Nonlinear Programming

THIRD EDITION

Dimitri P. Bertsekas

Massachusetts Institute of Technology

WWW site for book information and orders

http://www.athenasc.com

Athena Scientific, Belmont, Massachusetts

Athena Scientific
Belmont, Mass. 02478
U.S.A.

Email: info@athenasc.com
WWW: http://www.athenasc.com

Cover image: The graph of the two-dimensional cost function of a simple neural network training problem.

Publisher's Cataloging-in-Publication Data

Bertsekas, Dimitri P.
Nonlinear Programming: Second Edition
Includes bibliographical references and index
1. Nonlinear Programming. 2. Mathematical Optimization. I. Title.
T57.8.B47 2016 519.703

ISBN-10: 1-886529-05-1, ISBN-13: 978-1-886529-05-2

ABOUT THE AUTHOR

Dimitri Bertsekas studied Mechanical and Electrical Engineering at the National Technical University of Athens, Greece, and obtained his Ph.D. in system science from the Massachusetts Institute of Technology. He has held faculty positions with the Engineering-Economic Systems Department, Stanford University, and the Electrical Engineering Department of the University of Illinois, Urbana. Since 1979 he has been teaching at the Electrical Engineering and Computer Science Department of the Massachusetts Institute of Technology (M.I.T.), where he is currently the McAfee Professor of Engineering.

His teaching and research spans several fields, including deterministic optimization, dynamic programming and stochastic control, large-scale and distributed computation, and data communication networks. He has authored or coauthored numerous research papers and sixteen books, several of which are currently used as textbooks in MIT classes, including "Dynamic Programming and Optimal Control," "Data Networks," "Introduction to Probability," "Convex Optimization Theory," "Convex Optimization Algorithms," as well as the present book.

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

ATHENA SCIENTIFIC

OPTIMIZATION AND COMPUTATION SERIES

1. Nonlinear Programming, 3rd Edition, by Dimitri P. Bertsekas, 2016, ISBN 1-886529-05-1, 880 pages
2. Dynamic Programming and Optimal Control, Two-Volume Set, by Dimitri P. Bertsekas, 2017, ISBN 1-886529-08-6, 1270 pages
3. Convex Optimization Algorithms, by Dimitri P. Bertsekas, 2015, ISBN 978-1-886529-28-1, 576 pages
4. Abstract Dynamic Programming, by Dimitri P. Bertsekas, 2013, ISBN 978-1-886529-42-7, 256 pages
5. Convex Optimization Theory, by Dimitri P. Bertsekas, 2009, ISBN 978-1-886529-31-1, 256 pages
6. Introduction to Probability, 2nd Edition, by Dimitri P. Bertsekas and John N. Tsitsiklis, 2008, ISBN 978-1-886529-23-6, 544 pages
7. Convex Analysis and Optimization, by Dimitri P. Bertsekas, Angelia Nedić, and Asuman E. Ozdaglar, 2003, ISBN 1-886529-45-0, 560 pages
8. Network Optimization: Continuous and Discrete Models, by Dimitri P. Bertsekas, 1998, ISBN 1-886529-02-7, 608 pages
9. Network Flows and Monotropic Optimization, by R. Tyrrell Rockafellar, 1998, ISBN 1-886529-06-X, 634 pages
10. Introduction to Linear Optimization, by Dimitris Bertsimas and John N. Tsitsiklis, 1997, ISBN 1-886529-19-1, 608 pages
11. Parallel and Distributed Computation: Numerical Methods, by Dimitri P. Bertsekas and John N. Tsitsiklis, 1997, ISBN 1-886529-01-9, 718 pages
12. Neuro-Dynamic Programming, by Dimitri P. Bertsekas and John N. Tsitsiklis, 1996, ISBN 1-886529-10-8, 512 pages
13. Constrained Optimization and Lagrange Multiplier Methods, by Dimitri P. Bertsekas, 1996, ISBN 1-886529-04-3, 410 pages
14. Stochastic Optimal Control: The Discrete-Time Case, by Dimitri P. Bertsekas and Steven E. Shreve, 1996, ISBN 1-886529-03-5, 330 pages

Contents

Preface to the First Edition

Nonlinear programming is a mature field that has experienced major developments in the last ten years. The first such development is the merging of linear and nonlinear programming algorithms through the use of interior point methods. This has resulted in a profound rethinking of how we solve linear programming problems, and in a major reassessment of how we treat constraints in nonlinear programming. A second development, less visible but still important, is the increased emphasis on large-scale problems, and the associated algorithms that take advantage of problem structure as well as parallel hardware. A third development has been the extensive use of iterative unconstrained optimization to solve the difficult least squares problems arising in the training of neural networks. As a result, simple gradient-like methods and stepsize rules have attained increased importance.

The purpose of this book is to provide an up-to-date, comprehensive, and rigorous account of nonlinear programming at the beginning graduate student level. In addition to the classical topics, such as descent algorithms, Lagrange multiplier theory, and duality, some of the important recent developments are covered: interior point methods for linear and nonlinear programs, major aspects of large-scale optimization, and least squares problems and neural network training.

A further noteworthy feature of the book is that it treats Lagrange multipliers and duality using two different and complementary approaches: a variational approach based on the implicit function theorem, and a convex analysis approach based on geometrical arguments. The former approach applies to a broader class of problems, while the latter is more elegant and more powerful for the convex programs to which it applies.

The chapter-by-chapter description of the book follows:

Chapter 1: This chapter covers unconstrained optimization: main concepts, optimality conditions, and algorithms. The material is classic, but there are discussions of topics frequently left untreated, such as the behavior of algorithms for singular problems, neural network training, and discrete-time optimal control.

Chapter 2: This chapter treats constrained optimization over a convex set without the use of Lagrange multipliers. I prefer to cover this material before dealing with the complex machinery of Lagrange multipliers because I have found that students absorb easily algorithms such as conditional gradient, gradient projection, and coordinate descent, which can be viewed as natural extensions of unconstrained descent algorithms. This chapter contains also a treatment of the affine scaling method for linear programming.

Chapter 3: This chapter gives a detailed treatment of Lagrange multipliers, the associated necessary and sufficient conditions, and sensitivity analysis. The first three sections deal with nonlinear equality and inequality constraints. The last section deals with linear constraints and develops a simple form of duality theory for linearly constrained problems with differentiable cost, including linear and quadratic programming.

Chapter 4: This chapter treats constrained optimization algorithms that use penalties and Lagrange multipliers, including barrier, augmented Lagrangian, sequential quadratic programming, and primal-dual interior point methods for linear programming. The treatment is extensive, and borrows from my 1982 research monograph on Lagrange multiplier methods.

Chapter 5: This chapter provides an in-depth coverage of duality theory (Lagrange and Fenchel). The treatment is totally geometric, and everything is explained in terms of intuitive figures.

Chapter 6: This chapter deals with large-scale optimization methods based on duality. Some material is borrowed from my Parallel and Distributed Algorithms book (coauthored by John Tsitsiklis), but there is also an extensive treatment of nondifferentiable optimization, including subgradient, ϵ-subgradient, and cutting plane methods. Decomposition methods such as Dantzig-Wolfe and Benders are also discussed.

Appendixes: Four appendixes are given. The first gives a summary of calculus, analysis, and linear algebra results used in the text. The second is a fairly extensive account of convexity theory, including proofs of the basic polyhedral convexity results on extreme points and Farkas' lemma, as well the basic facts about subgradients. The third appendix covers one-dimensional minimization methods. The last appendix discusses an implementation of Newton's method for unconstrained optimization.

Inevitably, some coverage compromises had to be made. The subject of nonlinear optimization has grown so much that leaving out a number of important topics could not be avoided. For example, a discussion of variational inequalities, a deeper treatment of optimality conditions, and a more detailed development of Quasi-Newton methods are not provided. Also, a larger number of sample applications would have been desirable. I hope that instructors will supplement the book with the type of practical examples that their students are most familiar with.

The book was developed through a first-year graduate course that I taught at the Univ. of Illinois and at M.I.T. over a period of 20 years. The mathematical prerequisites are matrix-vector algebra and advanced calculus, including a good understanding of convergence concepts. A course in analysis and/or linear algebra should also be very helpful, and would provide the mathematical maturity needed to follow and to appreciate the mathematical reasoning used in the book. Some of the sections in the book may be ommited at first reading without loss of continuity. These sections have been marked by a star. The rule followed here is that the material discussed in a starred section is not used in a non-starred section.

The book can be used to teach several different types of courses.

(a) A two-quarter course that covers most sections of every chapter.

(b) A one-semester course that covers Chapter 1 except for Section 1.9, Chapter 2 except for Sections 2.4 and 2.5, Chapter 3 except for Section 3.4, Chapter 4 except for parts of Sections 4.2 and 4.3, the first three sections of Chapter 5, and a selection from Section 5.4 and Chapter 6. This is the course I usually teach at MIT.

(c) A one-semester course that covers most of Chapters 1, 2, and 3, and selected algorithms from Chapter 4. I have taught this type of course several times. It is less demanding of the students because it does not require the machinery of convex analysis, yet it still provides a fairly powerful version of duality theory (Section 3.4).

(d) A one-quarter course that covers selected parts of Chapters 1, 2, 3, and 4. This is a less comprehensive version of (c) above.

(e) A one-quarter course on convex analysis and optimization that starts with Appendix B and covers Sections 1.1, 2.1, 3.4, and Chapter 5.

There is a very extensive literature on nonlinear programming and to give a complete bibliography and a historical account of the research that led to the present form of the subject would have been impossible. I thus have not attempted to compile a comprehensive list of original contributions to the field. I have cited sources that I have used extensively, that provide important extensions to the material of the book, that survey important topics, or that are particularly well suited for further reading. I have also cited selectively a few sources that are historically significant, but the reference list is far from exhaustive in this respect. Generally, to aid researchers in the field, I have preferred to cite surveys and textbooks for subjects that are relatively mature, and to give a larger number of references for relatively recent developments.

Finally, I would like to express my thanks to a number of individuals for their contributions to the book. My conceptual understanding of the subject was formed at Stanford University while I interacted with David Luenberger and I taught using his books. This experience had a lasting influence on my thinking. My research collaboration with several colleagues, particularly Joe

Dunn, Eli Gafni, Paul Tseng, and John Tsitsiklis, were very useful and are reflected in the book. I appreciate the suggestions and insights of a number of people, particularly David Castanon, Joe Dunn, Terry Rockafellar, Paul Tseng, and John Tsitsiklis. I am thankful to the many students and collaborators whose comments led to corrections and clarifications. Steve Patek, Serap Savari, and Cynara Wu were particularly helpful in this respect. David Logan, Steve Patek, and Lakis Polymenakos helped me to generate the graph of the cover, which depicts the cost function of a simple neural network training problem. My wife Joanna cheered me up with her presence and humor during the long hours of writing, as she has with her companionship of over 30 years. I dedicate this book to her with my love.

Dimitri P. Bertsekas
November 1995

Preface to the Second Edition

The second edition has expanded by about 130 pages the coverage of the original. Nearly 40% of the new material represents miscellaneous additions scattered throughout the text. The remainder deals with three new topics. These are:

(a) A new section in Chapter 3 that focuses on a simple but far-reaching treatment of Fritz John necessary conditions and constraint qualifications, and also includes semi-infinite programming.

(b) A new section in Chapter 5 on the use of duality and Lagrangian relaxation for solving discrete optimization problems. This section describes several motivating applications, and provides a connecting link between continuous and discrete optimization.

(c) A new section in Chapter 6 on approximate and incremental subgradient methods. This material is the subject of ongoing joint research with Angelia Nedić, but it was thought sufficiently significant to be included in summary here.

One of the aims of the revision was to highlight the connections of nonlinear programming with other branches of optimization, such as linear programming, network optimization, and discrete/integer optimization. This should provide some additional flexibility for using the book in the classroom. In addition, the presentation was improved, the mathematical background material of the appendixes has been expanded, the exercises were reorganized, and a substantial number of new exercises were added.

A new internet-based feature was added to the book, which significantly extends its scope and coverage. Many of the theoretical exercises, quite a few of them new, have been solved in detail and their solutions have been posted in the book's www page

http://www.athenasc.com/nonlinbook.html

These exercises have been marked with the symbol **WWW**

The book's www page also contains links to additional resources, such as computer codes and my lecture slides from my MIT Nonlinear Programming class.

I would like to express my thanks to the many colleagues who contributed suggestions for improvement of the second edition. I would like to thank particularly Angelia Nedić for her extensive help with the internet-posted solutions of the theoretical exercises.

Dimitri P. Bertsekas
June 1999

Preface to the Third Edition

The third edition of the book is a thoroughly rewritten version of the 1999 second edition. New material was included, some of the old material was discarded, and a large portion of the remainder was reorganized or revised. The total number of pages has increased by about 10 percent.

Aside from incremental improvements, the changes aim to bring the book up-to-date with recent research progress, and in harmony with the major developments in convex optimization theory and algorithms that have occurred in the meantime. These developments were documented in three of my books: the 2015 book "Convex Optimization Algorithms," the 2009 book "Convex Optimization Theory," and the 2003 book "Convex Analysis and Optimization" (coauthored with Angelia Nedić and Asuman Ozdaglar). A major difference is that these books have dealt primarily with convex, possibly nondifferentiable, optimization problems and rely on convex analysis, while the present book focuses primarily on algorithms for possibly nonconvex differentiable problems, and relies on calculus and variational analysis.

Having written several interrelated optimization books, I have come to see nonlinear programming and its associated duality theory as the lynchpin that holds together deterministic optimization. I have consequently set as an objective for the present book to integrate the contents of my books, together with internet-accessible material, so that they complement each other and form a unified whole. I have thus provided bridges to my other works with extensive references to generalizations, discussions, and elaborations of the analysis given here, and I have used throughout fairly consistent notation and mathematical level.

Another connecting link of my books is that they all share the same style: they rely on rigorous analysis, but they also aim at an intuitive exposition that makes use of geometric visualization. This stems from my belief that success in the practice of optimization strongly depends on the intuitive (as well as the analytical) understanding of the underlying theory and algorithms.

Some of the more prominent new features of the present edition are:

(a) An expanded coverage of incremental methods and their connections to stochastic gradient methods, based in part on my 2000 joint work with Angelia Nedić; see Section 2.4 and Section 7.3.2.

(b) A discussion of asynchronous distributed algorithms based in large part on my 1989 "Parallel and Distributed Computation" book (coauthored

with John Tsitsiklis); see Section 2.5.

(c) A discussion of the proximal algorithm and its variations in Section 3.6, and the relation with the method of multipliers in Section 7.3.

(d) A substantial coverage of the alternating direction method of multipliers (ADMM) in Section 7.4, with a discussion of its many applications and variations, as well as references to my 1989 "Parallel and Distributed Computation" and 2015 "Convex Optimization Algorithms" books.

(e) A fairly detailed treatment of conic programming problems in Section 6.4.1.

(f) A discussion of the question of existence of solutions in constrained optimization, based on my 2007 joint work with Paul Tseng [BeT07], which contains further analysis; see Section 3.1.2.

(g) Additional material on network flow problems in Section 3.8 and 6.4.3, and their extensions to monotropic programming in Section 6.4.2, with references to my 1998 "Network Optimization" book.

(h) An expansion of the material of Chapter 4 on Lagrange multiplier theory, using a strengthened version of the Fritz John conditions, and the notion of pseudonormality, based on my 2002 joint work with Asuman Ozdaglar.

(i) An expansion of the material of Chapter 5 on Lagrange multiplier algorithms, with references to my 1982 "Constrained Optimization and Lagrange Multiplier Methods" book.

The book contains a few new exercises. As in the second edition, many of the theoretical exercises have been solved in detail and their solutions have been posted in the book's internet site

http://www.athenasc.com/nonlinbook.html

These exercises have been marked with the symbols **WWW**. Many other exercises contain detailed hints and/or references to internet-accessible sources. The book's internet site also contains links to additional resources, such as many additional solved exercises from my convex optimization books, computer codes, my lecture slides from MIT Nonlinear Programming classes, and full course contents from the MIT OpenCourseWare (OCW) site.

I would like to express my thanks to the many colleagues who contributed suggestions for improvement of the third edition. In particular, let me note with appreciation my principal collaborators on nonlinear programming topics since the 1999 second edition: Angelia Nedić, Asuman Ozdaglar, Paul Tseng, Mengdi Wang, and Huizhen (Janey) Yu.

Dimitri P. Bertsekas
June 2016

1
Unconstrained Optimization: Basic Methods

Contents

Mathematical models of optimization can be generally represented by a *constraint set* X and a *cost function* f that maps elements of X into real numbers. The set X consists of the available decisions x and the cost $f(x)$ is a scalar measure of undesirability of choosing decision x. We want to find an optimal decision, i.e., an $x^* \in X$ such that

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

In this book we focus on the case where each decision x is an n-dimensional vector; that is, x is an n-tuple of real numbers $(x_1, \ldots, x_n)$. Thus the constraint set X is a subset of $\Re^n$, the n-dimensional Euclidean space. (We refer to Appendix A for an account of our terminology and notational conventions.)

The optimization problem just stated is very broad and contains as special cases several important classes of problems that have widely differing structures. Our focus will be on nonlinear programming problems, so let us provide some orientation about the character of these problems and their relations with other types of optimization problems.

Continuous and Discrete Problems

Perhaps the most important characteristic of an optimization problem is whether it is *continuous* or *discrete*. Continuous problems are those where the constraint set X is infinite and has a "continuous" character. Typical examples of continuous problems are those where there are no constraints, i.e., where $X = \Re^n$, or where X is specified by some equations and inequalities involving continuous functions. Generally, continuous problems are analyzed using the mathematics of calculus and convexity.

Discrete problems are basically those that are not continuous, usually because of finiteness of the constraint set X. Typical examples arise in scheduling, route planning, and matching, among many others. An important type of discrete problems is *integer programming*, where there is a constraint that the optimization variables must take only integer values from some range (such as 0 or 1). Discrete problems are addressed with combinatorial and discrete mathematics, and other special methodology, some of which relates to continuous problems.

Nonlinear programming, the case where either the cost function f is nonlinear or the constraint set X is specified by nonlinear equations and inequalities, lies squarely within the continuous problem category. Several other important types of optimization problems have more of a hybrid character, but are strongly connected with nonlinear programming.

In particular, *linear programming* problems, the case where f is linear and X is a polyhedral set specified by linear inequality constraints, have many of the characteristics of continuous problems. However, they also have in part a combinatorial structure: according to a fundamental

theorem [Prop. B.20(c) in Appendix B], optimal solutions of a linear program can be found by searching among the (finite) set of extreme points of X. Thus the search for an optimum can be confined within this finite set, and indeed one of the most popular methods for linear programming, the simplex method, is based on this idea. We note, however, that other important linear programming methods, such as the interior point methods to be discussed in Section 5.1, and some of the duality-based methods in Chapters 6 and 7, rely on the continuous structure of linear programs and are based on nonlinear programming ideas.

Another major class of problems with a strongly hybrid character is *network optimization*. Here the constraint set X is a polyhedral set that is defined in terms of a graph consisting of nodes and directed arcs. The salient feature of this constraint set is that its extreme points have *integer components*, something that is not true for general polyhedral sets. As a result, important combinatorial or integer programming problems, such as for example some matching and shortest path problems, can be embedded and solved within a continuous network optimization framework.

Our objective in this book is to focus on nonlinear programming problems, their continuous character, and the associated mathematical analysis. However, we will maintain a view to other broad classes of problems that have in part a discrete character. In particular, we will consider extensively those aspects of linear programming that bear a close relation to nonlinear programming methodology, such as interior point methods and polyhedral convexity (see Section 5.1, and Sections B.3 and B.4 in Appendix B).

We will also discuss various aspects of network optimization problems that relate to both their continuous and their discrete character in Sections 3.1, 3.8, and 6.4.3. A far more extensive treatment, which straddles the boundary between continuous and discrete optimization, can be found in the author's network optimization textbook [Ber98].

Finally, we will discuss some of the major methods for integer programming and combinatorial optimization, such as branch-and-bound and Lagrangian relaxation. These methods rely on duality and the solution of continuous optimization subproblems (see Sections 6.5 and 7.5).

Let us also note that there is a methodological division within the class of continuous problems. On one hand we have problems where the cost function f is a differentiable, or even twice differentiable. This allows a calculus-based analysis, which will be the primary approach in our analysis of Chapters 1-5. On the other hand we have problems where f is nondifferentiable but is convex. This requires a line of analysis that relies on convexity (rather than differentiability). It will be our primary approach in Chapters 6 and 7. Of course differentiability can also play an important role within the context of convex problems. Moreover, nondifferentiable (convex or nonconvex) problems can often be fruitfully converted to differentiable ones by using smoothing transformations, as we will explain later (see Section 2.7).

Unconstrained Differentiable Optimization: An Outline

In this chapter and the next, we focus on unconstrained differentiable nonlinear programming problems. These are problems where f is at least once continuously differentiable and where $X = \Re^n$:

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in \Re^n. \end{aligned} \tag{UP}$$

The first and second derivatives of f play an important role in the characterization of optimal solutions via necessary and sufficient conditions, which are the main subject of Section 1.1. The first and second derivatives are also central in numerical algorithms for computing approximately optimal solutions. There is a broad range of such algorithms, with a rich theory, which is discussed in Sections 1.2-1.4, and in Chapter 2.

