, \eta_i, \xi_i) + L_{xx}(x_i, \eta_i, \xi_i)\delta x + f_x(x_i)^T \delta\eta + g_x(x_i)^T \delta\xi = 0, \\ g(x_i) + g_x(x_i)\delta x = 0, \\ \langle \eta_i + \delta\eta, f(x_i) \rangle + \langle \eta_i, f_x(x_i)\delta x \rangle = 0, \\ f(x_i) + f_x(x_i)\delta x \le 0, \\ \eta_i + \delta\eta \ge 0. \end{aligned}\right\} \tag{7f}$$

Comparing (7d) and (7f), we see that Wilson's method is just a second-order perturbation of the local Newton Algorithm 1.8.5 with the important distinction that the system of equations and inequalities (7d) always has a solution near a local minimizer of (1) satisfying second-order sufficient conditions, strict linear complementarity and linear independence of active gradients, while this is not true of the system of equations and inequalities (7f). Hence Algorithm 1.8.5 cannot be used to find a triplet $(\hat{x}, \hat{\eta}, \hat{\xi})$ satisfying the optimality relations in (7a), but Wilson's method can. As we will show at the end of this section, to reconcile this apparent paradox, Wilson's method must be viewed not as a perturbation of Algorithm 1.8.5, but as a local Newton method for solving a set of *nonsmooth equations* that are equivalent to the system of smooth equations and inequalities (7a).

2.9.2 Pang's Method

There are a several local SQP algorithms for solving (1). They emanate from the observation that one can set up a system of *nonsmooth* equations, with the number of equations equal to the number of variables, that are equivalent to the system of smooth equations and inequalities (7a). All of these local SQP algorithms are extensions of the local Newton method in (1.4.2b,c). As we will see, one can set up more than one system of nonsmooth equations that are equivalent to the equations and inequalities in (7a), each system leading to a different extension of Newton's local method. Such an extension of Newton's method was first described by Josephy [Jos.79] who also proved the quadratic convergence of a local algorithm based on (7c) in the vicinity of a Karush-Kuhn-Tucker triplet satisfying second-order sufficient conditions for (1).

An explanation of Wilson's method, as an extension of the local Newton method (1.4.2c), requires the introduction of a system of nonsmooth equations whose relationship to (7a) is not intuitive. Furthermore, a complete justification of a global algorithm using Wilson's formula requires inferences deduced from results scattered through a number of papers by various authors. Hence, at the end of this section, we will provide only a formal justification for using formula (7c) in conjunction with our global stabilization scheme.

Instead, we will present, in some detail, a rather transparent extension (due to Pang [Pan.91]) of Newton's local method (1.4.2c) to a set of nonsmooth equations whose relationship to (7a) is reasonably straightforward. Furthermore, all the omitted material can be found in the single source [Pan.91]. As we will see, Pang's formula for computing an update is similar, but not identical, to Wilson's.

To understand Pang's method, we must first rewrite (7a) as a system of $n+m+r$ equations, not all smooth. For this purpose, we let $z = (x, \eta, \xi)$ to be an element of $\mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$, and we define the function $G(\cdot)$, mapping $\mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$ into $\mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$, by

$$G(z) \triangleq \begin{bmatrix} \nabla_x L(x, \eta, \xi) \\ \max \{-\eta, f(x)\} \\ g(x) \end{bmatrix}, \tag{8a}$$

where the vector "max" operator produces a vector consisting of component maxima, i.e., given $u, v \in \mathbb{R}^q$, the vector $w = \max \{u, v\} \in \mathbb{R}^q$, has

components $w^j = \max\{u^j, v^j\}$, $j \in \mathbf{q}$.

Exercise 2.9.4. Show that $G(\hat{z}) = 0$ if and only if $\hat{z} = (\hat{x}, \hat{\eta}, \hat{\xi})$ satisfies (7a). □

For any $z = (x, \eta, \xi) \in \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$, we define three index sets:

$$\left.\begin{aligned} \alpha(z) &\triangleq \{j \in \mathbf{q} \mid -\eta^j < f^j(x)\} \\ \beta(z) &\triangleq \{j \in \mathbf{q} \mid -\eta^j = f^j(x)\} \\ \gamma(z) &\triangleq \{j \in \mathbf{q} \mid -\eta^j > f^j(x)\} \end{aligned}\right\}, \tag{8b}$$

so that $\alpha(z) \cup \beta(z) \cup \gamma(z) = \mathbf{q}$. If we adopt the notation that, given an index set $J \subset \mathbf{q}$, $f_J(x)$ is a vector consisting of the components $f^j(x)$, with $j \in J$, then we can expand (8a) as follows:

$$G(z) \triangleq \begin{bmatrix} \nabla_x L(x, \eta, \xi) \\ f_{\alpha(z)}(x) \\ \max\{-\eta_{\beta(z)}, f_{\beta(z)(x)}\} \\ -\eta_{\gamma(z)} \\ g(x) \end{bmatrix}. \tag{8c}$$

It follows from Corollary 5.4.6 that each component of $G(z)$ has a directional derivative. For any $d = (\delta x, \delta\eta, \delta\xi) \in \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$, let $G'(z; d) \in \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$ be a vector whose ith component is the directional derivative $dG^i(z; d)$, $i = 1, \ldots, n+q+r$. Then it is clear that

$$G'(z; d) = \begin{bmatrix} L_{xx}(x, \eta, \xi)\delta x + f_x(x)\delta\eta + g_x(x)\delta\xi \\ (f_{\alpha(z)})^x(x)\delta x \\ \max\{-\delta\eta_{\beta(z)}, (f_{\beta(z)})_x(x)\delta x\} \\ -\delta\eta_{\gamma(z)} \\ g_x(x)\delta x \end{bmatrix}, \tag{8d}$$

where, again, the "max" operator denotes the componentwise maximum.

The Pang extension of Newton's local method for finding a zero of $G(\cdot)$ is defined by the following recursion rule:

$$z_{i+1} = z_i + d_i, \tag{8e}$$

where d_i is such that

$$G(z_i) + G'(z_i, d_i) = 0. \tag{8f}$$

The relationship between Wilson's method and Pang's method can be seen from the following result:

Exercise 2.9.5. Let $z_i = (x_i, \eta_i, \xi_i) \in \mathbb{R}^{n+q+r}$. Show that if $(\delta x_i, \eta_{i+1}, \xi_{i+1}) \in \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$, with $(\eta_{i+1})_{\gamma(z_i)} = 0$, is a Karush-Kuhn-Tucker triplet for the quadratic program

$$\left.\begin{aligned} \min\{\langle \nabla f^0(x_i), \delta x\rangle + \tfrac{1}{2}\langle \delta x, L_{xx}(x_i, \eta_i, \xi_i)\delta x\rangle \mid \\ f_{\alpha(z_i)}(x_i) + (f_{\alpha(z_i)})_x(x_i)\delta x = 0, \\ f_{\beta(z_i)}(x_i) + (f_{\beta(z_i)})_x(x_i)\delta x \le 0, \\ g(x_i) + g_x(x_i)\delta x = 0\}, \end{aligned}\right. \tag{9}$$

and $\delta\eta_i \triangleq \eta_{i+1} - \eta_i$, $\delta\xi_i \triangleq \xi_{i+1} - \xi_i$, then $d_i = (\delta x_i, \delta\eta_i, \delta\xi_i)$ satisfies (8f). □

Exercise 2.9.6. Suppose that Assumption 2.9.1 is satisfied, that $\hat{z} = (\hat{x}, \hat{\eta}, \hat{\xi})$ is a Karush-Kuhn-Tucker triplet for problem (1), i.e., $G(\hat{z}) = 0$, and that the second-order sufficient conditions stated in Corollary 2.2.30 are satisfied, i.e., (with the Lagrangian defined by (7b)), there exists an $m > 0$ such that

$$\langle h, L_{xx}(\hat{x}, \hat{\eta}, \hat{\xi})h\rangle \ge m\|h\|^2, \ \forall\, h \in \mathcal{H}_{IE}(\hat{x}), \tag{10a}$$

where

$$\mathcal{H}_{IE}(\hat{x}) \triangleq \{h \in \mathbb{R}^n \mid (f_{\mathbf{q}_A(\hat{x})})_x(\hat{x})h = 0,\ g_x(\hat{x})h = 0\}. \tag{10b}$$

(i) Show that $\alpha(\hat{z}) = \mathbf{q}_A(\hat{x})$ and $\beta(\hat{z}) = \emptyset$, where $\emptyset$ denotes the empty set. Hence show that $G(\cdot)$ is differentiable in a ball about $\hat{z}$ and that

$$G_z(\hat{z}) = \begin{bmatrix} L_{xx}(\hat{x}, \hat{\eta}, \hat{\xi}) & f_x(\hat{x})^T & g_x(\hat{x})^T \\ (f_{\alpha(\hat{z})})_x(\hat{x}) & 0 & 0 \\ 0 & I_{\gamma(\hat{z})} & 0 \\ g_x(\hat{x}) & 0 & 0 \end{bmatrix}, \tag{10c}$$

where $I_{\gamma(\hat{z})}$ is an identity matrix with size equal to the cardinality of $\gamma(\hat{z})$.

(ii) Show that $G_z(\hat{z})$ is nonsingular. □

In view of Exercise 2.9.6, we see that near a Karush-Kuhn-Tucker triplet satisfying the assumptions stated in Exercise 2.9.6, Pang's method reduces to the Local Newton Algorithm (1.4.2c) for solving the equation $G(z) = 0$, and hence the following result is obvious.

Theorem 2.9.7. *Suppose that Assumption 2.9.1 is satisfied, that* $\hat{z} = (\hat{x}, \hat{\eta}, \hat{\xi})$ *is a Karush-Kuhn-Tucker triplet for problem (1), i.e., it satisfies* $G(\hat{z}) = 0$, *and that the second-order sufficient conditions stated in Corollary 2.2.30 are satisfied, i.e., (with the Lagrangian defined by (7b)), there exists an* $m > 0$ *such that (10a,b) hold. Then there exists a* $\rho > 0$ *such that, if* $z_0 = (x_0, \eta_0, \xi_0) \in B(\hat{z}, \rho)$, *then the sequence* $\{z_i\}_{i=0}^{\infty}$, *constructed according to the recursion rule*

$$z_{i+1} = z_i - G_z(z_i)^{-1} G(z_i), \; i \in \mathbb{N}, \tag{11a}$$

is well defined with $G_z(z_i)^{-1} G(z_i) = d_i$, *where* d_i *is uniquely defined via (9), as in Exercise 2.9.5. Furthermore, for some* $C \in (0, \infty)$,

$$\|z_{i+1} - \hat{z}\| \leq C \|z_i - \hat{z}\|^2, \; i \in \mathbb{N}. \tag{11b}$$

□

To obtain an algorithm that converges quadratically under somewhat less stringent conditions than those in Theorem 2.9.7, Pang modifies the sets $\alpha(z)$, $\beta(z)$, and $\gamma(z)$, as follows. First, for any $z = (x, \eta, \xi) \in \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$, he defines the sets

$$\left.\begin{aligned} \bar{\alpha}(z) &\triangleq \{j \in \alpha(z) \mid \eta^j \geq 0\} \\ \bar{\beta}(z) &\triangleq \beta(z) \cup \{j \in \alpha(z) \mid \eta^j < 0\} \cup \{j \in \gamma(z) \mid f^j(x) > 0\} \\ \bar{\gamma}(z) &\triangleq \{j \in \gamma(z) \mid f^j(x) \leq 0\} \end{aligned}\right\}, \tag{12}$$

so that $\bar{\alpha}(z) \cup \bar{\beta}(z) \cup \bar{\gamma}(z) = \mathfrak{q}$ with the cardinality of $\bar{\beta}(z)$ increased over the cardinality of $\beta(z)$ at the expense of indices removed from $\alpha(z)$ and $\gamma(z)$, and hence, with $\bar{\alpha}(z)$ and $\bar{\gamma}(z)$ having smaller cardinality than $\alpha(z)$ and $\gamma(z)$, respectively.

Looking at it heuristically, the set $\bar{\alpha}(z)$ contains the indices of the inequalities predicted to be active at a solution $\hat{z}$, the set $\bar{\gamma}(z)$ contains the indices of the inequalities predicted to be inactive at a solution $\hat{z}$, and the set $\bar{\beta}(z)$ contains the indices of the inequalities about which one is not prepared to make a prediction.

Definition (12) leads to the following local method for finding a zero of $G(\cdot)$, defined in (8c):

Local Pang Algorithm 2.9.8.

Data: $z_0 = (x_0, \eta_0, \xi_0) \in \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$.

Step 0. Set $i = 0$.

Step 1: Compute a Karush-Kuhn-Tucker triplet $(\delta x_i, \eta_{i+1}, \xi_{i+1})$, with $(\eta_{\bar{\gamma}(z_i)})_{i+1} = 0$, for the problem

$$\begin{aligned} \min \Big\{ &\langle \nabla f^0(x_i), \delta x \rangle + \tfrac{1}{2} \langle \delta x, L_{xx}(x_i, \eta_i, \xi_i) \delta x \rangle \mid \\ &f_{\bar{\alpha}(z_i)}(x_i) + (f_{\bar{\alpha}(z_i)})_x(x_i) \delta x = 0, \\ &f_{\bar{\beta}(z_i)}(x_i) + (f_{\bar{\beta}(z_i)})_x(x_i) \delta x \leq 0, \\ &g(x_i) + g_x(x_i) \delta x = 0 \Big\}. \end{aligned} \tag{13}$$

Step 2: Set $x_{i+1} = x_i + \delta x_i$, set $z_{i+1} = (x_{i+1}, \eta_{i+1}, \xi_{i+1})$, replace i by $i+1$, and go to Step 1.

This algorithm is quadratically convergent in the vicinity of a point $\hat{z} = (\hat{x}, \hat{\eta}, \hat{\xi})$ satisfying (7a), when $\hat{z}$ is a *Pang-regular point*, whose definition requires the following additional set:

$$\alpha_+(z) \triangleq \{j \in \bar{\alpha}(z) \mid \eta^j > 0\}. \tag{14}$$

Definition 2.9.9. *Let* $\hat{z} = (\hat{x}, \hat{\eta}, \hat{\xi}) \in \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$ *be such that (7a) is satisfied. Let*

$$A(\hat{z}) \triangleq \begin{bmatrix} L_{xx}(\hat{x}, \hat{\eta}, \hat{\xi}) & -(f_{\alpha_+(\hat{z})})_x(\hat{x}) & -g_x(\hat{x})^T \\ (f_{\alpha_+(\hat{z})})_x(\hat{x}) & 0 & 0 \\ g_x(\hat{x}) & 0 & 0 \end{bmatrix}, \tag{15a}$$

$$C(\hat{z}) \triangleq \begin{bmatrix} (f_{\beta(\hat{z})})_x(\hat{x}) \\ 0 \\ 0 \end{bmatrix}, \tag{15b}$$

$$B(\hat{z}) \triangleq C(\hat{z})^T A(\hat{z})^{-1} C(\hat{z}). \tag{15c}$$

We will say that $\hat{z}$ *is a* Pang-regular vector *for* $G(\cdot)$ *(defined in (8c)), if the matrix* $A(\hat{z})$ *is nonsingular and all the principal minors of* $B(\hat{z})$ *are positive.* □

The important property of a Pang-regular vector is that at a Pang-regular vector $\hat{z}$, the equation $G'(\hat{z}, d) = 0$ has a unique solution. It is shown in

[Pan.91] that any vector $\hat{z}$ satisfying the conditions in Theorem 2.9.7 is a Pang-regular vector. In, addition, in [Pan.91] we find proofs of the following two results.

Proposition 2.9.10. *Suppose that Assumption 2.9.1a is satisfied and that $\hat{z}$ is a Pang-regular vector for $G(\cdot)$, such that $G(\hat{z}) = 0$. Then there exists a $\rho > 0$ and a $K < \infty$ such that, for all $z_i \in B(\hat{z}, \rho)$, z_i is also a a Pang-regular vector for $G(\cdot)$, and the problem (13) has a unique Karush-Kuhn-Tucker triplet $\delta z_i \triangleq (\delta x_i, \eta_{i+1}, \xi_{i+1})$ with $(\eta_{-\gamma(z_i)})_{i+1} = 0$ satisfying $\|\delta z_i\| \leq K \|G(z_i)\|$.* □

Theorem 2.9.11. *Suppose that Assumption 2.9.1a is satisfied and that $\hat{z}$ is a Pang-regular vector for $G(\cdot)$, such that $G(\hat{z}) = 0$. Then there exists a $\rho > 0$ such that, if the Local Pang Algorithm 2.9.8 is initialized with a $z_0 \in B(\hat{z}, \rho)$, then z_i is well defined for all $i \geq 0$ and for some $C \in (0, \infty)$,*

$$\|z_{i+1} - \hat{z}\| \leq C \|z_i - \hat{z}\|^2, \quad \forall\, i \in \mathbb{N}. \tag{16}$$

□

2.9.3 The Local Maratos-Mayne-Polak Method for (2)

There are two major objections to using Wilson's method without modifications in solving problem (2). The first stems from the fact that (3i) is also the correct optimality condition for the problem $\max \{ f^0(x) \mid g(x) = 0 \}$. Hence the result of using local Newton's method can be much worse than the original guess (x_0, ζ_0). The second objection is that Newton's local method converges only locally.

To get around the above objections, it is customary to use exact penalty functions (first used in [Han.77]) or other merit functions both for globally stabilizing the local Newton method and for guarding against its convergence to a local maximizer. We will now present a simplified version of the Mayne-Maratos-Polak SQP algorithm [MaP79, MaP.82], which was the first version of an SQP algorithm, using an exact penalty function based step-size rule, that was demonstrably quadratically convergent. We begin with a version of this algorithm for problem (2), which we will analyze in full.

Consider the exact penalty function for problem (2), $f_\pi^0 : \mathbb{R}^n \to \mathbb{R}$, defined for any $\pi \geq 0$ by

$$f_\pi^0(x) \triangleq f^0(x) + \pi \|g(x)\|_\infty = f^0(x) + \pi \max_{j \in 2\mathbf{r}} g^j(x), \tag{17a}$$

where, as before, $g^{j+r}(x) \triangleq -g^j(x)$, $j \in \mathbf{r}$. Suppose that x_i, ζ_i are sufficiently near to a pair $(\hat{x}, \hat{\zeta})$ satisfying second-order sufficient conditions for problem (2), to ensure that (3e,f) has a solution $(\delta x_i, \delta\zeta_i)$. Furthermore, suppose that

$\pi > \sum_{j=1}^r |\zeta_{i+1}^j|$, where $\zeta_{i+1} = \zeta_i + \delta\zeta_{i+1}$. Then, recalling that, by our standard notation, $\hat{2\mathbf{r}}(x_i) \triangleq \{ j \in 2\mathbf{r} \mid g^j(x_i) = \|g(x_i)\|_\infty \}$, we conclude that

$$df_\pi^0(x_i; \delta x_i) = \langle \nabla f^0(x_i), \delta x_i \rangle + \pi \max_{\hat{2\mathbf{r}}(x_i)} \langle \nabla g^j(x_i), \delta x_i \rangle. \tag{17b}$$

Now, making use of (3e,f) we conclude that

$$\begin{aligned} \langle \nabla f^0(x_i), \delta x_i \rangle &= \langle \nabla_x L(x_i, \zeta_i) - g_x(x_i)^T \zeta_i, \delta x_i \rangle \\ &= -\langle \delta x_i, L_{xx}(x_i, \zeta_i)\delta x_i \rangle - \langle g_x(x_i)^T \zeta_{i+1}, \delta x_i) \rangle \\ &= -\langle \delta x_i, L_{xx}(x_i, \zeta_i)\delta x_i \rangle + \langle \zeta_{i+1}, g(x_i) \rangle. \end{aligned} \tag{17c}$$

Hence, since in view of (3g), $\max_{j \in \hat{2\mathbf{r}}(x_i)} \langle \nabla g^j(x_i), \delta x_i \rangle = -\|g(x_i)\|_\infty$, we obtain that

$$\begin{aligned} df_\pi^0(x_i; \delta x_i) &= -\langle \delta x_i, L_{xx}(x_i, \zeta_i)\delta x_i \rangle + \langle \zeta_{i+1}, g(x_i) \rangle - \pi \|g(x_i)\|_\infty \\ &\leq -\langle \delta x_i, L_{xx}(x_i, \zeta_i)\delta x_i \rangle - (\pi - \sum_{j=1}^l |\zeta_{i+1}^j|) \|g(x_i)\|_\infty. \end{aligned} \tag{17d}$$

Proposition 2.9.12. *Suppose that Assumption 2.9.1a is satisfied, that the pair $(\hat{x}, \hat{\zeta})$ satisfy the second-order sufficient conditions (5a,b) for problem (2), and that $\pi > \sum_{j=1}^r |\hat{\zeta}^j|$. Then there exists a $\rho^* > 0$ such that, for any pair $(x_i, \zeta_i) \neq (\hat{x}, \hat{\zeta})$ satisfying $x_i \in B(\hat{x}, \rho^*)$, $\zeta_i \in B(\hat{\zeta}, \rho^*)$, if δx_i and $\delta\zeta_i$ are determined by (3e,f), then*

$$df_\pi^0(x_i; \delta x_i) < 0. \tag{18}$$

Proof. First, by Proposition 2.9.11, there exists a $\hat{\rho} > 0$ such that, for all $x_i \in B(\hat{x}, \hat{\rho})$ and $\zeta_i \in B(\hat{\zeta}, \hat{\rho})$, δx_i and ζ_{i+1} are uniquely defined by (3f,g). Hence, by (17d), for all $x_i \in B(\hat{x}, \hat{\rho})$ and $\zeta_i \in B(\hat{\zeta}, \hat{\rho})$,

$$df_\pi^0(x_i; \delta x_i) \leq -\langle \delta x_i, L_{xx}(x_i, \zeta_i)\delta x_i \rangle - (\pi - \sum_{j=1}^l |\zeta_{i+1}^j|) \|g(x_i)\|_\infty. \tag{19a}$$

Next, since the function $w(\cdot, \cdot)$ defined in (6) is continuous, there exists a $\rho \in (0, \hat{\rho}]$ such that, for all $x_i \in B(\hat{x}, \rho)$ and $\zeta_i \in B(\hat{\zeta}, \rho)$,

$$\langle h', L_{xx}(x_i, \zeta_i)h' \rangle \geq \tfrac{1}{2} m \|h'\|^2, \ \forall\, h' \in \{ h' \mid g_x(x_i)h' = 0 \}. \tag{19b}$$

Let $\alpha \in (0, 1)$ be such that $\alpha\pi = \pi - \sum_{j=1}^r |\hat{\zeta}^j|$. By continuity, it follows from (3e,f) that there exists a $\rho' \in (0, \rho]$ such that, for all $x_i \in B(\hat{x}, \rho')$, $\zeta_i \in B(\hat{\zeta}, \rho')$, $\frac{1}{2}\alpha\pi \leq \pi - \sum_{j=1}^r |\zeta_{i+1}^j|$.