In Chapter 2, we will also discuss two types of special problem structures. The first involves an *additive cost function*,

$$f(x) = \sum_{i=1}^{m} f_i(x),$$

where the component functions f_i are differentiable. Many problems of interest, arising in signal processing, machine learning, and neural network training have this form. For such problems, an incremental algorithmic approach is often used, which involves sequential steps along the gradients of the functions f_i, with intermediate adjustment of x after processing each f_i. We will discuss algorithms of this type and their applications in Section 2.4. These methods include incremental versions of the gradient, Newton, and Gauss-Newton methods, discussed in Chapter 1. In many important contexts, some of the components f_i are nondifferentiable. Incremental methods for problems of this kind will also be developed in Chapter 7. In Section 2.5, we will also discuss various methods in a distributed asynchronous computation setting, involving multiple processors and communication delays between the processors.

The second type of special structure that we will discuss is *optimal control problems*, which involve a discrete-time dynamic system (see Section 2.6). These are problems of potentially very large dimension, whose structure can be exploited for the convenient implementation of gradient and Newton-like methods. An important characteristic of these problems is that the gradient and Newton directions can be computed economically, using the dynamic system structure.

Both additive cost and optimal control problems arise also in constrained settings, and on occasion we will pause to discuss constrained variants in subsequent chapters. A third type of special structure that arises primarily in a constrained setting is *network optimization problems*.

We will discuss these problems in Sections 3.1 and 3.8, after the development of the relevant constrained optimization algorithms.

Our analysis in this chapter will focus on explaining the basic properties of the various methods, and primarily their convergence and rate of convergence properties. Many of these properties can be adequately and intuitively explained using a quadratic problem. The rationale is that behavior of an algorithm for a positive definite quadratic cost function is typically a correct predictor of its behavior for a twice differentiable cost function in the neighborhood of a minimum where the Hessian matrix is positive definite. Since the gradient is zero at that minimum, the positive definite quadratic term dominates the other terms in the series expansion of f, and the asymptotic behavior of the method does not depend on terms of order higher than two. This line of analysis underlies some of the most widely used unconstrained optimization methods, such as Newton, Gauss-Newton, quasi-Newton, and conjugate direction methods. However, the rationale for these methods is weakened when the Hessian is singular at the minimum, since in this case third and higher order terms may become significant. Then it may be best to use first order methods and analysis that relies primarily on the first order differentiability of the cost function. Consistent with this idea, we will discuss both first and second order methods, in this and later chapters, and explain the circumstances under which each type of method is most suitable.

1.1 OPTIMALITY CONDITIONS

1.1.1 Variational Ideas

The main ideas underlying optimality conditions in nonlinear programming usually admit simple explanations although their detailed proofs are sometimes tedious. For this reason, we will first discuss informally these ideas in the present subsection, and leave detailed statements of results and proofs for the next subsection.

Local and Global Minima

A vector x^* is an *unconstrained local minimum* of a function $f : \Re^n \mapsto \Re$ if it is no worse than its neighbors; that is, if there exists an $\epsilon > 0$ such that†

$$f(x^*) \leq f(x), \qquad \forall\ x \in \Re^n \text{ with } \|x - x^*\| < \epsilon.$$

† Unless stated otherwise, we use the standard Euclidean norm $\|x\| = \sqrt{x'x}$. Appendix A describes in detail our mathematical notation and terminology.

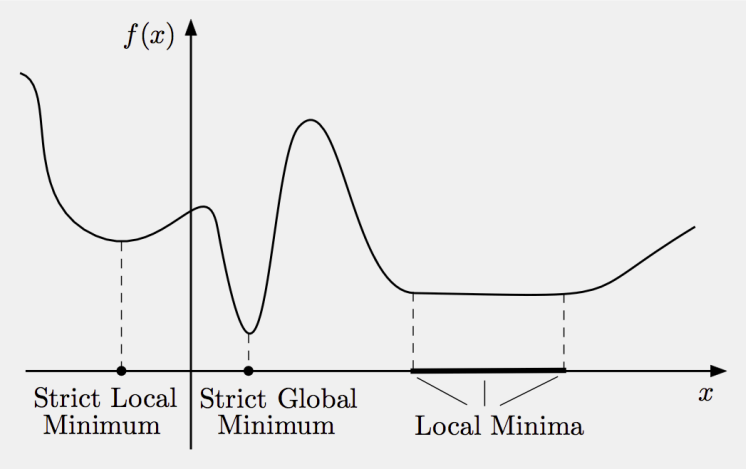

Figure 1.1.1. Unconstrained local and global minima in one dimension.

A vector x^* is an *unconstrained global minimum* of f if it is no worse than all other vectors; that is,

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

The unconstrained local or global minimum x^* is said to be *strict* if the corresponding inequality above is strict for $x \neq x^*$. Figure 1.1.1 illustrates these definitions.

The definitions of local and global minima can be extended to the case where minimization of f is subject to a constraint set $X \subset \Re^n$, the points of which are called *feasible*. In particular, we say that x^* is a *local minimum of f over X* if $x^* \in X$ and there is an $\epsilon > 0$ such that

$$f(x^*) \leq f(x), \qquad \forall\ x \in X \text{ with } \|x - x^*\| < \epsilon;$$

see Fig. 1.1.2. The definitions of a global and a strict minimum of f over X are analogous.

Local and global *maxima* are similarly defined. In particular, x^* is an unconstrained local (global) maximum of f, if x^* is an unconstrained local (global) minimum of the function $-f$.

Necessary Conditions for Optimality

If the cost function is differentiable, we can use gradients to compare the cost of a vector with the cost of its close neighbors. In particular, we consider small variations Δx from a given vector x^*, which approximately, up to first order, yield a cost variation

$$f(x^* + \Delta x) - f(x^*) \approx \nabla f(x^*)' \Delta x,$$

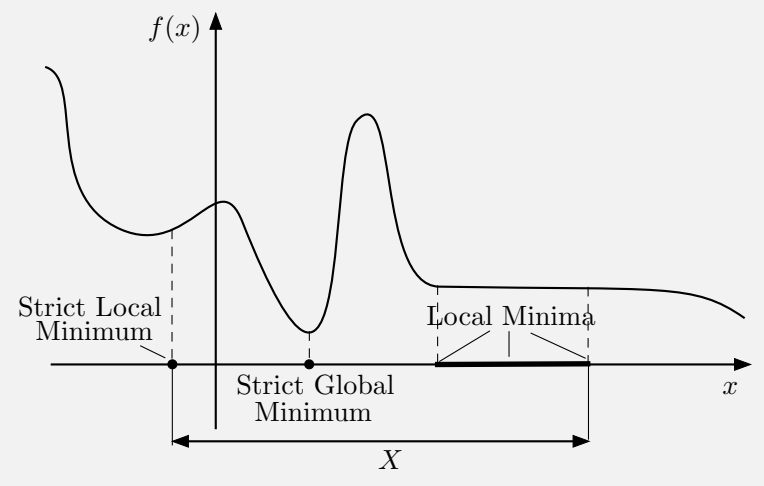

Figure 1.1.2. Local and global minima of f over the constraint set X.

and, up to second order, yield a cost variation

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

We expect that if x^* is an unconstrained local minimum, the first order cost variation due to a small variation Δx is nonnegative:

$$\nabla f(x^*)'\Delta x = \sum_{i=1}^{n} \frac{\partial f(x^*)}{\partial x_i}\Delta x_i \geq 0.$$

In particular, by taking Δx to be positive and negative multiples of the unit coordinate vectors (all coordinates equal to zero except for one which is equal to unity), we obtain $\partial f(x^*)/\partial x_i \geq 0$ and $\partial f(x^*)/\partial x_i \leq 0$, respectively, for all $i = 1, \ldots, n$. Equivalently, we have the necessary condition

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

[originally formulated by Fermat in 1637 in the short treatise "Methodus ad Disquirendam Maximam et Minimam" without proof (of course!)]. This condition is proved formally in Prop. 1.1.1, given in the next subsection.

The idea that at a local minimum x^*, the condition $\nabla f(x^*)'\Delta x \geq 0$ should hold for small variations Δx applies more broadly, including for problems with convex constraint sets X, when it takes the form

$$\nabla f(x^*)'(x - x^*) \geq 0, \qquad \forall\ x \in X.$$

This condition will be shown in Prop. 1.1.2 for the case of a convex cost function. In Chapter 3, it will become the basis for constrained versions of the computational methods of the present chapter.

We also expect that the second order cost variation due to a small variation Δx must also be nonnegative:

$$\nabla f(x^*)'\Delta x + \tfrac{1}{2}\Delta x'\nabla^2 f(x^*)\Delta x \geq 0.$$

Since $\nabla f(x^*)'\Delta x = 0$, we obtain

$$\Delta x'\nabla^2 f(x^*)\Delta x \geq 0, \qquad \forall\ \Delta x \in \Re^n,$$

which implies that

$$\nabla^2 f(x^*) : \text{positive semidefinite.}$$

We prove this necessary condition in the next subsection (Prop. 1.1.1). Appendix A reviews the definition and properties of positive definite and positive semidefinite matrices.

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

The Case of a Convex Cost Function

Convexity plays a very important role in nonlinear programming.† One reason is that when the cost function f is convex, there is no distinction between local and global minima; every local minimum is also global. The idea is illustrated in Fig. 1.1.3 and the formal proof is given in Prop. 1.1.2.

Another important fact is that the first order condition $\nabla f(x^*) = 0$ is also sufficient for optimality if f is convex. This is established in Prop. 1.1.3. The proof is based on a basic property of a convex function f: the linear approximation at a point x^* based on the gradient, i.e.,

$$f(x^*) + \nabla f(x^*)'(x - x^*),$$

underestimates $f(x)$, so if $\nabla f(x^*) = 0$, then $f(x^*) \leq f(x)$ for all x (see Prop. B.3 in Appendix B).

Sufficient Conditions for Optimality

If f is not convex, the first and second order necessary conditions can fail to guarantee local optimality of x^*. This is illustrated in Fig. 1.1.4. However, by strengthening the second order condition we obtain sufficient conditions

† The theory of convex sets and functions, particularly as it relates to optimization theory, is reviewed in Appendix B and is discussed extensively in the author's books [BNO03] and [Ber09].

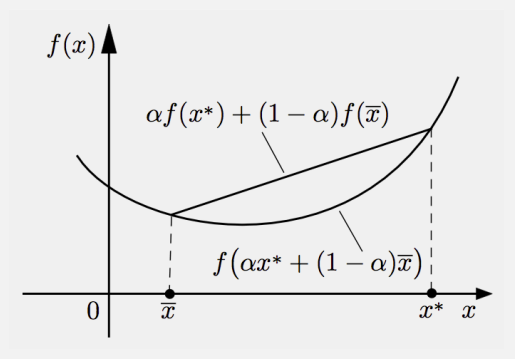

Figure 1.1.3. Illustration of why a local minimum of a convex function must also global. Suppose that f is convex and that x^* is not a global minimum, so that there exists $\bar{x}$ with $f(\bar{x}) < f(x^*)$. By convexity, for all $\alpha \in (0,1)$,

$$f\big(\alpha x^* + (1-\alpha)\bar{x}\big) \leq \alpha f(x^*) + (1-\alpha) f(\bar{x}) < f(x^*).$$

Thus, f has value strictly lower than $f(x^*)$ at every point on the line segment connecting x^* with $\bar{x}$, except x^*. Therefore x^* cannot be a local minimum.

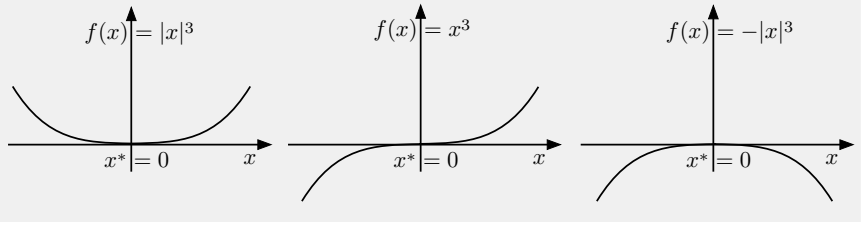

Figure 1.1.4. Illustration of the first order necessary optimality condition of zero slope [$\nabla f(x^*) = 0$] and the second order necessary optimality condition of nonnegative curvature [$\nabla^2 f(x^*) \geq 0$] for functions of one variable. The first order condition is satisfied not only by local minima, but also by local maxima and "inflection" points, such as the one on the middle figure above. In some cases [e.g. $f(x) = x^3$ and $f(x) = -|x|^3$] the second order condition is also satisfied by local maxima and inflection points. If the function f is convex, the condition $\nabla f(x^*) = 0$ is necessary and sufficient for global optimality of x^*.

for optimality. In particular, consider a vector x^* that satisfies the first order necessary optimality condition

$$\nabla f(x^*) = 0, \tag{1.1}$$

and also satisfies the following strengthened form of the second order necessary optimality condition

$$\nabla^2 f(x^*) : \text{positive definite}, \tag{1.2}$$

(i.e., the Hessian is positive definite rather than semidefinite). Then, for all $\Delta x \neq 0$ we have

$$\Delta x' \nabla^2 f(x^*) \Delta x > 0,$$

implying that at x^* the second order variation of f due to a small nonzero variation Δx is positive. Thus, f tends to increase strictly with small excursions from x^*, suggesting that the above conditions (1.1) and (1.2) are sufficient for local optimality of x^*. This is indeed established in Prop. 1.1.5.

Local minima that don't satisfy the positive definiteness sufficient condition (1.2) are called *singular*; otherwise they are called *nonsingular*. Singular local minima are harder to deal with for two reasons. First, in the absence of convexity of f, their optimality cannot be ascertained using easily verifiable sufficiency conditions. Second, in their neighborhood, the behavior of the most commonly used optimization algorithms tends to be slow and/or erratic, as we will see in the subsequent sections.

Quadratic Cost Functions

Consider the quadratic function

$$f(x) = \tfrac{1}{2} x' Q x - b' x,$$

where Q is a symmetric $n \times n$ matrix and b is a vector in $\Re^n$. If x^* is a local minimum of f, we must have, by the necessary optimality conditions,

$$\nabla f(x^*) = Q x^* - b = 0, \qquad \nabla^2 f(x^*) = Q : \text{ positive semidefinite.}$$

Thus, if Q is not positive semidefinite, f can have no local minima. If Q is positive semidefinite, f is convex [Prop. B.4(d) of Appendix B], so any vector x^* satisfying the first order condition $\nabla f(x^*) = Q x^* - b = 0$ is a global minimum of f. On the other hand there may not exist a solution of the equation $\nabla f(x^*) = Q x^* - b = 0$ if Q is singular. If, however, Q is positive definite (and hence invertible, by Prop. A.20 of Appendix A), the equation $Q x^* - b = 0$ can be solved uniquely and the vector $x^* = Q^{-1} b$ is the unique global minimum. This is consistent with Prop. 1.1.3(a) to be given shortly, which asserts that strictly convex functions can have at most one global minimum [f is strictly convex if and only if Q is positive definite; Prop. B.4(d) of Appendix B]. Figure 1.1.5 illustrates the various special cases considered.

Quadratic cost functions are important in nonlinear programming because they arise frequently in applications, but they are also important for another reason. From the second order expansion around a local minimum x^*,

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

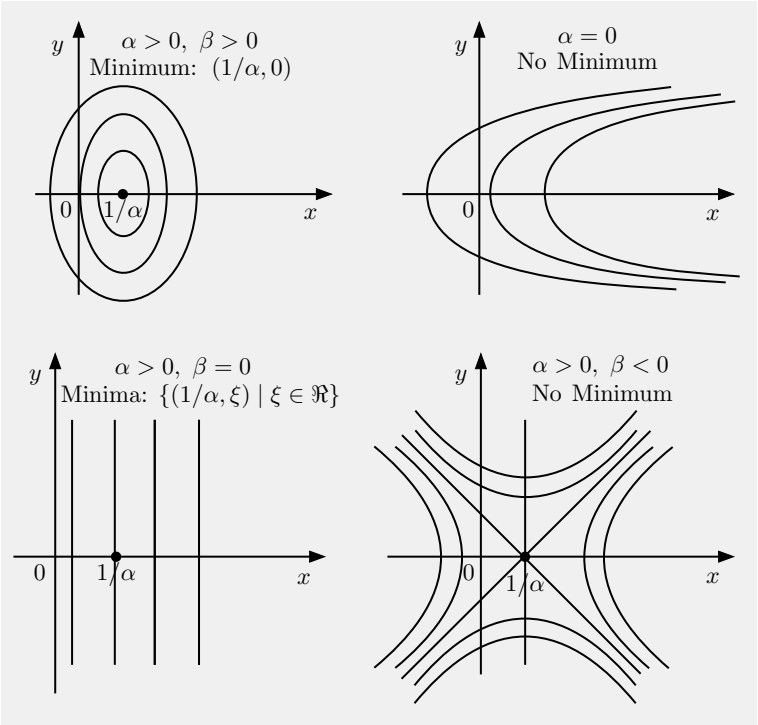

Figure 1.1.5. Illustration of the level sets $\{x \mid f(x) \leq c\}$ of the quadratic cost function $f : \Re^2 \mapsto \Re$ given by

$$f(x, y) = \tfrac{1}{2}\left(\alpha x^2 + \beta y^2\right) - x$$

for various values of α and β.

it is seen that a nonquadratic cost function can be approximated well by a quadratic function if x^* nonsingular [$\nabla^2 f(x^*)$: positive definite]. This means that we can carry out much of our analysis and experimentation with algorithms using positive definite quadratic functions and expect that the conclusions will largely carry over to more general cost functions near convergence to such local minima. However, for local minima near which the Hessian matrix either does not exist or is singular, the higher than second order terms in the series expansion are not negligible and an algorithmic analysis based on quadratic cost functions will likely be seriously flawed.

Existence of Optimal Solutions

In many cases it is useful to know that there exists at least one global

minimum of a function f over a set X. Generally, such a minimum need not exist. For example, the scalar functions $f(x) = x$ and $f(x) = e^x$ have no global minima over the set of real numbers. The first function decreases without bound to $-\infty$ as x tends toward $-\infty$, while the second decreases toward 0 as x tends toward $-\infty$ but always takes positive values.

Existence of at least one global minimum is guaranteed if f is a continuous function and X is a compact subset of $\Re^n$. This is the *Weierstrass theorem*; see Prop. A.8 in Appendix A. By a related result, also shown in Prop. A.8, existence of an optimal solution is guaranteed if f is continuous, X is closed, and f is coercive over X, i.e., $f(x^k) \to \infty$ for any sequence $\{x^k\} \subset X$ with $\|x^k\| \to \infty$. Section 3.1.2 presents a more advanced view of the existence question, where X is not required to be bounded.

Why do we Need Optimality Conditions?

Hardly anyone would doubt that optimality conditions are fundamental to the analysis of an optimization problem. In practice, however, optimality conditions play an important role in a variety of contexts, some of which may not be readily apparent.

The most straightforward method to use optimality conditions for solving an optimization problem, is as follows: First, find all points satisfying the first order necessary condition $\nabla f(x) = 0$; then (if f is not known to be convex), check the second order necessary condition ($\nabla^2 f$ is positive semidefinite) for each of these points, filtering out those that do not satisfy it; finally for the remaining candidates, check if $\nabla^2 f$ is positive definite, in which case we are sure that they are strict local minima.

A slightly different alternative is to find all points satisfying the necessary conditions, and to declare as global minimum the one with smallest cost value. However, here it is essential to know that a global minimum exists. As an example, for the one-dimensional function

$$f(x) = x^2 - x^4,$$

the points satisfying the necessary condition

$$\nabla f(x) = 2x - 4x^3 = 0$$

are 0, $1/\sqrt{2}$, and $-1/\sqrt{2}$, and of these, 0 gives the smallest cost value. Nonetheless, we cannot declare 0 as the global minimum, because we don't know if a global minimum exists. Indeed, in this example none of the points 0, $1/\sqrt{2}$, and $-1/\sqrt{2}$ is a global minimum, because f decreases to $-\infty$ as $|x| \to \infty$, and has no global minimum. Here is an example where the approach can be applied.

Example 1.1.1 (Arithmetic-Geometric Mean Inequality)

We want to show the following classical inequality [due to Cauchy (1821)]:

$$(x_1 x_2 \cdots x_n)^{1/n} \leq \frac{\sum_{i=1}^n x_i}{n}$$

for any set of positive numbers x_i, $i = 1, \ldots, n$. By making the change of variables

$$y_i = \ln(x_i), \qquad i = 1, \ldots, n,$$

we have $x_i = e^{y_i}$, so that this inequality is equivalently written as

$$e^{\frac{y_1 + \cdots + y_n}{n}} \leq \frac{e^{y_1} + \cdots + e^{y_n}}{n},$$

which must be shown for all scalars $y_1, \ldots, y_n$. Note that with this transformation, the nonnegativity requirements on the variables have been eliminated.

One approach to proving the above inequality is to minimize the function

$$\frac{e^{y_1} + \cdots + e^{y_n}}{n} - e^{\frac{y_1 + \cdots + y_n}{n}},$$

and to show that its minimal value is 0. An alternative, which works better if we use optimality conditions, is to minimize instead

$$e^{y_1} + \cdots + e^{y_n},$$

over all $y = (y_1, \ldots, y_n)$ such that

$$y_1 + \cdots + y_n = s$$

for an arbitrary scalar s, and to show that the optimal value is no less than $ne^{s/n}$.

To this end, we use an elimination technique, a common device to convert constrained optimization problems to unconstrained ones. In particular, we use the constraint $y_1 + \cdots + y_n = s$ to eliminate the variable y_n, thereby obtaining the equivalent *unconstrained* problem of minimizing

$$g(y_1, \ldots, y_{n-1}) = e^{y_1} + \cdots + e^{y_{n-1}} + e^{s - y_1 - \cdots - y_{n-1}},$$

over $y_1, \ldots, y_{n-1}$. The necessary conditions $\partial g / \partial y_i = 0$ yield the system of equations

$$e^{y_i} = e^{s - y_1 - \cdots - y_{n-1}}, \qquad i = 1, \ldots, n-1,$$

or

$$y_i = s - y_1 - \cdots - y_{n-1}, \qquad i = 1, \ldots, n-1.$$

This system has only one solution: $y_i^* = s/n$ for all i. The solution must be the unique global minimum if we can show that that there exists a global minimum. Indeed, it can be seen that the function $g(y_1, \ldots, y_{n-1})$ is coercive,

so it has an unconstrained global minimum (Prop. A.8, in Appendix A). Therefore, $(s/n, \ldots, s/n)$ is this minimum. Thus the optimal value of

$$e^{y_1} + \cdots + e^{y_n}$$

is $ne^{s/n}$, which as argued earlier, is sufficient to show the arithmetic-geometric mean inequality.

It is important to realize, however, that except under very favorable circumstances, using optimality conditions to obtain a solution as described above does *not* work. The reason is that solving for x the system of equations $\nabla f(x) = 0$ is usually nontrivial; algorithmically, it is typically as difficult as solving the original optimization problem.

The principal context in which optimality conditions become useful will not become apparent until we consider iterative optimization algorithms in subsequent sections. We will see that optimality conditions often provide the basis for the development and the analysis of algorithms. In particular, algorithms recognize solutions by checking whether they satisfy various optimality conditions and terminate when such conditions hold approximately. Furthermore, the behavior of various algorithms in the neighborhood of a local minimum often depends on whether various optimality conditions are satisfied at that minimum. Thus, for example, sufficiency conditions play a key role in assertions regarding the speed of convergence of various algorithms.

There is one other important context, prominently arising in microeconomic theory, where optimality conditions provide the basis for analysis. Here one is interested primarily not in finding an optimal solution, but rather in how the optimal solution is affected by changes in the problem data. For example, an economist may be interested in how the prices of some raw materials will affect the availability of certain goods that are produced by using these raw materials; the assumption here is that the amounts produced are the variables of a profit optimization problem, which is solved by the corresponding producers. This type of reasoning is known as *sensitivity analysis*, and is discussed next.

Sensitivity

Suppose that we want to quantify the variation of the optimal solution as a vector of parameters changes. In particular, consider the optimization problem

$$\begin{aligned} &\text{minimize} \quad f(x, a) \\ &\text{subject to} \quad x \in \Re^n, \end{aligned}$$

where $f : \Re^{m+n} \mapsto \Re$ is a twice continuously differentiable function involving the m-dimensional parameter vector a. Let $x(a)$ denote the global minimum corresponding to a, assuming for the moment that it exists, is

unique, and it is differentiable as a function of a. By the first order necessary condition we have

$$\nabla_x f\big(x(a), a\big) = 0, \qquad \forall\ a \in \Re^m,$$

and by differentiating this relation with respect to a, we obtain

$$\nabla x(a) \nabla^2_{xx} f\big(x(a), a\big) + \nabla^2_{xa} f\big(x(a), a\big) = 0,$$

where the elements of the $m \times n$ gradient matrix $\nabla x(a)$ are the first partial derivatives of the components of $x(a)$ with respect to the different components of a. Assuming that the inverse below exists, we have

$$\nabla x(a) = -\nabla^2_{xa} f\big(x(a), a\big) \left(\nabla^2_{xx} f\big(x(a), a\big)\right)^{-1}, \tag{1.3}$$

which gives the first order variation of the components of the optimal x with respect to the components of a.

For the preceding analysis to be precise, we must be sure that $x(a)$ exists and is differentiable with respect to a. The principal analytical framework for this is the implicit function theorem (Prop. A.25 in Appendix A). With the aid of this theorem, we can define $x(a)$ in some sphere around a minimum $\bar{x} = x(\bar{a})$ corresponding to a nominal parameter value $\bar{a}$, assuming that the Hessian matrix $\nabla^2_{xx} f\big(\bar{x}, \bar{a}\big)$ is positive definite. Thus, the preceding development and the formula (1.3) for the matrix $\nabla x(a)$ can be justified provided the nominal local minimum $\bar{x}$ is nonsingular.

We postpone further discussion of sensitivity analysis for Sections 4.2.3, 4.3.2, and 4.3.6, where we will show constrained versions of the expression (1.3) for $\nabla x(a)$.

1.1.2 Main Optimality Conditions

We now provide formal statements and proofs of the optimality conditions discussed in the preceding section.

Proposition 1.1.1: (Necessary Optimality Conditions) Let x^* be an unconstrained local minimum of $f : \Re^n \mapsto \Re$, and assume that f is continuously differentiable in an open set S containing x^*. Then

$$\nabla f(x^*) = 0. \qquad \text{(First Order Necessary Condition)}$$

If in addition f is twice continuously differentiable within S, then

$$\nabla^2 f(x^*) : \text{positive semidefinite.} \quad \text{(Second Order Necessary Condition)}$$

Proof: Fix some $d \in \Re^n$. Then, using the chain rule to differentiate the function $g(\alpha) = f(x^* + \alpha d)$ of the scalar α, we have

$$0 \leq \lim_{\alpha \downarrow 0} \frac{f(x^* + \alpha d) - f(x^*)}{\alpha} = \frac{dg(0)}{d\alpha} = d' \nabla f(x^*),$$

where the inequality follows from the assumption that x^* is a local minimum. Since d is arbitrary, the same inequality holds with d replaced by $-d$. Therefore, $d' \nabla f(x^*) = 0$ for all $d \in \Re^n$, which shows that $\nabla f(x^*) = 0$.

Assume that f is twice continuously differentiable, and let d be any vector in $\Re^n$. For all $\alpha \in \Re$, the second order expansion yields

$$f(x^* + \alpha d) - f(x^*) = \alpha \nabla f(x^*)' d + \frac{\alpha^2}{2} d' \nabla^2 f(x^*) d + o(\alpha^2).$$

Using the condition $\nabla f(x^*) = 0$ and the local optimality of x^*, we see that there is a sufficiently small $\epsilon > 0$ such that for all α with $\alpha \in (0, \epsilon)$,

$$0 \leq \frac{f(x^* + \alpha d) - f(x^*)}{\alpha^2} = \tfrac{1}{2} d' \nabla^2 f(x^*) d + \tfrac{o(\alpha^2)}{\alpha^2}.$$

Taking the limit as $\alpha \to 0$ and using the fact $\lim_{\alpha \to 0} o(\alpha^2)/\alpha^2 = 0$, we obtain $d' \nabla^2 f(x^*) d \geq 0$, showing that $\nabla^2 f(x^*)$ is positive semidefinite. **Q.E.D.**

The Convex Case

We will now consider the case where both the cost function f and the constraint set X are convex. The following proposition shows that a local minimum of f over X is also a global minimum over X. The proposition also deals with the cases where f is strictly convex and where it is strongly convex (see Appendix B for the definition and properties of strictly and strongly convex functions).

Proposition 1.1.2: If X is a convex subset of $\Re^n$ and $f : \Re^n \mapsto \Re$ is convex over X, then a local minimum of f over X is also a global minimum. If in addition f is strictly convex over X, then f has at most one global minimum over X. Moreover, if f is strongly convex and X is closed, then f has a unique global minimum over X.

Proof: Assume, to arrive at a contradiction, that x is a local minimum of f but not a global minimum. Then there exists some $y \neq x$ such that $f(y) < f(x)$. Using the convexity of f, we have

$$f\big(\alpha x + (1 - \alpha) y\big) \leq \alpha f(x) + (1 - \alpha) f(y) < f(x), \qquad \forall\ \alpha \in [0, 1).$$

This contradicts the assumption that x is a local minimum.

Assume, to arrive at a contradiction, that f is strictly convex, and two distinct global minima x and y exist. Then their average $(x+y)/2$ must belong to X, since X is convex, and the value of f at the average must be smaller than $\big(f(x)+f(y)\big)/2$, by the strict convexity of f. Since x and y are global minima, we obtain a contradiction.

A strongly convex function is coercive, so it has at least one minimum over the closed set X by Prop. A.8 in Appendix A. It is also strictly convex by Prop. B.5(a) of Appendix B, so the minimum is unique. **Q.E.D.**

The following proposition provides a simple necessary and sufficient condition for optimality; see Fig. 1.1.6.

Proposition 1.1.3: (Convex Case - Necessary and Sufficient Conditions) Let X be a convex set and let $f : \Re^n \mapsto \Re$ be a convex function over X.

(a) If f is continuously differentiable, then

$$\nabla f(x^*)'(x-x^*) \geq 0, \qquad \forall\ x \in X,$$

is a necessary and sufficient condition for a vector $x^* \in X$ to be a global minimum of f over X.

(b) If X is open and f is continuously differentiable over X, then $\nabla f(x^*) = 0$ is a necessary and sufficient condition for a vector $x^* \in X$ to be a global minimum of f over X.

Proof: (a) Using the convexity of f and Prop. B.3(a) of Appendix B, we have

$$f(x) \geq f(x^*) + \nabla f(x^*)'(x-x^*), \qquad \forall\ x \in X.$$

If the condition $\nabla f(x^*)'(x-x^*) \geq 0$ holds for all $x \in X$, then $f(x) \geq f(x^*)$ for all $x \in X$, so x^* minimizes f over X.

Conversely, assume to arrive at a contradiction that x^* minimizes f over X and that $\nabla f(x^*)'(x-x^*) < 0$ for some $x \in X$. Then, we have

$$\lim_{\alpha \downarrow 0} \frac{f\big(x^* + \alpha(x-x^*)\big) - f(x^*)}{\alpha} = \nabla f(x^*)'(x-x^*) < 0,$$

so $f\big(x^* + \alpha(x-x^*)\big)$ decreases strictly for sufficiently small $\alpha > 0$, contradicting the optimality of x^*.

(b) If $\nabla f(x^*) = 0$, the optimality of x^* follows as a special case of part (a). Conversely, assume to arrive at a contradiction that x^* minimizes f over X and that $\nabla f(x^*) \neq 0$. Then, since X is open and $x^* \in X$, there must

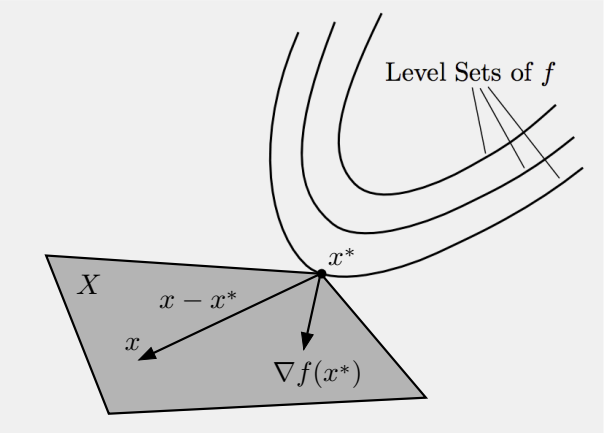

Figure 1.1.6. Illustration of the necessary and sufficient condition of Prop. 1.1.3(a) for x^* to minimize f over X. The gradient $\nabla f(x^*)$ makes an angle less or equal to $\pi/2$ with any vector of the form $x - x^*$, $x \in X$.

exist an open ball centered at x^* that is contained in X. Thus for some $x \in X$ we have

$$\nabla f(x^*)'(x - x^*) < 0.$$

The proof now proceeds as in part (a). **Q.E.D.**

The following is an illustration of the use of the preceding optimality conditions.

Example 1.1.2 (Arithmetic-Geometric Mean Inequality Revisited)

As an application of the preceding proposition, let us provide an alternative optimization-based proof of the arithmetic-geometric mean inequality, given in Example 1.1.1. We argued there that showing the inequality is equivalent to showing that for an arbitrary scalar s, the minimum value of the convex function

$$g(y) = e^{y_1} + \cdots + e^{y_n},$$

over all $y = (y_1, \ldots, y_n)$ such that

$$y_1 + \cdots + y_n = s$$

is no less than $ne^{s/n}$. We verified this by showing that the symmetric solution

$$y_1^* = \cdots = y_n^* = s/n$$

is optimal, via conversion of the problem to an unconstrained problem through elimination of one of the variables. Alternatively, we can prove the optimality

of this solution by checking the sufficiency condition of Prop. 1.1.3(a) (the cost function and constraint are easily verified to be convex). This condition is written as

$$\nabla g(y^*)'(y-y^*) = (\,e^{s/n},\ldots,e^{s/n}\,)' \begin{pmatrix} y_1 - s/n \\ \vdots \\ y_n - s/n \end{pmatrix} \geq 0,$$

or

$$e^{s/n}(y_1 + \cdots + y_n - s) \geq 0$$

for all y with $y_1 + \cdots + y_n = s$, which is evidently true.

Let us use Prop. 1.1.2 and the optimality condition of Prop. 1.1.3(a) to prove a basic theorem of analysis and optimization, which is illustrated in Fig. 1.1.7 and will be used frequently in this book.

Proposition 1.1.4: (Projection Theorem) Let X be a closed convex set and let $\|\cdot\|$ be the Euclidean norm.

(a) For every $x \in \Re^n$, there exists a unique vector that minimizes $\|y-x\|$ over all $y \in X$. This vector is called the *projection of x on X*, and is denoted by $[x]^+$, i.e.,

$$[x]^+ = \arg\min_{y\in X} \|y-x\|.$$

(b) Given some $x \in \Re^n$, a vector $x^* \in X$ is equal to $[x]^+$ if and only if

$$(y-x^*)'(x-x^*) \leq 0, \qquad \forall\ y \in X. \tag{1.4}$$

(c) The mapping $f : \Re^n \mapsto X$ defined by $f(x) = [x]^+$ is continuous and nonexpansive, i.e.,

$$\big\|[x]^+ - [y]^+\big\| \leq \|x-y\|, \qquad \forall\ x, y \in \Re^n.$$

Proof: (a) Given x, minimizing $\|y-x\|$ over $y \in X$ is equivalent to minimizing the convex and differentiable function

$$f(y) = \tfrac{1}{2}\|y-x\|^2$$

over X. Since, f is strongly convex, existence and uniqueness of the minimizing vector follows from Prop. 1.1.2.

(b) By Prop. 1.1.3, x^* minimizes f over X if and only if

$$\nabla f(x^*)'(y-x^*) \geq 0, \qquad \forall\ y \in X.$$

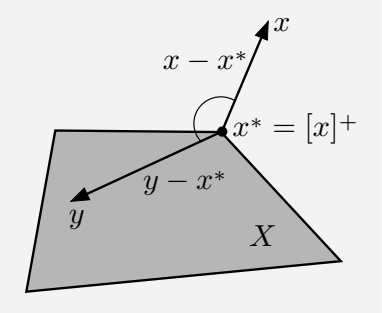

Figure 1.1.7. Illustration of the condition (1.4) of the projection theorem. For x^* to to be the projection of x on X, the vector $x - x^*$ should make an angle greater or equal to $\pi/2$ with any vector of the form $y - x^*$, $y \in X$.

Since $\nabla f(x^*) = x^* - x$, this condition is equivalent to Eq. (1.4).

(c) Let x and y be elements of $\Re^n$. From part (b), we have

$$\big(w - [x]^+\big)'\big(x - [x]^+\big) \leq 0, \qquad \forall\ w \in X.$$

Since $[y]^+ \in X$, we can use $w = [y]^+$ in the preceding relation and obtain

$$\big([y]^+ - [x]^+\big)'\big(x - [x]^+\big) \leq 0.$$

Exchanging the roles of x and y, we also obtain

$$\big([x]^+ - [y]^+\big)'\big(y - [y]^+\big) \leq 0.$$

Adding these two inequalities, we have

$$\big([y]^+ - [x]^+\big)'\big(x - [x]^+ - y + [y]^+\big) \leq 0.$$

By rearranging and by using the Schwarz inequality, we obtain

$$\big\|[y]^+ - [x]^+\big\|^2 \leq \big([y]^+ - [x]^+\big)'(y - x) \leq \big\|[y]^+ - [x]^+\big\| \cdot \|y - x\|,$$

showing that $[\cdot]^+$ is nonexpansive and *a fortiori* continuous. **Q.E.D.**

Sufficient Conditions without Convexity

In the absence of convexity, we have the following sufficiency conditions for local optimality.

Proposition 1.1.5: (Second Order Sufficient Optimality Conditions) Let $f : \Re^n \mapsto \Re$ be twice continuously differentiable over an open set S. Suppose that a vector $x^* \in S$ satisfies the conditions

$$\nabla f(x^*) = 0, \qquad \nabla^2 f(x^*) : \text{positive definite.}$$

Then, x^* is a strict unconstrained local minimum of f. In particular, there exist scalars $\gamma > 0$ and $\epsilon > 0$ such that

$$f(x) \geq f(x^*) + \frac{\gamma}{2}\|x - x^*\|^2, \qquad \forall\ x \ \text{ with } \|x - x^*\| < \epsilon. \tag{1.5}$$

Proof: Denote by λ the smallest eigenvalue of $\nabla^2 f(x^*)$. By Prop. A.20(b) of Appendix A, λ is positive since $\nabla^2 f(x^*)$ is positive definite. Furthermore, by Prop. A.18(b) of Appendix A,

$$d'\nabla^2 f(x^*)d \geq \lambda\|d\|^2, \qquad \forall\ d \in \Re^n.$$

Using this relation, the hypothesis $\nabla f(x^*) = 0$, and a second order expansion, we have for all d

$$\begin{aligned} f(x^* + d) - f(x^*) &= \nabla f(x^*)'d + \tfrac{1}{2}d'\nabla^2 f(x^*)d + o\big(\|d\|^2\big) \\ &\geq \frac{\lambda}{2}\|d\|^2 + o\big(\|d\|^2\big) \\ &= \left(\frac{\lambda}{2} + \frac{o(\|d\|^2)}{\|d\|^2}\right)\|d\|^2. \end{aligned}$$

Choose any $\epsilon > 0$ and $\gamma > 0$ such that

$$\frac{\lambda}{2} + \frac{o\big(\|d\|^2\big)}{\|d\|^2} \geq \frac{\gamma}{2}, \qquad \forall\ d \ \text{ with } \|d\| < \epsilon.$$

Then Eq. (1.5) is satisfied. **Q.E.D.**

E X E R C I S E S

1.1.1

For each value of the scalar β, find the set of all stationary points of the following function of the two variables x and y

$$f(x, y) = x^2 + y^2 + \beta xy + x + 2y.$$

Which of these stationary points are global minima?

1.1.2

In each of the following problems fully justify your answer using optimality conditions.

(a) Show that the 2-dimensional function $f(x, y) = (x^2 - 4)^2 + y^2$ has two global minima and one stationary point, which is neither a local maximum nor a local minimum.

(b) Find all local minima of the 2-dimensional function $f(x, y) = \frac{1}{2}x^2 + x\cos y$.

(c) Find all local minima and all local maxima of the 2-dimensional function $f(x, y) = \sin x + \sin y + \sin(x + y)$ within the set

$$\big\{(x, y) \mid 0 < x < 2\pi,\ 0 < y < 2\pi\big\}.$$

(d) Show that the 2-dimensional function $f(x, y) = (y - x^2)^2 - x^2$ has only one stationary point, which is neither a local maximum nor a local minimum.

(e) Consider the minimization of the function f in part (d) subject to no constraint on x and the constraint $-1 \leq y \leq 1$ on y. Show that there exists at least one global minimum and find all global minima.

1.1.3 [Hes75]

Let $f : \Re^n \mapsto \Re$ be a differentiable function. Suppose that a point x^* is a local minimum of f along every line that passes through x^*; that is, the function

$$g(\alpha) = f(x^* + \alpha d)$$

is minimized at $\alpha = 0$ for all $d \in \Re^n$.

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

(b) Show by example that x^* need not be a local minimum of f. *Hint*: Consider the function of two variables

$$f(y, z) = (z - py^2)(z - qy^2),$$

where $0 < p < q$; see Fig. 1.1.8. Show that $(0, 0)$ is a local minimum of f along every line that passes through $(0, 0)$. Moreover, if $p < m < q$, then

$$f(y, my^2) < 0 \qquad \forall\ y \neq 0,$$

while $f(0, 0) = 0$, so $(0, 0)$ is not a local minimum of f.

1.1.4

Use optimality conditions to show that for all $x > 0$ we have

$$\frac{1}{x} + x \geq 2.$$

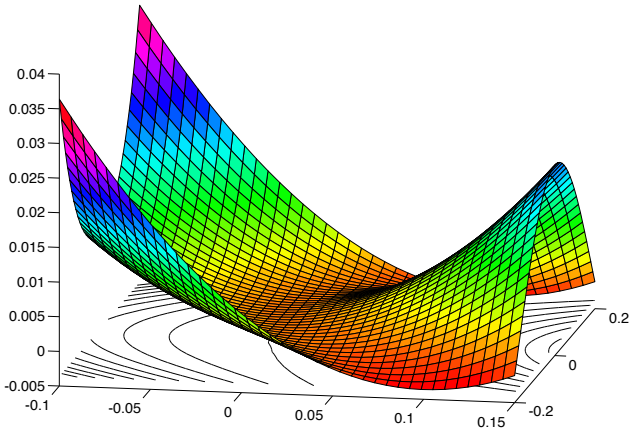

Figure 1.1.8. Three-dimensional graph of the function $f(y,z) = (z-py^2)(z-qy^2)$ for $p = 1$ and $q = 4$ (cf. Exercise 1.1.3). The origin is a local minimum with respect to every line that passes through it, but is not a local minimum of f.