Let $x_i \in B(\hat{x}, \rho)$ and $\zeta_i \in B(\hat{\zeta}, \rho)$ be arbitrary, let δx_i be determined by (3e,f), and let $h_i', h_i'' \in \mathbb{R}^n$ be such that $\delta x_i = h_i' + h_i''$, $g_x(x_i)h_i' = 0$, and $\langle h_i', h_i'' \rangle = 0$. Then h_i'' must be in the orthogonal complement of the null space of $g(x_i)$, i.e., for some $\eta \in \mathbb{R}^r$, $h_i'' = g_x(x_i)^T \eta_i$, and hence, because of (3g), we must have that

$$h_i'' = -g_x(x_i)^T [g_x(x_i) g_x(x_i)^T]^{-1} g(x_i) . \tag{19c}$$

Since, by assumption, $g(\cdot)$ is continuously differentiable and $g_x(x)$ has maximum row rank for all $x \in \mathbb{R}^n$, it follows from (19c) that there exists a $K < \infty$ such that, for all $x_i \in B(\hat{x}, \rho)$ $\zeta_i \in B(\hat{\zeta}, \rho)$ and δx_i determined by (3e,f),

$$\|h_i''\| \leq K \|g(x_i)\|_\infty . \tag{19d}$$

Hence, for all such (x_i, ζ_i),

$$\begin{aligned} df_\pi^0(x_i; \delta x_i) \leq & -\langle h_i', L_{xx}(x_i, \zeta_i) h_i' \rangle - \langle h_i'', L_{xx}(x_i, \zeta_i) h_i'' \rangle \\ & - 2\langle h_i', L_{xx}(x_i, \zeta_i) h_i'' \rangle - (\pi - \sum_{j=1}^{l} |\zeta_{i+1}^j|) \|g(x_i)\|_\infty \\ \leq & -\tfrac{1}{2} m \|h_i'\|^2 - \|h_i''\| [\alpha\pi/2K - (\|L_{xx}(x_i, \zeta_i)\|(\|h_i''\| + 2\|h_i'\|)] . \end{aligned} \tag{19e}$$

Since $(\delta x_i, \delta\zeta_i) \to 0$, as $(x_i, \zeta_i) \to (\hat{x}, \hat{\zeta})$, and consequently also $h_i', h_i'' \to 0$, there exists a $\rho^* \in (0, \rho']$ such that, for all $x_i \in B(\hat{x}, \rho^*)$ and $\zeta_i \in B(\hat{\zeta}, \rho^*)$, $df_\pi^0(x_i; \delta x_i) < 0$, which completes our proof. □

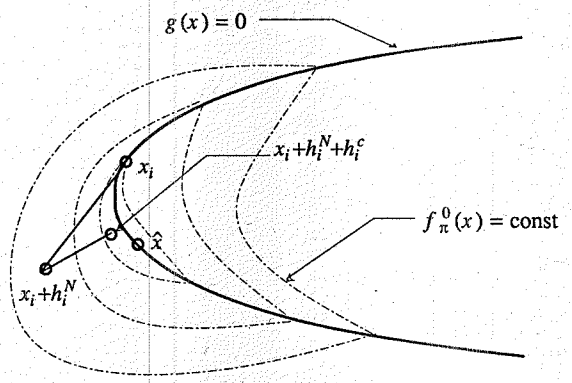


Fig. 2.9.1. The Maratos effect.

It was observed by Maratos [Mar.78] that although the directions δx_i generated by Newton's method in solving (3i) are descent directions for $f_\pi^0(\cdot)$, for π sufficiently large, as shown above, $f_\pi^0(x_i + \delta x_i) > f_\pi^0(x_i)$ may hold for an infinite number of times. A possible scenario is shown in Fig. 2.9.1 for the case where $n = 2$ and a particular value of π, when x_i is such that $g(x_i) = 0$ and hence $h_i^N = \delta x_i$ determined by (3e,f) is tangent to the curve $C = \{x \in \mathbb{R}^2 \mid g(x) = 0\}$ (since $g_x(x_i)\delta x_i = 0$ must hold by (3g)). Note that the equi-cost contours $\{x \in \mathbb{R}^2 \mid f_\pi^0(x) = const\}$ have corners on the curve C, and that the effect of increasing π on these contours is to make them bunch up closer to the curve C. Hence for all x_i such that $g(x_i) = 0$, because $\|g(x_{i+1})\|_\infty \geq \|g(x_i)\|_\infty$, if $f_\pi^0(x_i + \delta x_i) > f_\pi^0(x_i) = f^0(x_i)$, for some $\pi \geq 0$, then $f_{\pi'}^0(x_i + \delta x_i) > f_{\pi'}^0(x_i) = f^0(x_i)$ for any $\pi' \geq \pi$. Therefore it follows that, if one uses exact penalty functions of the form $f_\pi^0(\cdot)$ in (17a) as descent functions in conjunction with search directions generated by Newton's method in solving (3i), irrespective of the value of $\pi < \infty$ used, a step-size of 1 may fail to be accepted an infinite number of times, destroying the quadratic rate of convergence of Newton's method. This observation has become known as the *Maratos Effect* and is explained by the fact that when $g(x_i) \simeq 0$, the increase in the penalty term from $\pi\|g(x_i)\|_\infty$ to $\pi\|g(x_i + \delta x_i)\|_\infty$ cannot be overcome by the decrease in cost from $f^0(x_i)$ to $f^0(x_i + \delta x_i)$. To make it possible to use exact penalty functions as descent functions, Newton's method has to be modified without destroying its quadratic convergence. There are several modified Newton methods in the literature, all of which are referred to as SQP algorithms because of Wilson's original formulation, that can use exact penalty functions as descent functions. The following local Maratos-Mayne-Polak (MMP) algorithm for solving (3i) is the first method of this type. It adds a small correction h_i^c to the Newton direction h_i^N to reduce the constraint violation, as shown in Fig. 2.9.1.

Local MMP Algorithm 2.9.13.

Data. $z_0 = (x_0, \zeta_0) \in \mathbb{R}^n \times \mathbb{R}^r$.

Step 0. Set $i = 0$.

Step 1. Compute the *Newtonian search direction* (h_i^N, v_i^N) for (3i) by solving the system of equations (3e,f), so that

$$\nabla_x L(x_i, \zeta_i) + L_{xx}(x_i, \zeta_i) h_i^N + g_x(x_i)^T v_i^N = 0 , \tag{20a}$$

$$g(x_i) + g_x(x_i) h_i^N = 0 . \tag{20b}$$

Step 2. Compute the Newtonian *feasibility correction direction* h_i^c for (3i), using formula (1.8.9a) in Algorithm 1.8.5, so that

$$h_i^c = \arg\min \{ \|h\|^2 \mid g(x_i + h_i^N) + g_x(x_i)h = 0 \}$$

$$= -g_x(x_i)^T [g_x(x_i)g_x(x_i)^T]^{-1} g(x_i + h_i^N). \tag{20c}$$

Step 3. Set $x_{i+1} = x_i + h_i^N + h_i^c$, $\zeta_{i+1} = \zeta_i + v_i^N$, $z_{i+1} = (x_{i+1}, \zeta_{i+1})$, replace i by $i+1$, and go to Step 1.

Lemma 2.9.14. *Suppose that Assumption 2.9.1a is satisfied. If $\hat{x} \in \mathbb{R}^n$ is a strict local minimizer of (2) which together with $\hat{\zeta} \in \mathbb{R}^r$ satisfies the second-order sufficient conditions (5a,b), then there exists a $\hat{\rho} > 0$ such that*

(a) if $x_i \in B(\hat{x}, \hat{\rho})$ and $\zeta_i \in B(\hat{\zeta}, \hat{\rho})$, then h_i^N, h_i^c, and v_i^N are all well defined, and

(b) there exists a constant $K < \infty$ such that

$$\|g(x_i + h_i^N)\|_\infty \le K \|h_i^N\|^2, \tag{21a}$$

$$\|g(x_i + h_i^N + h_i^c)\|_\infty \le K \|h_i^N\|^3, \tag{21b}$$

and

$$\|h_i^c\| \le K \|h_i^N\|^2. \tag{21c}$$

Proof. *(a)* This part follows directly from Proposition 2.9.11 and Theorem 1.8.7.

(b) Suppose that $x_i \in B(\hat{x}, \hat{\rho})$ and $\zeta_i \in B(\hat{\zeta}, \hat{\rho})$. Making use of the first-order expansion formula (5.1.18a), the Lipschitz continuity of $g_x(\cdot)$, the boundedness of (h_i^N, v_i) for all $x_i \in B(\hat{x}, \hat{\rho})$ and $\zeta_i \in B(\hat{\zeta}, \hat{\rho})$, the fact that $\|x\|_\infty \le \|x\|$, for all $x \in \mathbb{R}^n$, and of (20b), we obtain

$$\|g(x_i + h_i^N)\|_\infty = \|g(x_i) + g_x(x_i)h_i^N + \int_0^1 [g_x(x_i + sh_i^N) - g_x(x_i)]\, ds\, h_i^N\|_\infty$$

$$\le \tfrac{1}{2}K' \|h_i^N\|^2, \tag{22a}$$

where $K' < \infty$ is a local Lipschitz constant for $g_x(\cdot)$ (with respect to the $\|\cdot\|_\infty$ norm), which proves (21a), with $K = \frac{1}{2}K'$. It now follows directly from (22a) and (20c) that (21c) must hold for some $K = K'' \in (K', \infty)$.

Again resorting to first-order expansions, the equivalence of norms on $\mathbb{R}^n$, and using (20c), we find that

$$\|g(x_i + h_i^N + h_i^c)\|_\infty = \|g(x_i + h_i^N) + g_x(x_i)h_i^c$$

$$+ \int_0^1 [g_x(x_i + h_i^N + sh_i^c) - g_x(x_i)]ds h_i^c\|_\infty$$

$$\le K' \|h_i^c\|(\|h_i^N\| + \tfrac{1}{2}\|h_i^c\|)$$

$$\le K'K''(1 + \tfrac{1}{2}K''\|h_i^N\|)\|h_i^N\|^3, \tag{22b}$$

which, in view of the boundedness of h_i^N, establishes (21b) for

$$K = K''' \triangleq \max_{x_i \in B(\hat{x}, \hat{\rho}), \zeta_i \in B(\hat{\zeta}, \hat{\rho})} K'K''(1 + \tfrac{1}{2}K''\|h_i^N\|) < \infty. \tag{22c}$$

Hence (21a,b,c) hold with $K = \max\{K', K'', K'''\}$. □

Theorem 2.9.15. *Suppose that Assumption 2.9.1a is satisfied and that $\hat{x} \in \mathbb{R}^n$ is a strict local minimizer of (2) which together with $\hat{\zeta} \in \mathbb{R}^r$ satisfies the second-order sufficient conditions (5a,b). Then there exists a $\hat{\rho} > 0$ such that, if $x_0 \in B(\hat{x}, \hat{\rho})$ and $\zeta_0 \in B(\hat{\zeta}, \hat{\rho})$, then the sequence $\{(x_i, \zeta_i)\}_{i=0}^\infty$ constructed by the Local MMP Algorithm 2.9.13 in solving (3i) is well defined and converges Q-quadratically to $(\hat{x}, \hat{\zeta})$.*

Proof. Again, let $z = (x, \zeta) \in \mathbb{R}^n \times \mathbb{R}^r$, and let $G : \mathbb{R}^{n \times r} \to \mathbb{R}^{n \times r}$ be defined as in (3j). Then $G(\cdot)$ is Lipschitz continuously differentiable on bounded sets. Also, if we set $\hat{z} = (\hat{x}, \hat{\zeta})$, then $G(\hat{z}) = 0$ and, in view of Lemma 2.9.10, $G_z(\hat{z})$ is nonsingular. With this notation, we find that the Local MMP Algorithm 2.9.13 has the form

$$z_{i+1} = z_i - G_z(z_i)^{-1}G(z_i) + u_i, \tag{23a}$$

where $u_i = (h_i^c, 0)$. In view of Proposition 2.9.11 and Lemma 2.9.14 (see (21c)), there exists a $\rho > 0$ such that, for all $z_i \in B(\hat{z}, \rho)$, z_{i+1} is well defined, and $\|u_i\| = \|h_i^c\| \le K\|h_i^N\|^2 \le K\|G_z(z_i)^{-1}G(z_i)\|^2$, for some $K < \infty$. Clearly, both $G_z(z)$ and $G_z(z)^{-1}$ are bounded on the ball $B(\hat{x}, \rho)$. Hence, because $G(\hat{z}) = 0$, expanding $G(z_i)$ around $\hat{z}$, using formula (5.1.18a), we deduce that

$$\|u_i\| \le K\|G_z(z_i)^{-1}\int_0^1 G_z(\hat{z} + s(z_i - \hat{z}))ds\,(z_i - \hat{z})\|^2 \le K'\|z_i - \hat{z}\|^2, \tag{23b}$$

for some $K' < \infty$. Next, making use of the first-order expansion formula (5.1.18a), we conclude from (23a) that, for all $z_i \in B(\hat{z}, \rho)$,

$$\|z_{i+1} - \hat{z}\| - \|u_i\| \le \|(z_{i+1} - \hat{z}) - u_i\|$$

$$\le \|G_z(z_i)^{-1}\| \int_0^1 \|G_z(\hat{z} + s(z_i - \hat{z}))\|ds\, \|z_i - \hat{z}\|$$

$$\leq \tfrac{1}{2}bL\,\|z_i - \hat{z}\|^2, \tag{23c}$$

where $b \triangleq \max_{z \in B(\hat{z},\rho)} \|G_z(z)^{-1}\|$ and $L < \infty$ is a Lipschitz constant for $G_z(\cdot)$ on $B(\hat{z},\rho)$. Combining the results in (23b,c), we conclude that

$$\|z_{i+1} - \hat{z}\| \leq K\,\|z_i - \hat{z}\|^2, \tag{23d}$$

where $K \triangleq \tfrac{1}{2}bL + K'$. Let $\alpha \in (0,1)$, and let $\hat{\rho} \triangleq \min\{\rho, \alpha/K\}$. Then we see that if $z_0 \in B(\hat{z},\hat{\rho})$, then, for all $i > 0$, $z_i \in B(\hat{z},\hat{\rho})$ and relation (23c) holds. The desired result now follows from Theorem 1.2.41*(b)*. □

We will need the following result in proving Lemma 2.9.17 below which shows that the Local MMP Algorithm 2.9.13 produces search directions $h_i = h_i^N + h_i^c$ satisfying $f_\pi(x_i + h_i) < f_\pi(x_i)$.

Proposition 2.9.16. *Suppose that Assumption 2.9.1a is satisfied and that $\hat{x} \in \mathbb{R}^n$ is a strict local minimizer of (2), which together with $\hat{\zeta} \in \mathbb{R}^r$ satisfies the second-order sufficient conditions (5a,b). For any $\pi \geq 0$ and $(x,\zeta) \in \mathbb{R}^n \times \mathbb{R}^r$, let*

$$L_\pi(x,\zeta) \triangleq L(x,\zeta) + \tfrac{1}{2}\pi\|g(x)\|^2. \tag{24a}$$

Then there exist a $\hat{\pi} > 0$ and an $m' \in (0,m]$, where $m > 0$ is as in (5b) such that, for any $\pi \geq \hat{\pi}$, there exists a $\rho > 0$, such that, for all $x \in B(\hat{x},\rho)$ and $\zeta \in B(\hat{\zeta},\rho)$,

$$\langle h, L_{\pi xx}(x,\zeta)h\rangle \geq \tfrac{1}{2}m'\|h\|^2, \quad \forall\, h \in \mathbb{R}^n. \tag{24b}$$

Proof. It follows from Proposition 2.8.1 that there exist a $\hat{\pi} > 0$ and an $m' \in (0,m]$ such that

$$\langle h, L_{\hat{\pi}xx}(\hat{x},\hat{\zeta})h\rangle \geq m'\|h\|^2, \quad \forall\, h \in \mathbb{R}^n. \tag{25a}$$

Let $\pi \geq \hat{\pi}$ be given. Then we see that

$$\langle h, L_{\pi xx}(\hat{x},\hat{\zeta})h\rangle = \langle h, L_{\hat{\pi}xx}(\hat{x},\hat{\zeta})h\rangle + (\pi - \hat{\pi})\|g_x(\hat{x})h\|^2 \geq m'\|h\|^2 \tag{25b}$$

also holds for all $h \in \mathbb{R}^n$. Clearly, if $\underline{\sigma}_\pi(\hat{x},\hat{\zeta})$ is the smallest eigenvalue of $L_{\pi xx}(\hat{x},\hat{\zeta})$, then we must have that $\underline{\sigma}_\pi(\hat{x},\hat{\zeta}) \geq m'$. Since $\underline{\sigma}_\pi(\cdot,\cdot)$ is continuous, it follows that there exists a $\rho > 0$ such that $\underline{\sigma}_\pi(\hat{x},\hat{\zeta}) \geq \tfrac{1}{2}m'$ for all $(x,\zeta) \in B(\hat{x},\rho) \times B(\hat{\zeta},\rho)$, which implies that (24b) holds for for all $(x,\zeta) \in B(\hat{x},\rho) \times B(\hat{\zeta},\rho)$. □

Lemma 2.9.17. *Suppose that Assumption 2.9.1a is satisfied, that $\omega \in (0,1)$ is given, and that $\hat{x} \in \mathbb{R}^n$ is a strict local minimizer of (2) which, together with $\hat{\zeta} \in \mathbb{R}^r$, satisfies the second-order sufficient conditions (5a,b). Let $\hat{\pi} \geq 0$ be as in Proposition 2.9.16, let $\pi \geq \hat{\pi}$ be such that $\pi - \sum_{l=1}^r |\hat{\zeta}^l| = 2\varepsilon > 0$, let $z = (x,\zeta) \in \mathbb{R}^n \times \mathbb{R}^r$, and let $G : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^n \times \mathbb{R}^r$ be defined by (3j) and $f_\pi^0 : \mathbb{R}^n \to \mathbb{R}$ by (17a). If $\{(x_i,\zeta_i)\}_{i=0}^\infty$ is a sequence constructed by the Local MMP Algorithm 2.9.13, in solving (3i), that converges to $(\hat{x},\hat{\zeta})$, then there exists an i^* such that, for all $i \geq i^*$ with $z_i = (x_i,\zeta_i)$ for all i,*

$$\|G(z_{i+1})\| \leq \omega\|G(z_i)\| \tag{26a}$$

and

$$f_\pi^0(x_{i+1}) < f_\pi^0(x_i). \tag{26b}$$

Proof. First we will prove (26a). Making use of the first-order expansion formula (5.1.18a) and the fact that $G(\hat{z}) = 0$, we conclude that for all i,

$$G(z_i) = \int_0^1 G_z(\hat{z} + s(z_i - \hat{z}))\,ds\,(z_i - \hat{z}). \tag{27a}$$

Since there exists a $K' < \infty$ such that

$$\|\textstyle\int_0^1 G_z(\hat{z} + s(z_i - \hat{z})\,ds\| \leq K', \tag{27b}$$

for all i, we conclude, using (23c) in the proof of Theorem 2.9.15, that there must exist an i_1 such that, for all $i \geq i_1$,

$$\|G(z_{i+1})\| \leq K'\|z_{i+1} - \hat{z}\| \leq K'K\|z_i - \hat{z}\|^2. \tag{27c}$$

Next, there must exist an $i_2 \geq i_1$ and a $K'' < \infty$ such that the matrix $\int_0^1 G_z(\hat{z} + s(z_i - \hat{z}))\,ds$ is nonsingular and $\|(\int_0^1 G_z(\hat{z} + s(z_i - \hat{z}))\,ds)^{-1}\| \leq K''$. Hence it follows from (27a) and (27c) that, for all $i \geq i_2$,

$$\|G(z_{i+1})\| \leq K'K\|z_i - \hat{z}\|^2 \leq K''^2K'K\|G(z_i)\|^2. \tag{27d}$$

Since $\|G(z_i)\| \to 0$, as $i \to \infty$, it follows that there exists an $i_3 \geq i_2$ such that $K''^2K'K\|G(z_i)\| \leq \omega$, for all $i \geq i_3$, and hence (26a) holds for all $i \geq i_3$.

Next we turn to (26b). First, it follows from Lemma 2.9.14 and Proposition 2.9.16 that there exist an $i_4 \geq i_3$ and a $\hat{\pi} > 0$ such that for all $i \geq i_4$, both (24b) and (21a,b,c) hold with $x = x_i$ and $\zeta = \zeta_i$. In addition, we assume that i_4 is sufficiently large to ensure that

$$\pi - \sum_{l=1}^r |\zeta_i^l| \geq \varepsilon > 0 \tag{27e}$$

holds for all $i \geq i_4$ with $2\varepsilon \triangleq \sum_{l=1}^r |\hat{\zeta}^l|$. Let $L_\pi(x,\zeta)$ be defined as in (24a).

Then, for all $i \geq i_4$,

$$\begin{aligned} f_\pi^0(x_i + h_i^N + h_i^c) - f_\pi^0(x_i) &\leq f_\pi^0(x_i + h_i^N + h_i^c) - f_\pi^0(x_i) \\ &\quad + L_\pi(x_i + h_i^N + h_i^c, \zeta_i) - L(x_i + h_i^N + h_i^c, \zeta_i) \\ &= \{f_\pi^0(x_i + h_i^N + h_i^c) - L(x_i + h_i^N + h_i^c, \zeta_i)\} \\ &\quad + \{[L_\pi(x_i + h_i^N + h_i^c, \zeta_i) - L_\pi(x_i, \zeta_i)\} \\ &\quad - \{f_\pi^0(x_i) - L_\pi(x_i, \zeta_i)\}. \end{aligned} \tag{27f}$$

We will now examine (27f) term by term. First, making use of (21b), we find that for the first term in braces,

$$\begin{aligned} [f_\pi^0(x_i + h_i^N + h_i^c) - L(x_i + h_i^N + h_i^c, \zeta_i)] &\leq [\pi + \sum_{l=1}^{r} |\zeta_i^l|] \|g(x_i + h_i^N + h_i^c)\|_\infty \\ &\leq K_1 \|h_i^N\|^3, \end{aligned} \tag{27g}$$

where $K_1 \in (0, \infty)$ is such that $K[\pi + \sum_{l=1}^{r} |\zeta_i^l|] \leq K_1$ for all $i \geq i_4$.