1.1.5

Find the rectangular parallelepiped of unit volume that has the minimum surface area. *Hint*: By eliminating one of the dimensions, show that the problem is equivalent to the minimization over $x > 0$ and $y > 0$ of

$$f(x,y) = xy + \frac{1}{x} + \frac{1}{y}.$$

Show that the sets $\big\{(x,y) \mid f(x,y) \leq \gamma,\ x > 0,\ y > 0\big\}$ are compact for all scalars γ.

1.1.6 (The Weber Point of a Set of Points)

We want to find a point x in the plane whose sum of weighted distances from a given set of points $y_1, \ldots, y_m$ is minimized. Mathematically, the problem is

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{m} w_i \|x - y_i\| \\ &\text{subject to} && x \in \Re^n, \end{aligned}$$

where $w_1, \ldots, w_m$ are given positive scalars.

(a) Show that there exists a global minimum for this problem and that it can be realized by means of the mechanical model shown in Fig. 1.1.9.

(b) Is the optimal solution always unique?

(c) Show that an optimal solution minimizes the potential energy of the mechanical model of Fig. 1.1.9, defined as $\sum_{i=1}^{m} w_i h_i$, where h_i is the height of the ith weight, measured from some reference level.

Note: This problem stems from Weber's work [Web29], which is generally viewed as the starting point of location theory.

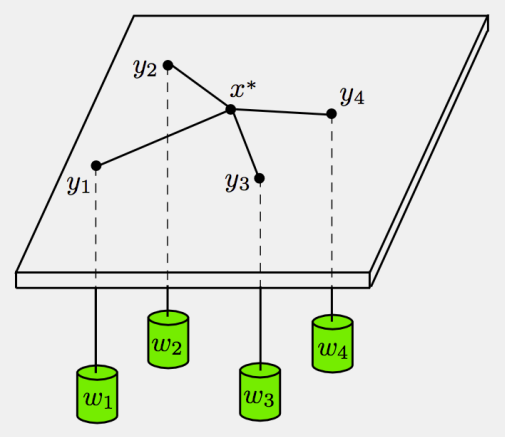

Figure 1.1.9. Mechanical model (known as the Varignon frame) associated with the Weber problem (Exercise 1.1.6). It consists of a board with a hole drilled at each of the given points y_i. Through each hole, a string is passed with the corresponding weight w_i attached. The other ends of the strings are tied with a knot as shown. In the absence of friction or tangled strings, the forces at the knot reach equilibrium when the knot is located at an optimal solution x^*.

1.1.7 (Fermat-Torricelli-Viviani Problem)

Given a triangle in the plane, consider the problem of finding a point whose sum of distances from the vertices of the triangle is minimal. Show that such a point is either a vertex, or else it is such that each side of the triangle is seen from that point at a 120 degree angle (this is known as the Torricelli point). *Note*: This problem, whose detailed history is traced in [BMS99], was suggested by Fermat to Torricelli who solved it. Viviani also solved the problem a little later and proved the following generalization: Suppose that x_i, $i = 1, \ldots, m$, are points in the plane, and x is a point in their convex hull such that $x \neq x_i$ for all i, and the angles $x_i x \bar{x}_{i+1}$, $i < m$, and $x_m \bar{x} x_1$ are all equal to $2\pi/m$. Then x minimizes $\sum_{i=1}^{m} \|z - x_i\|$ over all z in the plane (show this as an exercise by using sufficient optimality conditions; compare with the preceding exercise). Fermat is credited with being the first to study systematically optimization problems in geometry.

1.1.8 (Diffraction Law in Optics)

Let p and q be two points on the plane that lie on opposite sides of a horizontal axis. Assume that the speed of light from p and from q to the horizontal axis is v and w, respectively, and that light reaches a point from other points along paths of minimum travel time. Find the path that a ray of light would follow from p to q.

1.1.9 (www)

Let $f : \Re^n \mapsto \Re$ be a twice continuously differentiable function that satisfies

$$m\|y\|^2 \leq y'\nabla^2 f(x)y \leq M\|y\|^2, \qquad \forall\ x,\, y \in \Re^n,$$

where m and M are some positive scalars. Show that f has a unique global minimum x^*, which satisfies

$$\frac{1}{2M}\big\|\nabla f(x)\big\|^2 \leq f(x) - f(x^*) \leq \frac{1}{2m}\big\|\nabla f(x)\big\|^2, \qquad \forall\ x \in \Re^n,$$

and

$$\frac{m}{2}\|x - x^*\|^2 \leq f(x) - f(x^*) \leq \frac{M}{2}\|x - x^*\|^2, \qquad \forall\ x \in \Re^n.$$

Hint: Use a second order expansion and the relation

$$\min_{y \in \Re^n}\left\{\nabla f(x)'(y - x) + \frac{\alpha}{2}\|y - x\|^2\right\} = -\frac{1}{2\alpha}\big\|\nabla f(x)\big\|^2, \qquad \forall\ \alpha > 0.$$

1.1.10 (Nonconvex Level Sets [Dun87])

Let $f : \Re^2 \mapsto \Re$ be the function

$$f(x) = x_2^2 - ax_2\|x\|^2 + \|x\|^4,$$

where $0 < a < 2$ (see Fig. 1.1.10). Show that $f(x) > 0$ for all $x \neq 0$, so that the origin is the unique global minimum. Show also that there exists a $\bar{\gamma} > 0$ such that for all $\gamma \in (0, \bar{\gamma}]$, the level set $L_\gamma = \big\{x \mid f(x) \leq \gamma\big\}$ is not convex. *Hint*: Show that for $\gamma \in (0, \bar{\gamma}]$, there is a $p > 0$ and a $q > 0$ such that the vectors $(-p, q)$ and (p, q) belong to L_γ, but $(0, q)$ does not belong to L_γ.

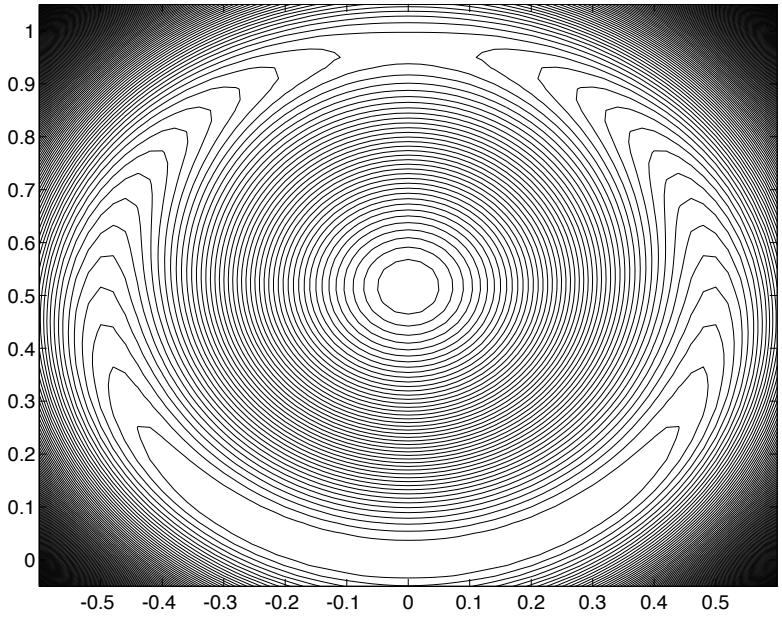

Figure 1.1.10. Level sets of the function f of Exercise 1.1.10 for the case where $a = 1.98$. The unique global minimum is the origin, but the level sets of f are nonconvex.

1.1.11 (Singular Strict Local Minima [Dun87]) (www)

Show that if x^* is a nonsingular strict local minimum of a twice continuously differentiable function $f : \Re^n \mapsto \Re$, then x^* is an isolated stationary point; that is, there is a sphere centered at x^* such that x^* is the only stationary point of f within that sphere. Use the following example function $f : \Re \mapsto \Re$ to show that this need not be true if x^* is a singular strict local minimum:

$$f(x) = \begin{cases} x^2 \left(\sqrt{2} - \sin \left(\dfrac{5\pi}{6} - \sqrt{3} \ln \left(x^2 \right) \right) \right) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

In particular, show that $x^* = 0$ is the unique (singular) global minimum, while the sequence $\{x^k\}$ of nonsingular local minima, where

$$x^k = e^{\frac{(1-8k)\pi}{8\sqrt{3}}},$$

converges to x^* (cf. Fig. 1.1.11). Verify also that there is a sequence of nonsingular local maxima that converges to x^*.

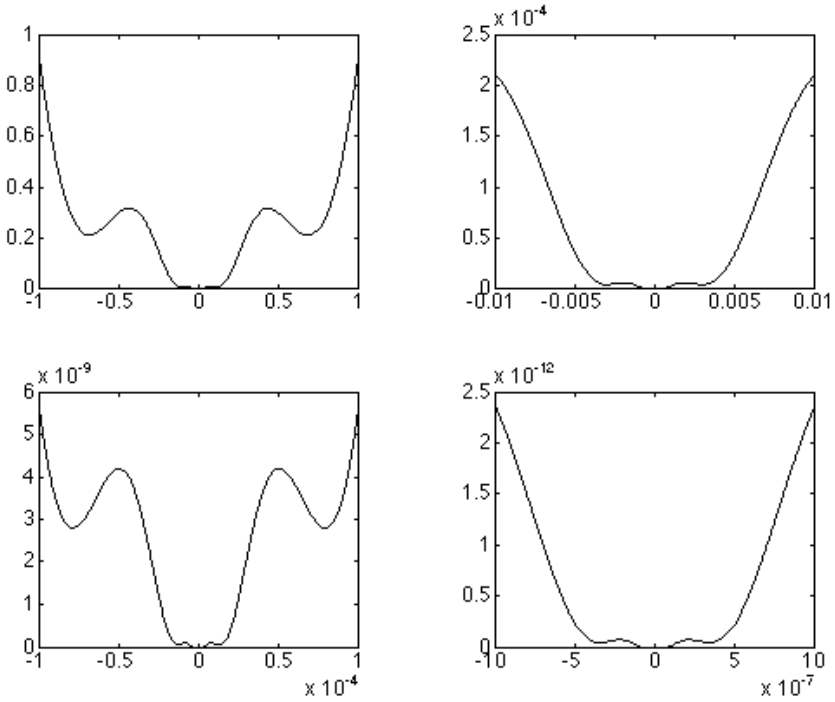

Figure 1.1.11. Illustration of the function f of Exercise 1.1.11 in progressively finer scale. The point $x^* = 0$ is the unique (singular) global minimum, but there are sequences of nonsingular local minima and local maxima that converge to x^*.

1.1.12 (Stability) www

We are often interested in whether optimal solutions change radically when the problem data are slightly perturbed. This issue is addressed by *stability analysis*, to be contrasted with sensitivity analysis, which deals with *how much* optimal solutions change when problem data change. An unconstrained local minimum x^* of a function f is said to be *locally stable* if there exists a $\delta > 0$ such that all sequences $\{x^k\}$ with

$$f(x^k) \to f(x^*), \qquad ||x^k - x^*|| < \delta, \quad \forall\ k \geq 0,$$

converge to x^*. Suppose that f is a continuous function and let x^* be a local minimum of f.

(a) Show that x^* is locally stable if and only if x^* is a strict local minimum.

(b) Let g be a continuous function. Show that if x^* is locally stable, there exists a $\delta > 0$ such that for all sufficiently small $\epsilon > 0$, the function $f(x) + \epsilon g(x)$ has an unconstrained local minimum x_ϵ that lies within the sphere centered at x^* with radius δ. Furthermore, $x_\epsilon \to x^*$ as $\epsilon \to 0$.

1.1.13 (Sensitivity) www

Let $f : \Re^n \mapsto \Re$ and $g : \Re^n \mapsto \Re$ be twice continuously differentiable functions, and let x^* be a nonsingular local minimum of f. Show that there exists an $\bar{\epsilon} > 0$ and a $\delta > 0$ such that for all $\epsilon \in [0, \bar{\epsilon})$ the function

$$f(x) + \epsilon g(x)$$

has a unique local minimum x_ϵ within the sphere $\big\{x \mid \|x - x^*\| < \delta\big\}$, and we have

$$x_\epsilon = x^* - \epsilon\big(\nabla^2 f(x^*)\big)^{-1}\nabla g(x^*) + o(\epsilon).$$

Hint: Use the implicit function theorem (Prop. A.25 in Appendix A).

1.2 GRADIENT METHODS – CONVERGENCE

We now start our development of computational methods for unconstrained optimization. The conceptual framework of this section is fundamental in nonlinear programming and applies to constrained optimization methods as well, as we will see in Chapter 3.

1.2.1 Descent Directions and Stepsize Rules

As in the case of optimality conditions, the main ideas of unconstrained optimization methods have simple geometrical explanations, but the corresponding convergence analysis is often complex. Thus, for pedagogical reasons, we first discuss informally the methods and their behavior in the present subsection, and we substantiate our conclusions with rigorous analysis in Section 1.2.2.

Consider the problem of unconstrained minimization of a continuously differentiable function $f : \Re^n \mapsto \Re$. Most of the interesting algorithms for this problem rely on an important idea, called *iterative descent* that works as follows: We start at some point x^0 (an initial guess) and successively generate vectors $x^1, x^2, \ldots$, such that f is decreased at each iteration, i.e.,

$$f(x^{k+1}) < f(x^k), \qquad k = 0, 1, \ldots,$$

(cf. Fig. 1.2.1). In doing so, we successively improve our current solution estimate and we hope to decrease f all the way to its minimum. In this section, we introduce a general class of algorithms based on iterative descent, and we analyze their convergence to local minima. In Section 1.3 we examine their rate of convergence properties.

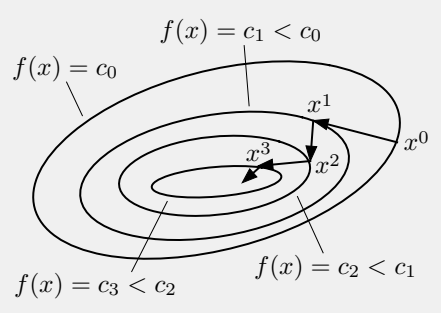

Figure 1.2.1. Iterative descent for minimizing a function f. Each vector in the generated sequence has a lower cost than its predecessor.

Gradient Methods

Given a vector $x \in \Re^n$ with $\nabla f(x) \neq 0$, consider the half line of vectors

$$x_\alpha = x - \alpha \nabla f(x), \qquad \forall\ \alpha \geq 0.$$

From the first order expansion around x we have

$$\begin{aligned} f(x_\alpha) &= f(x) + \nabla f(x)'(x_\alpha - x) + o\big(\|x_\alpha - x\|\big) \\ &= f(x) - \alpha \big\|\nabla f(x)\big\|^2 + o\big(\alpha \|\nabla f(x)\|\big), \end{aligned}$$

so we can write

$$f(x_\alpha) = f(x) - \alpha \big\|\nabla f(x)\big\|^2 + o(\alpha).$$

The term

$$\alpha \big\|\nabla f(x)\big\|^2$$

dominates $o(\alpha)$ for α near zero, so for positive but sufficiently small α, $f(x_\alpha)$ is smaller than $f(x)$ as illustrated in Fig. 1.2.2.

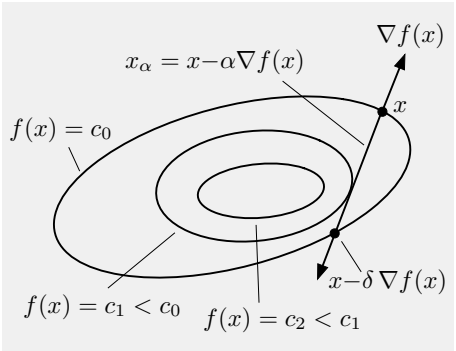

Figure 1.2.2. If $\nabla f(x) \neq 0$, there is an interval $(0, \delta)$ of stepsizes such that

$$f(x - \alpha \nabla f(x)) < f(x)$$

for all $\alpha \in (0, \delta)$.

Carrying this idea one step further, consider the half line of vectors

$$x_\alpha = x + \alpha d, \qquad \forall\ \alpha \geq 0,$$

where the direction vector $d \in \Re^n$ makes an angle with $\nabla f(x)$ that is greater than 90 degrees, i.e.,

$$\nabla f(x)'d < 0.$$

Again we have

$$f(x_\alpha) = f(x) + \alpha \nabla f(x)'d + o(\alpha).$$

For α near zero, the term $\alpha \nabla f(x)'d$ dominates $o(\alpha)$ and as a result, for positive but sufficiently small α, $f(x+\alpha d)$ is smaller than $f(x)$ as illustrated in Fig. 1.2.3.

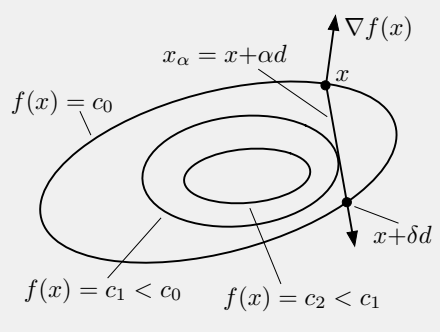

Figure 1.2.3. If the direction d makes an angle with $\nabla f(x)$ that is greater than 90 degrees, i.e., $\nabla f(x)'d < 0$, there is an interval $(0, \delta)$ of stepsizes such that

$$f(x + \alpha d) < f(x)$$

for all $\alpha \in (0, \delta)$.

The preceding observations form the basis for the broad and important class of algorithms

$$x^{k+1} = x^k + \alpha^k d^k, \qquad k = 0, 1, \ldots, \tag{1.6}$$

where, if $\nabla f(x^k) \neq 0$, the direction d^k is chosen so that

$$\nabla f(x^k)'d^k < 0, \tag{1.7}$$

and the stepsize α^k is chosen to be positive. If $\nabla f(x^k) = 0$, the method stops, i.e., $x^{k+1} = x^k$ (equivalently we choose $d^k = 0$). In view of the relation (1.7) of the direction d^k and the gradient $\nabla f(x^k)$, we call algorithms of this type *gradient methods*. [There is no universally accepted name for these algorithms; some authors reserve the name "gradient method" for the special case where $d^k = -\nabla f(x^k)$, and we will occasionally follow the same practice, when no confusion can arise.] The majority of the gradient methods that we will consider are also descent algorithms; that is, the stepsize α^k is selected so that

$$f(x^k + \alpha^k d^k) < f(x^k), \qquad k = 0, 1, \ldots. \tag{1.8}$$

However, there are some exceptions, which will be described shortly.

There is a large variety of possibilities for choosing the direction d^k and the stepsize α^k in a gradient method. Indeed there is no single gradient method that can be recommended for all or even most problems. Otherwise said, given any one of the numerous methods and variations thereof that we will discuss, there are interesting types of problems for which this method is well-suited. Our principal analytical aim is to develop a few guiding principles for understanding the performance of broad classes of methods and for appreciating the practical contexts in which their use is most appropriate.

Selecting the Descent Direction

Many gradient methods are specified in the form

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k), \tag{1.9}$$

where D^k is a positive definite symmetric matrix. Since $d^k = -D^k \nabla f(x^k)$, the descent condition $\nabla f(x^k)' d^k < 0$ is written as

$$\nabla f(x^k)' D^k \nabla f(x^k) > 0,$$

and holds thanks to the positive definiteness of D^k.

Here are some examples of choices of the matrix D^k, resulting in methods that are widely used:

Steepest Descent

$$D^k = I, \qquad k = 0, 1, \ldots,$$

where I is the $n \times n$ identity matrix. This is the simplest choice but it often leads to slow convergence, as we will see in Section 1.3. The difficulty is illustrated in Fig. 1.2.4 and motivates the methods of the subsequent examples. The name "steepest descent" is derived from an interesting property of the (normalized) negative gradient direction

$$d^k = -\frac{\nabla f(x^k)}{\big\|\nabla f(x^k)\big\|}.$$

Among all directions $d \in \Re^n$ that are normalized so that $\|d\| = 1$, it is the one that minimizes the slope $\nabla f(x^k)' d$ of the cost $f(x^k + \alpha d)$ along the direction d at $\alpha = 0$. Indeed, by the Schwarz inequality (Prop. A.2 in Appendix A), we have for all d with $\|d\| = 1$,

$$\nabla f(x^k)' d \geq -\big\|\nabla f(x^k)\big\| \cdot \|d\| = -\big\|\nabla f(x^k)\big\|,$$

and it is seen that equality is attained above for d equal to $-\nabla f(x^k)/\big\|\nabla f(x^k)\big\|$.

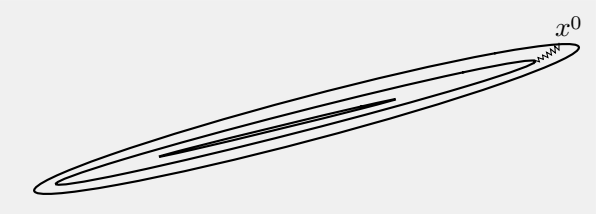

Figure 1.2.4. Slow convergence of the steepest descent method

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

when the equal cost surfaces of f are very "elongated." The difficulty is that the gradient direction is almost orthogonal to the direction that leads to the minimum. As a result the method is zig-zagging without making fast progress.

Newton's Method

$$D^k = \left(\nabla^2 f(x^k)\right)^{-1}, \qquad k = 0, 1, \ldots,$$

provided $\nabla^2 f(x^k)$ is positive definite. If $\nabla^2 f(x^k)$ is not positive definite, some modification is necessary as will be explained in Section 1.4. The idea in Newton's method is to minimize at each iteration the quadratic approximation of f around the current point x^k given by

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

(see Fig. 1.2.5). By setting the derivative of $f^k(x)$ to zero,

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

we obtain the next iterate x^{k+1} as the minimum of $f^k(x)$:

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k).$$

This is the pure Newton iteration. It is the special case of the more general iteration

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

where the stepsize $\alpha^k = 1$. Note that Newton's method finds the global minimum of a positive definite quadratic function in a single iteration (assuming $\alpha^k = 1$). For other cost functions, Newton's method typically converges very fast asymptotically and does not exhibit the zig-zagging behavior of steepest descent, as we will show in Section 1.4. For this reason many other methods try to emulate Newton's method. Some examples will be given shortly.

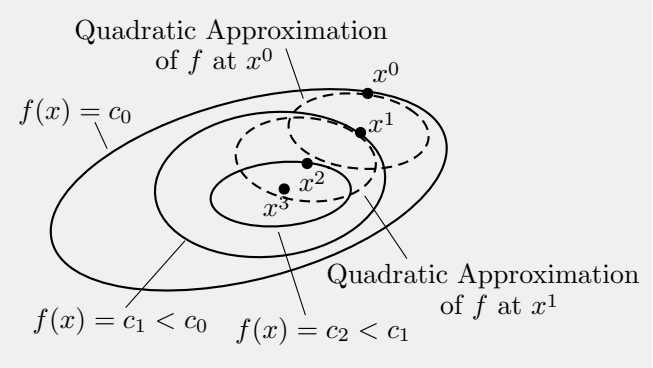

Figure 1.2.5. Illustration of the fast convergence rate of Newton's method with a stepsize $\alpha^k = 1$. Given x^k, the method obtains x^{k+1} as the minimum of a quadratic approximation of f based on a second order expansion around x^k.

Diagonally Scaled Steepest Descent

$$D^k = \begin{pmatrix} d_1^k & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & d_2^k & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & 0 & d_{n-1}^k & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & d_n^k \end{pmatrix}, \qquad k = 0, 1, \ldots,$$

where d_i^k are positive scalars, thus ensuring that D^k is positive definite. A popular choice, resulting in a method known as a *diagonal approximation to Newton's method*, is to take d_i^k to be an approximation to the inverted second partial derivative of f with respect to x_i, i.e.,

$$d_i^k \approx \left(\frac{\partial^2 f(x^k)}{(\partial x_i)^2} \right)^{-1}$$

(making sure of course that $d_i^k > 0$).

Modified Newton's Method

$$D^k = \left(\nabla^2 f(x^0) \right)^{-1}, \qquad k = 0, 1, \ldots,$$

provided $\nabla^2 f(x^0)$ is positive definite. This method is the same as Newton's method except that to economize on overhead, the Hessian matrix is not recalculated at each iteration. A related method is obtained when the Hessian is recomputed every $p > 1$ iterations.

Discretized Newton's Method

$$D^k = \big(H(x^k)\big)^{-1}, \qquad k = 0, 1, \ldots,$$

where $H(x^k)$ is a positive definite symmetric approximation of $\nabla^2 f(x^k)$, formed by using finite difference approximations of the second derivatives, based on first derivatives or values of f.

Gauss-Newton Method

This method applies to the problem of minimizing the sum of squares of real-valued functions $g_1, \ldots, g_m$, a problem often encountered in statistical data analysis and in the context of neural network training (see Section 1.4.4). By denoting $g = (g_1, \ldots, g_m)$, the problem is written as

$$\begin{aligned} &\text{minimize} \quad f(x) = \tfrac{1}{2}\big\|g(x)\big\|^2 = \tfrac{1}{2}\textstyle\sum_{i=1}^m \big(g_i(x)\big)^2 \\ &\text{subject to} \quad x \in \Re^n. \end{aligned}$$

We choose

$$D^k = \big(\nabla g(x^k)\nabla g(x^k)'\big)^{-1}, \qquad k = 0, 1, \ldots,$$

assuming the matrix $\nabla g(x^k)\nabla g(x^k)'$ is invertible. The latter matrix is always positive semidefinite, and it is positive definite and hence invertible if and only if the matrix $\nabla g(x^k)$ has rank n (Prop. A.20 in Appendix A). Since

$$\nabla f(x^k) = \nabla g(x^k) g(x^k),$$

the Gauss-Newton method takes the form

$$x^{k+1} = x^k - \alpha^k \big(\nabla g(x^k)\nabla g(x^k)'\big)^{-1} \nabla g(x^k) g(x^k). \tag{1.10}$$

We will see in Section 1.4.4 that the Gauss-Newton method may be viewed as an approximation to Newton's method, particularly when the optimal value of $\big\|g(x)\big\|^2$ is small.

Other choices of D^k yield the class of *Quasi-Newton methods* discussed in Section 2.2. There are also some interesting descent methods where the direction d^k is not usually expressed as $d^k = -D^k \nabla f(x^k)$. Important examples are the *conjugate gradient method* and the *coordinate descent methods*, which are discussed in Sections 2.1.1 and 2.3.1, respectively.

Stepsize Selection

There are a number of rules for choosing the stepsize α^k in a gradient method. We give some that are used widely in practice:

Minimization Rule

Here α^k is such that the cost function is minimized along the direction d^k, i.e., α^k satisfies

$$f(x^k + \alpha^k d^k) = \min_{\alpha \geq 0} f(x^k + \alpha d^k). \tag{1.11}$$

Limited Minimization Rule

This is a version of the minimization rule, which is more easily implemented in many cases. A fixed scalar $s > 0$ is selected and α^k is chosen to yield the greatest cost reduction over all stepsizes in the interval $[0, s]$, i.e.,

$$f(x^k + \alpha^k d^k) = \min_{\alpha \in [0,\, s]} f(x^k + \alpha d^k).$$

The minimization and limited minimization rules must typically be implemented with the aid of one-dimensional line search algorithms (see Appendix C). In general, the minimizing stepsize cannot be computed exactly, and in practice, the line search is stopped once a stepsize α^k satisfying some termination criterion is obtained. Some stopping criteria are discussed in Exercise 1.2.15.

Generally, compared with other stepsize rules, the minimization rules tend to require more function and/or gradient evaluations per iteration. However, their use tends to reduce the number of required iterations for practical convergence, because of the greater cost reduction per iteration that they achieve. The minimization rules are also favored in cases where the structure of the problem can be exploited to economize on the associated computations. A prominent example arises when the cost function has the form

$$f(x) = h(Ax),$$

where A is a matrix such that the calculation of the vector $y = Ax$ for a given x is far more expensive than the calculation of $h(y)$ and its gradient and Hessian (assuming it exists). In this case, calculation of values, first, and second derivatives of the function $g(\alpha) \equiv f(x + \alpha d) = h(Ax + \alpha Ad)$ requires just two expensive operations: the one-time calculation of the matrix-vector products Ax and Ad.

Successive Stepsize Reduction – Armijo Rule

To avoid the often considerable computation associated with the line minimization rules, it is natural to consider rules based on successive stepsize reduction. In the simplest rule of this type an initial stepsize s is chosen, and

if the corresponding vector $x^k + sd^k$ does not yield an improved value of f, i.e., $f(x^k + sd^k) \geq f(x^k)$, the stepsize is reduced, perhaps repeatedly, by a certain factor, until the value of f is improved. While this method often works in practice, it is theoretically unsound because the cost improvement obtained at each iteration may not be substantial enough to guarantee convergence to a minimum. This is illustrated in Fig. 1.2.6.

The Armijo rule is essentially the successive reduction rule just described, suitably modified to eliminate the theoretical convergence difficulty shown in Fig. 1.2.6. Here, fixed scalars s, β, and σ, with $0 < \beta < 1$, and $0 < \sigma < 1$ are chosen, and we set $\alpha^k = \beta^{m_k} s$, where m_k is the first nonnegative integer m for which

$$f(x^k) - f(x^k + \beta^m sd^k) \geq -\sigma\beta^m s\nabla f(x^k)'d^k. \tag{1.12}$$

In other words, the stepsizes $\beta^m s$, $m = 0, 1, \ldots$, are tried successively until the above inequality is satisfied for $m = m_k$. Thus, the cost improvement must not be just positive; it must be sufficiently large as per the test (1.12). Figure 1.2.7 illustrates the rule.

Usually σ is chosen close to zero, for example, $\sigma \in [10^{-5}, 10^{-1}]$. The reduction factor β is usually chosen from 1/2 to 1/10 depending on the confidence we have on the quality of the initial stepsize s. We can always take $s = 1$ and multiply the direction d^k by a scaling factor. Many methods, such as Newton-like methods, incorporate some type of implicit scaling of the direction d^k, which makes $s = 1$ a good stepsize choice (see the discussion on rate of convergence in Section 1.3). If a suitable scaling factor for d^k is not known, one may use various ad hoc schemes to determine one. For example, a simple possibility is based on quadratic interpolation of the function

$$g(\alpha) = f(x^k + \alpha d^k),$$

which is the cost along the direction d^k, viewed as a function of the stepsize α. In this scheme, we select some stepsize $\bar{\alpha}$, evaluate $g(\bar{\alpha})$, and perform the quadratic interpolation of g on the basis of $g(0) = f(x^k)$, $dg(0)/d\alpha = \nabla f(x^k)'d^k$, and $g(\bar{\alpha})$. If $\hat{\alpha}$ minimizes the quadratic interpolation, we replace d^k by $\hat{d}^k = \hat{\alpha}d^k$, and we use an initial stepsize $s = 1$. Of course some safeguards are needed when implementing heuristics of this type; for example if $g(\alpha)$ is linear or concave in the interval $[0, \bar{\alpha}]$, the quadratic interpolation scheme just described will fail; see also Exercise 1.2.15.

Constant Stepsize

Here a fixed stepsize $s > 0$ is selected and

$$\alpha^k = s, \qquad k = 0, 1, \ldots.$$

The constant stepsize rule is very simple. However, if the stepsize is too large, divergence will occur, while if the stepsize is too small, the rate of convergence may be very slow. Thus, the constant stepsize rule is useful only for problems

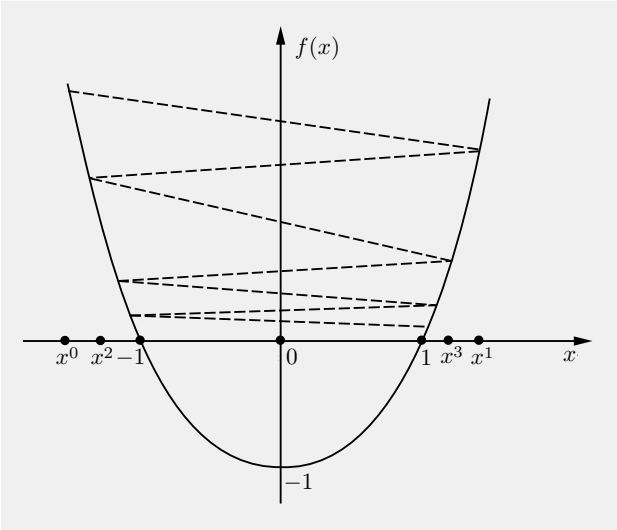

Figure 1.2.6. Example of failure of the successive stepsize reduction rule for the one-dimensional function

$$f(x) = \begin{cases} \dfrac{3(1-x)^2}{4} - 2(1-x) & \text{if } x > 1, \\ \dfrac{3(1+x)^2}{4} - 2(1+x) & \text{if } x < -1, \\ x^2 - 1, & \text{if } -1 \le x \le 1. \end{cases}$$

The gradient of f is given by

$$\nabla f(x) = \begin{cases} \dfrac{3x}{2} + \frac{1}{2} & \text{if } x > 1, \\ \dfrac{3x}{2} - \frac{1}{2} & \text{if } x < -1, \\ 2x, & \text{if } -1 \le x \le 1. \end{cases}$$

It is seen that f is strictly convex, continuously differentiable, and is minimized at $x^* = 0$. Furthermore, $f(x) < f(\bar{x})$ if and only if $|x| < |\bar{x}|$. For $x > 1$, we have

$$x - \nabla f(x) = x - \frac{3x}{2} - \tfrac{1}{2} = -\left(1 + \tfrac{x-1}{2}\right),$$

from which it can be verified that $|x - \nabla f(x)| < |x|$, so that $f\big(x - \nabla f(x)\big) < f(x)$ and $x - \nabla f(x) < -1$. Similarly, for $x < -1$, we have $f\big(x - \nabla f(x)\big) < f(x)$ and $x - \nabla f(x) > 1$. Consider now the steepest descent iteration where the stepsize is successively reduced from an initial stepsize $s = 1$ until descent is obtained. Let the starting point satisfy $|x^0| > 1$. From the preceding equations, it follows that $f\big(x^0 - \nabla f(x^0)\big) < f(x^0)$ and the stepsize $s = 1$ will be accepted by the method. Thus, the next point is $x^1 = x^0 - \nabla f(x^0)$, which satisfies $|x^1| > 1$. By repeating the preceding argument, we see that the generated sequence $\{x^k\}$ satisfies $|x^k| > 1$ for all k, and cannot converge to the unique stationary point $x^* = 0$. In fact, it can be shown that $\{x^k\}$ will have two limit points, $\bar{x} = 1$ and $\bar{x} = -1$, for every x^0 with $|x^0| > 1$.

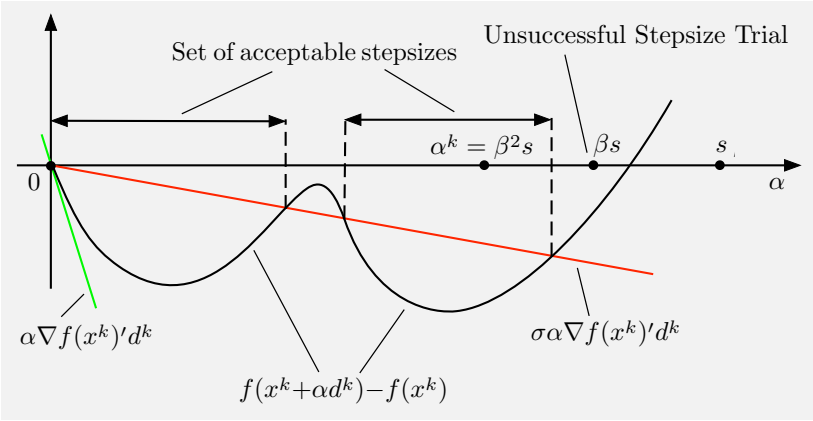

Figure 1.2.7. Line search by the Armijo rule. We start with the trial stepsize s and continue with $\beta s, \beta^2 s, ...$, until the first time that $\beta^m s$ falls within the set of stepsizes α satisfying the inequality

$$f(x^k) - f(x^k + \alpha d^k) \geq -\sigma\alpha\nabla f(x^k)'d^k.$$

While this set need not be an interval, it will always contain an interval of the form $[0, \delta]$ with $\delta > 0$, provided $\nabla f(x^k)'d^k < 0$. For this reason the stepsize α^k chosen by the Armijo rule is well defined and will be found after a finite number of trial evaluations of f at the points $(x^k + sd^k), (x^k + \beta sd^k), \ldots$

where an appropriate constant stepsize value is known or can be determined fairly easily.

For the case where f is convex, there are methods that attempt to determine automatically an appropriate value of stepsize; see Exercise 1.2.19 and also [Ber15a], Section 6.1. In these methods an initial value of stepsize is selected, and using the results of the computation over several iterations, the stepsize is reduced to a level that eventually stays constant. Still the convergence of a gradient method using a constant or eventually constant stepsize requires that the gradient ∇f satisfies a Lipschitz condition (see the subsequent Prop. 1.2.2 and the discussion that follows it). By contrast the line minimization rules and the Armijo rule do not require this restriction.

Diminishing Stepsize

Here the stepsize converges to zero,

$$\alpha^k \to 0.$$

This stepsize rule is different than the preceding ones in that it does not guarantee descent at each iteration, although descent becomes more likely as the stepsize diminishes. One difficulty with a diminishing stepsize is that it

may become so small that substantial progress cannot be maintained, even when far from a stationary point. For this reason, we require that

$$\sum_{k=0}^{\infty} \alpha^k = \infty.$$

The last condition guarantees that $\{x^k\}$ does not converge to a nonstationary point. Indeed, if $x^k \to \bar{x}$, then for any large indexes m and n ($m > n$) we have

$$x^m \approx x^n \approx \bar{x}, \qquad x^m \approx x^n - \left(\sum_{k=n}^{m-1} \alpha^k \right) \nabla f(\bar{x}),$$

which is a contradiction when $\bar{x}$ is nonstationary and $\sum_{k=n}^{m-1} \alpha^k$ can be made arbitrarily large. Generally, the diminishing stepsize rule has good theoretical convergence properties (see Prop. 1.2.3, and Exercises 1.2.12 and 1.2.13). The associated convergence rate tends to be slow, so this stepsize rule is used primarily in situations where slow convergence is inevitable; for example, in singular problems or when the gradient is calculated with error (see the discussion later in this section).

The preceding stepsize rules are based on cost function reduction (or eventual cost function reduction, in the case of a diminishing stepsize). There are also some other rules, often called *nonmonotonic*, which do not explicitly try to enforce cost function descent and have achieved some success, but are based on ideas that we will not discuss in this book; see [BaB88], [GLL91], [Ray93], [Ray97], [BMR00], [DHS06].

Convergence Issues

Let us now delineate the type of convergence issues that we would like to clarify. We will first discuss informally these issues, and we will state and prove the associated convergence results in Section 1.2.2.

Given a gradient method, ideally we would like the generated sequence $\{x^k\}$ to converge to a global minimum. Unfortunately, however, this is too much to expect, at least when f is not convex, because of the presence of local minima that are not global. Indeed a gradient method is guided downhill by the form of f near the current iterate, while being oblivious to the global structure of f, and thus, can easily get attracted to any type of minimum, global or not. Furthermore, if a gradient method starts or lands at any stationary point, including a local maximum, it stops at that point. Thus, the most we can expect from a gradient method is that it converges to a stationary point. Such a point is a global minimum if f is convex, but this need not be so for nonconvex problems. It must therefore be recognized that gradient methods can be quite inadequate, particularly if little is known about the location and/or other properties of global minima. For such problems one should either try an often difficult and frustrating

process of running a gradient method from multiple starting points, or else resort to a fundamentally different approach.

Generally, depending on the nature of the cost function f, the sequence $\{x^k\}$ generated by a gradient method need not have a limit point; in fact $\{x^k\}$ is typically unbounded if f has no local minima. If, however, we know that the level set $\{x \mid f(x) \leq f(x^0)\}$ is bounded, and the stepsize is chosen to enforce descent at each iteration, then the sequence $\{x^k\}$ must be bounded since it belongs to this level set. It must then have at least one limit point; this is because every bounded sequence has at least one limit point (see Prop. A.5 of Appendix A).

Even if $\{x^k\}$ is bounded, convergence to a single limit point may not be easy to guarantee. However, it can be shown that local minima, which are isolated stationary points (they are unique stationary points within some open sphere), tend to attract most types of gradient methods, i.e., once a gradient method gets sufficiently close to such a local minimum, it converges to it. This is the subject of a simple and remarkably powerful result, the *capture theorem*, which is given in the next subsection (Prop. 1.2.4).

Another single limit convergence result for the steepest descent method is given in Exercise 1.2.12 for the case where f is convex and the constant or diminishing stepsize rule is used. Generally, if there are multiple global minima, it is possible that $\{x^k\}$ has multiple limit points (see Exercise 1.2.17 from [Zou76]; also [Gon00]).

We now address the question whether each limit point of a sequence $\{x^k\}$ generated by a gradient method is a stationary point. From the first order expansion

$$f(x^{k+1}) = f(x^k) + \alpha^k \nabla f(x^k)' d^k + o(\alpha^k),$$

we see that if the slope of f at x^k along the direction d^k, which is $\nabla f(x^k)' d^k$, has "substantial" magnitude, the rate of progress of the method will also tend to be substantial. If on the other hand, the directions d^k tend to become asymptotically orthogonal to the gradient direction,

$$\frac{\nabla f(x^k)' d^k}{\|\nabla f(x^k)\| \|d^k\|} \to 0,$$

as x^k approaches a nonstationary point, there is a chance that the method will get "stuck" near that point. To ensure that this does not happen, we consider rather technical conditions on the directions d^k, which are either naturally satisfied or can be easily enforced in most algorithms of interest.