Next, we examine the second term in braces. Upon second-order expansion using formula (5.1.18b), we obtain

$$\begin{aligned} &L_\pi(x_i + h_i^N + h_i^c, \zeta_i) - L_\pi(x_i, \zeta_i) \\ &= \langle \nabla_x L_\pi(x_i, \zeta_i), (h_i^N + h_i^c) \rangle + \tfrac{1}{2} \langle (h_i^N + h_i^c), L_{\pi xx}(x_i, \zeta_i)(h_i^N + h_i^c) \rangle \\ &\quad + \int_0^1 (1-s) \langle (h_i^N + h_i^c), [L_{\pi xx}(x_i + s(h_i^N + h_i^c), \zeta_i) - L_{\pi xx}(x_i, \zeta_i)](h_i^N + h_i^c) \rangle ds \\ &\leq \langle \nabla_x L_\pi(x_i, \zeta_i), h_i^N + h_i^c \rangle + \tfrac{1}{2} \langle (h_i^N + h_i^c), L_{\pi xx}(x_i, \zeta_i)(h_i^N + h_i^c) \rangle \\ &\quad + (1/6) K_2 \|h_i^N + h_i^c\|^3, \end{aligned} \tag{27h}$$

where $K_2 < \infty$ is a local Lipschitz constant for $L_{\pi xx}(\cdot, \cdot)$. Hence, in view of (21c) and the fact that $L_{\pi xx}(x_i, \zeta_i)$ and $\|h_i^N\|$ are bounded, there exists a $K_3 < \infty$ such that

$$\begin{aligned} L_\pi(x_i + h_i^N + h_i^c, \zeta_i) - L_\pi(x_i, \zeta_i) &\leq \langle \nabla_x L_\pi(x_i, \zeta_i), h_i^N \rangle \\ &\quad + \tfrac{1}{2} \langle h_i^N, L_{\pi xx}(x_i, \zeta_i) h_i^N \rangle \\ &\quad + K \|\nabla_x L_\pi(x_i, \zeta_i)\| \|h_i^N\|^2 + K_3 \|h_i^N\|^3. \end{aligned} \tag{27i}$$

Since $\nabla_x L_\pi(x_i, \zeta_i) = \nabla_x L(x_i, \zeta_i) + \pi g_x(x_i)^T g(x_i)$, it follows from (20a,b) that

$$\begin{aligned} \langle \nabla_x L_\pi(x_i, \zeta_i), h_i^N \rangle &= \langle \nabla_x L(x_i, \zeta_i), h_i^N \rangle + \pi \langle g(x_i), g_x(x_i) h_i^N \rangle \\ &= -\langle h_i^N, L_{xx}(x_i, \zeta_i) h_i^N \rangle + \langle g(x_i), (\zeta_{i+1} - \zeta_i) \rangle \\ &\quad - \pi \|g_x(x_i) h_i^N\|^2 \\ &\leq -\langle h_i^N, L_{\pi xx}(x_i, \zeta_i) h_i^N \rangle + \|g(x_i)\|_\infty \omega(x_i) \|h_i^N\|^2 \\ &\quad + \langle g(x_i), (\zeta_{i+1} - \zeta_i) \rangle, \end{aligned} \tag{27j}$$

where $\omega(x) \triangleq \pi \sum_{l=1}^{r} \|g_{xx}^l(x)\|$. Hence it follows from (24b) that

$$\begin{aligned} L_\pi(x_i + h_i^N + h_i^c, \zeta_i) - L_\pi(x_i, \zeta_i) &\leq -\tfrac{1}{4} m' \|h_i^N\|^2 + \big(\|g(x_i)\|_\infty \omega(x_i) \\ &\quad + K \|\nabla_x L_\pi(x_i, \zeta_i)\|\big) \|h_i^N\|^2 \\ &\quad + K_3 \|h_i^N\|^3 + \sum_{l=1}^{r} (\zeta_{i+1}^l - \zeta_i^l) g^l(x_i). \end{aligned} \tag{27k}$$

Finally, the last term in braces in (27f) expands as follows:

$$f_\pi^0(x_i) - L_\pi(x_i, \zeta_i) = \pi \|g(x_i)\|_\infty - \sum_{l=1}^{r} \zeta_i^l g^l(x_i) - \frac{\pi}{2} \|g(x_i)\|^2. \tag{27l}$$

Combining (27g), (27k) and (27l), we find that

$$\begin{aligned} f_\pi^0(x_i + h_i^N + h_i^c) - f_\pi^0(x_i) &\leq -\|h_i^N\|^2 \big[\tfrac{1}{4} m' - (K_1 + K_3) \|h_i^N\| \\ &\quad - K \|\nabla_x L_\pi(x_i, \zeta_i)\| - \|g(x_i)\|_\infty \omega(x_i)\big] \\ &\quad - \pi \|g(x_i)\|_\infty + \sum_{l=1}^{r} \zeta_{i+1}^l g^l(x_i) + \frac{\pi}{2} \|g(x_i)\|^2. \end{aligned} \tag{27m}$$

Next for all $i \geq i \geq i_4$, it follows from (27e) that

$$-\pi \|g(x_i)\|_\infty + \sum_{l=1}^{r} \zeta_{i+1}^l g^l(x_i) \leq -(\pi - \sum_{l=1}^{r} |\zeta_{i+1}^l|) \|g(x_i)\|_\infty \leq -\varepsilon \|g(x_i)\|_\infty. \tag{27n}$$

For all $i \in \mathbb{N}$, let

$$\beta_i \triangleq (\tfrac{1}{4} m' - (K_1 + K_3) \|h_i^N\| - \|g(x_i)\|_\infty \omega(x_i) - K \|\nabla_x L_\pi(x_i, \zeta_i)\|). \tag{27o}$$

Then (27m) leads to

$$f_\pi^0(x_i + h_i^N + h_i^c) - f_\pi^0(x_i) \leq -\beta_i \|h_i^N\|^2 - \varepsilon \|g(x_i)\|_\infty + \frac{\pi}{2} \|g(x_i)\|^2. \tag{27p}$$

Because $h_i^N \to 0$, $g(x_i) \to 0$, and $\nabla_x L_\pi(x_i, \zeta_i) \to 0$, as $i \to \infty$, it follows that there must be an $i_5 \geq i_4$ such that, for all $i \geq i_5$, $\beta_i > 0$ and $-\varepsilon \|g(x_i)\|_\infty + \frac{\pi}{2} \|g(x_i)\|^2 \leq 0$. Hence we see that (26b) holds for all $i \geq i_5$, which completes our proof. □

2.9.4 Global MMP Algorithm for (2)

We will now combine the Local MMP Algorithm 2.9.13 with the Exact Penalty Function Algorithm 2.7.17 to produce a global algorithm. This algorithm computes Lagrange multipliers in two ways. The first is as in Algorithm 2.7.17 and uses the function $\zeta : \mathbb{R}^n \to \mathbb{R}^r$, defined by

$$\zeta(x) \triangleq -[g_x(x)g_x(x)^T]^{-1}g_x(x)\nabla f^0(x), \tag{28}$$

which is valid under Assumption 2.9.1a. The second way of computing Lagrange multipliers is as in the Local MMP Algorithm 2.9.13.

The multipliers computed by means of (28) are used in the test for penalty adjustment and for initializing the Local MMP Algorithm. The tests used in the algorithm below, for accepting the update formula of the Local MMP Algorithm 2.9.13, are derived from those in the Stabilized Newton-Armijo Algorithm 1.4.15 and the results in Lemma 2.9.17.

In Lemma 2.9.17, we needed that, with $\pi = \pi_i$, $\pi - \sum_{l=1}^r |\zeta^l(x_i)| \geq \varepsilon > 0$, which can be replaced by the requirement that $\pi_i \geq \sigma \sum_{l=1}^r |\zeta^l(x_i)|$, with $\sigma > 1$. This requirement forms the basis for the the construction in Step 1 below. We also needed that $\langle h_i, L_\pi(x_i, \zeta_i)h_i \rangle \geq \frac{1}{2}m' \|h_i\|^2$ hold for some $m' > 0$. This requirements is addressed by the construction in Step 5†, where the parameter $\pi_{\max}$ is used to prevent the penalty π_i from being driven to infinity when second-order sufficient conditions may not hold at a local minimizer to which the algorithm below is converging.

Finally, the stabilization tests used in Algorithm 2.9.18 below follow the pattern of those in the Stabilized Newton-Armijo Algorithm 1.4.15.

Stabilized MMP Algorithm 2.9.18.

Parameters. $\alpha, \beta, \omega \in (0,1)$, τ, $\sigma > 1$, δ, $\pi_{-1} > 0$, $0 < \varepsilon_0 \ll 1$, $\pi_{\max} > \pi_{-1}$, $c_1, c_2 \gg 1$.

Data. $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$, $\zeta_0 = \zeta(x_0)$, $z_0 = (x_0, \zeta_0)$.

† The reason for using the test in (29e), rather than the test $\langle h_i, L_{\pi,xx}(x_i, \zeta_i)h \rangle \geq \varepsilon_i \|h_i\|^2$, is that the ball, where (24b) holds, may shrink as π increases, which might result in π_i being quickly driven to its upper bound, while this is not necessarily true of the less stringent test in (29e).

Step 1. Compute the smallest integer $j_i \geq 0$ such that

$$\tau^{j_i}\pi_{i-1} - \sigma \sum_{l=1}^r |\zeta^l(x_i)| \geq 0, \tag{29a}$$

and set $\pi^* = \tau^{j_i}\pi_{i-1}$.

Step 2. Construct the matrix

$$G_z(z_i) \triangleq \begin{bmatrix} L_{xx}(x_i, \zeta_i) & g_x(x_i)^T \\ g_x(x_i) & 0 \end{bmatrix}. \tag{29b}$$

If $\operatorname{cond}[G_z(z_i)] > c_1$ or $\|G_z(z_i)^{-1}\| > c_2$, set $\pi_i = \pi^*$, and go to Step 8‡.

Else, go to Step 3.

Step 3. Compute the *Newtonian search direction* (h_i^N, v_i^N) by solving the system of equations

$$\begin{bmatrix} L_{xx}(x_i, \zeta_i) & g_x(x_i)^T \\ g_x(x_i) & 0 \end{bmatrix} \begin{bmatrix} h_i^N \\ v_i^N \end{bmatrix} = \begin{bmatrix} -\nabla_x L(x_i, \zeta_i) \\ -g(x_i) \end{bmatrix}. \tag{29c}$$

Step 4. Compute the Newtonian *feasibility correction direction* h_i^c for (3i), using formula (1.8.9a) in Algorithm 1.8.5, so that

$$h_i^c = \arg\min \{ \|h\|^2 \mid g(x_i + h_i^N) + g_x(x_i)h = 0 \},$$

so that

$$h_i^c = -g_x(x_i)^T[g_x(x_i)g_x(x_i)^T]^{-1}g(x_i + h_i^N). \tag{29d}$$

Step 5. Set $h^* = h_i^N + h_i^c$.

If $\pi^* < \pi_{\max}$ and

$$\langle h^*, L_{xx}(x_i, \zeta(x_i))h^* \rangle + \pi^*\|g_x(x_i)h^*\|^2 < \varepsilon_i \|h^*\|^2, \tag{29e}$$

set $\pi_i = \tau\pi^*$, $\varepsilon_{i+1} = \varepsilon_i/2$, and go to Step 6.

Else, set $\pi_i = \pi^*$, set $\varepsilon_{i+1} = \varepsilon_i$, and go to Step 6.

Step 6. Set $z^* = (x_i + h_i^N + h_i^c, \zeta_i^N + v_i^N)$.

If

$$\|G(z^*)\| \leq \omega \|G(z_i)\|, \tag{29f}$$

‡ The condition number of a square matrix is the ratio of its largest singular value to its smallest singular value.

and

$$f^0_{\pi_i}(x_i + h_i^N + h_i^c) < f^0_{\pi_i}(x_i), \tag{29g}$$

go to Step 7.

Else, go to Step 8.

Step 7. Update: Set

$$x_{i+1} = x_i + h_i^N + h_i^c, \quad \zeta_{i+1} = \zeta_i + v_i^N, \quad z_{i+1} = (x_{i+1}, \zeta_{i+1}), \tag{29h}$$

replace i by $i+1$, and go to Step 1.

Step 8. Compute the PPP *search direction*

$$h_i \triangleq \arg\min_{h \in \mathbb{R}^n} \left\{ \tfrac{1}{2}\|h\|^2 + \langle \nabla f^0(x_i), h \rangle + \max_{l \in 2\mathbf{r}} \pi_i \left[g^l(x_i) - \|g(x_i)\|_\infty + \langle \nabla g^l(x_i), h \rangle \right] \right\} \tag{29i}$$

and the value of the *optimality function*

$$\theta_i \triangleq \min_{h \in \mathbb{R}^n} \left\{ \tfrac{1}{2}\|h\|^2 + \langle \nabla f^0(x_i), h \rangle + \max_{l \in 2\mathbf{r}} \pi_i \left[g^l(x_i) - \|g(x_i)\|_\infty + \langle \nabla g^l(x_i), h \rangle \right] \right\}. \tag{29j}$$

Step 9. Compute the *step-size*

$$\lambda_i = \arg\max_{k \in \mathbb{N}} \{\beta^k \mid f^0_{\pi_i}(x_i + \beta^k h_i) - f^0_{\pi_i}(x_i) - \beta^k \alpha \theta_i \le 0\}. \tag{29k}$$

Step 10. Update: Set

$$x_{i+1} = x_i + \lambda_i h_i, \quad \zeta_{i+1} = \zeta(x_{i+1}), \quad z_{i+1} = (x_{i+1}, \zeta_{i+1}), \tag{29l}$$

replace i by $i+1$, and go to Step 1.

Lemma 2.9.19. *Suppose that Assumption 2.9.1a is satisfied and that $\{x_i\}_{i=0}^\infty$ is a sequence constructed by Algorithm 2.9.13. Let $K \subset \mathbb{N}$ be such that, for each $i \in K$, $j_i > 0$ in (29a), so that $\pi_{i+1} > \pi_i$ for all $i \in K$. If the subsequence $\{x_i\}_{i \in K}$ is infinite, then it has no accumulation points.*

Proof. Let σ, τ, and π_{-1} be as in Algorithm 2.9.18. For $j \in \mathbb{N}$, let $t_j(x) \triangleq -(\tau^j \pi_{-1} - \sigma \sum_{l=1}^r |\zeta^l(x)|)$. Hence, since $\zeta(\cdot)$ is continuous, we see that $t_j(\cdot)$ is continuous.

We will now follow the proof of Theorem 2.3.7 *(c)*. Thus, for the sake of contradiction, suppose that $\{x_i\}_{i \in K}$ is infinite and that it has an accumulation point x^*. Then there must exist an infinite subset $K' \subset K$ such that $x_i \to^{K'} x^*$, as $i \to \infty$, and, because $\zeta(\cdot)$ is continuous, there exists a $\rho^* > 0$ and a $j^* \in \mathbb{N}$ such that $t_j(x) \le 0$ for all $x \in B(x^*, \rho^*)$ and $j \ge j^*$. Clearly, there exists an $i^* \in \mathbb{N}$ such that in the test in Step 1 of Algorithm 2.9.18, $\tau^{j_i}\pi_{i-1} \ge \tau^{j^*}\pi_{-1}$ for all $i \ge i^*$, and therefore $t_{j_i}(x) \le 0$ for all $i \ge i^*$ and $x \in B(x^*, \rho^*)$. Hence, for all $i \ge i^*$, $i \in K$, $x_i \notin B(x^*, \rho^*)$, since otherwise, we would not have $\pi_{i+1} > \pi_i$, and therefore x^* cannot be an accumulation point of $\{x_i\}_{i \in K}$. This completes our proof. □

It should be clear from Lemma 2.9.19 that if the sequence $\{x_i\}_{i=0}^\infty$ constructed by Algorithm 2.9.18 is bounded, then π_i is bounded.

The following theorem has the same form as Theorem 2.7.22, which established the convergence properties of Algorithm 2.7.20.

Theorem 2.9.20. *Consider problem (2), and suppose that Assumption 2.9.1a is satisfied. If Algorithm 2.9.18 constructs a bounded sequence $\{z_i\}_{i=0}^\infty$, then any accumulation point $\hat{z}$ of $\{z_i\}_{i=0}^\infty$ satisfies the first-order optimality condition for (2), i.e.,*

$$G(\hat{z}) = 0. \tag{30}$$

Proof. Since by assumption, the sequence $\{z_i\}_{i=0}^\infty$ is bounded, it follows from Lemma 2.9.19 that there exist an i_0 and $\pi' < \infty$ such that, for all $i \ge i_0$, $\pi_i = \pi'$. Since the sequence $\{z_i\}_{i=0}^\infty$ must have at least one accumulation point $\hat{z}$, and since it follows from (29g,k) that the sequence $\{f^0_{\pi'}(x_i)\}_{i=i_0}^\infty$ is monotone decreasing, it follows from Proposition 5.1.16, that the sequence $\{f^0_{\pi'}(x_i)\}_{i=i_0}^\infty$ converges to $f^0_{\pi'}(\hat{x})$. Now we must consider three cases.

Case 1. There exists an $i_1 \ge i_0$ such that, for all $i \ge i_1$, z_i is constructed by the PPP algorithm, i.e., by Steps 8 - 10. Then the desired result follows directly from Theorem 2.7.22.

Case 2. There exists an $i_2 \ge i_0$ such that, for all $i \ge i_2$, z_i is constructed by the local MMP algorithm, i.e., by Steps 3, 4, and 7. Then, by construction, for all $i \ge i_2$, the following three statements hold:

$$z_{i+1} - z_i = -G_z(z_i)^{-1} G(z_i) + u_i, \tag{31a}$$

where $u_i = (h_i^c, 0)$,

$$\|G(z_{i+1})\| \le \omega \|G(z_i)\|, \tag{31b}$$

and

$$\|G_z(z_i)^{-1}\| \le c_2. \tag{31c}$$

First, it follows from (31c) that for all $i \ge i_2$, $\|h_i^N\| \le c_2 \|G(z_i)\|$. Next, since we

are dealing with a bounded sequence, there must exist constants $0 < c' \le c'' < \infty$ such that for all $i \ge i_2$,

$$\|u_i\| = \|h_i^c\| \le c' \|h_i^N\|^2 \le c'' [\|g(x_i)\| + \|h_i^N\|] \le c'' \|G(z_i)\|^2, \tag{31d}$$

where we have used the fact that, because of the constraints imposed on the construction of (h_i^N, h_i^c), (21c) is valid in this case, and the fact that $\|h_i^N\| \le c_2 \|G(z_i)\|$. Hence, there exists a constant $c_3 < \infty$ such that for, all $i \ge i_2$,

$$\|z_{i+1} - z_i\| \le c_3 \|G(z_i)\| \le c_3 \|G(z_{i_2})\| \omega^{i-i_2}. \tag{31e}$$

Therefore, for any $i \ge i_2$ and integer $k \ge 1$,

$$\|z_{i+k} - z_i\| \le \sum_{j=i}^{\infty} \|z_{j+1} - z_j\| \le c_3 \|G(z_{i_2})\| \frac{\omega^{i-i_2}}{(1-\omega)}, \tag{31f}$$

which shows that the sequence $\{z_i\}_{i=i_2}^{\infty}$ is Cauchy and hence that it converges to some $\hat{z} = (\hat{x}, \hat{\zeta})$. Clearly, $\hat{z}$ satisfies $G(\hat{z}) = 0$.

Case 3. The sequence $\{z_i\}_{i=0}^{\infty}$ contains an infinite number of points constructed by the PPP algorithm and an infinite number of points constructed by the local MMP algorithm. Suppose that there exist a $\hat{z} \in \mathbb{R}^n \times \mathbb{R}^r$ and an infinite subset $K \subset \mathbb{N}$, such that $z_i \to^K \hat{z}$, as $i \to \infty$. Now we will consider two possibilities.

(a) There exists an infinite subset $K' \subset K$ such that, for all $i \in K'$, z_{i+1} is constructed by the PPP algorithm (Steps 8-10). Then we must have that $0 \in \partial f_{\pi'}^0(\hat{x})$. Because, if not, by the properties of the PPP algorithm, there exists an i_3 and a $\hat{\delta} > 0$ such that, for all $i \ge i_3$, $i \in K'$,

$$f_{\pi'}^0(x_{i+1}) - f_{\pi'}^0(x_i) \le -\hat{\delta}, \tag{31g}$$

which contradicts the fact that $f_{\pi'}^0(x_i) \to f_{\pi'}^0(\hat{x})$, as $i \to \infty$. Since it follows from the test in Step 1 that $\pi' \ge \delta \sum_{l=1}^r |\hat{\zeta}^l|$, we conclude that $G(\hat{z}) = 0$.

(b) For all $i \in K$, z_{i+1} is constructed by the local MMP algorithm, i.e., Steps 3, 4 and 7. Let $k : \mathbb{N} \to \mathbb{N}$, be defined by the rule that $k(i)$ is the largest integer such that $k(i) < i$ and $z_{k(i)+1}$ was constructed by the PPP algorithm. Since the sequence $\{z_{k(i)}\}_{i \in K}$ is bounded, it must have at least one accumulation point $z^* = (x^*, \zeta^*)$. By the arguments used for *(a)* above, we conclude that $G(z^*) = 0$. Thus, suppose that $K'' \subset K$ is such that $z_{k(i)} \to^{K''} z^*$, as $i \to \infty$, and consider the sequence $\{z_{k(i)+1}\}_{i \in K''}$. Since, by continuity of the PPP search direction function, $h_{k(i)} \to^{K''} 0$, as $i \to \infty$, we conclude from (29k,l) that $\|x_{k(i)+1} - x_{k(i)}\| \to^{K''} 0$, as $i \to \infty$. Hence, $x_{k(i)+1} \to^{K''} x^*$, as $i \to \infty$. Next, because $\zeta(\cdot)$ is continuous and because $\zeta_{k(i)+1} = \zeta(x_{k(i)+1})$, $\zeta_{k(i)+1} \to^{K''} \zeta(x^*)$, as $i \to \infty$. Since the multiplier ζ^* that satisfies $G(x^*, \zeta^*) = 0$ is unique and is

given by $\zeta^* = \zeta(x^*)$, we conclude that $\zeta_{k(i)+1} \to^{K''} \zeta^*$, as $i \to \infty$. Hence $z_{k(i)+1} \to^{K''} z^*$, as $i \to \infty$, and therefore $G(z_{k(i)+1}) \to^{K''} G(z^*) = 0$, as $i \to \infty$.

Now, for every $i \in K$, either $i = k(i) + 1$ or else z_i was constructed by the local MMP algorithm in j_i iterations, starting from $z_{k(i)+1}$. Referring to our derivation of (31d,e), we conclude that for all $i \in K$,

$$\|z_i - z_{k(i)+1}\| \le \frac{c_3}{1-\omega} \|G(z_{k(i)+1})\|. \tag{31h}$$

Since $G(z_{k(i)+1}) \to^{K''} 0$, as $i \to \infty$, and $z_{k(i)+1} \to^{K''} z^*$, as $i \to \infty$, we conclude from (31h) that $z_i \to^{K''} z^*$, as $i \to \infty$. However, by assumption, because $K'' \subset K$, $z_i \to^{K''} \hat{z}$, as $i \to \infty$ which implies that $z^* = \hat{z}$. Thus we have shown again that $G(\hat{z}) = 0$, which completes our proof. □

The following result is a direct consequence of Lemma 2.9.17 and Theorem 2.9.15.

Corollary 2.9.21. *Consider problem (2) and suppose that Assumption 2.9.1a is satisfied. Suppose that Algorithm 2.9.18 constructs a bounded sequence $\{z_i\}_{i=0}^{\infty}$ which has an accumulation point $\hat{z}$ that satisfies the second-order sufficient conditions (5a,b). Let $K \subset \mathbb{N}$ be such that $x_i \to^K \hat{x}$, as $i \to \infty$. If there exists an $i_0 \in \mathbb{N}$ such that $\text{cond}[G_z(z_i)] \le c_1$ and $\|G_z(z_i)^{-1}\| \le c_2$ for all $i \ge i_0$, $i \in K$ and, in addition, $\pi_i < \pi_{\max}$ for all $i \in \mathbb{N}$, then $\{z_i\}_{i=0}^{\infty}$ converges to $\hat{z}$ Q-quadratically.*

Proof. Since there must be an $i_1 \ge i_0$ such that $\pi_i = \pi^* < \pi_{\max}$, for all $i \ge i_1$, it follows that (29e) is satisfied for all $i \ge i_1$, with $\varepsilon_i = \varepsilon^* > 0$. Since

$$L_{\pi^* xx}(x_i, \zeta_i) = L_{xx}(x_i, \zeta_i) + \pi^* [g_x(x_i)^T g_x(x_i) + \textstyle\sum_{l=1}^r g^l(x_i) g_{xx}^l(x_i)] \tag{32a}$$

and $g(x_i) \to 0$, as $i \to \infty$, it follows from (29e) that there exists an $i_2 \ge i_1$ such that, for some $m' \ge 0$,

$$\langle h_i, L_{\pi^* xx}(x_i, \zeta_i) h_i \rangle \ge \tfrac{1}{2} m' \|h_i\|^2 \tag{32b}$$

holds for all $i \ge i_2$. Hence it follows from Lemma 2.9.17 that, for all $i \ge i_2$, the tests (29f,g) are satisfied and hence that z_{i+1} is constructed according to (29h). It now follows directly from Theorem 2.9.15 that the sequence $\{z_i\}_{i=0}^{\infty}$ converges to $\hat{z}$ Q-quadratically. □

2.9.5 The Maratos-Mayne-Polak-Pang Method for (1)

We will now present, without proofs, the extensions of Algorithm 2.9.13 and Algorithm 2.9.18 to problem (1). The rationale for the extensions is provided by the fact that under Assumption 2.9.1, near a Karush-Kuhn-Tucker triplet

$(\hat{x}, \hat{\eta}, \hat{\xi})$ for (1) satisfying the second-order sufficient conditions stated in Corollary 2.2.30†, the function $G(\cdot)$ defined by (8c) is locally Lipschitz continuously differentiable and $G_z(z)^{-1}$ exists. Furthermore, near such a Karush-Kuhn-Tucker triplet, the Rockafellar Lagrangian defined by (2.8.13i) is twice continuously differentiable, and, as we have already stated, such a Karush-Kuhn-Tucker triplet is a Pang-regular point. Hence we construct our extensions formally by simply replacing the function $G: \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^n \times \mathbb{R}^r$, defined by (3j), that was used in Algorithm 2.9.13 and Algorithm 2.9.18 for problem (2), (in the form (3i)), with the function $G: \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$, defined by (8c).

The obvious extension of Algorithm 2.9.13 to problem (1) is as follows:

The Local MMPP Algorithm 2.9.22.

Data. $z_0 = (x_0, \eta_0, \xi_0) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^r$.

Step 0. Set $i = 0$.

Step 1. Compute the *Pang search direction* $h_i^N = \delta x_i$ by solving the quadratic program (13), together with the corresponding multipliers η_{i+1} and ξ_{i+1}.