One such condition for the case where

$$d^k = -D^k \nabla f(x^k),$$

is to assume that the eigenvalues of the positive definite symmetric matrix D^k are bounded above and bounded away from zero, i.e., for some positive

scalars c_1 and c_2, we have

$$c_1\|z\|^2 \leq z'D^k z \leq c_2\|z\|^2, \qquad \forall\ z \in \Re^n,\ k = 0, 1, \ldots \tag{1.13}$$

It can then be seen that

$$\left|\nabla f(x^k)'d^k\right| = \left|\nabla f(x^k)'D^k\nabla f(x^k)\right| \geq c_1\left\|\nabla f(x^k)\right\|^2.$$

It follows that as long as $\nabla f(x^k)$ does not tend to zero, $\nabla f(x^k)$ and d^k cannot become asymptotically orthogonal.

We will introduce another "nonorthogonality" type of condition, which is more general than the "bounded eigenvalues" condition (1.13). Let us consider the sequence $\{x^k, d^k\}$ generated by a given gradient method. We say that the direction sequence $\{d^k\}$ is *gradient related* to $\{x^k\}$ if the following property can be shown:

> *For any subsequence $\{x^k\}_{k\in\mathcal{K}}$ that converges to a nonstationary point, the corresponding subsequence $\{d^k\}_{k\in\mathcal{K}}$ is bounded and satisfies*
>
> $$\limsup_{k\to\infty,\, k\in\mathcal{K}} \nabla f(x^k)'d^k < 0. \tag{1.14}$$

In particular, if $\{d^k\}$ is gradient related to $\{x^k\}$, it follows that if a subsequence $\left\{\nabla f(x^k)\right\}_{k\in\mathcal{K}}$ tends to a nonzero vector, the corresponding subsequence of directions d^k is bounded and does not tend to be orthogonal to $\nabla f(x^k)$. Roughly, this means that d^k does not become "too small" or "too large" relative to $\nabla f(x^k)$, and that the angle between d^k and $\nabla f(x^k)$ does not get "too close" to 90 degrees.

We can often guarantee *a priori* that $\{d^k\}$ is gradient related. In particular, if $d^k = -D^k\nabla f(x^k)$ and the eigenvalues of D^k are bounded as in the "bounded eigenvalues" condition (1.13), it can be seen that $\{d^k\}$ is gradient related. [The boundedness requirement of $\{d^k\}_{k\in\mathcal{K}}$ holds in view of the relation

$$\|d^k\|^2 = \left|\nabla f(x^k)'(D^k)^2\nabla f(x^k)\right| \leq c_2^2\left\|\nabla f(x^k)\right\|^2,$$

which follows from Eq. (1.13), since c_2 is no less than the largest eigenvalue of D^k, and the eigenvalues of $(D^k)^2$ are equal to the squares of the corresponding eigenvalues of D^k (Props. A.18 and A.13 in Appendix A).]

Two other examples of conditions that, if satisfied for some scalars $c_1 > 0$, $c_2 > 0$, $p_1 \geq 0$, $p_2 \geq 0$, and all k, guarantee that $\{d^k\}$ is gradient related are

(a)
$$c_1\left\|\nabla f(x^k)\right\|^{p_1} \leq -\nabla f(x^k)'d^k, \qquad \|d^k\| \leq c_2\left\|\nabla f(x^k)\right\|^{p_2}.$$

(b)
$$d^k = -D^k\nabla f(x^k),$$

with D^k a positive definite symmetric matrix satisfying

$$c_1\big\|\nabla f(x^k)\big\|^{p_1}\,\|z\|^2 \leq z'D^k z \leq c_2\big\|\nabla f(x^k)\big\|^{p_2}\,\|z\|^2, \qquad \forall\ z \in \Re^n.$$

This condition generalizes the "bounded eigenvalues" condition (1.13), which is obtained for $p_1 = p_2 = 0$.

An important convergence result is that if $\{d^k\}$ is gradient related and the minimization rule, or the limited minimization rule, or the Armijo rule is used, then all limit points of $\{x^k\}$ are stationary. This is shown in Prop. 1.2.1, given in the next subsection. When a constant stepsize is used, convergence can be proved assuming that the stepsize is sufficiently small and that f satisfies some further conditions (cf. Prop. 1.2.2).

There is a common line of proof for these convergence results. The main idea is that the cost function is improved at each iteration and that, based on our assumptions, the improvement is "substantial" near a nonstationary point, i.e., it is bounded away from zero. We then argue that the algorithm cannot approach a nonstationary point, since in this case the total cost improvement would accumulate to infinity.

Termination of Gradient Methods

Generally, gradient methods are not finitely convergent, so it is necessary to have criteria for terminating the iterations with some assurance that we are reasonably close to at least a local minimum. A typical approach is to stop the computation when the norm of the gradient becomes sufficiently small, i.e., when a point x^k is obtained with

$$\big\|\nabla f(x^k)\big\| \leq \epsilon,$$

where ϵ is a small positive scalar. Unfortunately, it is not known a priori how small one should take ϵ in order to guarantee that the final point x^k is a "good" approximation to a stationary point. The appropriate value of ϵ depends on how the problem is scaled. In particular, if f is multiplied by some scalar, the appropriate value of ϵ is also multiplied by the same scalar. It is possible to correct this difficulty by replacing the criterion $\big\|\nabla f(x^k)\big\| \leq \epsilon$ with

$$\frac{\big\|\nabla f(x^k)\big\|}{\big\|\nabla f(x^0)\big\|} \leq \epsilon.$$

Still, however, the gradient norm $\big\|\nabla f(x^k)\big\|$ depends on all the components of the gradient, and depending on how the optimization variables are scaled, the preceding termination criterion may not work well. In particular, some components of the gradient may be naturally much smaller than others, thus requiring a smaller value of ϵ than the other components.

Assuming that the direction d^k captures the relative scaling of the optimization variables, it may be appropriate to terminate computation when the norm of the direction d^k becomes sufficiently small, i.e.,

$$\|d^k\| \leq \epsilon.$$

Still the appropriate value of ϵ may not be easy to guess, and it may be necessary to experiment prior to settling on a reasonable termination criterion for a given problem. Sometimes, other problem-dependent criteria are used, in addition to or in place of $\big\|\nabla f(x^k)\big\| \leq \epsilon$ and $\|d^k\| \leq \epsilon$.

When $\nabla^2 f(x)$ is positive definite, the condition $\big\|\nabla f(x^k)\big\| \leq \epsilon$ yields bounds on the distance from local minima. In particular, if x^* is a local minimum of f and there exists $m > 0$ such that for all x in a sphere S centered at x^* we have

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

then every $x \in S$ satisfying $\big\|\nabla f(x)\big\| \leq \epsilon$ also satisfies

$$\|x - x^*\| \leq \frac{\epsilon}{m}, \qquad f(x) - f(x^*) \leq \frac{\epsilon^2}{m},$$

(see Exercise 1.2.9).

In the absence of positive definiteness conditions on $\nabla^2 f(x)$, it may be very difficult to infer the proximity of the current iterate to the optimal solution set by just using the gradient norm. We will return to this point when we will discuss singular local minima in the next section.

Spacer Steps

Often, optimization problems are solved with complex descent algorithms in which the rule used to determine the next point may depend on several previous points or on the iteration index k. Some of the conjugate direction algorithms discussed in Section 2.1 are of this type. Other algorithms consist of a combination of different methods and switch from one method to the other in a manner that may either be prespecified or may depend on the progress of the algorithm. Such combinations are usually introduced in order to improve speed of convergence or reliability. However, their convergence analysis can become extremely complicated.

It is thus often valuable to know that if in such algorithms one inserts, perhaps irregularly but infinitely often, an iteration of a convergent algorithm such as the gradient methods of this section, then the theoretical convergence properties of the overall algorithm are quite satisfactory. Such an iteration is known as a *spacer step*. The related convergence result is given in Prop. 1.2.5. The only requirement imposed on the iterations of the algorithm other than the spacer steps is that they do not increase the cost; these iterations, however, need not strictly decrease the cost, and this allows for flexibility in the design of algorithms.

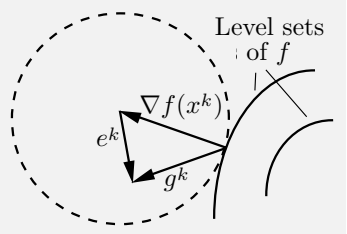

Figure 1.2.8. Illustration of the descent property of the direction $g^k = \nabla f(x^k) + e^k$ of the steepest descent method with error. If the error e^k has smaller norm than the gradient $\nabla f(x^k)$, then g^k lies strictly within the sphere centered at $\nabla f(x^k)$ with radius $\big\|\nabla f(x^k)\big\|$, and thus makes an angle less than 90 degrees with $\nabla f(x^k)$.

Gradient Methods with Random and Nonrandom Errors

Frequently in optimization problems the gradient $\nabla f(x^k)$ is not computed exactly. Instead, one has available

$$g^k = \nabla f(x^k) + e^k,$$

where e^k is an uncontrollable error vector. There are several potential sources of error; roundoff error, and discretization error due to finite difference approximations to the gradient are two possibilities, but there are others that will be discussed in more detail later in Chapter 2, in the context of incremental and asynchronous algorithms. Let us focus for concreteness on the steepest descent method with errors,

$$x^{k+1} = x^k - \alpha^k g^k,$$

and let us consider several qualitatively different cases:

(a) e^k **is small relative to the gradient**, i.e.,

$$\|e^k\| < \big\|\nabla f(x^k)\big\|, \qquad \forall\ k.$$

Then, assuming $\nabla f(x^k) \neq 0$, $-g^k$ is a direction of cost improvement, i.e., $\nabla f(x^k)'g^k > 0$. This is illustrated in Fig. 1.2.8, and is verified by the calculation

$$\begin{aligned} \nabla f(x^k)'g^k &= \big\|\nabla f(x^k)\big\|^2 + \nabla f(x^k)'e^k \\ &\geq \big\|\nabla f(x^k)\big\|^2 - \big\|\nabla f(x^k)\big\|\,\|e^k\| \\ &= \big\|\nabla f(x^k)\big\|\big(\|\nabla f(x^k)\| - \|e^k\|\big) \\ &> 0. \end{aligned} \tag{1.15}$$

In this case convergence results that are analogous to Props. 1.2.3 and 1.2.4 can be shown.

(b) $\{e^k\}$ **is bounded**, i.e.,

$$\|e^k\| \leq \delta, \qquad \forall\ k,$$

where δ is some scalar. Then by the preceding calculation (1.15), the method operates like a descent method within the region

$$\big\{x \mid \|\nabla f(x)\| > \delta\big\}.$$

In the complementary region where $\big\|\nabla f(x)\big\| \leq \delta$, the method can behave quite unpredictably. For example, if the errors e^k are constant, say $e^k \equiv e$, then since $g^k = \nabla f(x^k) + e$, the method will essentially be trying to minimize $f(x) + e'x$ and will typically converge to a point $\bar{x}$ with $\nabla f(\bar{x}) = -e$. If the errors e^k vary substantially, the method will tend to oscillate within the region where $\big\|\nabla f(x)\big\| \leq \delta$ (see Exercise 1.2.16 and also Exercise 1.3.5 in the next section). The precise behavior will depend on the precise nature of the errors, and on whether a constant or a diminishing stepsize is used (see also the following cases).

(c) $\{e^k\}$ **is proportional to the stepsize**, i.e.,

$$\|e^k\| \leq \alpha^k q, \qquad \forall\ k,$$

where q is some scalar. If the stepsize is constant, we come under case (b), while if the stepsize is diminishing, the behavior described in case (b) applies, but with $\delta \to 0$, so the method will tend to converge to a stationary point of f. Important situations where the condition $\|e^k\| \leq \alpha^k q$ holds will be encountered in the context of incremental methods in Section 2.4. A more general condition under which similar behavior occurs is

$$\|e^k\| \leq \alpha^k\big(q + p\|\nabla f(x^k)\|\big), \qquad \forall\ k,$$

where q and p are some scalars. Generally, under this condition and with a diminishing stepsize, the convergence behavior is similar to the case where there are no errors; see the following Prop. 1.2.3 (also Exercise 1.2.20, whose solution is posted online).

(d) $\{e^k\}$ **are independent zero mean random vectors with finite variance**. An important special case where such errors arise is when f is of the form

$$f(x) = E_w\big\{F(x,w)\big\}, \tag{1.16}$$

where $F : \Re^{m+n} \to \Re$ is some function, w is a random vector in $\Re^m$, and $E_w\{\cdot\}$ denotes expected value. Under very mild assumptions it can be shown that if F is continuously differentiable, the same is true of f and furthermore,

$$\nabla f(x) = E_w\big\{\nabla_x F(x,w)\big\}.$$

Often an approximation g^k to $\nabla f(x^k)$ is computed by simulation or by using a limited number of samples of $\nabla F(x, w)$, with potentially substantial error resulting. In an extreme case, we have

$$g^k = \nabla_x F(x^k, w^k),$$

where w^k is a single sample value corresponding to x^k. Then the error

$$e^k = \nabla_x F(x^k, w^k) - \nabla f(x^k) = \nabla_x F(x^k, w^k) - E_w\big\{\nabla_x F(x^k, w)\big\}$$

need not diminish with $\big\|\nabla f(x^k)\big\|$, but has zero mean, and under appropriate conditions, its effects are "averaged out." What is roughly happening here is that the descent condition $\nabla f(x^k)' g^k > 0$ holds *on the average* at nonstationary points x^k. It is still possible that for some sample values of e^k, the direction g^k is "bad", but with a diminishing stepsize, the occasional use of a bad direction cannot deteriorate the cost enough for the method to oscillate, given that on the average the method uses "good" directions (see also the discussion of incremental gradient methods in Section 2.4.1). The detailed analysis of gradient methods with random errors (also called *stochastic gradient* methods) is beyond the scope of this text, and properly belongs to the algorithmic field of *stochastic approximation* (see e.g. [BeT89], [BeT96], [BeT00], [KuC78], [KuY97], [LPW92], [Pfl96], [PoT73a], [Pol87], [TBA86]). We mention one representative convergence result from [BeT00], which parallels the following Prop. 1.2.3 that deals with a gradient method without errors: if in the iteration

$$x^{k+1} = x^k - \alpha^k\big(\nabla f(x^k) + e^k\big)$$

the random variables $e^0, e^1, \ldots$ are independent, zero mean, with bounded variance, the stepsize is diminishing and satisfies

$$\alpha^k \to 0, \qquad \sum_{k=0}^{\infty} \alpha^k = \infty, \qquad \sum_{k=0}^{\infty} (\alpha^k)^2 < \infty,$$

and the gradient ∇f is Lipschitz continuous, then with probability one, we either have $f(x^k) \to -\infty$ or else $\nabla f(x^k) \to 0$. Furthermore, every limit point of $\{x^k\}$ is a stationary point of f.

The Role of Convergence Analysis

The following subsection gives a number of mathematical propositions relating to the convergence properties of gradient methods. The meaning of these propositions is usually quite intuitive but their statement often requires complicated mathematical assumptions. Furthermore, their proof often involves tedious ϵ-δ arguments, so at first sight students may wonder whether "we really have to go through all this."

When Euclid was faced with a similar question from king Ptolemy of Alexandria, he replied that "there is no royal road to geometry." In our case, however, the answer is not so simple because we are not dealing with a purely mathematical subject such as geometry that may be developed without regard for its practical application. In the eyes of most people, the value of an analysis or algorithm in nonlinear programming is judged primarily by its practical utility in solving various types of problems. It is therefore important to give some thought to the interface between convergence analysis and its practical application. To this end it is useful to consider two extreme viewpoints; most workers in the field find themselves somewhere between the two.

In the first viewpoint, convergence analysis is considered primarily a mathematical subject. The properties of an algorithm are quantified to the extent possible through mathematical statements. General and broadly applicable assertions, and simple and elegant proofs are at a premium here. The rationale is that simple statements and proofs are more readily understood at an intuitive level, and general statements apply not only to the problems at hand but also to other problems that are likely to appear in the future. On the negative side, one may remark that simplicity is not always compatible with relevance, and broad applicability is often achieved through assumptions that are hard to verify or appreciate.

The second viewpoint largely discounts the role of mathematical analysis. The rationale here is that the validity and the properties of an algorithm for a given class of problems must be verified through practical experimentation anyway, so if an algorithm looks promising on intuitive grounds, why bother with a convergence analysis. Furthermore, there are a number of important practical questions that are hard to address analytically, such as roundoff error, multiple local minima, and a variety of finite termination and approximation issues. The main criticism of this viewpoint is that mathematical analysis often reveals (and explains) fundamental flaws of algorithms that experimentation may miss. These flaws often point the way to better algorithms or modified algorithms that are tailored to the type of practical problem at hand. Similarly, analysis may be more effective than experimentation in delineating the types of problems for which particular algorithms are well-suited.

Our own mathematical approach is tempered by practical concerns, but we note that the balance between theory and practice in nonlinear programming is particularly delicate, subjective, and problem dependent. Aside from the fact that the mathematical proofs themselves often provide valuable insight into algorithms, here are some of our reasons for insisting on a rigorous convergence analysis:

(a) We want to delineate the range of applicability of various methods. In particular, we want to know for what type of cost function (once or twice differentiable, convex or nonconvex, with singular or nonsingu-

lar minima) each algorithm is best suited. If the cost function violates the assumptions under which a given algorithm can be proved to converge, it is reasonable to suspect that the algorithm is unsuitable for this cost function.

(b) We want to understand the qualitative behavior of various methods. For example, we want to know whether convergence of the method depends on the availability of a good starting point, whether the iterates x^k or just the function values $f(x^k)$ are guaranteed to converge, etc. This information, supplemented by theoretical examples and counterexamples, may guide the computational experimentation.

(c) We want to provide guidelines for choosing a few algorithms for further experimentation out of the often bewildering array of candidate algorithms that are applicable for the solution of a given type of problem. One of the principal means for this is the rate of convergence analysis to be given in Section 1.3. Note here that while an algorithm may provably converge, in practice it may be entirely inappropriate for a given problem because it converges very slowly. Experience has shown that without a good understanding of the rate of convergence properties of algorithms it may be difficult to exclude bad candidates from consideration without costly experimentation.

At the same time one should be aware of some of the limitations of the mathematical results that we will provide. For example, some of the assumptions under which an algorithm will be proved convergent may be hard to verify for a given type of problem. Furthermore, our convergence rate analysis of Section 1.3 is largely asymptotic; that is, it applies near the eventual limit of the generated sequence. It is possible, that an algorithm has a good asymptotic rate of convergence but it works poorly in practice for a given type of problem because it is very slow in its initial phase. Finally, some lines of mathematical analyses can be so dry and lacking in intuition that they are of little use to all but a small number of theorists.

There is still another viewpoint, which is worth addressing because it is often adopted by the casual user of nonlinear programming algorithms. This user is interested in a particular application of nonlinear programming in his/her special field, and is counting on an existing code or package to solve the problem (several such packages are commercially or publicly available). Since the package will do most of the work, the user may hope that a superficial acquaintance with the properties of the algorithms underlying the package will suffice. This hope is sometimes realized but unfortunately in many cases it is not. There are a number of reasons for this. First, there are many packages implementing a lot of different methods, and to choose the right package, one needs to have insight into the suitability of different methods for the special features of the application at hand. Second, to use a package one must often know how to suitably formulate the

problem, how to set various parameters (e.g. termination criteria, stepsize parameters, etc.), and how to interpret the results of the computation (particularly when things don't work out as hoped initially, which is often the case). For this, one needs considerable insight into the inner workings of the algorithm underlying the package. Finally, for a challenging practical optimization problem (e.g. one of large dimension), it may be essential to exploit its special structure, and packages often do not have this capability. As a result the user may have to modify the package or write an altogether new code that is tailored to the application at hand. Both of these require an intimate understanding of the convergence properties and other characteristics of the relevant nonlinear programming algorithms.

1.2.2 Convergence Results

We now provide an analysis of the convergence behavior of gradient methods. The following proposition is the main convergence result.

Proposition 1.2.1: (Stationarity of Limit Points for Gradient Methods) Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$, and assume that $\{d^k\}$ is gradient related [cf. Eq. (1.14)] and α^k is chosen by the minimization rule, or the limited minimization rule, or the Armijo rule. Then every limit point of $\{x^k\}$ is a stationary point.

Proof: Consider first the Armijo rule and let $\bar{x}$ be a limit point of $\{x^k\}$. Since $\{f(x^k)\}$ is monotonically nonincreasing, $\{f(x^k)\}$ either converges to a finite value or diverges to $-\infty$. Since f is continuous, $f(\bar{x})$ is a limit point of $\{f(x^k)\}$, so it follows that the entire sequence $\{f(x^k)\}$ converges to $f(\bar{x})$, and

$$f(x^k) - f(x^{k+1}) \to 0. \tag{1.17}$$

Moreover, by the definition of the Armijo rule, we have

$$f(x^k) - f(x^{k+1}) \geq -\sigma\alpha^k \nabla f(x^k)' d^k, \tag{1.18}$$

so the right-hand side in the above relation tends to 0.

Let $\{x^k\}_{\mathcal{K}}$ be a subsequence converging to $\bar{x}$, and assume to arrive at a contradiction that $\bar{x}$ is nonstationary. Since $\{d^k\}$ is gradient related, we have

$$\limsup_{\substack{k\to\infty \\ k\in\mathcal{K}}} \nabla f(x^k)' d^k < 0,$$

and therefore from Eqs. (1.17) and (1.18),

$$\{\alpha^k\}_{\mathcal{K}} \to 0.$$

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

$$f(x^k) - f\big(x^k + (\alpha^k/\beta)d^k\big) < -\sigma(\alpha^k/\beta)\nabla f(x^k)'d^k, \qquad \forall\ k \in \mathcal{K},\ k \geq \bar{k}, \tag{1.19}$$

i.e., the initial stepsize s will be reduced at least once for all $k \in \mathcal{K}$, $k \geq \bar{k}$. Since $\{d^k\}$ is gradient related, $\{d^k\}_{\mathcal{K}}$ is bounded, so there exists a subsequence $\{d^k\}_{\bar{\mathcal{K}}}$ of $\{d^k\}_{\mathcal{K}}$ such that

$$\{d^k\}_{\bar{\mathcal{K}}} \to \bar{d},$$

where $\bar{d}$ is some vector. From Eq. (1.19), we have

$$\frac{f(x^k) - f(x^k + \bar{\alpha}^k d^k)}{\bar{\alpha}^k} < -\sigma\nabla f(x^k)'d^k, \qquad \forall\ k \in \bar{\mathcal{K}},\ k \geq \bar{k}, \tag{1.20}$$

where $\bar{\alpha}^k = \alpha^k/\beta$. By using the mean value theorem, this relation is written as

$$-\nabla f(x^k + \tilde{\alpha}^k d^k)'d^k < -\sigma\nabla f(x^k)'d^k, \qquad \forall\ k \in \bar{\mathcal{K}},\ k \geq \bar{k},$$

where $\tilde{\alpha}^k$ is a scalar in the interval $[0, \bar{\alpha}^k]$. Taking limits in the above relation we obtain

$$-\nabla f(\bar{x})'\bar{d} \leq -\sigma\nabla f(\bar{x})'\bar{d}$$

or

$$0 \leq (1-\sigma)\nabla f(\bar{x})'\bar{d}.$$

Since $\sigma < 1$, it follows that

$$0 \leq \nabla f(\bar{x})'\bar{d}, \tag{1.21}$$

which contradicts the assumption that $\{d^k\}$ is gradient related. This proves the result for the Armijo rule.

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

$$f(x^k) - f(x^{k+1}) \geq f(x^k) - f(\tilde{x}^{k+1}) \geq -\sigma\tilde{\alpha}^k\nabla f(x^k)'d^k. \tag{1.22}$$

By repeating the arguments of the earlier proof following Eq. (1.18), replacing α^k by $\tilde{\alpha}^k$, we can obtain a contradiction. In particular, we have

$$\{\tilde{\alpha}^k\}_{\mathcal{K}} \to 0,$$

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

$$f(x^k) - f\big(x^k + (\tilde{\alpha}^k/\beta)d^k\big) < -\sigma(\tilde{\alpha}^k/\beta)\nabla f(x^k)'d^k, \qquad \forall\ k \in \mathcal{K},\ k \geq \bar{k},$$

[cf. Eq. (1.19)]. Proceeding as earlier, we obtain Eqs. (1.20) and (1.21) (with $\bar{\alpha}^k = \tilde{\alpha}^k/\beta$), and a contradiction of Eq. (1.21).

The line of argument just used [cf. Eq. (1.22)] establishes that any stepsize rule that gives a larger reduction in cost at each iteration than the Armijo rule inherits the convergence properties of the latter. This also proves the proposition for the limited minimization rule. **Q.E.D.**

The next proposition establishes, among other things, convergence for the case of a constant stepsize. The idea is that if the rate of growth of the gradient of f is bounded from above (i.e., the curvature of f is bounded), then one can construct a quadratic function that majorizes f; see Fig. 1.2.9. Given x^k and d^k, an appropriate constant stepsize α^k can then be obtained within an interval around the scalar $\bar{\alpha}^k$ that minimizes this quadratic function along the direction d^k. The proposition requires that for some constant $L > 0$, we have

$$\big\|\nabla f(x) - \nabla f(y)\big\| \leq L\|x - y\|, \qquad \forall\ x, y \in \Re^n, \tag{1.23}$$

which insures the boundedness of the curvature of f in every direction. This is called *Lipschitz continuity* of ∇f.

Proposition 1.2.2: (Constant Stepsize) Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$, where $\{d^k\}$ is gradient related. Assume that the Lipschitz condition (1.23) holds, and that for all k we have $d^k \neq 0$ and

$$\epsilon \leq \alpha^k \leq (2 - \epsilon)\bar{\alpha}^k, \tag{1.24}$$

where

$$\bar{\alpha}^k = \frac{\big|\nabla f(x^k)' d^k\big|}{L\|d^k\|^2},$$

and $\epsilon \in (0, 1]$ is a fixed scalar. Then every limit point of $\{x^k\}$ is a stationary point of f.

Proof: By using the descent lemma (Prop. A.24 of Appendix A), we obtain

$$\begin{aligned} f(x^k) - f(x^k + \alpha^k d^k) &\geq -\alpha^k \nabla f(x^k)' d^k - \tfrac{1}{2}(\alpha^k)^2 L\|d^k\|^2 \\ &= \alpha^k \big(|\nabla f(x^k)' d^k| - \tfrac{1}{2}\alpha^k L\|d^k\|^2\big). \end{aligned} \tag{1.25}$$

The right-hand side of Eq. (1.24) yields

$$\big|\nabla f(x^k)' d^k\big| - \tfrac{1}{2}\alpha^k L\|d^k\|^2 \geq \tfrac{1}{2}\epsilon\big|\nabla f(x^k)' d^k\big|.$$

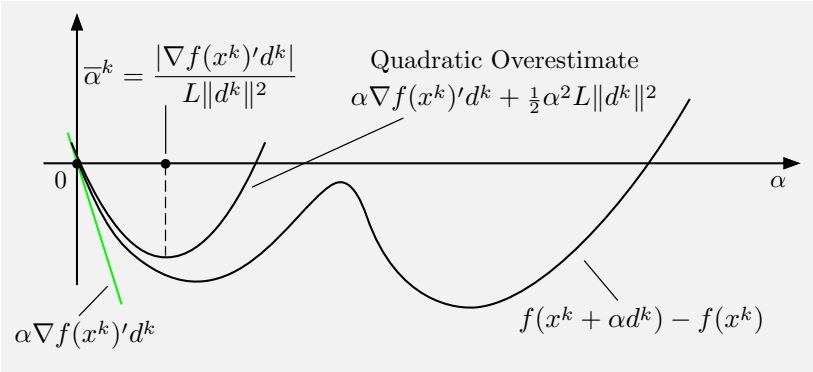

Figure 1.2.9. The idea of the proof of Prop. 1.2.2. Given x^k and the descent direction d^k, the cost difference $f(x^k + \alpha d^k) - f(x^k)$ is majorized by

$$\alpha\nabla f(x^k)'d^k + \tfrac{1}{2}\alpha^2 L\|d^k\|^2$$

(see the proof of Prop. 1.2.2). Minimization of this function over α yields the stepsize

$$\bar{\alpha}^k = \frac{\left|\nabla f(x^k)'d^k\right|}{L\|d^k\|^2},$$

which reduces the cost function f as well (see the proof of Prop. 1.2.2).

Using this relation together with the condition $\alpha^k \geq \epsilon$ in the inequality (1.25), we obtain the following bound on the cost improvement obtained at iteration k:

$$f(x^k) - f(x^k + \alpha^k d^k) \geq \tfrac{1}{2}\epsilon^2\left|\nabla f(x^k)'d^k\right|.$$

Now if a subsequence $\{x^k\}_{\mathcal{K}}$ converges to a nonstationary point $\bar{x}$, we must have, as in the proof of Prop. 1.2.1, $f(x^k) - f(x^{k+1}) \to 0$, and the preceding relation implies that $\left|\nabla f(x^k)'d^k\right| \to 0$. This contradicts the assumption that $\{d^k\}$ is gradient related. Hence, every limit point of $\{x^k\}$ is stationary. **Q.E.D.**

In the case of steepest descent $[d^k = -\nabla f(x^k)]$, the condition (1.24) on the stepsize becomes

$$\epsilon \leq \alpha^k \leq \frac{2-\epsilon}{L}.$$

Thus a constant stepsize roughly in the middle of the interval $[0, 2/L]$ guarantees convergence. This is a classical convergence result.

Note that the existence of an ϵ satisfying Eq. (1.24) is guaranteed by standard conditions that imply the gradient related assumption. In particular, if $\{d^k\}$ is such that there exist positive scalars c_1, c_2 such that

for all k,

$$c_1\big\|\nabla f(x^k)\big\|^2 \leq -\nabla f(x^k)'d^k, \qquad \|d^k\|^2 \leq c_2\big\|\nabla f(x^k)\big\|^2, \tag{1.26}$$

then Eq. (1.24) is satisfied if for all k we have

$$\epsilon \leq \alpha^k \leq \frac{c_1(2-\epsilon)}{Lc_2}. \tag{1.27}$$

Furthermore, if d^k has the form

$$d^k = -D^k\nabla f(x^k)$$

with D^k positive definite symmetric and having eigenvalues in an interval $[\gamma, \Gamma]$, the condition (1.26) can be seen to hold with

$$c_1 = \gamma, \qquad c_2 = \Gamma^2.$$

Exercise 1.2.3 provides an example showing that the Lipschitz condition (1.23) is essential for the validity of Prop. 1.2.2. This condition requires roughly that the "curvature" of f is no more than L at all points and in all directions. In particular, it is possible to show that this condition is satisfied, if f is twice differentiable and the eigenvalues of the Hessian $\nabla^2 f$ are bounded over $\Re^n$ by L. Unfortunately, however, it is generally difficult to obtain an estimate of L, so in most cases the range of stepsizes that guarantee convergence [cf. Eq. (1.24) or (1.27)] is unknown, and experimentation may be necessary to obtain appropriate stepsize values.

Even worse, many types of cost function f, while twice differentiable, have Hessian $\nabla^2 f$ that is unbounded over $\Re^n$ [this is so for any function $f(x)$ that grows faster than a quadratic as $x \to \infty$, such as $f(x) = \|x\|^3$]. Fortunately, the Lipschitz condition can be significantly weakened, as shown in Exercise 1.2.5. In particular, it is sufficient that it holds for all x, y in the level set

$$\big\{x \mid f(x) \leq f(x^0)\big\},$$

in which case, however, the range of stepsizes that guarantee convergence depends on the starting point x^0. Let us also note that, for the case where f is convex, there are methods that attempt to determine automatically a constant stepsize [one such method is given in Exercise 1.2.19 (with solution posted online), and another is described in [Ber15a], Section 6.1].

The Lipschitz continuity condition also essentially guarantees convergence for a diminishing stepsize, as shown by the following proposition.

Proposition 1.2.3: (Diminishing Stepsize) Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$. Assume that the Lipschitz condition (1.23) holds, and that there exist positive scalars c_1, c_2 such that for all k we have

$$c_1\big\|\nabla f(x^k)\big\|^2 \leq -\nabla f(x^k)'d^k, \qquad \|d^k\|^2 \leq c_2\big\|\nabla f(x^k)\big\|^2. \tag{1.28}$$

Suppose also that

$$\alpha^k \to 0, \qquad \sum_{k=0}^{\infty} \alpha^k = \infty.$$

Then either $f(x^k) \to -\infty$ or else $\big\{f(x^k)\big\}$ converges to a finite value and $\nabla f(x^k) \to 0$. Furthermore, every limit point of $\{x^k\}$ is a stationary point of f.

Proof: Combining Eqs. (1.25) and (1.28), we have

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \alpha^k \left(\tfrac{1}{2}\alpha^k L\|d^k\|^2 - |\nabla f(x^k)'d^k|\right) \\ &\leq f(x^k) - \alpha^k \left(c_1 - \tfrac{1}{2}\alpha^k c_2 L\right)\big\|\nabla f(x^k)\big\|^2. \end{aligned}$$

Since the linear term in α^k dominates the quadratic term in α^k for sufficiently small α^k, and $\alpha^k \to 0$, we have for some positive constant c and all k greater than some index $\bar{k}$,

$$f(x^{k+1}) \leq f(x^k) - \alpha^k c\big\|\nabla f(x^k)\big\|^2. \tag{1.29}$$

From this relation, we see that for $k \geq \bar{k}$, $\big\{f(x^k)\big\}$ is monotonically decreasing, so either $f(x^k) \to -\infty$ or $\big\{f(x^k)\big\}$ converges to a finite value. In the latter case, by adding Eq. (1.29) over all $k \geq \bar{k}$, we obtain

$$c\sum_{k=\bar{k}}^{\infty} \alpha^k\big\|\nabla f(x^k)\big\|^2 \leq f(x^{\bar{k}}) - \lim_{k\to\infty} f(x^k) < \infty.$$

We see that there cannot exist an $\epsilon > 0$ such that

$$\big\|\nabla f(x^k)\big\|^2 > \epsilon$$

for all k greater than some $\hat{k}$, since this would contradict the assumption $\sum_{k=0}^{\infty} \alpha^k = \infty$. Therefore, we must have

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

To show that $\nabla f(x^k) \to 0$, assume the contrary; that is,

$$\limsup_{k\to\infty} \big\|\nabla f(x^k)\big\| \geq \epsilon > 0. \tag{1.30}$$

Let $\{m_j\}$ and $\{n_j\}$ be sequences of indexes such that

$$m_j < n_j < m_{j+1},$$

$$\frac{\epsilon}{3} < \big\|\nabla f(x^k)\big\| \qquad \text{for } m_j \leq k < n_j, \tag{1.31}$$

$$\big\|\nabla f(x^k)\big\| \leq \frac{\epsilon}{3} \qquad \text{for } n_j \leq k < m_{j+1}. \tag{1.32}$$

Let also $\bar{j}$ be sufficiently large so that

$$\sum_{k=m_{\bar{j}}}^{\infty} \alpha^k \big\|\nabla f(x^k)\big\|^2 < \frac{\epsilon^2}{9L\sqrt{c_2}}. \tag{1.33}$$

For any $j \geq \bar{j}$ and any m with $m_j \leq m \leq n_j - 1$, we have

$$\begin{aligned}
\big\|\nabla f(x^{n_j}) - \nabla f(x^m)\big\| &\leq \sum_{k=m}^{n_j-1} \big\|\nabla f(x^{k+1}) - \nabla f(x^k)\big\| \\
&\leq L \sum_{k=m}^{n_j-1} \|x^{k+1} - x^k\| \\
&= L \sum_{k=m}^{n_j-1} \alpha^k \|d^k\| \\
&\leq L\sqrt{c_2} \sum_{k=m}^{n_j-1} \alpha^k \big\|\nabla f(x^k)\big\| \\
&\leq \frac{3L\sqrt{c_2}}{\epsilon} \sum_{k=m}^{n_j-1} \alpha^k \big\|\nabla f(x^k)\big\|^2 \\
&\leq \frac{3L\sqrt{c_2}}{\epsilon} \frac{\epsilon^2}{9L\sqrt{c_2}} \\
&= \frac{\epsilon}{3},
\end{aligned}$$

where the last two inequalities follow using Eqs. (1.31) and (1.33). Thus

$$\big\|\nabla f(x^m)\big\| \leq \big\|\nabla f(x^{n_j})\big\| + \frac{\epsilon}{3} \leq \frac{2\epsilon}{3}, \qquad \forall\ j \geq \bar{j},\ m_j \leq m \leq n_j - 1.$$

Thus, using also Eq. (1.32), we have for all $m \geq m_{\bar{j}}$

$$\big\|\nabla f(x^m)\big\| \leq \frac{2\epsilon}{3}.$$

This contradicts Eq. (1.30), implying that $\lim_{k\to\infty} \nabla f(x^k) = 0$.

Finally, if $\bar{x}$ is a limit point of x^k, then $f(x^k)$ converges to the finite value $f(\bar{x})$. Thus we have $\nabla f(x^k) \to 0$, implying that $\nabla f(\bar{x}) = 0$. **Q.E.D.**

Under the assumptions of the preceding proposition, descent is not guaranteed in the initial iterations. However, if the stepsizes are all sufficiently small [e.g., they satisfy the right-hand side inequality of Eq. (1.27)], descent is guaranteed at all iterations. In this case, it is sufficient that the Lipschitz condition $\big\|\nabla f(x) - \nabla f(y)\big\| \leq L\|x - y\|$ holds for all x, y in the set $\big\{x \mid f(x) \leq f(x^0)\big\}$ (see Exercise 1.2.5); otherwise the Lipschitz condition must hold over a set larger than $\big\{x \mid f(x) \leq f(x^0)\big\}$ to guarantee convergence (see Exercise 1.2.14 for an example).

The following proposition explains to some extent why sequences generated by gradient methods tend in practice to have unique limit points. It essentially states that local minima which are "isolated" tend to attract gradient methods: once the method gets close enough to such a minimum it remains close and converges to it.

Proposition 1.2.4: (Capture Theorem) Let f be continuously differentiable and let $\{x^k\}$ be a sequence satisfying $f(x^{k+1}) \leq f(x^k)$ for all k and generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$, which is convergent in the sense that every limit point of sequences that it generates is a stationary point of f. Assume that there exist scalars $s > 0$ and $c > 0$ such that for all k there holds

$$\alpha^k \leq s, \qquad \|d^k\| \leq c\,\big\|\nabla f(x^k)\big\|.$$

Let x^* be a local minimum of f, which is the only stationary point of f within some open set. Then there exists an open set S containing x^* such that if $x^{\bar{k}} \in S$ for some $\bar{k} \geq 0$, then $x^k \in S$ for all $k \geq \bar{k}$ and $\{x^k\} \to x^*$. Furthermore, given any scalar $\bar{\epsilon} > 0$, the set S can be chosen so that $\|x - x^*\| < \bar{\epsilon}$ for all $x \in S$.

Proof: Suppose that $\rho > 0$ is such that

$$f(x^*) < f(x), \qquad \forall\ x \text{ with } \|x - x^*\| \leq \rho.$$

Define for $t \in [0, \rho]$

$$\phi(t) = \min_{\{x \mid t \leq \|x - x^*\| \leq \rho\}} f(x) - f(x^*),$$

and note that ϕ is a monotonically nondecreasing function of t, and that $\phi(t) > 0$ for all $t \in (0, \rho]$. Given any $\epsilon \in (0, \rho]$, let $r \in (0, \epsilon]$ be such that

$$\|x - x^*\| < r \quad \Rightarrow \quad \|x - x^*\| + sc\big\|\nabla f(x)\big\| < \epsilon. \tag{1.34}$$

Consider the open set

$$S = \big\{x \mid \|x - x^*\| < \epsilon,\, f(x) < f(x^*) + \phi(r)\big\}.$$

We claim that if $x^k \in S$ for some k, then $x^{k+1} \in S$.

Indeed if $x^k \in S$, from the definition of ϕ and S we have

$$\phi\big(\|x^k - x^*\|\big) \le f(x^k) - f(x^*) < \phi(r).$$

Since ϕ is monotonically nondecreasing, the above relation implies that $\|x^k - x^*\| < r$, so that by Eq. (1.34),

$$\|x^k - x^*\| + sc\big\|\nabla f(x^k)\big\| < \epsilon.$$

We also have by using the hypotheses $\alpha^k \le s$ and $\|d^k\| \le c\big\|\nabla f(x^k)\big\|$

$$\|x^{k+1} - x^*\| \le \|x^k - x^*\| + \|\alpha^k d^k\| \le \|x^k - x^*\| + sc\big\|\nabla f(x^k)\big\|,$$

so from the last two relations it follows that $\|x^{k+1} - x^*\| < \epsilon$. Since $f(x^{k+1}) < f(x^k)$, we also obtain $f(x^{k+1}) - f(x^*) < \phi(r)$, so we conclude that $x^{k+1} \in S$.

By using induction it follows that if $x^{\bar{k}} \in S$ for some $\bar{k}$, we have $x^k \in S$ for all $k \ge \bar{k}$. Let $\bar{S}$ be the closure of S. Since $\bar{S}$ is compact, the sequence $\{x^k\}$ will have at least one limit point, which by assumption must be a stationary point of f. Now the only stationary point of f within $\bar{S}$ is the point x^* (since we have $\|x - x^*\| \le \rho$ for all $x \in \bar{S}$). Hence $x^k \to x^*$. Finally given any $\bar{\epsilon} > 0$, we can choose $\epsilon \le \bar{\epsilon}$ in which case we have $\|x - x^*\| < \bar{\epsilon}$ for all $x \in S$. **Q.E.D.**

Note that in the preceding proposition, the conditions $f(x^{k+1}) \le f(x^k)$ and $\alpha^k \le s$ are satisfied for the Armijo rule and the limited minimization rule. They are also satisfied for a constant and a diminishing stepsize under conditions that guarantee descent at each iteration (see the proofs of Props. 1.2.2 and 1.2.3). The condition $\|d^k\| \le c\big\|\nabla f(x^k)\big\|$ is satisfied if $d^k = -D^k \nabla f(x^k)$ with the eigenvalues of D^k bounded from above.

Finally, we state a result that deals with the convergence of algorithms involving a combination of different methods. It shows that for convergence it is enough to insert, perhaps irregularly but infinitely often, an iteration of a convergent gradient algorithm, provided that the other iterations do

not degrade the value of the cost function. The proof is similar to the one of Prop. 1.2.1, and is left for the reader.