Step 2. Compute the Newtonian *feasibility correction direction* h_i^c for constraints in (1), using formula (1.8.9a) in Algorithm 1.8.5, so that

$$h_i^c = \arg\min \{ \|h\|^2 \mid f(x_i + h_i^N) + f_x(x_i)h \leq 0, \\ g(x_i + h_i^N) + g_x(x_i)h = 0 \}. \tag{33}$$

Step 2. Set $x_{i+1} = x_i + h_i^N + h_i^c$, $z_{i+1} = (x_{i+1}, \eta_{i+1}, \xi_{i+1})$, replace i by $i+1$, and go to Step 1.

With regard to Algorithm 2.9.22, Lemma 2.9.14 generalizes as follows:

Lemma 2.9.23. *Suppose that Assumption 2.9.1(i,ii) is satisfied and that the Karush-Kuhn-Tucker triplet $(\hat{x}, \hat{\eta}, \hat{\xi})$ for (1) (satisfying $G(\hat{z}) = 0$, with $G(\cdot)$ as in (8c)) is a Pang-regular point. Then there exists a $\hat{\rho} > 0$ such that*

(a) if $x_i \in B(\hat{x}, \hat{\rho})$ and $\xi_i \in B(\hat{\xi}, \hat{\rho})$, then h_i^N, h_i^c, and v_i^N are all well defined, and

† Strictly speaking, we mean that the multipliers $\hat{\mu}^0 \triangleq 1/(1 + \sum_{j=1}^q \hat{\eta}^j)$, $\hat{\mu}^j \triangleq \hat{\eta}^j \hat{\mu}^0$, $j \in \mathbf{q}$, and $\hat{\zeta} = \hat{\xi}\hat{\mu}^0$ satisfy the conditions in Corollary 2.2.30.

(b) there exists a constant $K < \infty$ such that, with $\Psi: \mathbb{R}^n \to \mathbb{R}$ defined by

$$\Psi(x) \triangleq \max \{ \|f(x)_+\|_\infty, \|g(x)\|_\infty \}, \tag{34a}$$

$$\|\Psi(x_i + h_i^N)\|_\infty \leq K \|h_i^N\|^2, \tag{34b}$$

$$\|\Psi(x_i + h_i^N + h_i^c)\|_\infty \leq K \|h_i^N\|^3, \tag{34c}$$

and

$$\|h_i^c\| \leq K \|h_i^N\|^2. \tag{34d}$$

□

Exercise 2.9.24. Prove Lemma 2.9.23. *Hint*: Use the fact that $(a+b)_+ \leq a_+ + b_+$, Corollary 1.8.4, and the fact that $\|h_i^c\|$ must be smaller than or equal to the norm of any solution of the system

$$f_x(x_i)h \leq -f(x_i)_+, \tag{35a}$$

$$g_x(x_i)h = -g(x_i), \tag{35b}$$

where $f(x)_+ \triangleq (f^1(x)_+, \ldots, f^q(x)_+)$. □

Exercise 2.9.25. Suppose that Assumption 2.9.1*(i,ii)* is satisfied and that $\hat{z} = (\hat{x}, \hat{\eta}, \hat{\xi}) \in \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$ is a Pang-regular vector for the function $G(\cdot)$, defined in (8c), such that $G(\hat{z}) = 0$. Show that there is a $\rho > 0$ such that if any $z_0 = (x_0, \eta_0, \xi_0) \in B(\hat{z}, \rho)$ is used to initialize the Local MMPP Algorithm 2.9.22, then the sequence $\{z_i\}_{i=0}^\infty$, with $z_i = (x_i, \eta_i, \xi_i)$ constructed by the Local MMPP Algorithm 2.9.22, satisfies

$$\|z_{i+1} - \hat{z}\| \leq C \|z_i - \hat{z}\|^2, \tag{36}$$

for some $C \in (0, \infty)$. □

To obtain a generalization of Lemma 2.9.17, we need the following generalization of Proposition 2.9.16:

Exercise 2.9.26. For any $\pi \geq 0$, let $L_\pi: \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r \to \mathbb{R}$ denote the Rockafellar Lagrangian in (2.8.13j), i.e.,

$$L_\pi(x, \eta, \xi) \triangleq f^0(x) + \langle \xi, g(x) \rangle + \frac{\pi}{2} \|g(x)\|^2 \\ + \frac{1}{2\pi} \sum_{j=1}^q [\pi(f^j(x) + \eta^j)_+^2 - (\eta^j)^2]. \tag{37a}$$

Suppose that Assumption 2.9.1 is satisfied and that $\hat{z} = (\hat{x}, \hat{\eta}, \hat{\xi}) \in \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$ is a Karush-Kuhn-Tucker triplet for (1) satisfying the second-order sufficient conditions stated in Corollary 2.2.30.

Show that, under the above assumptions, there exist a $\hat{\pi} > 0$ and an $m' > 0$ such that, for any $\pi \geq \hat{\pi}$, there exists a $\rho > 0$ such that $L_{\pi xx}(x, \eta, \xi)$ is well defined for all $(x, \eta, \xi) \in B(\hat{x}, \rho) \times B(\hat{\eta}, \rho) \times B(\hat{\xi}, \rho)$, and

$$\langle h, L_{\pi xx}(x, \eta, \xi)h \rangle \geq \tfrac{1}{2}m' \|h\|^2, \quad \forall\, h \in \mathbb{R}^n. \tag{37b}$$

□

Using the above results, one can prove the following generalization of Lemma 2.9.17:

Lemma 2.9.27. *Suppose that Assumption 2.9.1 is satisfied and that $\hat{z} = (\hat{x}, \hat{\eta}, \hat{\xi}) \in \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$ is a Karush-Kuhn-Tucker triplet for (1) satisfying the second-order sufficient conditions in Corollary 2.2.30.*

Let $\omega \in (0, 1)$ be given, let $\hat{\pi} \geq 0$ be as in Exercise 2.9.26, let $\pi \geq \hat{\pi}$ be such that $\pi - \sum_{j=1}^{q} \hat{\eta}^j - \sum_{l=1}^{r} |\hat{\xi}^l| \triangleq 2\varepsilon > 0$, let $z = (x, \eta, \xi) \in \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$, let $G : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^n \times \mathbb{R}^r$ be defined by (8a), and let $f_\pi^0 : \mathbb{R}^n \to \mathbb{R}$ be defined by

$$f_\pi^0(x) \triangleq f^0(x) + \pi \max\{\|f(x)_+\|_\infty, \|g(x)\|_\infty\}. \tag{38a}$$

If $\{(x_i, \eta_i, \xi_i)\}_{i=0}^{\infty}$ is a sequence converging to $(\hat{x}, \hat{\eta}, \hat{\xi})$, constructed by the Local MMPP Algorithm 2.9.22, then there exists an i^* such that for all $i \geq i^*$,

$$\|G(z_{i+1})\| \leq \omega \|G(z_i)\|, \tag{38b}$$

where $z_i = (x_i, \eta_i, \xi_i)$, and

$$f_\pi^0(x_{i+1}) < f_\pi^0(x_i). \tag{38c}$$

□

To construct an extension of the Stabilized MMP Algorithm 2.9.18 to problem (1), we replace the exact penalty function $f_\pi^0(\cdot)$ defined in (17a) by the exact penalty function $f_\pi^0 : \mathbb{R}^n \to \mathbb{R}$, with $\pi \geq 0$ defined by (38a), and we replace the function $\zeta(\cdot)$ defined in (28) by the pair of functions $(\mathrm{H}, \Xi) : \mathbb{R}^n \to 2^{\mathbb{R}^q \times \mathbb{R}^r}$ defined by (2.8.13k), i.e.,

$$(\mathrm{H}(x), \Xi(x)) \triangleq \arg\min\{\|\nabla_x L(x, \eta, \xi)\|^2 + \langle \eta, F(x)\eta \rangle \mid \mu^j \geq 0,\ j \in \mathbf{q}\} \tag{39}$$

where $F(x) \triangleq \mathrm{diag}([f^1(x)]^2, \ldots, [f^q(x)]^2)$, i.e., it is a $q \times q$ diagonal matrix with elements $[f^j(x)]^2$. We recall that it was shown in Proposition 2.8.16 that, under Assumption 2.9.1, the functions $\mathrm{H}(\cdot)$ and $\Xi(\cdot)$ are point-valued continuous functions near points $\hat{x}$ that are part of Karush-Kuhn-Tucker triplets satisfying second-order sufficient conditions and that, if $(\hat{x}, \hat{\eta}, \hat{\xi})$ is such a Karush-Kuhn-Tucker triplet for (1), then $\mathrm{H}(\hat{x}) = \{\hat{\eta}\}$ and $\Xi(\hat{x}) = \{\hat{\xi}\}$.

With these results in place and assuming that Assumption 2.9.1 holds, the extension of the Stabilized MMP Algorithm 2.9.18 to problem (1) assumes the following form:

Stabilized MMPP Algorithm 2.9.28.

Parameters. $\alpha, \beta, \omega \in (0, 1)$, τ, $\sigma > 1$, δ, $\pi_{-1} > 0$, $\pi_{\max} > \pi_{-1}$, $0 < \varepsilon_0 \ll 1$, $c \gg 1$.

Data. $x_0 \in \mathbb{R}^n$, $\pi_{-1} > 0$.

Step 0. Set $i = 0$, compute $(\eta_0, \xi_0) \in (\mathrm{H}(x_0), \Xi(x_0))$, and set $z_0 = (x_0, \eta_0, \xi_0)$.

Step 1. Compute a pair $(\eta_*, \xi_*) \in (\mathrm{H}(x_i), \Xi(x_i))$ and the smallest integer $j_i \geq 0$ such that

$$\tau^{j_i}\pi_{i-1} - \sigma\Big(\textstyle\sum_{j=1}^{q} |\mu_*^j| + \sum_{l=1}^{r} |\xi_*^l|\Big) \geq 0, \tag{40a}$$

and set $\pi^* = \tau^{j_i}\pi_{i-1}$.

Step 2. If the Pang quadratic program

$$\min\Big\{ \langle \nabla f^0(x_i), \delta x \rangle + \tfrac{1}{2}\langle \delta x, L_{xx}(x_i, \eta_i, \xi_i)\delta x \rangle \mid f_{\bar{\alpha}(z_i)}(x_i) + (f_{\bar{\alpha}(z_i)})_x(x_i)\delta x = 0,\ f_{\bar{\beta}(z_i)}(x_i) + (f_{\bar{\beta}(z_i)})_x(x_i)\delta x \leq 0,\ g(x_i) + g_x(x_i)\delta x = 0 \Big\} \tag{40b}$$

has a feasible solution, compute a Karush-Kuhn-Tucker triplet $(\delta x_i, \eta_{i+1}, \xi_{i+1})$ for it, with $(\eta_{i+1})_\gamma(z_i) = 0$, set $h_i^N = \delta x_i$, and go to Step 3.

Else go to Step 8.

Step 3. If $\|h_i^N\| > c$, go to Step 8.

Else, go to Step 4.

Step 4. If the quadratic program

$$\min\{\|h\|^2 \mid f(x_i + h_i^N) + f_x(x_i)h \leq 0,\ g(x_i + h_i^N) + g_x(x_i)h = 0\}, \tag{40c}$$

is infeasible, go to Step 8.

Else, solve it for the Newtonian *feasibility correction direction* h_i^c, and go to Step 5.

Step 5. Set $h^* = h_i^N + h_i^c$.

If $\pi^* < \pi_{\max}$ and

$$\langle h^*, L_{xx}(x_i, \eta_i \xi_i) h^* \rangle + \pi^* (\|f_x(x_i) h^*\|^2 + \|g_x(x_i) h^*\|^2) < \varepsilon_i \|h^*\|^2, \quad (40d)$$

set $\pi_i = \tau\pi^*$, $\varepsilon_{i+1} = \varepsilon_i/2$, and go to Step 6.

Else, set $\pi_i = \pi^*$, set $\varepsilon_{i+1} = \varepsilon_i$, and go to Step 6.

Step 6. Set $z^* = (x_i + h_i^N + h_i^c, \eta_{i+1}, \xi_{i+1})$.

If

$$\|G(z^*)\| \leq \omega \|G(z_i)\|, \quad (40e)$$

where $G(\cdot)$ is defined by (8c), and

$$f^0_{\pi_i}(x_i + h_i^N + h_i^c) < f^0_{\pi_i}(x_i), \quad (40f)$$

go to Step 7.

Else, go to Step 8.

Step 7. Update: Set

$$x_{i+1} = x_i + h_i^N + h_i^c, \quad z_{i+1} = (x_{i+1}, \eta_{i+1}, \xi_{i+1}), \quad (40g)$$

replace i by $i+1$, and go to Step 1.

Step 8. Compute the PPP *search direction* (see (2.7.33))

$$h_i \triangleq \arg\min_{h \in \mathbb{R}^n} \left\{ \tfrac{1}{2}\|h\|^2 + \langle \nabla f^0(x_i), h \rangle \right.$$

$$\left. + \max\{\|[f(x_i) + f_x(x_i)h]_+\|_\infty, \|g(x_i) + g_x(x_i)h\|_\infty\} \right\} - \pi\Psi(x_i), \quad (40h)$$

and the value of the value *optimality function*

$$\theta_i \triangleq \min_{h \in \mathbb{R}^n} \left\{ \tfrac{1}{2}\|h\|^2 + \langle \nabla f^0(x_i), h \rangle \right.$$

$$\left. + \max\{\|(f(x_i) + f_x(x_i)h)_+\|_\infty, \|g(x_i) + g_x(x_i)h\|_\infty\} \right\} - \pi\Psi(x_i), \quad (40i)$$

Step 9. Compute the *step-size*

$$\lambda_i = \arg\max_{k \in \mathbb{N}} \{\beta^k \mid f^0_{\pi_i}(x_i + \beta^k h_i) - f^0_{\pi_i}(x_i) - \beta^k \alpha \theta_i \leq 0\}. \quad (40j)$$

Step 10. Update: Set

$$x_{i+1} = x_i + \lambda_i h_i, \quad (40k)$$

compute

$$(\eta_{i+1}, \xi_{i+1}) \in (H(x_{i+1}), \Xi(x_{i+1})), \quad (40l)$$

set $z_{i+1} = (x_{i+1}, \eta_{i+1}, \xi_{i+1})$, replace i by $i+1$, and go to Step 1.

In view of the developments up to this point, we have reason to believe that the following analog of Theorem 2.9.20 and Corollary 2.9.21 is valid:

Theorem 2.9.29. *Consider problem (1), and suppose that Assumption 2.9.1(i,ii) is satisfied. If Algorithm 2.9.28 constructs a bounded sequence $\{z_i\}_{i=0}^\infty$, then any accumulation point $\hat{z}$ of $\{z_i\}_{i=0}^\infty$ satisfies the first-order optimality condition for (1), i.e., $G(\hat{z}) = 0$, with $G(\cdot)$ defined by (8a).* □

Corollary 2.9.30. *Consider problem (1), and suppose that Assumption 2.9.1 is satisfied. If Algorithm 2.9.28 constructs a bounded sequence $\{z_i\}_{i=0}^\infty$ which has an accumulation point $\hat{z}$ that satisfies the second-order sufficient conditions stated in Corollary 2.2.30, then $\{z_i\}_{i=0}^\infty$ converges to $\hat{z}$ Q-quadratically.* □

To obtain an SQP algorithm with the same structure as Algorithm 2.9.28, using the Wilson quadratic search problem (7c), we only need to replace (40b) by (7c) and redefine the function $G: \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r \to \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$ in (40e), as follows:

$$G(z) \triangleq \begin{bmatrix} \nabla_x L(x, [\eta + f(x)]_+, \xi) \\ f(x) - [\eta + f(x)]_- \\ g(x) \end{bmatrix}, \quad (41a)$$

where, for any vector $y \in \mathbb{R}^q$, $y_+ \triangleq (y_+^1, \ldots, y_+^q)$ and $y_- \triangleq y - y_+$. The justification for this is as follows. First, it is easy to verify that $\hat{z} = (\hat{x}, \hat{\eta}, \hat{\xi})$ satisfies $G(\hat{z}) = 0$, with $G(\cdot)$ defined by (41a), if and only if $\hat{z}$ satisfies (7a). Next, if we introduce an artificial variable $y \in \mathbb{R}^q$ to replace $\eta + f(x)$ in (41a), we can define a new function $\mathcal{G}: \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r \to \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^r$ as follows:

$$\mathcal{G}(z) \triangleq \begin{bmatrix} \nabla_x L(x, y_+, \xi) \\ f(x) - y_- \\ g(x) \end{bmatrix}, \quad (41b)$$

where $z \triangleq (x, y, \xi)$. Then we can obtain a first-order expansion of $\mathcal{G}(z)$ about a point z_i, as follows:

$$\mathcal{G}(z) \approx \mathcal{G}(z_i) + \mathcal{G}'(z_i; \delta z), \quad (41c)$$

where $\delta z \triangleq z - z_i$ and

$$\mathcal{G}'(z_i, \delta z) \triangleq \begin{bmatrix} L_{xx}(x_i, (y_i)_+, \xi_i)\delta x + f_x(x_i)^T[y_+ - (y_i)_+] + g_x(x_i)^T \delta\xi \\ f(x_i) + f_x(x_i)\delta x - y_- \\ g(x_i) + g_x(x_i)\delta x \end{bmatrix} . \quad (41d)$$

Exercise 2.9.31. *(a)* Show that, if $\hat{z} = (\hat{x}, \hat{\eta}, \hat{\xi})$ satisfies (7a), then $\hat{z} = (\hat{x}, \hat{y}, \hat{\xi})$ with $\hat{y} \triangleq \hat{\eta} + f(\hat{x})$ satisfies $\mathcal{G}(\hat{z}) = 0$.

Conversely, suppose that $\hat{z} = (\hat{x}, \hat{y}, \hat{\xi})$ is such that $\mathcal{G}(\hat{z}) = 0$. Show that $(\hat{x}, \hat{\eta}, \hat{\xi})$ with $\hat{\eta} = \hat{y}_+$ satisfies (7a).

(b) Show that, if $(\delta x_i, \eta_{i+1}, \xi_{i+1})$ is a Karush-Kuhn-Tucker triplet for (7c), i.e., it satisfies (7d), and we define $z_i \triangleq (x_i, y_i, \xi_i)$ with $(y_i)_+ \triangleq \eta_i$, then

$$z_{i+1} \triangleq (x_i + \delta x_i, \eta_{i+1} + [f(x_i) + f_x(x_i)\delta x_i], \xi_{i+1}) \quad (42a)$$

satisfies

$$\mathcal{G}(z_i) + \mathcal{G}'(z_i; z_{i+1} - z_i) = 0 \quad (42b)$$

Conversely, suppose that $z_{i+1} = (x_{i+1}, y_{i+1}, \xi_{i+1})$ satisfies

$$\mathcal{G}(z_i) + \mathcal{G}'(z_i; z_{i+1} - z_i) = 0 . \quad (42c)$$

Show that, for $z_i \triangleq (x_i, \eta_i, \xi_i)$ with $\eta_i = (y_i)_+$, and $\delta x_i \triangleq x_{i+1} - x_i$, $(\delta x_i, (y_{i+1})_+, \xi_{i+1})$ satisfies (7d). □

In view of the above, it is clear that the Wilson update-generating problem (7c) produces a solution to the problem $\mathcal{G}(z_i) + \mathcal{G}'(z_i; z) = 0$ and hence results in a Newton-like method for solving the equation $\mathcal{G}(z) = 0$. It can be deduced from the literature that Theorem 2.9.29 and Corollary 2.9.30 remain valid when the function $G(\cdot)$ defined by (8c) is replaced by the function $G(\cdot)$ defined by (41a) and the update-generating problem (40b) is replaced by the Wilson update-generating problem (8c).

2.9.6 Notes

Josephy [Jos.79] was the first to relate Wilson's method to a method for solving nonsmooth equations and to show that it converges quadratically in the vicinity of a Karush-Kuhn-Tucker triplet satisfying second-order sufficient conditions for (1). Since his work was never published in the open literature, a number of his important results were summarized in [Rob.93]. There is now considerable literature on nonsmooth equations; see, e.g., [Koj.80, Pan.91, QiL.93, Rob.93, Rob.94, Ral.94]. Most of the important applications are to complementarity and variational inequality problems. As we have seen, the transcription of the necessary conditions for (1) into a nonsmooth equation, such as (8a), is not unique. The transcription in (41a) can be found in [Koj.80, Rob.93]; their results must be employed to justify the use of (7c) in conjunction with our global stabilization scheme.

Han [Han.77] was the first to use the widely accepted exact nondifferentiable penalty function as a merit function. Another possibility is to use a nondifferentiable augmented Lagrangian, as in [CDT.84, CGT.88, CGT.88a, Bon.89, Lec.91], or an exact differentiable penalty function, as in [PoY.86].

The feasibility correction scheme proposed in [MaP.79, MaP.82] and appearing in (20c), is not the only one designed for dealing with the Maratos effect; see, for example [Gab.82, CLP.82, CGT.88a, Lec.91, Fle.82, Fle.87, Yua.85, BSS.87, PaT.91, BPT.92]. For a nice discussion of the Maratos effect with illustrating examples, see [Pow.86].

As we have mentioned earlier in this section, many of the SQP algorithms in the literature admit BFGS type approximations to the Hessian of the Lagrangian. Monotone and nonmonotone line searches as well as trust regions are used for global stabilization. For a range of possibilities, see [Bur.89, BuH.89, Bur.92, BSS.87, CDT.84, CGT.88, CLP.82, Fle.87, Han.77, Lec.91, MaP.82, Pan.91, PoY.91, PaT.91, BPT.92, QiL.93, Ral.94, Yua.85].

Of particular interest to engineers are the SQP methods reported in [Her.86, PTH.88, PaT.91, PaT.93, BPT.92], for problems with inequality constraints only, because once a feasible point is found, feasibility is preserved for all the subsequent iterations. These have led to the public domain SQP package CFSQP, available from the web site http://www.isr.umd.edu/Labs/CACSE/FSQP/fsqp.html. The algorithms in [Her.86, PTH.88] are QP-free, i.e., they solve linear systems of equations rather than quadratic programs at each iteration.

Finally, the Stanford SQP software code described in [GMS.86] is among the most efficient currently available.

Then define $\tilde{p}^{\eta}_{N,k} : [0, k/N] \to \mathbb{R}^n$ by

$$\tilde{p}^{\eta}_{N,k}(t) \triangleq \sum_{j=0}^{k-1} p^{\eta}_{N,k}((j+1)/N)\hat{\pi}_{N,j}(t), \tag{46b}$$

and rewrite (35a) in the form

$$\begin{aligned} D^2 f_{N,k}(\eta ; \delta\eta, \delta\eta) &= \langle F_{\xi\xi}(\xi, x^{\eta}_N(k/N))\delta\xi, \delta\xi\rangle \\ &+ 2\langle F_{\xi x}(\xi, x^{\eta}_N(k/N))\delta\xi, Dx^{\eta}_N(k/N ; \delta\eta)\rangle \\ &+ \int_0^{(k-1)/N} \langle \mathcal{H}_{xx}(x^{\eta}_N(s), u(s), \tilde{p}^{\eta}_{N,k}(s))Dx^{\eta}_N(s ; \delta\eta), Dx^{\eta}_N(j/N ; \delta\eta)\rangle\, ds \\ &+ 2\int_0^{(k-1)/N} \langle \mathcal{H}_{xu}(x^{\eta}_N(s), u(s), \tilde{p}^{\eta}_{N,k}(s))Dx^{\eta}(s ; \delta\eta), \delta u(s)\rangle\, ds \\ &+ \int_0^{(k-1)/N} \langle \mathcal{H}_{uu}(x^{\eta}_N(s), u(s), \tilde{p}^{\eta}_{N,k}(s))\delta u(s), \delta u(s)\rangle\, ds . \end{aligned} \tag{46c}$$

□

5.6.5 Notes

It is difficult to find books that deal with all the material in this section. Existence and uniqueness results can be found in every textbook dealing with differential equations; see e.g., [Pon.62]. A more general version of the Picard Lemma 5.6.3 that deals with differential inclusions can be found in [Cla.83]; see Theorem 3.1.6. Differentiability of solutions with respect to parameters in $\mathbb{R}^m$ can be found in [Pon.62]. The earliest results on differentiability of solutions of differential equations with respect to controls can be found in [McS.44], though this book is very hard to read. More accessible is the treatment in [Lan.83] and [ATF.87]. Approximation results dealing with Euler's method of integration as well as other numerical integration techniques can be found in [Atk.89, But.87].

The space $L^m_{\infty,2}[0, 1]$ was introduced by the author, around 1989, in [PoH.89] and in an early draft [BaP.90] of the paper [BaP.94], where eventually it was abandoned in favor of additional growth assumptions on the function $h(\cdot, \cdot)$ in (1a). It was used again in [PoH.92] and in [Pol.93]. Apart from making it easy to relate the approximations to optimal control problems defined on finite-dimensional Hilbert spaces to the original optimal control problems, it eliminates the need for the use of two norms in deriving second-order conditions for optimal control problems, as developed in [Iof.79, Mau.81].

Bibliography

[Al-.85] M. Al-Baali, "Descent property and global convergence of the Fletcher-Reeves method with inexact line searches," *IMA J. Numerical Analysis.*, Vol. 5, pp. 121-124, 1985.

[AlP.88] Q. Al-Hassan and A. Poore, "The expanded Lagrangian system for constrained optimization problems," *SIAM J. Control and Optimization*, Vol. 26, No. 2, pp. 417-427, 1988.

[AMR.88] U. Ascher, R. Mattheij, and R. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[Ani.89] V. V. Anisovich, "Design of optimal control in a minimax problem, *Automation and Remote Control USSR*, Vol. 50, No. 7, pp. 904-907, 1989.

[Ani.90] V. V. Anisovich, "Synthesis of optimal control in nonlinear minimax problem," *Automation and Remote Control USSR*, Vol. 51, No. 10, pp. 1322-1326, 1990.

[AnL.89] E. J. Anderson and A. S. Lewis, "An extension of the simplex algorithm for semi-infinite linear programming," *Mathematical Programming*, Vol. 44, No. 3, pp. 247-269, 1989.

[AnP.94] E. J. Anderson and A. B. Philpott, "On the solutions of a class of continuous linear programs," *SIAM J. Control and Optimization*, Vol. 32, No. 5, pp. 1289-1296, 1994.

[Apo.60] T. M. Apostol, *Mathematical Analysis*, Addison-Wesley, Reading, MA, 1960.

[Arm.66] L. Armijo, "Minimization of functions having Lipschitz continuous first-partial derivatives," *Pacific J. Mathematics*, Vol.16, pp. 1-3, 1966.

[AsB.86] U. Ascher and G. Bader, "Stability of collocation at Gaussian points," *SIAM J. Numerical Analysis*, Vol. 23, No. 2, pp. 412-422, 1986.

[AsK.88] M. D. Asic and V. V. Kovacevic-Vujcic, "An interior semi-infinite programming method," *J. Optimization Theory and Applications*, Vol. 59, No. 3, pp. 353-67, 1988.

[ATF.87] B. M. Alekseev, V. M. Tikhomirov, and S. V. Fomin, *Optimal Control*, Consultants Bureau, New York and London, 1987.

[Atk.89] K. E. Atkinson, *An Introduction to Numerical Analysis*, (2nd ed.), John Wiley & Sons, New York, 1989.

[Att.79] H. Attouch, "Familles d'opérateurs maximaux monotones et mesurabilité," *Annales Mathematiques Pures et Appliqueès*, Vol. 120, pp. 35-111, 1979.

[AtW.81] H. Attouch and R. J-B Wets, "Approximation and convergence in nonlinear optimization," in *Nonlinear Programming 4*, O. Mangasarian, R. R. Meyer, and S. Robinson, eds., Academic Press, New York, pp. 367-394, 1981.

[AuF.90] J. P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhauser, Boston, 1990.

[Aus.72] A. Auslender, *Problemes de Minimax via l'Analyse Convexe et les Inegalites Variationnelles: Theorie et Algorithmes*, Springer-Verlag, Berlin, New York, 1972.

[BaB.82] J. Barzilai and A. Ben-Tal, "Nonpolynomial and inverse interpolation for line searches," *SIAM J. Numerical Analysis*, Vol. 19, pp. 1263-1277, 1982.

[BaG.82] M. Bazara and J. Good, "Sufficient conditions for a globally exact penalty function without convexity," *Mathematical Programming Study 19*, pp. 1-15, 1982.

[BaP.90] T. E. Baker and E. Polak, "An Algorithm for Optimal Slewing of Flexible Structures," University of California, Electronics Research Laboratory, Memo UCB/ERL M89/37, 11 April 1989, Revised, 4 June 1990.

[BaP.94] T. E. Baker and E. Polak, "On the optimal control of systems described by evolution equations," *SIAM J. Control and Optimization*, Vol. 32, No. 1, pp. 224-260, 1994.

[Bar.93] E. N. Barron, "Averaging in Lagrange and minimax problems of optimal control," *SIAM J. Control and Optimization*, Vol. 31, No. 6, pp. 1630-1652, 1993.

[BaS.77] P. Baptist and J. Stoer, "On the relation between quadratic termination and convergence properties of minimization algorithms, Part II, Applications," *Numerische Mathematik*, Vol. 28, pp. 367-392, 1977.

[BaS.92] R. G. Bartle, and D. R. Sherbert, *Introduction to Real Analysis*, 2nd ed., John Wiley & Sons, New York, 1992.

[BBS.69] B. M. Budak, E. M. Berkovich and E. N. Solov'eva, "Difference approximations in optimal control problems," *SIAM J. Control*, Vol. 7, pp. 18-31,1969.

[BeH.93] J. T. Betts and W. P. Huffman, "Path-constrained trajectory optimization using sparse sequential quadratic programming." *J. Guidance, Control, and Dynamics*, Vol. 16, No. 1, pp. 59-68, 1993.

[Bel.57] R. E. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

[Bel.90] B. M. Bell, "Global convergence of a semi-infinite optimization method," *J. Applied Mathematics and Optimization*, Vol. 21, No.1, pp. 69-88, 1990.

[Ber.63] C. Berge, *Topological Spaces*, Macmillan Co., New York, 1963.

[Ber.74] D. P. Bertsekas, "Partial conjugate gradient methods for a class of optimal control problems," *IEEE Trans. on Automatic Control*, Vol. 19, pp. 209-217, 1974.

[Ber.75] D. P. Bertsekas, "Necessary and sufficient conditions for a penalty method to be exact," *Mathematical Programming*, Vol. 9, pp. 87-99, 1975.

[Ber.82] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[Ber.82a] D. P. Bertsekas, "Projected Newton methods for optimization problems with simple constraints," *SIAM J. Control and Optimization*, Vol. 20, pp. 221-246, 1982.

[Ber.95] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1995.

[BLN.87] R. H. Byrd, D. C. Liu, and J. Nocedal, "Global convergence of a class of quasi-Newton methods on convex problems," *SIAM J. Numerical Analysis*, Vol. 24, pp. 1171-1189, 1987.

[BMP.91] R. Bulrisch, F. Montrose, and H. J. Pesch, "Abort landing in the presence of windshear as a minimax optimal control problem Part 2: Multiple shooting and homotopy," *J. Optimization Theory and Applications*, Vol. 70, No. 2, pp. 223-254, 1991.

[BNP.93] R. Bulirsch, E. Nerz, H. J. Pesch, and O. Stryk, "Combining direct and indirect methods in optimal control: range maximization of a hang glider," *International Series of Numerical Mathematics*, Vol. 111, pp. 273-288, 1993.

[BNY.87] R. H. Byrd, J. Nocedal, and Y. Yuan, "Global convergence of a class of quasi-Newton methods on convex problems," *SIAM J. Numerical Analysis*, Vol. 24, pp. 1171-1189, 1987.

[Bon.89] J. F. Bonnans, "Asymptotic admissibility of the unit step size in exact penalty methods," *SIAM J. Control and Optimization*, Vol. 27, pp. 631-641, 1989.

[Bon.90] J. F. Bonnans, "Theorie de la penalization exacte," *Mathematical Modeling and Numerical Analysis*, Vol. 24, pp. 197-210, 1990.

[Boo.75] W. M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.

[Bou.32] G. Bouligand, "Sur la semi-continuife d'inclusions et quelque sujets connexes," *Enseignement Mathematique*, Vol. 31, pp. 14-22, 1932.

[BPT.92] J. F. Bonnans, E. R. Panier, A. L. Tits and J. L. Zhou, "Avoiding the Maratos effect by means of a nonmonotone line search II. Inequality constrained problems - feasible iterates," *SIAM J. Numerical Analysis*, Vol. 29, No. 4, pp. 1187-1202, 1992.

[Bre.72] R. P. Brent, *Algorithms for Minimization without Derivatives*, Prentice-Hall, Englewood Cliffs, NJ, 1972.

[Bre.83] H. Brezis, *Analyse Fonctionelle: Theorie et Applications*, 3rd ed., Masson, Paris, 1983.

[Bro.65] C. G. Broyden, "A class of methods for solving nonlinear simultaneous equations," *Mathematics of Optimization*, Vol. 19, pp. 577-593, 1965.

[Bro.67] C. G. Broyden, "Quasi-Newton methods and their application to function minimization," *Mathematics of Optimization*, Vol. 21, pp. 368-81, 1967.

[Bro.70] C. G. Broyden, "The convergence of a class of double-rank minimization algorithms," Parts I and II, *J. Institute of Mathematics and its Applications*, Vol. 6, pp. 76-90 and 222-236, 1970.

[BSS.87] R. H. Byrd, R. B. Schnabel, and G. A. Shultz, "A trust region algorithm for nonlinearly constrained optimization," *SIAM J. Numerical Analysis*, Vol. 24, pp. 1152-1170, 1987.

[BuH.66] Bui-Trong-Liu and P. Huard, "La methode des centres dans les espaces topologiques," *Numerische Mathematik*, Vol. 8, pp. 65-67, 1966.

[BuH.89] J. V. Burke and S. P. Han, "A robust sequential quadratic programming method," *Mathematical Programming*, Vol. 43, pp. 277-303, 1989.

[BuL.83] A. Buckley and A. LeNir, "QN-like variable storage conjugate gradients," *Mathematical Programming*, Vol. 27, pp. 155-175, 1983.

[Bur.89] J. V. Burke, "A sequential quadratic programming method for potentially infeasible mathematical programs," *J. Mathematical Analysis and Applications*, Vol. 139, pp. 319-351, 1989.

[Bur.91] J. V. Burke, "An exact penalization viewpoint of constrained optimization," *SIAM J. Control and Optimization*, Vol. 29, No. 2, pp. 493-497, 1991.

[Bur.91a] J. V. Burke, "Calmness and exact penalization," *SIAM J. Control and Optimization*, Vol. 29, pp. 968-998, 1991.

[Bur.92] J. V. Burke, "A robust trust region method for constrained nonlinear programming problems," *SIAM J. Optimization*, Vol. 2, pp. 277-303, 1992.

[But.87] A J. C. Butcher, *The Numerical Analysis of Ordinary Differential Equations*, John Wiley & Sons, London, 1987.

[ByN.89] R. H. Byrd and J. Nocedal, "A tool for the analysis of quasi-Newton methods with applications to unconstrained optimization," *SIAM J. Numerical Analysis*, Vol. 26, No.3, pp. 727-739, 1989.

[CaC.68] M. D. Canon and C. D. Cullum, "A tight upper bound on the rate of convergence of the Frank-Wolfe algorithm," *SIAM J. Control*, Vol. 6, No. 4, 1968.

[Cau.847] A. Cauchy, "Methode génerale pour la résolution des systèms d'equations simultanées', *C. R. Acad. Sci. Paris*, Vol. 25, pp. 536-538, 1847.

[CCP.66] M. D. Canon, C. D. Cullum, and E. Polak, "Constrained minimization problems in finite dimensional spaces," *SIAM J. Control*, Vol. 4, No. 3, pp. 528-547, 1966.

[CCP.70] M. D. Canon, C. D. Cullum, and E. Polak, *Theory of Optimal Control and Mathematical Programming*, McGraw-Hill Co., New York, 1970.

[CDL.93] G. Contaldi, G. Di Pillo and S. Lucidi, "A continuously differentiable exact penalty function for nonlinear programming problems with

unbounded feasible set," *Operations Research Letters*, Vol. 14, No.3, pp. 153-161, 1993.

[CDT.84] M. R. Celis, J. E. Dennis, and R. Tapia, "A trust region strategy for nonlinear equality constrained optimization," in *Numerical Optimization*, P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., SIAM, Philadelphia, PA, pp. 71-82, 1984.

[CGT.88] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, "Global convergence of a class of trust region algorithms for optimization with simple bounds," *SIAM J. Numerical Analysis*, Vol. 25, pp. 433-460, 1988.

[CGT.88a] A. R. Conn, N. I. M. Gould and Ph. L. Toint, "Testing a class of methods for solving minimization problems with simple bounds on variables," *J. Mathematics of Computation*, Vol. 50, pp. 399-430, 1988.

[CGT.91] A. R. Conn, N. I. M. Gould and Ph. L. Toint, "Convergence of quasi-Netwon matrices generated by the symmetric rank one update," *Mathematical Programming*, Vol. 50, pp. 177-195, 1991.

[ChC.78] C. Charalambous and A. R. Conn, "An efficient method to solve the minimax problem directly," *SIAM J. Numerical Analysis*, Vol. 15, pp. 162-187, 1978.

[ChG.59] E. W. Cheney and A. A. Goldstein, "Newton's Method for convex programming and Tchebycheff approximation," *Numerische Mathematik*, Vol. I, pp. 253-268, 1959.

[Cla.83] F. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.

[CLP.82] R. M. Chamberlain, C. Lemarechal, H. C. Pederson, and M. J. D. Powell, "The watchdog technique for forcing convergence in constrained optimization," *Mathematical Programming Study 16*, pp. 1-17, 1982.

[CoC.82] T. F. Coleman and A. R. Conn, "Nonlinear programming via exact penalty functions: asymptotic analysis," *Mathematical Programming*, Vol. 24, pp. 123-136, 1982.

[CoG.87] A. R. Conn and N. I. M. Gould, "An exact penalty function for semi-infinite programming," *Mathematical Programming*, Vol. 37, pp. 19-40, 1987.

[Coh.72] A. I. Cohen, "Rate of convergence of several conjugate gradient algorithms". *SIAM J. Numerical Analysis*, Vol. 9, pp. 248-259, 1972.

[Con.73] A. R. Conn, "Constrained optimization using a nondifferentiable penalty function," *SIAM J. Numerical Analysis*, Vol. 10, pp. 769-784, 1973.

[Con.81] A. R. Conn, "Penalty function methods," in *Nonlinear Optimization*, M. J. D. Powell, ed., Academic Press, New York, pp. 235-242, 1981.

[CoP.77] A. R. Conn and T. Pietrzykowski, "A penalty function method converging directly to a constrained optimum," *SIAM J. Numerical Analysis*, Vol. 14, pp. 348-378, 1977.

[Cou.43] R. Courant, "Variational methods for the solution of problems of equilibrium and vibrations," *Bull. American Mathematical Society*, Vol. 49, pp. 1-23, 1943.

[CoW.87] I. D. Coop and G. A. Watson, "A projected Lagrangian algorithm for semi-infinite programming," *Mathematical Programming*, Vol. 32, pp. 257-376, 1987.

[CrW.72] H. P. Crowder and P. Wolfe, "Linear convergence of the conjugate gradient method," *IBM J. Research and Development*, Vol. 16, pp. 431-433, 1972.

[CuB.79] J. Cullum and R. K. Brayton, "Some remarks on the symmetric rank-one update," *J. Optimization Theory and Applications*, Vol. 29, No. 4, pp. 493-519, 1979.

[CuB.87] J. E. Cuthrell and L. T. Biegler, "On the optimization of differential-algebraic process systems." *AIChE Journal*, Vol. 3, No. 1/2, pp. 1257-1270, 1987.

[CuB.89] J. E. Cuthrell and L. T. Biegler, "Simultaneous optimization and solution methods for batch reactor control profiles." *Computers and Chemical Engineering*, Vol. 13, pp. 49-62, 1989.

[Cul.69] J. Cullum, "Discrete approximations to continuous optimal control problems," *SIAM J. Control*, Vol. 7, pp. 32-49, 1969.

[Cul.71] J. Cullum, "An explicit procedure for discretizing continuous, optimal control problems," *J. Optimization Theory and Applications*, Vol. 8, pp. 15-35, 1971.

[Cul.72] J. Cullum, "Finite-dimensional approximations of state-constrained continuous optimal control problems," *SIAM J. Control*, Vol. 10, pp. 649-670, 1972.

[DaM.81] V. A. Daugavet and V. N. Malozemov, "Quadratic rate of convergence of a linearization method for solving discrete minimax problems," *J U.S.S.R. Computational Mathematics and Mathematical Physics*, Vol. 21, No. 4, pp. 19-28, 1981.

[Dan.66] J. M. Danskin, "The theory of minmax with applications," *SIAM J. Applied Mathematics*, Vol. 14, pp. 641-655, 1966.

[Dan.67] J. M. Danskin, *The Theory of Max-Min and its Application to Weapons Allocation Problems*, Springer-Verlag, Berlin, New York, 1967.

[Dan.69] J. W. Daniel, "On the approximate minimization of functionals," *Mathematics of Computation*, Vol. 23, pp. 573-582, 1969.

[Dan.71] J. W. Daniel, *The Approximation Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[Dan.73] J. W. Daniel, "The Ritz-Galerkin method for abstract optimal control problems." *SIAM J. Control*, Vol. 11, No. 1, pp. 53-63, 1973.

[Dav.59] W. C. Davidon, "Variable metric methods for minimization," AEC Research and Development Rept ANL 5990 (Rev.), 1959.

[Dav.68] W. C. Davidon, "Variance algorithms for minimization," *Computer Journal*, Vol. 10, pp. 406-410, 1968.

[DeF.75] E. DeGiorgi and T. Franzoni, "Su un tipo di convergenza variazionale," *Atti Acad. Naz. Lincei Rend. Cl. Sc. Fis. Mat. Natur.*, Vol. 58, No.8, pp. 842-850, 1975.

[Dem.66] V. F. Demyanov, "On the solution of certain minimax problems I," *Cybernetics*, Vol. 2, No. 6, pp. 47-53, 1966.

[Dem.67] V. F. Demyanov, "On the solution of certain minimax problems II," *Cybernetics*, Vol. 3, No. 3, pp. 51-54, 1967.

[Dem.68] V. F. Demyanov, "Differentiability of a maxmin function," *U.S.S.R. Computational Mathematics and Mathematical Physics*, Vol. 8, no. 6, pp. 1-15, 1968.

[DeM.74] V. F. Demyanov and B. N. Malozemov, *Introduction to Minimax*, John Wiley & Sons, New York, 1974.

[DeM.74a] J. E. Dennis and J. J. Moré, "A characterization of superlinear convergence and its application to quasi-Newton methods," *Mathematics of Computation*, Vol. 28, pp. 549-560, 1974.

[DeM.77] J. E. Dennis and J. J. Moré, "Quasi-Newton methods, motivation and theory," *SIAM Review*, No. 19, pp. 46-84, 1977.

[DeS.83] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[DeS.89] J. E. Dennis, Jr. and R. B. Schnabel, "A view of unconstrained optimization," in *Handbooks in Operations Research and Management Science, Vol. 1: Optimization*, G. L. Nemhauser, A. H. G. Rinnooy Kan and M. J. Todd, eds., North Holland, Amsterdam, pp. 1-66, 1989.

[DeV.81] V. F. Demyanov and L. V. Vasilyev, *Nondifferentiable Optimization* (in Russian), Nauka, Moscow, 1981.

[Die.60] J. Dieudonne, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[DGL.83] G. Di Pillo, L. Grippo and F. Lampariello, "A class of structured quasi-Newton algorithms for optimal control problems." *Proc. IFAC Appl. of Nonlinear Prog. to Optim. and Control*, Palo Alto, CA, 1983.

[DGL.93] G. Di Pillo, L. Grippo and S. Lucidi, "A smooth method for the finite minimax problem," *Mathematical Programming*, Vol. 60, No. 2, pp. 187-214, 1993.

[DHP.95] A. S. Dontchev, W. W. Hager, A. B. Poore, and B. Yang, "Optimality, stability and convergence in nonlinear control," preprint, June 1992. *J. Applied Mathematics and Optimization*, Vol. 31, pp. 297-326, 1995.

[DiG.85] G. Di Pillo and L. Grippo, "A continuously differentiable exact penalty function for nonlinear programming with with inequality constraints," *SIAM J. Control and Optimization*, Vol. 23, pp. 72-84, 1985.

[DiG.86] G. Di Pillo and L. Grippo, "An exact penalty function method with global convergence properties for nonlinear programming problems," *Mathematical Programming*, Vol. 36, pp. 1-18, 1986.

[DiG.89] G. Di Pillo and L. Grippo, "Exact penalty functions in constrained optimization," *SIAM J. Control and Optimization*, Vol. 27, No. 6, pp. 1333-1360, 1989.

[DiP.94] G. Di Pillo, "Exact penalty methods," in *Algorithms for Continuous Optimization: the State of the Art*, E. Spedicato, ed., Kluwer Academic Publishers, Dordrecht, 1994.

[Dix.72] L. C. D. Dixon, "Quasi-Newton algorithms generate identical points," Parts I and II, *Mathematical Programming*, Vol.2 pp. 383-387, Vol. 3. pp. 345-358, 1972.

[Dix.75] L. C. D. Dixon, "Conjugate gradient algorithms: finite termination without linear searches," *J. Institute of Mathematics and its Applications*, Vol. 15, pp. 9-18, 1975.

[Dmi.94] A. V. Dmitruk, "Second-order optimality conditions for singular extremals," *International Series of Numerical Mathematics*, Vol. 115, pp. 71-81, 1994.

[DoZ.93] A. S. Dontchev and T. Zolezzi, *Well-Posed Optimization Problems*, Springer-Verlag, Berlin, New York, 1993.

[DuB.89] J. C. Dunn and D. P. Bertsekas, "Efficient dynamic programming implementations of Newton's method for unconstrained optimal control," *J. Optimization Theory and Applications*, Vol. 83, No. 1, pp. 23-38, 1989.

[Dun.88] J. C. Dunn, "A projected Newton method for minimization problems with nonlinear inequality constraints," *Numerische Mathematik*, Vol. 53, pp. 377-409, 1988.

[Dun.95] J. C. Dunn, "Second-order optimality conditions in sets of L^∞ functions with range in a polyhydron," *SIAM J. Control and Optimization*, Vol. 33, No. 5, pp. 1603-1635, 1995.

[DuT.92] J. C. Dunn and T. Tian, "Variants of Kuhn-Tucker sufficient conditions in cones of nonnegative functions," *SIAM J. Control and Optimization*, Vol. 30, pp. 1361-1384, 1992.

[EaZ.71] B. C. Eaves and W. I. Zangwill, "Generalized cutting plane algorithms," *SIAM J. Control and Optimization*, Vol. 5, pp. 529-542, 1971.

[EdP.76] E. R. Edge and W. F. Powers, "Function-space quasi-Newton algorithms for optimal control problems with bounded controls and singular arcs," *J. Optimization Theory and Applications*, Vol. 20, pp. 455-479, 1976.

[EkT.76] I. Ekeland and R. Temam, *Convex Analysis and Variational problems*, American Elsevier Pub. Co., Inc., New York, 1976.

[ElM.75] J. Elzinga and T. G. Moore, "A central cutting plane algorithm for the convex programming problem," *Mathematical Programming*, Vol. 8, pp. 134-145, 1975.

[ElR.95] G.N. Elnagar and M. Razzaghi, "Solution of linear two-point boundary value problems via a collocation method and application to optimal control," *International J. Computer Mathematics*, Vol. 55, No. 1-2, pp. 105-11, 1995.

[Ere.66] I. I. Eremin, "The penalty method in convex programming," *Soviet Mathematical Doklady*, Vol. 8, pp. 459-462, 1966.

[Fan.52] K. Fan, "Fixed point and minimax theorems in locally convex topological linear spaces," *Proc. Nat. Acad. Sci. USA*, Vol. 38, pp. 121-126, 1952.

[Fan.53] K. Fan, "Minimax theorems," *Proc. Nat. Acad. Sci. USA*, Vol. 39, pp. 42-47, 1953.

[Fan.64] K. Fan, "Sur an theorem minimax," *C. R. Acad. Sci. Paris*, Vol. 259, pp. 3925-3928, 1964.

[FeP.89] M. C. Ferris and A. B. Philpott, "An interior point method of semi-infinite programming," *Mathematical Programming*, Vol. 43, pp. 257-276, 1989.

[FeP.92] M. C. Ferris and A. B. Philpott, "On affine scaling and semi-infinite programming," *Mathematical Programming*, Vol. 56, No. 3, pp. 361-364, 1992.

[FiM.67] A. V. Fiacco and G. P. McCormick, "The sequential unconstrained minimization technique without parameters," *Operations Research*, Vol. 15, pp. 820-227, 1967.

[FiM.68] A. Fiacco, and G. McCormack, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley & Sons, New York, 1968.

[Fle.65] W. H. Fleming, *Functions of Several Variables*, Addison-Wesley, Reading, MA, 1965.

[Fle.70] R. Fletcher, "A class of methods for nonlinear programming with termination and convergence properties," in *Integer and Nonlinear Programming*, J. Abadie, ed., North Holland, Amsterdam, 1970.

[Fle.70a] R. Fletcher, "A new approach to variable metric algorithms," *Computer Journal*, Vol.13, pp. 357-372, 1970.

[Fle.72] R. Fletcher, "A class of methods for nonliner programming III: Rate of convergence," in *Numerical Methods for Nonlinear Optimization*, F.A. Lootsma, ed., Academic Press, New York, 1972.

[Fle.73] R. Fletcher, "An exact penalty function for nonlinear programming with inequality constraints," *Mathematical Programming*, Vol 5, pp. 129-150, 1973.

[Fle.82] R. Fletcher, "Second-order correction for nondifferentiable optimizaton," in *Numerical Analysis*, G. A. Watson, ed., Springer-Verlag, Berlin, pp. 85-114, 1982.

[Fle.82a] R. Fletcher, "A model algorithm for composite nondifferentiable optimization," *Mathematical Programming Study 17*, pp. 67-76, 1982.

[Fle.84] R. Fletcher, "An l_1 penalty method for nonlinear constraints" in *Numerical Optimization 1984*, P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., SIAM, Philadelphia, pp. 26-40, 1984.

[Fle.87] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., John Wiley & Sons, New York, 1987.

[Fle.94] R. Fletcher, "An overview of unconstrained optimization," Dundee Numerical Analysis Report NA/149, 1993, and in *Algorithms for Continuous Optimization: the State of the Art* (Proc. NATO ASI (Il Ciocco), Sept. 1993.), E Spedicato, ed., Kluwer Academic Publications, Dordrecht, pp. 109-144, 1994.

[FlL.71] R. Fletcher and S. Lill, "A class of methods for nonlinear programming II: Computational experience," in *Nonlinear Programming*, J. B. Rosen, O. L. Mangasarian and K. Ritter, eds., Academic Press, New York, 1971.

[FlP.63] R. Fletcher and M. J. D. Powell, "A rapidly convergent descent method for minimization," *Computer Journal*, Vol. 8, pp. 163-168, 1963.

[FlR.64] R. Fletcher and C. M. Reeves, "Function minimization by conjugate gradients, *Computer Journal*, Vol.7, No.2, pp. 149-154, 1964.

[FlW.80] R. Fletcher and G. A. Watson, "First and second-order conditions for a class of nondifferentiable optimization problems," *Mathematical Programming*, Vol. 18, pp. 291-308, 1980.

[FMM.77] G. E. Forsythe, M. A. Malcom, and C. B. Moler, *Computer Methods for Mathematical Computations*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

[FoS.95] R. Fotouhi and W. Szyszkowski, "A numerical approach for time-optimal control of double arms robot." *Proc. 4th IEEE Conference on Control Applications*, Albany, NY, pp. 1128-1133, Sept. 1995.

[Fri.70] A. Friedman, *Foundations of Modern Analysis*, Holt, Rinehart, and Winston, New York, 1970.

[FrW.56] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval Research Logistics Quarterly*, Vol. 3, pp. 95-110, 1956.

[Fuk.83] M. Fukushima, "An outer approximations algorithm for solving general convex problems," *Operations Research*, Vol. 31, pp. 101-113, 1983.

[Fuk.84] M. Fukushima, "On the convergence of a class of outer approximations algorithms for convex programs," *J. Computational Applied Mathematics*, Vol. 10, No. 2, pp. 147-156, 1984.

[Fuk.86] M. Fukushima, "A successive quadratic programming algorithm with global and superlinear convergence properties," *Mathematical Programming*, Vol. 35, No. 3, pp. 253-264, 1986.

[Gab.82] D. Gabay, "Reduced quasi-Newton methods with feasibility improvements for nonlinearly constrained optimization," *Mathematical Programming Study 16*, pp. 18-44, 1982.

[Gan.81] W. Gander, "Least squares with a quadratic constraint," *Numerische Mathematik*, Vol. 36, pp. 266-290, 1981.

[GHS.95] G. Gramlich, R. Hettich and E. Sachs, "Local convergence of SQP-methods for semi-infinite programming," *SIAM J. Optimization*, Vol. 5, pp. 641--658, 1995.

[GiL.89] J. C. Gilbert and C. Lemarechal, "Some numerical experiments with variable storage quasi-Newton algorithms," *Mathematical Programming*, Vol. 45, pp. 407-436, 1989.

[GiM.79] P. E. Gill and W. Murray, "Conjugate gradient methods for large-Scale nonlinear optimization," Tech. Rept. SOL 79-15, Department of Operations Research, Stanford University, Stanford, CA, 1979.

[GiN.92] J. C. Gilbert and J. Nocedal, "Global convergence properties of conjugate gradient methods of optimization," *SIAM J. Optimization*, Vol.2, No.1, pp. 21-42, 1992.

[GLL.86] L. Grippo, F. Lampariello, and S. Lucidi, "A Nonmonotone line search technique for Newton's method," *SIAM J. Numerical Analysis*, Vol. 23, No. 4, pp. 707-716, 1986.

[GlP.79] T. Glad and E. Polak, "A multiplier method with automatic limitation of penalty growth," *Mathematical Programming*, Vol. 17, No. 2, pp. 140-156, 1979.

[GMS.86] P. E. Gill, W. M. Murray, M. A. Saunders and M. H. Wright, "User's Guide for NPSOL (Version 4.0): A Fortran Package for Nonlinear Programming," Technical Report SOL 86-2, Department of Operations Research, Stanford University, Stanford, CA, 1986.

[GMW.81] P. E. Gill, W. Murray and M. H. Wright, *Practical optimization*, Academic Press, London, New York, 1981.

[GMW.91] P. E. Gill, W. Murray and M. H. Wright, *Numerical linear algebra and optimization*, Addison-Wesley, Redwood City, CA, 1991.

[Gol.70] D. Goldfarb, A family of variable metric methods derived by variational means," *Mathematics of Optimization*, Vol. 24, pp. 23-26, 1970.

[Gol.65] A. A. Goldstein, "On steepest descent," *SIAM J. Control*, Vol.3, pp. 147-151, 1965.

[Gol.67] A. A. Goldstein, *Constructive Real Analysis*, Harper, New York, 1967.

[Gol.74] A. A. Goldstein, "On gradient projection," *Proc. 12th Allerton Conference*, Allerton, IL, pp. 38-40, 1974.

[GoL.92] M. A. Goberna and M. A. Lopez, "Conditions for the uniquencss of the optimal solution in linear semi-infinite programming," *J. Optimization Theory and Applications*, Vol. 72, No. 2, pp. 225-246, 1992.

[GoM.91] G. H. Golub and U. von Matt, "Quadratically constrained least squares and quadratic problems," *Numerische Mathematik*, Vol. 59, pp. 561-580, 1991.

[GoP.67] A. A. Goldstein and J. F. Price, "An effective algorithm for minimization," *Numerische Mathematik*, Vol. 10, pp. 184-189, 1967.

[GoP.79] C. Gonzaga and E. Polak, "On constraint dropping schemes and optimality functions for a class of outer approximations algorithms," *SIAM J. Control and Optimization*, Vol. 17, No. 4, pp. 477-493, 1979.

[GPT.80] C. Gonzaga, E. Polak, and R. Trahan, "An improved algorithm for optimization problems with functional inequality constraints," *IEEE Trans. on Automatic Control*, Vol. AC-25, No. 1, pp. 49-54, 1980.

[HaB.70] P. C. Haarhoff and J. D. Buys, "A new method for the optimization of a nonlinear function subject to nonlinear constraints," *Computer Journal*, Vol. 13, pp. 178-184, 1970.

[Hag.75] W. W. Hager, "The Ritz-Trefftz method for state and control constrained optimal control problem." *SIAM J. Numerical Analysis*, Vol. 12, No. 6, pp. 854-867, 1975.

[Hag.76] W. W. Hager, "Rates of convergence for discrete approximations to unconstrained control problems," *SIAM J. Numerical Analysis*, Vol. 13, pp. 449-472, 1976.

[Hag.89] W. W. Hager, "A derivative-based bracketing scheme for univariate minimization and the conjugate gradient method," *Computers and Mathematics with Applications.*, Vol. 18, No. 9 , pp. 779-795, 1989.

[Hag.90] W. W. Hager, "Multiplier methods for nonlinear optimal control." *SIAM J. Numerical Analysis*, Vol. 27, No. 4, pp. 1061-1080, 1990.

[HaI.84] W. W. Hager and G. D. Ianculescu, "Dual approximations in optimal control." *SIAM J. Control and Optimization*, Vol. 22, No. 3, pp. 423-465, 1984.

[HaM.79] S. P. Han and O. L. Mangasarian, "Exact penalty functions in nonlinear programming". *Mathematical Programming*. Vol. 17, pp. 251-269, 1979.

[HaM.81] J. Hald and K. Madsen, "Combined LP and quasi-Newton methods for minimax optimization," *Mathematical Programming*, Vol. 20, pp. 49-62, 1981.

[Han.77] S. P. Han, "A globally convergent method for nonlinear programming," *J. Optimization Theory and Applications*, Vol. 22, pp. 297-309, 1977.

[Han.78] S.P. Han, "Superlinear convergence of a minimax method," Report No. TR78-336, Department of Computer Science, Cornell University, Ithaca, NY, 1978.

[Han.81] S. P. Han, "Variable metric methods for minimizing a class of nondifferentiable functions," *Mathematical Programming*, Vol. 20, pp. 1-13, 1981.

[HaP.87] C. R. Hargraves and S. W. Paris, "Direct trajectory optimization using nonlinear programming and collocation," *J. Guidance, Control, and Dynamics*, Vol. 10, pp. 338-342, 1987.

[HeC.92] A. L. Herman and B.A. Conway, "Optimal spacecraft attitude control using collocation and nonlinear programming," *J. Guidance, Control, and Dynamics*, Vol. 15, No. 5, pp. 1287-9, 1992.

[HeG.90] R. Hettich and G. Gramlich, "A note on an implementation of a method for quadratic semi-infinite programming," *Mathematical Programming*, Vol. 46, No.2, pp. 249-254, 1990.

[HeH.79] R. Hettich and W. van Honstede, "On quadratically convergent methods for semi-infinite programming," in *Semi-Infinite Programming*, R. Hettich, ed., Springer-Verlag, Berlin, pp. 97-111, 1979.

[HeK.93] R. Hettich and K. O. Kortanek, "Semi-infinite programming: theory methods, and applications," *SIAM Review*, Vol. 35, No. 3, pp. 380-429, 1993.

[HeK.94] R. Henrion and D. Klatte, "Metric regularity of the feasible set mapping in semi-infinite optimization," *Applied Mathematics and Optimization*, Vol. 30, No. 1, pp. 103-109, 1994.

[HeP.90] L. He and E. Polak, "Effective diagonalization strategies for the solution of a class of optimal design problems," *IEEE Trans on Automatic Control*, Vol. 35, No.3, pp. 258-267, 1990.

[Her.86] J. N. Herskovits, "A two-stage feasible directions algorithm for nonlinear constrained optimization," *Mathematical Programming*, Vol. 36, No. 1, pp. 19-38, 1986.

[HeS.52] M. R. Hestenes and E. L. Stiefel, "Methods of conjugate gradients for solving linear systems," *J. Res. Bureau National Standards*, Section B, Vol. 49, pp. 409-436, 1952.

[Hes.69] M. R. Hestenes, "Multiplier and gradient methods," *J. Optimization Theory and Applications*, Vol. 4, pp. 303-320, 1969.

[Hes.75] M. R. Hestenes, *Optimization Theory: The Finite Dimensional Case*, John Wiley & Sons, New York, 1975.

[Hes.80] M. R. Hestenes, *Conjugate Directions Methods in Optimization*, Springer-Verlag, Berlin, 1980.

[Het.86] R. Hettich, "An implementation of a discretization method for semi-infinite programming," *Mathematical Programming*, Vol. 34, pp. 354-361, 1986.

[HiL.93] J-B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, New York, 1993.

[HiP.90] J. E. Higgins and E. Polak, "Minimizing pseudo-convex functions on convex compact sets," *J. Optimization Theory and Applications*, Vol.65, No.1, pp. 1-28, 1990.

[HiP.91] J. E. Higgins and E. Polak, "An ε-active barrier function method for solving minimax problems," *Applied Mathematics and Optimization*, Vol. 23, pp. 275-297, 1991.

[Hog.73] W. W. Hogan, "Applications of a general convergence theory for outer approximations algorithms," *Mathematical Programming*, Vol. 5, pp. 151-168, 1973.

[Hua.67] P. Huard, "Method of centers," in *Nonlinear Programming*, J. Abadie, ed., North Holland, Amsterdam, 1967.

[Hua.68] P. Huard, "Programmation mathematique convexe," *Revue Francaise d'Informatique et de Recherche Operationelle*, Vol. 7, pp. 43-59, 1968.

[Hua.75] P. Huard, "Optimization algorithms and point-to-set maps," *Mathematical Programming*, Vol. 8, pp. 308-331, 1975.

[Hua.79] P. Huard, "Extensions of Zangwill's theorem," *Mathematical Programming Study 10*, pp. 98-103, 1979.

[HuH.90] H. Hu , "A one-phase algorithm for semi-infinite linear programming," *Mathematical Programming*, Vol. 46, No. 1, pp. 85-103, 1990.

[Iof.79] A. D. Ioffe, "Necessary and sufficient conditions for a local minimum 3: Second order conditions and augmented duality," *SIAM J. Control and Optimization*, Vol. 17, pp. 266-288, 1979.

[Iof.88] A. D. Ioffe, "On some recent developments in the theory of second-order optimality conditions," *Lecture Notes in Mathematics* 1405, S. Dolecki, ed., Springer-Verlag, New York, 1988.

[Jac.68] D. H. Jacobson, "Second-order and second variation methods for determining optimal control: a comparative analysis using differential dynamic programming" *International J. Control*, Vol. 7 No. 2, pp. 175-196, 1968.

[JaM.70] D. H. Jacobson and D. Q. Mayne, *Differential Dynamic Programming*, American Elsevier Press, New York, 1970.

[JaP.90] C. Jansch and M. Paus, "Aircraft trajectory optimization with direct collocation using movable gridpoints," *Proc. American Control Conference*, San Diego, pp. 262-267, 1990.

[Jar.88a] F. Jarre, "Convergence of the method of analytic centers for generalized convex programs," Schwerpunktprogramm der Deutschen Forschungsgemeinschaft - Anwendungsbezogene Optimierung und Steurerung, Institut fur Angewandte Mathematik und Statistik, Universitat Wurzburg, Report No. 67, 1988.

[Jar.88b] F. Jarre, "An implementation of the method of analytic centers," *Proc. 8th Conference on Analysis and Optimization of Systems*, INRIA, Antibes, France, June 1988.

[Joh.48] F. John, "Extremum problems with inequalities as side conditions," in *Studies and Essays: Courant Anniversary Volume*, K. O. Friedrichs, O. W. Neugebauer, and J. J. Stoker, eds., pp. 187-204, Interscience Publishers, Inc., New York, 1948.

[Jos.79] N. H. Josephy, "Newton's method for generalized equations," Technical Summary Report No. 1965, Mathematics Research Center, University of Wisconsin, Madison, WI, 1979.

[JRW.94] H.T. Jongen, J. J. Ruckmann, and G. W. Weber, "One-parametric semi-infinite optimization - on the stability of the feasible set," *SIAM J. Optimization*, Vol. 4. No. 3, pp. 637-648, 1994.

[JTW.92] H.T. Jongen, F. Twilt and G. W. Weber, "Semi-infinite optimization - structure and stability of the feasible set," *J. Optimization Theory and Applications*, Vol. 72, No. 3, pp. 529-552, 1992.

[JWZ.87] T. Jongen, W. Wetterling, and G. Zwier, "On sufficient conditions for local optimality in semi-infinite programming," *Optimization*, Vol. 18, No. 2, pp. 165-78, 1987.

[KaA.59] L. Kantorovich and G. Akilov, *Functional Analysis in Normed Spaces*, Fizmatgiz, Moscow, 1959; translated by D. Brown and A. Robertson, Pergamon Press, Oxford, 1964.

[Kar.39] W. Karush, *Minima of functions of several variables with inequalities as side conditions*, M.Sc. Thesis, Dept. of Mathematics, University of Chicago, Chicago IL, 1939.

[KaT.92] A. A. Kaplan and R. Tichatschke, "Adaptive method for solving incorrect problems of semi-infinite convex programming," *Doklady Akademii Nauk SSSR*, Vol. 322, No. 3, pp. 460-464, 1992.

[KBS.93] H. F. Khalfan, R. H. Byrd and R. B. Schnabel, "A Theoretical and experimental study of the symmetric rank-one update," *SIAM J. Optimization*, Vol. 3, No. 1, pp. 1-24, 1993.

[Kel.60] J. E. Kelley, Jr., "The cutting plane method for solving convex programs," *SIAM Journal*, Vol. 8, pp. 703-712, 1960.

[Kel.68] H. B. Keller, *Numerical Methods for Two-Point Boundary Value Problems*, Blaisdel, London, 1968.

[KeS.87] C. T. Kelley and E. W. Sachs, "Quasi-Newton methods and unconstrained optimal control problems," *SIAM J. Control and Optimization*, Vol. 25, pp. 1503-1516, 1987.

[KhS.90] K. M. Khoda and C. Storey, "A generalized Polak-Ribière algorithm," Technical Report A128, Dept. of Mathematical Sciences, Loughborough University of Technology, England, 1990.

[Kie.53] J. Kiefer, "Sequential minimax search for a maximum," *Proc. American Mathematical Society*, 4 pp. 503-506, 1953.

[KiP.91] C. Kirjner Neto and E. Polak, "A secant method based on cubic interpolation for solving one dimensional optimization problems," University of California, Berkeley, Electronics Research Laboratory Memo No. UCB/ERL M91/91, 15 October 1991.

[Kiw.85] K. C. Kiwiel, *Methods of Descent for Nonsmooth Optimization*, Springer-Verlag, Berlin, 1985.

[KlP.70] R. Klessig, *Implementation of Conceptual Algorithms*, Ph.D. Dissertation, University of California, Berkeley, 1970.

[KlP.72] R. Klessig and E. Polak, "Efficient implementations of the Polak-Ribière conjugate gradient algorithm," *SIAM J. Control*, Vol. 10, No. 3, pp. 524-549, 1972.

[KlP.73] R. Klessig and E. Polak, "An adaptive algorithm for unconstrained optimization with applications to optimal control," *SIAM J. Control*, Vol. 11, No. 1, pp. 80-94, 1973.

[KLS.92] K. M. Khoda, Y. Liu, and C. Storey, "Optimized software for a generalized Polak-Ribie´re algorithm in unconstrained optimization," Technical Report A156, Department of Mathematical Sciences, Loughborough University of Technology, England, 1992.

[KlT.84] E. Klein and A. C. Thompson, *Theory of Correspondences: Including Applications to Mathematical Economics*, John Wiley & Sons, New York, 1984.

[Koj.80] M. Kojima, "Strongly stable stationary solutions in nonlinear programs," in *Analysis and Computation of Fixed Points*, S. M. Robinson, ed., Academic Press, New York, pp. 93-138, 1980.

[KoF.75] A. N. Kolmogorov and S. V. Fomin, *Introductory Real Analysis*, translated and edited by R. A. Silverman, Rev. English ed., Dover Publications, New York, 1975,

[KoO.68] J. Kowalik and M. R. Osborne, *Methods for Unconstrained Optimization Problems*, American Elsevier Pub. Co., New York, 1968.

[Kur.58] K. Kuratowski, *Topologie*, Vols. 1 and 2, 4th ed., Corrected, Panstwowe Wyd. Nauk, Warszawa 1958 (Academic Press, New York, 1966).

[KuT.51] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," *Proc. Second Berkeley Symp. Mathematics, Statistics, and Probability*, University of California Press, Berkeley, CA, pp. 481-492, 1951.

[KSW.91] C. T. Kelley, E. W. Sachs, and B. Watson, "Pointwise quasi-Newton method for unconstrained optimal control problems.," *J. Optimization Theory and Applications*, Vol. 71, No. 3, pp. 535-547, 1991.

[LaL.61] J. P. Lasalle and S. Lefschetz, *Stability by Lyapunov's Direct Method with Applications*, Academic Press, New York, 1961.

[Lan.83] S. Lang, *Real Analysis*, 2nd ed., Addison-Wesley, Reading, MA, 1983.

[Las.70] L. S. Lasdon, "Conjugate direction methods for optimal control," *IEEE Trans. on Automatic Control*, pp. 267-268, 1970.

[Leb.75] G. Lebourg, "Valeur moyenne pour gradient generalisé," *Contes Rendus Academie des Sciences*, Paris, Vol. 281, pp. 795-797, 1975.

[Lec.91] M. Lecrenier, "Convergence of trust region algorithms for optimization with bounds when strict complementarity does not hold," *SIAM J. Numerical Analysis*, Vol. 28, pp. 476-495, 1991.

[Lem.81] C. Lemarechal, "A view of line searches," in *Optimization and Optimal Control*, Lecture Notes in Control and Information Science 30, A. Auslander, W. Oettli, and J. Stoer, eds., Springer-Verlag, Berlin, pp. 59-78, 1981.

[LeP.66] E. S. Levitin and B. T. Polyak, "Constrained minimization methods," *U.S.S.R. Computational Mathematics and Mathematical Physics*, Vol. 6, No. 5, pp. 1-50, 1966.

[Lev.44] K. Levengberg, "A method for the solution of certain nonlinear problems in least squares," *Quart. Applied Mathematics*, Vol. 2, pp. 321-328, 1944.

[LeV.94] T. Leon and E. Vercher, "New descent rules for solving the linear semi-infinite programming problem," *Operations Research Letters*, Vol. 15, No. 2, pp. 105-114, 1994.

[LiN.89] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, Vol. 45, pp. 503-528, 1989.

[LiX.92] X. Li, "An entropy-based aggregate method for minimax optimization," *Engineering Optimization*, Vol. 18, pp. 277-285, 1992.

[LMW.67] L. S. Lasdon, S. K. Mitter and A. D. Waren, "The conjugate gradient method for optimal control problems," *IEEE Trans. on Automatic Control*, Vol. 12, pp. 132-138, 1967.

[Loo.68] F. A. Lootsma, "Constrained optimization via penalty functions," M.S. 5814, *Phillips Res. Repts.*, Eindhoven, Holland, 1968.

[LPY.90] B. N. Lundberg, A. Poore, and B. Yang, "Smooth penalty functions and continuation methods for constrained optimization," in *Lectures in Applied Mathematics Series*, Volume 26, E. L. Allgower and K. Georg, eds., American Mathematical Society, Providence, RI., pp. 389-412, 1990.

[Lue.84] D. G. Luenberger, *Introduction to Linear and Nonlinear Programming*, 2nd ed., Addison-Wesley, New York, 1984.

[LuP.91] B. N. Lundberg and A. Poore, "Variable order Adams-Bashforth predictors with error-stepsize control for continuation methods," *SIAM J. Scientific and Statistical Computing*, Vol. 12, No. 3, pp. 895-723, 1991.

[LZT.94] C. Lawrence, J. L. Zhou, and A. L. Tits, "User's guide for CFSQP version 2.1: A C Code for solving (large scale) constrained nonlinear (minimax) optimization problems," Institute for Systems Research, University of Maryland technical report TR-94-16rl, 1994.

[Mac.88] K. C. P. Machielsen, *Numerical Solution of Optimal Control Problems with State Constraints by Sequential Quadratic Programming in Function Space*, CWI-Tract, Vol. 53, Centrum voor Wiskunde en Informatica, Amsterdam, 1988. *Chemical Engineering*, Vol. 13, pp. 49-62, 1989.

[Mad.75] K. Madsen, "An algorithm for minimax solution of overdetermined systems of non-linear equations," *J. Institute of Mathematics and its Applications*, Vol. 16, pp. 321-328, 1975.

[MaF.67] O. L. Mangasarian and S. Fromowitz, "The Fritz John necessary optimality conditions in the presence of equality constraints," *J. Mathematical Analysis and Applications*, Vol. 17, pp. 34-47, 1967.

[Mal.92] K. Malanowski, "Second-order conditions and constraint qualifications in stability and sensitivity analysis of solutions to optimization problems in Hilbert spaces," *Applied Mathematics and Optimization.*, Vol. 25, pp. 51-79, 1992.

[Man.82] O. L. Mangasarian, *Nonlinear Programming*, R.E. Krieger, Malabar, FL, 1982.

[MaP.75] D. Q. Mayne and E. Polak, "First-order, strong variations algorithms for optimal control," *J. Optimization Theory and Applications*, Vol. 16, No. 3/4, pp. 277-301, 1975.

[MaP.76] D. Q. Mayne and E. Polak, "Feasible directions algorithms for optimization problems with equality and inequality constraints," *Mathematical Programming*, Vol. 11, pp. 87-80, 1976.

[MaP.79] D. Q. Mayne and E. Polak, "A Superlinearly Convergent Algorithm for Constrained Optimization Problems," University of California, Berkeley, Electronics Research Laboratory Memo. No. UCB/ERL M79/13, Jan. 1979.

[MaP.80] D. Q. Mayne and E. Polak "An exact penalty function algorithm for optimal control problems with control and terminal equality constraints, Part 1," *J. Optimization Theory and Applications*, Vol. 32, No. 2, pp. 211-246, 1980.

[MaP.80a] D. Q. Mayne and E. Polak "An exact penalty function algorithm for optimal control problems with control and terminal equality constraints, Part 2," *J. Optimization Theory and Applications*, Vol. 32, No. 3, pp. 345-363, 1980.

[MaP.82] D. Q. Mayne and E. Polak, "A superlinearly convergent algorithm for constrained optimization problems," *Mathematical Programming Study 16*, pp. 45-61, 1982.

[MaP.82a] D. Q. Mayne and E. Polak, "A quadratically convergent algorithm for solving infinite dimensional inequalities," *J. Applied Mathematics and Optimization*, Vol. 9, pp. 25-40, 1982.

[MaP.84] D. Q. Mayne and E. Polak, "Outer approximations algorithm for nondifferentiable optimization Problems," *J. Optimization Theory and Applications*, Vol. 42, No. 1, pp. 19-30, 1984.

[Mar.63] D. W. Marquardt, "An algorithm for least squares estimation of nonlinear parameters," *SIAM J. Applied Mathematics*, Vol. 11, pp. 431-441, 1963.

[Mar.78] N. Maratos, *Exact Penalty Function Algorithms for Finite Dimensional and Optimization Problems*, Ph.D. Dissertation, Imperial College of Science and Technology, University of London, 1978.

[MaR.78] F. H. Mathis and G. W. Reddien, "Difference approximations to control problems with functional arguments," *SIAM J. Control and Optimization*, Vol. 16, pp. 436-449, 1978.

[MaS.78] A K. Madsen and H. Schjaer-Jacobsen "Linearly constrained minimax optimization," *Mathematical Programming*, Vol. 14, pp. 208-223, 1978.

[Mau.81] H. Maurer, "First- and second-order sufficient optimality conditions in mathematical programming and optimal control," *Mathematical Programming Study 14*, pp. 163-177, 1981.

[May.65] D. Q. Mayne, "A gradient method for determining optimal control of nonlinear stochastic systems," *Proc. IFAC Symposium, Theory of Self-Adaptive Control Systems*, P. H. Hammond, ed., Plenum Press, New York, pp. 19-27, 1965.

[May.66] D. Q. Mayne, "A second-order gradient method for determining optimal trajectories of nonlinear discrete-time systems," *International J. Control*, Vol. 3, pp. 85-95, 1966.

[MaZ.79] H. Maurer and J. Zowe, "First- and second-order necessary and sufficient optimality conditions for infinite dimensional programming problems," *Mathematical Programming*, Vol. 16, pp. 98-110, 1979.

[McC.67] G. P. McCormick, "Second-order conditions for constrained minima," *SIAM J. Applied Mathematics*, Vol. 15, pp. 641-652, 1967.

[McS.44] E. J. McShane, *Integration*, Princeton University Press, Princeton, NJ, 1944.

[Mey.76] R. R. Meyer, "Sufficient conditions for the convergence of monotonic mathematical programming algorithms," *J. Computer and Systems Science*, Vol. 12, pp. 108-121, 1976.

[Mey.77] R. R. Meyer, "A comparison of the forcing function and point-to-set mapping approaches to convergence analysis," *SIAM J. Control and Optimization*, Vol. 15, pp. 699-715, 1977.

[Mif.76] R. Mifflin, "Rates of convergence for a method of centers algorithm," *J. Optimization Theory and Applications*, Vol. 18, No. 2, pp. 199-228, 1976.

[Mif.84] R. Mifflin, "Stationarity and superlinear convergence of an algorithm for univariate locally Lipschitz constrained minimization," *Mathematical Programming*, Vol. 28, pp. 50-71, 1984.

[Mit.66] S. K. Mitter, "Successive approximation methods for the solution of optimal control problems," *Automatica*, Vol. 3, pp. 135-149, 1966.

[MMV.82] A. Miele, B. P. Mohanty, P. Venkataraman, and Y. M. Kuo, "Numerical solution of minimax problems of optimal control, I," *J. Optimization Theory and Applications*, Vol.38, No. 1, pp. 97-109, 1982.

[MMV.82a] A. Miele, B. P. Mohanty, P. Venkataraman, and Y. M. Kuo, "Numerical solution of minimax problems of optimal control, II," *J. Optimization Theory and Applications*, Vol.38, No. 1, pp. 111-135, 1982.

[MoL.91] D.J. Mook and J. Lew, "Multiple shooting algorithms for jump-discontinuous problems in optimal control and estimation," *IEEE Trans. on Automatic Control*, Vol. 36, pp. 979-983, 1991.

[Mor.78] B. Sh. Mordukhovich, "On difference approximations of optimal control systems," *Applied Mathematics and Mechanics*, Vol. 42, pp. 452-461, 1978.

[Mor.83] J. J. Moré, "Recent developments in algorithms and software for trust region methods," in *Mathematical Programming, the State of the Art*, A. Bachem, M. Grotschel and B. Korte, eds., Springer-Verlag, New York, pp. 258-287, 1983.

[Mor.88] B. Sh. Mordukhovich, *Methods of Approximation in Optimal Control Problems* (in Russian), Nauka, Moscow, 1988.

[MoS.83] J. J. Moré and D. C. Sorensen, "Computing a trust region step," *SIAM J. Scientific and Statistical Computing*, Vol. 4, pp. 553-572, 1983.

[MoT.94] J. J. Moré and D. J. Thuente, "On line search algorithms with guaranteed sufficient decrease," *ACM Trans. Mathematical Software*, Vol. 20, pp. 286--307, 1994.

[MuO.80] W. Murray and M. L. Overton, "A Projected Lagrangian algorithm for nonlinear minimax optimization," *SIAM J. Scientific and Statistical Computing*, Vol. 1, pp. 201-223, 1980.

[MuP.75] H. Mukai and E. Polak, "A quadratically convergent primal-dual algorithm with global convergence properties for solving optimization problems with equality constraints," *Mathematical Programming*, Vol. 9, No. 3, pp. 336-350, 1975.

[MuP.78a] H. Mukai and E. Polak, "A second-order algorithm for unconstrained optimization," *J. Optimization Theory and Applications*, Vol. 26, No. 4, 1978.

[MuP.78b] H. Mukai and E. Polak, "A second-order algorithm for the general nonlinear programming problem," *J. Optimization Theory and Applications*, Vol. 26, No. 4, 1978.

[MuS.69] B. A. Murtagh and R. W. H. Sargent, "A constrained minimization method with quadratic convergence," in *Optimization*, R. Fletcher, ed., Academic Press, London, 1969.

[MuY.84] D. M. Murray and S. Yakowitz, "Differential dynamic programming and Newton's method for discrete-time optimal control problems," *J. Optimization Theory and Applications*, Vol. 43, pp. 395-414, 1984.

[Mye.68] G. E. Myers, "Properties of the conjugate gradient and Davidon methods," *J. Optimization Theory and Applications*, Vol. 2, No.4, pp. 209-219, 1968.

[NaN.82] J. L. Nazareth and J. Nocedal, "Conjugate direction methods with variable storage," *Mathematical Programming*, Vol. 23, pp. 326-340, 1982.

[Naz.76] J. L. Nazareth, "Generation of conjugate directions for unconstrained minimization without derivatives," *Mathematics of Computation*, Vol. 30, pp. 115-131, 1976.

[Naz.79] J. L. Nazareth, "A Conjugate direction algorithm for unconstrained minimization without line searches," *J. Optimization Theory and Applications*, Vol. 23, pp. 373-387, 1979.

[Naz.86] J. L. Nazareth, "Conjugate direction algorithms less dependent on conjugacy," *SIAM Review*, Vol. 28, pp. 501-511, 1986.

[Naz.94] J. L. Nazareth, *The Newton-Cauchy Framework: A Unified Approach to Unconstrained Nonlinear Optimization*, Springer-Verlag, Berlin, 1994.

[Ned.84] N. B. Nedeljković, "New algorithms for discrete-time optimal control problems. Part 2." *J. Australian Mathematical Society, Series B (Applied Mathematics)*, Vol. 26, pp. 129-145, 1984.

[NeY.83] A. S. Nemirovsky and D. B. Yudin, *Slozhnost' Zadach i Effektivnost' Metodov Optimizatsii*, English title *Problem Complexity and Method Efficiency in Optimization*, translated by E. R. Dawson, John Wiley & Sons, Chichester, New York, 1983.

[New.720] Sir Isaac Newton, *Universal Arithmetick: or, A Treatise of Arithmetical Composition and Resolution.* Translated from the Latin by the late Mr. Ralphson; and rev. and corr. by Mr. Cunn, to which is added Dr. Halley's "Method of finding the roots of equations arithmetically," Printed for J. Senex et al, London, 1720.

[Obe.86] H.J. Oberle, "Numerical solution of minimax optimal control problems by multiple shooting technique," *J. Optimization Theory and Applications*, Vol.50, No.2, pp. 331-57, 1986.

[OrR.70] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, London, 1970.

[Osb.69] M. R. Osbourne, "On shooting methods for boundary value problems," *J. Mathematical Analysis and Applications*, Vol. 27, pp. 417-433, 1969.

[Osm.95] N. P. Osmolovskii, "Quadratic conditions for nonsingular extremals in optimal control (theoretical treatment)," *Russian J. Mathematical Physics*, Vol. 2, pp. 487-516, 1995.

[Out.84] J. V. Outrata, "Minimization of nonsmooth nonregular functions: application to discrete-time optimal control problems." *Problems of Control and Information Theory*, Vol. 13, No. 6, pp. 413-424, 1984.

[PaM.91] J. F. A. Pantoja and D. Q. Mayne, "Sequential quadratic programming algorithm for discrete optimal control problems with control inequality constraints," *International J. Control*, Vol. 53, pp. 823-836, 1991.

[Pan.91] J.-S. Pang, "A B-differentiable equation based, globally, and locally quadratically convergent algorithm for nonlinear programs, complementarity, and variational inequality problems," *Mathemetical Programming*, Vol. 51, pp. 101-131, 1991.

[Pan.88] J. Pantoja, "Differential dynamic programming and optimal control," *International J. Control*, Vol. 47, pp. 1539-1553, 1988.

[PaT.89] E. R. Panier and A. L. Tits, "A globally converging algorithm with adaptively refined discretization for semi-infinite optimization problems arising in engineering design," *IEEE Trans. on Automatic Control*, Vol. AC. 34, pp. 903-908, 1989.

[PaT.91] E. R. Panier and A. L. Tits, "Avoiding the Maratos effect by means of a nonmonotone line search, general constrained problems," *SIAM J. Numerical Analysis*, Vol. 28, No. 4, pp. 1183-1195, 1991.

[PaT.93] E. R. Panier and A. L. Tits, "On combining feasibility, descent and superlinear convergence in inequality constrained optimization," *Mathematical Programming*, Vol. 59, No. 2, pp. 261-276, 1993.

[PaZ.94] Z. Pales and V. Zeidan, "First- and second-order necessary conditions for control problems with constraints," *Trans. American Mathematical Society*, Vol. 346, pp. 421-453, 1994.

[PBG.62] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mischenko, *The Mathematical Theory of Optimal Processes*, Wiley-Interscience, New York, 1962.

[Pes.89] H. J. Pesch, "Real-time computation of feedback controls for constrained optimal control problems. Part I: neighbouring extremals," *Optimal Control Applications and Methods*, Vol. 10, pp. 129-145, 1989.

[Pes.89a] H. J. Pesch, "Real-time computation of feedback controls for constrained optimal control problems. Part II: "A correction method based on multiple shooting," *Optimal Control Applications and Methods*, Vol. 10, pp. 147-171, 1989.

[Phi.94] A. B. Philpott, "Continuous-time shortest path problems and linear programming," *SIAM J. Control and Optimization*, Vol. 32, No. 2, pp. 538-552, 1994.

[PHM.92] E. Polak, J. Higgins and D. Q. Mayne, "A barrier function method for minimax problems," *Mathematical Programming.*, Vol. 54, No.2, pp. 155-176, 1992.

[Pie.69] T. Pietrzykowski, "An exact potential method for constrained maxima," *SIAM J. Numerical Analysis*, No.6, pp. 299-304, 1969.

[PiP.72] O. Pironneau and E. Polak, "On the rate of convergence of certain methods of centers," *Mathematical Programming*, Vol. 2, No. 2, pp. 230-258, 1972.

[PiP.73] O. Pironneau and E. Polak, "Rate of convergence of a class of methods of feasible directions," *SIAM J. Numerical Analysis*, Vol. 10, No. 1, pp. 161-174, 1973.

[PMH.91] E. Polak, D. Q. Mayne and J. Higgins, "A superlinearly convergent algorithm for min-max problems," *J. Optimization Theory and Applications*, Vol. 89, No.3, pp. 407-439, 1991.

[PMH.92] E. Polak, D. Q. Mayne, and J. Higgins, "On the extension of Newton's method to semi-infinite minimax problems," *SIAM J. Control and Optimization*, Vol. 30, No.2, pp. 376-389, 1992.

[PMT.79] E. Polak, D. Q. Mayne and R. Trahan, "An outer approximations algorithm for computer aided design problems," *J. Optimization Theory and Applications*, Vol. 28, No. 3, pp. 331-352, 1979.

[PTH.88] E. R. Panier, A. L. Tits and J. N. Herskovits, "A QP-free, globally convergent, locally superlinearly convergent algorithm for inequality constrained optimization," *SIAM J. Control and Optimization*, Vol.26, No.4, pp. 788-811, 1988.

[PMW.83] E. Polak, D.Q. Mayne and Y. Wardi, "On the extension of constrained optimization algorithms from differentiable to nondifferentiable problems," *SIAM J. Control and Optimization*, Vol. 21, No. 2, pp. 179-204, 1983.

[PoH.89] E. Polak and L. He, "Rate-preserving discretization strategies for semi-infinite programming and optimal control," University of California, Electronics Research Laboratory, Memo UCB/ERL M89/112, 25 September 1989.

[PoH.91] E. Polak and L. He, "Finite-termination schemes for solving semi-infinite satisfycing problems," *J. Optimization Theory and Applications*, Vol. 70, No. 3, pp. 429-466, 1991.

[PoH.91a] E. Polak and L. He, "A unified steerable phase I - phase II method of feasible directions for semi-infinite optimization," *J. Optimization Theory and Applications*, Vol. 69, No.1, pp. 83-107, 1991.

[PoH.92] E. Polak and L. He, "Rate-preserving discretization strategies for semi-infinite programming and optimal control," *SIAM J. Control and Optimization*, Vol. 30, No. 3, pp. 548-572, 1992.

[Pol.69] E. Polak, "On the convergence of optimization algorithms," *Revue Francaise d'Informatique et de Recherche Operationelle, Serie Rouge*, No. 16, pp. 17-34, 1969.

[Pol.70] E. Polak, "On the implementation of conceptual algorithms," in *Nonlinear Programming*, O. L. Mangasarian, R. Ritter, and J. B. Rosen, eds., Academic Press, New York, pp. 275-291, 1970.

[Pol.71] E. Polak, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.

[Pol.71a] E. Polak, "On the use of models in the synthesis of optimization algorithms," in *Differential Games and Related Topics*, H. Kuhn and G. Szegö, eds., North Holland, Amsterdam, pp. 263-279, 1971.

[Pol.74] E. Polak, "A modified secant method for unconstrained optimization," *Mathematical Programming*, Vol. 8, No. 3, pp. 264-280, 1974.

[Pol.76] E. Polak, "On the global stabilization of locally convergent algorithms for optimization and root finding," *Automatica*, Vol. 12, pp. 337-342, 1976.

[Pol.83] B. T. Polyak, *Introduction to Optimization* (in Russian), Nauka, Moscow, 1983.

[Pol.87] E. Polak, "On the mathematical foundations of nondifferentiable optimization in engineering design," *SIAM Review*, Vol. 29, No. 1., pp. 21-91, 1987.

[Pol.88] R. A. Polyak, "Smooth optimization method for minimax problems," *SIAM J. Control and Optimization*, Vol. 26, pp. 1274-1286, 1988.

[Pol.89] E. Polak, "Basics of minimax algorithms," in *Nonsmooth Optimization and Related Topics*, F. H. Clarke, V. F. Dem'yanov and F. Giannessi, eds., Plenum Press, New York, pp. 343-367, 1989.

[Pol.93] E. Polak, "On the use of consistent approximations in the solution of semi-infinite optimization and optimal control problems," *Mathematical Programming*, Series B, Vol. 82, No.2, pp. 385-414, 1993.

[PoM.75] E. Polak and D. Q. Mayne, "First-order, strong variations algorithms for optimal control problems with terminal inequality constraints," *J. Optimization Theory and Applications*, Vol. 16, No. 3/4, pp. 303-325, 1975.

[PoM.76] E. Polak and D. Q. Mayne, "An algorithm for optimization problems with functional inequality constraints," *IEEE Trans. on Automatic Control*, Vol. AC-21, No. 2, pp. 184-193, 1976.

[PoM.81] E. Polak and D. Q. Mayne, "On the solution of singular value inequalities over a continuum of frequencies" *IEEE Transactions on Automatic Control*, Vol. AC-26, No. 3, pp. 890-695, 1981.

[Pon.62] L. S. Pontryagin, *Ordinary Differential Equations*, Addison-Wesley, Reading, MA, 1962.

[PoR.69] E. Polak and G. Ribière, "Note sur la convergence de methodes de directions conjugueés," *Revue Francaise d'Informatique et de Recherche Operationelle, Serie Rouge*, No. 16, pp. 35-43, 1969.

[PoT.80] E. Polak and A. Tits, "A globally convergent implementable multiplier method with automatic penalty limitation," *J. Applied Mathematics and Optimization*, Vol. 6, pp. 335-360, 1980.

[PoT.82] E. Polak and A. Tits, " A recursive quadratic programming algorithm for semi-infinite optimization problems," *J. Applied Mathematics and Optimization*, Vol. 8, pp. 325-349, 1982.

[Pow.69] M. J. D. Powell, "A method for nonlinear constraints in minimization problems," in *Optimization*, R. Fletcher, ed., Academic Press, London, pp. 283-298, 1969.

[Pow.75] M. J. D. Powell, "Convergence properties of a class of minimization algorithms," in *Nonlinear Programming 2*, O. L. Mangasarian, R. R. Meyer, S. M. Robinson, eds., Academic Press, New York, 1975.

[Pow.75a] M. J. D. Powell, "Some global convergence properties of a variable metric algorithm without exact line searches," R. W. Cottle and C. E. Lemke, eds., *Nonlinear Programming*, SIAM-AMS Proceedings, Vol. IX, American Mathematical Society, Providence, RI, 1975.

[Pow.76] M. J. D. Powell, "Some convergence properties of the conjugate gradient method," *Mathematical Programming*, Vol. 11, pp. 42-49, 1976.

[Pow.77] M. J. D. Powell, "Restart procedures for the conjugate gradient method," *Mathematical Programming*, Vol. 12, pp. 241-254, 1977.

[PoW.82] E. Polak and Y. Wardi, "A nondifferentiable optimization algorithm for the design of control systems subject to singular value inequalities over a frequency range," *Automatica*, Vol. 18, No. 3, pp. 267-283, 1982.

[Pow.84] M. J. D. Powell, "On the global convergence of trust region algorithms for unconstrained optimization," *Mathematical Programming*, Vol. 29, No. 3, pp. 297-303, 1984.

[Pow.84a] M. J. D. Powell, "Nonconvex minimization calculations and the conjugate gradient method," *Lecture Notes in Mathematics* 1066, Springer-Verlag, Berlin, pp. 122-131, 1984.

[PoW.84b] E. Polak and Y. Y. Wardi, "A study of minimizing sequences," *SIAM J. Control and Optimization*, Vol. 22, No. 4, pp. 599-609, 1984.

[Pow.86] M. J. D. Powell, "Convergence properties of algorithms for nonlinear optimization," *SIAM Review*, Vol. 28, No. 4, pp. 487-500, 1986.

[PoW.90] E. Polak and E. J. Wiest, "A variable metric technique for the solution of affinely parametrized nondifferentiable optimal design problems," *J. Optimization Theory and Applications*, Vol. 86, No. 3, pp. 391-414, 1990.

[PoY.86] M. J. D. Powell and Y. Yuan, "A recursive quadratic programming algorithm that uses exact differentiable penalty functions," *Mathematical Programming*, Vol. 35, No. 3, pp. 265-279, 1986.

[PoY.91] M. J. D. Powell and Y. Yuan, "A trust region algorithm for equality constrained optimization," *Mathematical Programming*, Vol. 49, pp. 189-211, 1991.

[PsD.75] B. N. Pshenichnyi, and Yu. M. Danilin, *Numerical Methods in Extremal Problems (Chislennye Metody v Ekstremal'nykh Zadachakh)*, Nauka, Moscow, 1975.

[Psh.70] B. N. Pshenichnyi, "Newton's method for the solution of equalities and inequalities," *Mathematical Notes Acad. Sc. USSR*, Vol. 8, pp. 827-830, 1970.

[PSS.74] E. Polak, R. W. H.Sargent and D. J. Sebastian, "On the convergence of sequential minimization algorithms," *J. Optimization Theory and Applications*, Vol. 14, No. 4, pp. 439-442, 1974.

[PTM.79] E. Polak, R. Trahan and D. Q. Mayne, "Combined phase I - phase II methods of feasible directions," *Mathematical Programming*, Vol. 17, No. 1, pp. 32-61, 1979.

[PYM.93] E. Polak, T. Yang, and D. Q. Mayne, "A method of centers based on barrier functions for solving optimal control problems with continuum state and control constraints," *SIAM J. Control and Optimization*, Vol.31, pp. 159-179, 1993.

[PyM.86] R. Pytlak and K Malinowski, "Numerical methods for minimax dynamic optimal control problem in discrete time," in *Analysis and Optimization of Systems*, (Proc. Seventh International Conference on the Analysis and Optimization of Systems, Antibes, France, 25-27 June 1986), A. Bensoussan and J. L. Lions, eds., Springer-Verlag, Berlin, pp. 81-92, 1986.

[Pyt.87] R. Pytlak, "On the algorithms based on the maximum principle for minimax optimal control problems," in *Automatic Control - World Congress, 1987.* Selected Papers from the 10th Triennial World Congress of the International Federation of Automatic Control, Munich, Germany, 27-31 July 1987, R. Isermann, ed., Pergamon, Oxford, Vol.8, pp. 51-57, 1988.

[QiL.93] L. Qi, "Convergence analysis of some algorithms for solving nonsmooth equations," *Mathematics of Operations Research*, Vol. 18, pp. 227-244, 1993.

[Ral.94] D. Ralph, "Global convergence of damped Newton's method for nonsmooth equations via the path search," *Mathematics of Operations Reseach*, Vol. 19. No. 2, pp. 352-389, 1994.

[RaS.90] A. Rakshit and S. Sen, "Sequential rank-one/rank-two updates for quasi-Newton differential dynamic programming," *Optimal Control Applications and Methods*, Vol. 11, No. 1, pp. 95-101, 1990.

[Red.79] G. W. Reddien, "Collocation at Gauss points as a discretization in optimal control," *SIAM J. Control and Optimization*, Vol. 17, No. 2, pp. 298-306, 1979.

[Ree.91] R. Reemtsen, "Discretization methods for the solution of semi-infinite programming problems," *J. Optimization Theory and Applications*, Vol. 71, No. 1, pp. 85-103, 1991.

[Ree.92] R. Reemtsen, "A cutting plane method for solving minimax problems in the convex plane," *Numerical Algorithms*, Vol. 2, pp. 409-436, 1992.

[Rit.79] K. Ritter, "Local and superlinear convergence of a class of variable metric methods," *Computing*, Vol. 23, pp. 287-297, 1979.

[Rit.81] K. Ritter, "Global and superlinear convergence of a class of variable metric methods," *Mathematical Programming Study 14*, pp. 178-205, 1981.

[Rob.72] S. M. Robinson, "Extension of Newton's method to mixed systems of nonlinear equations and inequalities," *Numerische Mathematik*, Vol. 19, pp. 341-347, 1972.

[Rob.74] S. M. Robinson, "Perturbed Kuhn-Tucker points and rates of convergence for a class of nonlinear-programming algorithms," *Mathematical Programming*, Vol. 7, pp. 1-16, 1974.

[Rob.83] S. M. Robinson, "Generalized equations," in *Mathematical Programming: The State of the Art, Bonn 1982*, A. Bachem, M. Grotschel and B. Korte, eds., Springer-Verlag, Berlin, pp. 346-367, 1983.

[Rob.93] S. M. Robinson, "Normal maps induced by linear transformations," *Mathematics of Operations Research*, Vol. 17, pp. 691-714, 1993.

[Rob.94] S. M. Robinson, "Newton's method for a class of nonsmooth functions," *Set-Valued Analysis*, Vol. 2, pp. 291-305, 1994.

[Roc.70] R. T. Rockafellar, *Convex Analysis*, Vol.28 of Princeton Mathematics Series, Princeton University Press, Princeton, NJ, 1970.

[Roc.73] R. T. Rockafellar, "The multiplier method of Hestenes and Powell applied to convex programming," *J. Optimization Theory and Applications*, Vol. 12, pp. 555-562, 1973.

[Roc.74] R. T. Rockafellar, "Augmented Lagrange multiplier functions and duality in nonconvex programming," *SIAM J. Control*, Vol. 12, pp. 268-285, 1974.

[Roc.83] R. T. Rockafellar, "Lagrange multipliers and optimality," *SIAM Review*, Vol. 35, No. 2, pp. 183-238, 1983.

[RoW.97] R. T. Rockafellar and R. J-B. Wets, *Variational Analysis*, book in progress to be published by Springer-Verlag, New York, 1997.

[Roy.88] H. L. Royden, *Real Analysis*, 3rd ed., Macmillan, New York, 1988.

[Rud.76] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed., McGraw-Hill, New York, 1976.

[Rux.93] D. J. W. Ruxton, "Differential dynamic programming applied to continuous optimal control problems with state variable inequality constraints," *Dynamics and Control*, Vol.3, No.2, pp. 175-85, 1993.

[SaW.77] G. Salinetti and R. J-B. Wets, "On the relation between two types of convergence for convex functions," *J. Mathematical Analysis and Applications*, Vol. 60, pp. 211-226, 1977.

[Sch.81] K. Schittkowski, "The Nonlinear programming method of Wilson, Han, and Powell with an augmented Lagrangian type of line search function, Part I: Convergence analysis," *Numerische Mathematik*, Vol. 38, pp. 83-114, 1981.

[Sch.96] A. L. Schwartz, *Theory and Implementation of Numerical Methods Based on Runge-Kutta Integration for Solving Optimal Control Problems*, PhD dissertation, University of California, Berkeley, 1996.

[ScP.95] A. Schwartz and E. Polak, "Runge-Kutta discretization of optimal control problems," *Proc. 10th IFAC Workshop on Control Applications of Optimization*, Haifa, Israel, Dec. 19-21, 1995.

[ScP.96] A. Schwartz and E. Polak, "Consistent approximations for optimal control problems based on Runge-Kutta integration," *SIAM J. Control and Optimization*, Vol. 34, No. 4, pp. 1235-1269, 1996.

[ScP.97] A. Schwartz and E. Polak, "A family of projected descent methods for optimization problems with simple bounds," *J. Optimization Theory and Applications*, Vol. 92, No. 1, pp. 1-32, 1997.

[SeD.93] S. Sen and K. B. Datta, "Time-optimal control algorithm for two-time-scale discrete systems - a multirate approach." *Control Theory and Advanced Technology*, Vol. 9, No. 3, pp. 733-43, 1993.

[SeY.87] S. Sen and S. J. Yakowitz, A quasi-Newton differential dynamic programming algorithm for discrete-time optimal control," *Automatica*, Vol.23, No.6, pp. 749-52 1987.

[Sha.67] V. E. Shamanskii, "On a modification of the Newton method," *Ukrainski Matematicheski Zhurnal*, Vol. 19, No. 1, pp. 133-138, 1967.

[Sha.70] D. F. Shanno, "Conditioning of quasi-Newton methods for function minimization," *Mathematics of Optimization*, Vol. 24, pp. 847-656, 1970.

[Sha.78] D. F. Shanno, "Conjugate gradient methods with inexact line searches," *Mathematics of Operations Reseach*, Vol. 3, pp. 244-256, 1978.

[ShI.94] K. Shimizu and S. Ito, "Constrained optimization in Hilbert space and a generalized dual quasi-Newton algorithm for state-constrained optimal control problems." *IEEE Trans. on Automatic Control*, Vol. 39, No. 5, pp. 982-986, 1994.

[Sho.85] N. Z. Shor, *Minimization Methods for Nondifferentiable Functions*, Springer-Verlag, Berlin, 1985.

[Sio.58] M. Sion, "On general min-max theorems," *Pacific J. Mathematics*, Vol. 8, pp. 171-176, 1958.

[Son.87] G. Sonnevend, "New algorithms in convex programming based on a notion of centre (for systems of analytic inequalities) and on rational extrapolation," in *Trends in Mathematical Optimization*, K.-H. Hoffmann et al., eds., ISNM, Vol. 84, Birkhauser Verlag, Stuttgart, pp. 311-327, 1987.

[Sor.82] D. C. Sorensen, "Newton's method with a model trust region modification," *SIAM J. Numerical Analysis*, Vol. 19, pp. 409-426, 1982.

[Sor.92] D. C. Sorensen, "Implicit application of polynomial filters in a k–step Arnoldi method," *SIAM J. Matrix Analysis Applications*, Vol. 13, pp. 357-385, 1992.

[Sor.94] D. C. Sorensen, "Minimization of a large scale quadratic function subject to an elliptical constraint," Rice University, Dept. Computational and Applied Mathematics Tech. Rept. TR94-27, Houston, TX, 1994.

[SoS.88] G. Sonnevend and J. Stoer, "Global ellipsoidal approximations and homotopy methods for solving convex analytic programs," Schwerpunktprogramm der Deutschen Forschungsgemeinschaft - Anwendungsbezogene Optimierung und Steurerung, Institut fur Angewandte Mathematik und Statistik, Universitat Wurzburg, Report No. 40, Wurzburg, 1988.

[StB.92] O. Stryk and R. Bulirsch, "Direct and indirect methods for trajectory optimization." *Annals of Operations Research*, Vol. 37, pp. 357-373, 1992.

[Ste.70] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton Mathematics Series, No. 30, Princeton University Press, Princeton, NJ, 1970.

[Sti.89] D. M. Stimler, "Scheduling turbofan engine control set points by semi-infinite optimization," *Automatica*, Vol. 25, No. 3, pp. 413-420, 1989.

[Sto.77] J. Stoer, "On the relation between quadratic termination and convergence properties of minimization algorithms, Part I, Theory," *Numerische Mathematik*, Vol. 28, pp. 343-366, 1977.

[Sto.83] J. Stoer, "Solution of large linear systems of equations by conjugate gradient type methods," in *Mathematical Programming: The State of the Art, Bonn 1982*, A. Bachem, M. Grotschel and B. Korte, eds., Springer-Verlag, Berlin, pp. 357-373, 1983.

[Str.93] H. Strauss, "Uniqueness in linear semi-infinite optimization," *J. Approximation Theory*, Vol. 75, No. 2, pp. 198-213, 1993.

[Str.93a] O. Stryk, "Numerical solution of optimal control problems by direct collocation"", *International Series of Numerical Mathematics*, Vol. 111, pp. 129-143, 1993.

[Str.95] O. Stryk, *Numerische Losung Optimaler Steuerungsprobleme: Diskretisierung, Parameteroptimierung und Erechnung der Adjungierten Variablen*, Diploma-Math, Munchen University of Technology, VDI Verlag No. 441, Germany, 1995.

[SWF.95] R. L. Sheu, S. Y. Wu, and S. C. Fang, "Primal-dual infeasible-interior-point algorithm for linear semi-infinite programming," *Computers and Mathematics with Applications*, Vol. 29, No. 8, pp. 7-18, 1995.

[Tam.76] A. Tamir, "Line search techniques based on interpolating polynomials using function values only," *Management Science*, Vol. 5, pp. 576-586, 1976.

[Tam.79] A. Tamir, "Rates of convergence of a one-dimensional search based on interpolating polynomials," *J. Optimization Theory and Applications*, Vol. 27, pp. 187-203, 1979.

[TaS.89] D. Talwar and R. Sivan, "An efficient numerical algorithm for the solution of a class of optimal control problems." *IEEE Trans. Automatic Control*, Vol. 34, No. 12, pp. 1308-1311, 1989.

[TFI.88] Y. Tanaka, M. Fukushima, and U. Ibaraki, "A globally convergent SQP method for semi-infinite nonlinear optimization," *J. Computational and Applied Mathematics*, Vol. 23, pp. 141-153, 1988.

[Tod.94] M. J. Todd, "Interior-point algorithms for semi-infinite programming," *Mathematical Programming*, Vol. 65, No. 2, pp. 217-245, 1994.

[Top.70] D. M. Topkis, "Cutting plane methods without nested constraint sets," *Operations Research*, Vol. 18, pp. 404-413, 1970.

[TPK.79] S. Tishyadhigama, E. Polak, and R. Klessig, "A comparative study of several general convergence conditions for algorithms modeled by point to set maps," *Mathematical Programming Study 10*, pp. 172-190, 1979.

[Tra.64] J. F. Traub, *Iterative Methods for the Solution of Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[Tre.68] R. Tremolieres, *La Method des Centres a Troncature Variable*, These Docteur d'Etat, University of Paris, 1968.

[TrN.70] S. S. Tripathi and K. S. Narendra, "Optimization using conjugate gradient methods," *IEEE Trans. on Automatic Control*, Vol. AC-15, No.2, pp. 268-270, 1970.

[TuH.80] P. R. Turner and E. Huntley, "Self-scaling variable metric methods in Hilbert space with applications to control problems," *Optimal Control Applications and Methods*, Vol. 1, pp. 155-166, 1980.

[Vai.74] M. M. Vainberg, *Variational Method and Method of Monotone Operators in the Theory of Nonlinear Equations*, John Wiley & Sons, New York, 1974.

[vHo.77] B. von Hohenbalken, "Simplicial decomposition in nonlinear programming algorithms," *Mathematical Programming*, Vol. 13, pp. 49-68, 1977.

[vNe.28] J. von Neumann, "Zur Theorie der Gesellsaftsspiele," *Math. Annals*, Vol. 100, pp. 922-320, 1928.

[vNM.44] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behaviour*, Princeton University Press, Princeton, NJ, 1944.

[War.72] J. Warga, *Optimal Control of Differential Equations and Functional Equations*, Academic Press, New York, 1972.

[War.77] J. Warga, "Steepest descent with relaxed controls," *SIAM J. Control and Optimization*, Vol. 15, pp. 674-682, 1977.

[War.82] J. Warga, "Iterative procedures for constrained and unilateral optimization problems," *SIAM J. Control and Optimization*, Vol. 20, pp. 360-367, 1982.

[War.84] J. Warga, "Iterative optimization with equality constraints," *Mathematics of Operations Reseach*, Vol. 9, pp. 592-605, 1984.

'Wer.78] J. Werner, "Uber die globale Konvergenz von Variable-Metric Verfahren mit nichtexakter Schritweitenbestimmung," *Numerische Mathematik*, Vol. 31, pp. 321-334, 1978.

[Wet.80] R. J-B. Wets, "Convergence of convex functions, variational inequalities and convex optimization problems," in *Variational Inequalities and Complementarity Problems*, R. Cottle, F. Giannessi, and J.L. Lions, eds., John Wiley & Sons, Chichester, England, pp. 405-419, 1980.

[Wil.63] R. B. Wilson, *A Simplicial Algorithm for Concave Programming*, Ph.D. dissertation, Harvard University, Cambridge, MA, 1963.

[Wil.73] L. J. Williamson, *Convergence Properties of Optimal Control Algorithms*, Ph.D. dissertation, University of California, Berkeley 1973.

[WiP.76] L. J. Williamson and E. Polak, "Relaxed controls and the convergence of optimal control algorithms," *SIAM J. Control*, Vol. 14, No. 4, pp. 737-757, 1976.

[WiP.91] E. J. Wiest and E. Polak, "On the rate of convergence of two minimax algorithms," *J. Optimization Theory and Applications*, Vol. 71, No. 1, pp. 1-30, 1991.

[WoF.86] R. S. Womersley, and R. Fletcher, "An algorithm for composite nonsmooth optimization problems," *J. Optimization Theory and Applications*, Vol. 48, pp. 493-523, 1986.

[Wol.68] Ph. Wolfe, "Another variable metric method," working paper, 1968.

[Wol.69] Ph. Wolfe, "Convergence conditions for ascent methods," *SIAM Review*, Vol. 11, pp. 226-235, 1969.

[Wol.71] Ph. Wolfe, "Convergence conditions for ascent methods II: Some corrections," *SIAM Review*, Vol. 13, pp. 185-188, 1971.

[Wol.76] Ph. Wolfe, "Finding the nearest point in a polytope," *Mathematical Programming*, Vol. 11, pp. 128-149, 1976.

[Wom.82] R. S. Womersley, "Optimality conditions for piecewise smooth functions," *Mathematical Programming Study 17*, pp. 13-27, 1982.

[Wri.90] S. J. Wright, "Solution of discrete-time optimal control problems on parallel computers." *Parallel Computing*, Vol.16, No.2-3, pp. 221-37, Dec. 1990.

[Wri.91] S. J. Wright, "Structured interior point methods for optimal control." *Proc. 30th IEEE Conference on Decision and Control*, Brighton, England, Vol. 2, pp. 1711-1716, Dec. 1991.

[YeY.87] Y. Ye, *Interior Algorithms for Linear, Quadratic, and Linearly Constrained Convex Programming*, Ph.D. Thesis, Department of Engineering-Economic Systems, Stanford University, Stanford, CA, 1987.

[You.69] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

[Yua.85] Y. Yuan, "On the superlinear convergence of a trust region algorithm for nonsmooth optimization," *Mathematical Programming*, Vol. 31, pp. 269-285, 1985.

[Zan.67] W. I. Zangwill, "Nonlinear programming via penalty functions," *Management Science*, Vol. 13, pp. 344-358, 1967.

[Zan.69] W. I. Zangwill, "Convergence conditions for nonlinear programming algorithms," *Management Science*, Vol. 16, pp. 1-13, 1969.

[Zan.69a] W. I. Zangwill, *Nonlinear Programming; a Unified Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1969.

[Zan.80] I. Zang, "A smoothing technique for min-max optimization" *Mathematical Programming*, Vol.19, pp. 61-77, 1980.

[Zei.84] V. Zeidan, "First- and second-order sufficient conditions for optimal control and the calculus of variations," *J. Applied Mathematics and Optimization*, Vol. 11, pp. 209-226, 1984.

[Zei.94] V. Zeidan, "The Riccati equation for optimal control problems with mixed state-control constraints: necessity and sufficiency," *SIAM J. Control and Optimization*, Vol. 32, pp. 1297-1321, 1994.

[Zha.95] Q. X. Zhang, "Optimality conditions and duality for arcwise semi-infinite programming with parametric inequality constraints," *J. Mathematical Analysis and Applications*, Vol. 196, No. 3, pp. 998-1007, 1995.

[ZhT.93] J. L. Zhou and A. L. Tits, "Nonmonotone line search for minimax problems," *J. Optimization Theory and Applications*, Vol. 76, No. 3, pp. 455-476, 1993.

[ZhT.96] J. L. Zhou and A A. L. Tits, "An SQP algorithm for finely discretized continuous minimax problems and other minimax problems with many objective functions," *SIAM J. Optimization*, Vol. 6, No. 2, pp. 461-487, 1996.

[Zou.60] G. Zoutendijk, *Methods of Feasible Directions*, Elsevier, Amsterdam, 1960.

[Zou.70] G. Zoutendijk, "Nonlinear programming, computational methods," in *Integer and Nonlinear Programming*, J. Abadie, ed., North Holland, Amsterdam, pp. 37-86, 1970.

[Zou.76] G. Zoutendijk, *Mathematical Programming Methods*, North Holland, Amsterdam, 1976.

Index

Applied Mathematical Sciences

(continued from page ii)

61. *Sattinger/Weaver:* Lie Groups and Algebras with Applications to Physics, Geometry, and Mechanics.
62. *LaSalle:* The Stability and Control of Discrete Processes.
63. *Grasman:* Asymptotic Methods of Relaxation Oscillations and Applications.
64. *Hsu:* Cell-to-Cell Mapping: A Method of Global Analysis for Nonlinear Systems.
65. *Rand/Armbruster:* Perturbation Methods, Bifurcation Theory and Computer Algebra.
66. *Hlavácek/Haslinger/Necasl/Lovísek:* Solution of Variational Inequalities in Mechanics.
67. *Cercignani:* The Boltzmann Equation and Its Applications.
68. *Temam:* Infinite-Dimensional Dynamical Systems in Mechanics and Physics, 2nd ed.
69. *Golubitsky/Stewart/Schaeffer:* Singularities and Groups in Bifurcation Theory, Vol. II.
70. *Constantin/Foias/Nicolaenko/Temam:* Integral Manifolds and Inertial Manifolds for Dissipative Partial Differential Equations.
71. *Catlin:* Estimation, Control, and the Discrete Kalman Filter.
72. *Lochak/Meunier:* Multiphase Averaging for Classical Systems.
73. *Wiggins:* Global Bifurcations and Chaos.
74. *Mawhin/Willem:* Critical Point Theory and Hamiltonian Systems.
75. *Abraham/Marsden/Ratiu:* Manifolds, Tensor Analysis, and Applications, 2nd ed.
76. *Lagerstrom:* Matched Asymptotic Expansions: Ideas and Techniques.
77. *Aldous:* Probability Approximations via the Poisson Clumping Heuristic.
78. *Dacorogna:* Direct Methods in the Calculus of Variations.
79. *Hernández-Lerma:* Adaptive Markov Processes.
80. *Lawden:* Elliptic Functions and Applications.
81. *Bluman/Kumei:* Symmetries and Differential Equations.
82. *Kress:* Linear Integral Equations.
83. *Bebernes/Eberly:* Mathematical Problems from Combustion Theory.
84. *Joseph:* Fluid Dynamics of Viscoelastic Fluids.
85. *Yang:* Wave Packets and Their Bifurcations in Geophysical Fluid Dynamics.
86. *Dendrinos/Sonis:* Chaos and Socio-Spatial Dynamics.
87. *Weder:* Spectral and Scattering Theory for Wave Propagation in Perturbed Stratified Media.
88. *Bogaevski/Povzner:* Algebraic Methods in Nonlinear Perturbation Theory.
89. *O'Malley:* Singular Perturbation Methods for Ordinary Differential Equations.
90. *Meyer/Hall:* Introduction to Hamiltonian Dynamical Systems and the N-body Problem.
91. *Straughan:* The Energy Method, Stability, and Nonlinear Convection.
92. *Naber:* The Geometry of Minkowski Spacetime.
93. *Colton/Kress:* Inverse Acoustic and Electromagnetic Scattering Theory.
94. *Hoppensteadt:* Analysis and Simulation of Chaotic Systems.
95. *Hackbusch:* Iterative Solution of Large Sparse Systems of Equations.
96. *Marchioro/Pulvirenti:* Mathematical Theory of Incompressible Nonviscous Fluids.
97. *Lasota/Mackey:* Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics, 2nd ed.
98. *de Boor/Höllig/Riemenschneider:* Box Splines.
99. *Hale/Lunel:* Introduction to Functional Differential Equations.
100. *Sirovich (ed):* Trends and Perspectives in Applied Mathematics.
101. *Nusse/Yorke:* Dynamics: Numerical Explorations.
102. *Chossat/Iooss:* The Couette-Taylor Problem.
103. *Chorin:* Vorticity and Turbulence.
104. *Farkas:* Periodic Motions.
105. *Wiggins:* Normally Hyperbolic Invariant Manifolds in Dynamical Systems.
106. *Cercignani/Illner/Pulvirenti:* The Mathematical Theory of Dilute Gases.
107. *Antman:* Nonlinear Problems of Elasticity.
108. *Zeidler:* Applied Functional Analysis: Applications to Mathematical Physics.
109. *Zeidler:* Applied Functional Analysis: Main Principles and Their Applications.
110. *Diekmann/van Gils/Verduyn Lunel/Walther:* Delay Equations: Functional-, Complex-, and Nonlinear Analysis.
111. *Visintin:* Differential Models of Hysteresis.
112. *Kuznetsov:* Elements of Applied Bifurcation Theory.
113. *Hislop/Sigal:* Introduction to Spectral Theory: With Applications to Schrödinger Operators.
114. *Kevorkian/Cole:* Multiple Scale and Singular Perturbation Methods.
115. *Taylor:* Partial Differential Equations I, Basic Theory.
116. *Taylor:* Partial Differential Equations II, Qualitative Studies of Linear Equations.
117. *Taylor:* Partial Differential Equations III, Nonlinear Equations.

(continued on next page)

Applied Mathematical Sciences

(continued from previous page)

118. *Godlewski/Raviart:* Numerical Approximation of Hyperbolic Systems of Conservation Laws.
119. *Wu:* Theory and Applications of Partial Functional Differential Equations.
120. *Kirsch:* An Introduction to the Mathematical Theory of Inverse Problems.
121. *Brokate/Sprekels:* Hysteresis and Phase Transitions.
122. *Gliklikh:* Global Analysis in Mathematical Physics: Geometric and Stochastic Methods.
123. *Le/Schmitt:* Global Bifurcation in Variational Inequalities: Applications to Obstacle and Unilateral Problems.
124. *Polak:* Optimization: Algorithms and Consistent Approximations.
125. *Arnold/Khesin:* Topological Methods in Hydrodynamics.
126. *Hoppensteadt/Izhikevich:* Weakly Connected Neural Networks.