Proposition 1.2.5: (Convergence for Spacer Steps) Consider a sequence $\{x^k\}$ such that

$$f(x^{k+1}) \leq f(x^k), \qquad k = 0, 1, \ldots$$

Assume that there exists an infinite set $\mathcal{K}$ of integers for which

$$x^{k+1} = x^k + \alpha^k d^k, \qquad \forall\ k \in \mathcal{K},$$

where $\{d^k\}_{\mathcal{K}}$ is gradient related and α^k is chosen by the minimization rule, or the limited minimization rule, or the Armijo rule. Then every limit point of the subsequence $\{x^k\}_{\mathcal{K}}$ is a stationary point.

E X E R C I S E S

1.2.1

Consider the problem of minimizing the function of two variables $f(x, y) = 3x^2 + y^4$.

(a) Apply one iteration of the steepest descent method with $(1, -2)$ as the starting point and with the stepsize chosen by the Armijo rule with $s = 1$, $\sigma = 0.1$, and $\beta = 0.5$.

(b) Repeat (a) using $s = 1$, $\sigma = 0.1$, $\beta = 0.1$ instead. How does the cost of the new iterate compare to that obtained in (a)? Comment on the tradeoffs involved in the choice of β.

(c) Apply one iteration of Newton's method with the same starting point and stepsize rule as in (a). How does the cost of the new iterate compare to that obtained in (a)? How about the amount of work involved in finding the new iterate?

1.2.2

Describe the behavior of the steepest descent method with constant stepsize s for the function $f(x) = \|x\|^{2+\beta}$, where $\beta \geq 0$. For which values of s and x^0 does

the method converge to $x^* = 0$. Relate your answer to the assumptions of Prop. 1.2.2.

1.2.3

Consider the function $f : \Re^n \mapsto \Re$ given by

$$f(x) = \|x\|^{3/2},$$

and the method of steepest descent with a constant stepsize. Show that for this function, the Lipschitz condition $\big\|\nabla f(x) - \nabla f(y)\big\| \leq L\|x - y\|$ for all x and y is not satisfied for any L. Furthermore, for any value of constant stepsize, the method either converges in a *finite* number of iterations to the minimizing point $x^* = 0$ or else it does not converge to x^*.

1.2.4

Apply the steepest descent method with constant stepsize α to the function f of Exercise 1.1.11. Show that the gradient ∇f satisfies the Lipschitz condition

$$\big\|\nabla f(x) - \nabla f(y)\big\| \leq L\|x - y\|, \qquad \forall\ x, y \in \Re,$$

for some constant L. Write a computer program to verify that the method is a descent method for $\alpha \in (0, 2/L)$. Do you expect to get in the limit the global minimum $x^* = 0$?

1.2.5 (www)

Suppose that the Lipschitz condition

$$\big\|\nabla f(x) - \nabla f(y)\big\| \leq L\|x - y\|, \qquad \forall\ x, y \in \Re^n,$$

[cf. Eq. (1.23)] is replaced by the condition that for every bounded set $A \subset \Re^n$, there exists some constant L such that

$$\big\|\nabla f(x) - \nabla f(y)\big\| \leq L\|x - y\|, \qquad \forall\ x, y \in A. \tag{1.35}$$

Show that:

(a) Condition (1.35) is always satisfied if the level sets $\big\{x \mid f(x) \leq c\big\}$, $c \in \Re$, are bounded, and f is twice continuously differentiable.

(b) The convergence result of Prop. 1.2.2 remains valid provided that the level set

$$A = \big\{x \mid f(x) \leq f(x^0)\big\}$$

is bounded and the stepsize is allowed to depend on the choice of the initial vector x^0. *Hint*: The key idea is to show that x^k stays in the set A, and

to use a stepsize α^k that depends on the constant L corresponding to this set. Let

$$R = \max\big\{\|x\| \mid x \in A\big\},$$

$$G = \max\big\{\|\nabla f(x)\| \mid x \in A\big\},$$

and

$$B = \big\{x \mid \|x\| \leq R + 2G\big\}.$$

Using condition (1.35), there exists some constant L such that $\big\|\nabla f(x) - \nabla f(y)\big\| \leq L\|x - y\|$, for all $x, y \in B$. Suppose the stepsize α^k satisfies

$$0 < \epsilon \leq \alpha^k \leq (2 - \epsilon)\gamma^k \min\{1, 1/L\},$$

where

$$\gamma^k = \frac{\big|\nabla f(x^k)' d^k\big|}{\|d^k\|^2}.$$

Let $\beta^k = \alpha^k(\gamma^k - L\alpha^k/2)$, which can be seen to satisfy $\beta^k \geq \epsilon^2\gamma^k/2$ by our choice of α^k. Show by induction on k that with such a choice of stepsize, we have $x^k \in A$ and

$$f(x^{k+1}) \leq f(x^k) - \beta^k\|d^k\|^2, \qquad \forall\ k \geq 0.$$

1.2.6

Suppose that f is quadratic and of the form $f(x) = \frac{1}{2}x'Qx - b'x$, where Q is positive definite and symmetric.

(a) Show that the Lipschitz condition $\big\|\nabla f(x) - \nabla f(y)\big\| \leq L\|x - y\|$ is satisfied with L equal to the maximal eigenvalue of Q.

(b) Consider the gradient method $x^{k+1} = x^k - sD\nabla f(x^k)$, where D is positive definite and symmetric. Show that the method converges to $x^* = Q^{-1}b$ for every starting point x^0 if and only if $s \in (0, 2/\bar{L})$, where $\bar{L}$ is the maximum eigenvalue of $D^{1/2}QD^{1/2}$.

1.2.7

An electrical engineer wants to maximize the current I between two points A and B of a complex network by adjusting the values x_1 and x_2 of two variable resistors, where $0 \leq x_1 \leq R_1$, $0 \leq x_2 \leq R_2$, and R_1, R_2 are given. The engineer does not have an adequate mathematical model of the network and decides to adopt the following procedure. She keeps the value x_2 of the second resistor fixed and adjusts the value of the first resistor until the current I is maximized. She then keeps the value x_1 of the first resistor fixed and adjusts the value of the second resistor until the current I is maximized. She then repeats the procedure until no further progress can be made. She knows *a priori* that during this procedure, the values x_1 and x_2 can never reach their extreme values 0, R_1, and R_2. Explain whether there is a sound theoretical basis for the engineer's procedure. *Hint*: Consider how the steepest descent method works for two-dimensional problems.

1.2.8

Consider the gradient method $x^{k+1} = x^k + \alpha^k d^k$, where α^k is chosen by the Armijo rule or the line minimization rule and

$$d^k = -\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{\partial f(x^k)}{\partial x_i} \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

where i is the index for which $\left|\partial f(x^k)/\partial x_j\right|$ is maximized over $j = 1, \ldots, n$. Show that every limit point of $\{x^k\}$ is stationary.

1.2.9 (www)

Let f be twice continuously differentiable. Suppose that x^* is a local minimum such that for all x in an open sphere S centered at x^*, we have, for some $m > 0$,

$$m\|d\|^2 \leq d'\nabla^2 f(x)d, \qquad \forall\ d \in \Re^n.$$

Show that for every $x \in S$, we have

$$\|x - x^*\| \leq \frac{\left\|\nabla f(x)\right\|}{m}, \qquad f(x) - f(x^*) \leq \frac{\left\|\nabla f(x)\right\|^2}{2m}.$$

Hint: Use the relation

$$\nabla f(y) = \nabla f(x) + \int_0^1 \nabla^2 f\big(x + t(y-x)\big)(y-x)dt.$$

See also Exercise 1.1.9.

1.2.10 (Alternative Assumptions for Convergence) (www)

Consider the gradient method $x^{k+1} = x^k + \alpha^k d^k$. Instead of $\{d^k\}$ being gradient related, assume *one* of the following two conditions:

(i) It can be shown that for any subsequence $\{x^k\}_{k\in\mathcal{K}}$ that converges to a nonstationary point, the corresponding subsequence $\{d^k\}_{k\in\mathcal{K}}$ is bounded and satisfies

$$\liminf_{k\to\infty,\, k\in\mathcal{K}} \nabla f(x^k)'d^k < 0.$$

(ii) α^k is chosen by the minimization rule, and for some $c > 0$ and all k, we have

$$\left|\nabla f(x^k)'d^k\right| \geq c\left\|\nabla f(x^k)\right\| \|d^k\|.$$

Show that the conclusion of Prop. 1.2.1 holds.

1.2.11 (Behavior of Steepest Descent Near a Saddle Point)

Let $f(x) = (1/2)x'Qx$, where Q is symmetric, invertible, and has at least one negative eigenvalue. Consider the steepest descent method with constant stepsize and show that unless the starting point x^0 belongs to the subspace spanned by the eigenvectors of Q corresponding to the nonnegative eigenvalues, the generated sequence $\{x^k\}$ diverges.

1.2.12 (Convergence of Steepest Descent to a Single Limit) (www)

Consider the steepest descent method $x^{k+1} = x^k - \alpha^k \nabla f(x^k)$, and assume that f is convex, has at least one minimizing point, and for all x, y, and some $L > 0$, satisfies

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

Show that $\{x^k\}$ converges to a minimizing point of f under each of the following two stepsize rule conditions:

(i) For some $\epsilon > 0$, we have

$$\epsilon \leq \alpha^k \leq \frac{2 - \epsilon}{L}, \qquad \forall\ k.$$

(ii) $\alpha^k \to 0$ and $\sum_{k=0}^{\infty} \alpha^k = \infty$.

Note: This result is due to [BGI95], who also show convergence to a single limit for a variant of the Armijo rule.

1.2.13 (Steepest Descent with Diminishing Stepsize [CoL94]) (www)

Consider the steepest descent method $x^{k+1} = x^k - \alpha^k \nabla f(x^k)$, and assume that the function f is convex.

(a) Use the convexity of f to show that for any $y \in \Re^n$, we have

$$\|x^{k+1} - y\|^2 \leq \|x^k - y\|^2 - 2\alpha^k \big(f(x^k) - f(y)\big) + \big(\alpha^k \|\nabla f(x^k)\|\big)^2.$$

(b) Assume that

$$\sum_{k=0}^{\infty} \alpha^k = \infty, \qquad \alpha^k \big\|\nabla f(x^k)\big\|^2 \to 0.$$

Show that $\liminf_{k\to\infty} f(x^k) = \inf_{x\in\Re^n} f(x)$. *Hint*: Argue by contradiction. Assume that for some $\delta > 0$, there exists y with $f(y) < f(x^k) - \delta$ for all k sufficiently large. Use part (a).

(c) Assume that

$$\alpha^k = \frac{s^k}{\big\|\nabla f(x^k)\big\|},$$

where

$$\sum_{k=0}^{\infty} s^k = \infty, \qquad \sum_{k=0}^{\infty} (s^k)^2 < \infty.$$

Show that $\liminf_{k\to\infty} f(x^k) = \inf_{x\in\Re^n} f(x)$, and that if f has at least one global minimum, then $\{x^k\}$ converges to some global minimum. *Hint*: In part (a), set y to some x^* such that $f(x^*) < f(x^k)$ for all k (if no such x^* exists, we are done). Show that the relation

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 + (s^k)^2$$

implies that $\{x^k\}$ is bounded and hence also that $\{\nabla f(x^k)\}$ is bounded. Use part (b).

1.2.14 (Divergence with Diminishing Stepsize)

Consider the one-dimensional function

$$f(x) = \frac{2}{3}|x|^3 + \frac{1}{2}x^2$$

and the method of steepest descent with stepsize $\alpha^k = \gamma/(k+1)$, where γ is a positive scalar.

(a) Show that for $\gamma = 1$ and $|x^0| \geq 1$ the method diverges. In particular, show that $|x^k| \geq k+1$ for all k.

(b) Characterize as best as you can the set of pairs (γ, x^0) for which the method converges to $x^* = 0$.

(c) How do you reconcile the results of (a) and (b) with Prop. 1.2.3.

1.2.15 (Wolfe Conditions for Line Search Accuracy)

There are several criteria for implementing approximately the minimization rule in a gradient method. An example of such a criterion is that α^k satisfies simultaneously

$$f(x^k) - f(x^k + \alpha^k d^k) \geq -\sigma\alpha^k \nabla f(x^k)' d^k, \tag{1.36}$$

$$\nabla f(x^k + \alpha^k d^k)' d^k \geq \beta \nabla f(x^k)' d^k, \tag{1.37}$$

where α and β are some scalars with $\sigma \in (0, 1/2)$ and $\beta \in (\sigma, 1)$. If α^k is indeed a minimizing stepsize, then $\nabla f(x^k + \alpha^k d^k)' d^k = dg(\alpha^k)/d\alpha = 0$, where g is the function $g(\alpha) = f(x^k + \alpha d^k)$, so Eq. (1.37) is in effect a test on the accuracy of the minimization (see Fig. 1.2.10).

(a) Show that if conditions (1.36) and (1.37) are satisfied by a gradient method at each iteration and the direction sequence is gradient related, then all limit points of the generated sequence $\{x^k\}$ are stationary points of f.

(b) Assume that there is a scalar b such that $f(x) \geq b$. Show that there exists an interval $[c_1, c_2]$ with $0 < c_1 < c_2$, such that every $\alpha \in [c_1, c_2]$ satisfies Eqs. (1.36) and (1.37).

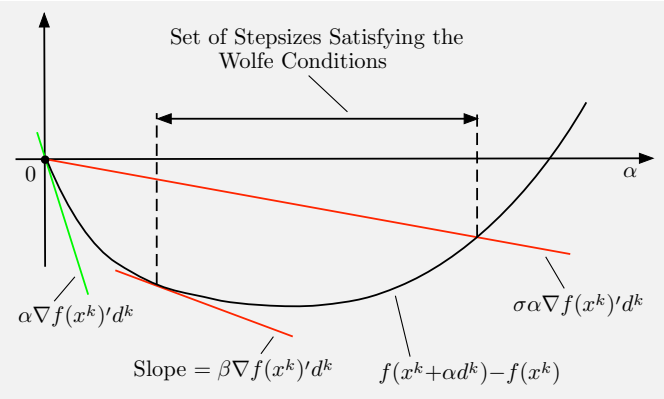

Figure 1.2.10. Illustration of the stepsize selection criterion based on the Wolfe conditions.

1.2.16 (Steepest Descent with Errors) www

Consider the steepest descent method $x^{k+1} = x^k - \alpha^k\big(\nabla f(x^k) + e^k\big)$, where e^k is an error satisfying $\|e^k\| \leq \delta$ for all k. Assume that ∇f is Lipschitz continuous. Show that for any $\delta' > \delta$, there exists a range of positive stepsizes $[\underline{\alpha}, \bar{\alpha}]$ such that if $\alpha^k \in [\underline{\alpha}, \bar{\alpha}]$ for all sufficiently large k, then either $f(x^k) \to -\infty$ or $\big\|\nabla f(x^k)\big\| < \delta'$ for infinitely many values of k. *Hint*: Use the reasoning of Prop. 1.2.2.

1.2.17 (Multiple Limit Points for Steepest Descent [Zou76])

Consider the two-dimensional function

$$f(x) = \begin{cases} (r-1)^2 - \frac{1}{2}(r-1)^2 \cos\left(\frac{1}{r-1} - \phi\right) & \text{if } r \neq 1, \\ 0 & \text{if } r = 1, \end{cases}$$

where

$$r = \sqrt{x_1^2 + x_2^2}, \qquad \phi = \arctan(x_1/x_2).$$

This function is minimized at each point of the circle where $r = 1$. Consider a nonoptimal starting point and the method of steepest descent where x^{k+1} is set equal to the first local minimum along the line $\big\{x^k - \alpha\nabla f(x^k) \mid \alpha \geq 0\big\}$. Show that this method follows a spiral path that comes arbitrarily close to every point of the circle of optimal points. *Note:* For another example of convergence to multiple limit points, which involves a convex differentiable cost function, see [Gon00]. In this example the steepest descent method with the exact line minimization rule produces a sequence with four limit points.

1.2.18 (Simplified Steepest Descent) www

(a) Consider the unconstrained minimization of a function f of the form

$$f(x) = F\big(x, g(x)\big),$$

where $g : \Re^m \to \Re^n$ is continuously differentiable and $F(x, y)$ is a continuously differentiable function of the two arguments $x \in \Re^n$ and $y \in \Re^m$. It is sometimes convenient to approximate the gradient of $F\big(x, g(x)\big)$ by neglecting the dependence on g. This leads to the method

$$x^{k+1} = x^k - \alpha^k \nabla_x F\big(x^k, g(x^k)\big),$$

where α^k is chosen by the minimization rule or the Armijo rule on the function f. (Such a method makes sense when $\nabla_x F$ is much easier to compute than $\nabla g \nabla_y F$.) Show that if there exists $\gamma \in (0, 1)$ such that

$$\big\|\nabla g(x) \nabla_y F\big(x, g(x)\big)\big\| \leq \gamma \big\|\nabla_x F\big(x, g(x)\big)\big\|, \qquad \forall\, x \in \Re^n,$$

then the method is convergent in the sense that all limit points of the sequences that it generates are stationary points of f.

(b) Consider the constrained minimization problem

$$\begin{aligned} &\text{minimize} \quad f(x, y) \\ &\text{subject to} \quad h(x, y) = 0 \end{aligned}$$

where $f : \Re^{n+m} \to \Re$ and $h : \Re^{n+m} \to \Re^m$ are continuously differentiable functions of the two arguments $x \in \Re^n$ and $y \in \Re^m$. Consider also a method of the form

$$x^{k+1} = x^k - \alpha^k \nabla_x f\big(x^k, y^k\big),$$

where y^k is a solution of $h(x^k, y) = 0$, viewed as a system of m equations in the unknown vector y, and α^k is chosen by the minimization rule or the Armijo rule. Formulate conditions that guarantee that this method is convergent.

1.2.19 (A Stepsize Reduction Rule for Convex Problems) www

Suppose that the cost function f is convex, and consider a gradient method $x^{k+1} = x^k + \alpha^k d^k$ where the assumptions of Prop. 1.2.3 are satisfied, except that the stepsize α^k is determined by the following rule:

$$\alpha^{k+1} = \begin{cases} \alpha^k & \text{if } \nabla f(x^{k+1})' d^k \leq 0, \\ \beta \alpha^k & \text{otherwise,} \end{cases}$$

where $\beta \in (0,1)$ is a fixed scalar and α^0 is any positive scalar.

(a) Show that the stepsize is reduced after iteration k if and only if the interval I^k connecting x^k and x^{k+1} contains in its interior all the vectors $\bar{x}^k$ that minimize $f(x)$ over $x \in I^k$.

(b) Show that the stepsize will be constant after a finite number of iterations.

(c) Show that either $f(x^k) \to -\infty$ or else $\{f(x^k)\}$ converges to a finite value and $\nabla f(x^k) \to 0$. Furthermore, every limit point of $\{x^k\}$ is a global minimum of f.

1.2.20 (Convergence of Gradient Method with Errors [BeT96], [BeT00]) (www)

Let $\{x^k\}$ be a sequence generated by the gradient method with errors

$$x^{k+1} = x^k + \alpha^k(d^k + e^k),$$

where the following hold:

(1) ∇f satisfies the Lipschitz assumption of Prop. 1.2.3.

(2) d^k satisfies

$$c_1\big\|\nabla f(x^k)\big\|^2 \leq -\nabla f(x^k)'d^k, \quad \|d^k\| \leq c_2\big(1 + \|\nabla f(x^k)\|\big), \qquad \forall\ k,$$

where c_1 and c_2 are some scalars.

(3) The stepsizes α^k satisfy

$$\sum_{k=0}^{\infty} \alpha^k = \infty, \qquad \sum_{k=0}^{\infty} (\alpha^k)^2 < \infty.$$

(4) The errors e^k satisfy

$$\|e^k\| \leq \alpha^k\big(q + p\|\nabla f(x^k)\|\big), \qquad \forall\ k,$$

where q and p are some scalars.

Show that either $f(x^k) \to -\infty$ or else $\big\{f(x^k)\big\}$ converges to a finite value and $\nabla f(x^k) \to 0$. Furthermore, every limit point of $\{x^k\}$ is a stationary point of f. *Hint*: Show that for sufficiently large k, we have

$$f(x^{k+1}) \leq f(x^k) - \alpha^k b_1\big\|\nabla f(x^k)\big\|^2 + (\alpha^k)^2 b_2$$

for some constants b_1 and b_2. Use the line of argument of Prop. 1.2.3.

1.3 GRADIENT METHODS – RATE OF CONVERGENCE

The second major issue regarding gradient methods relates to the rate (or speed) of convergence of the generated sequences $\{x^k\}$. The mere fact that $\{x^k\}$ converges to a stationary point x^* will be of little practical value unless the points x^k are reasonably close to x^* after relatively few iterations. Thus, the study of the rate of convergence provides what are often the dominant criteria for selecting one algorithm in favor of others for solving a particular problem.

Approaches for Rate of Convergence Analysis

There are several approaches towards quantifying the rate of convergence of nonlinear programming algorithms. We will discuss briefly three possibilities and then concentrate on the third.

(a) *Computational complexity approach*: Here we try to estimate the number of elementary operations needed by a given method to find an optimal solution exactly or within an ϵ-tolerance. Usually, this approach provides worst-case estimates, that is, upper bounds on the number of required operations over a class of problems of given dimension and type (e.g. linear, convex, etc.). These estimates may also involve parameters such as the distance of the starting point from the optimal solution set, etc.

(b) *Informational complexity approach*: One difficulty with the computational complexity approach is that for a diverse class of problems, it is often difficult or meaningless to quantify the amount of computation needed for a single function or gradient evaluation. For example, in estimating the computational complexity of the gradient method applied to the entire class of differentiable convex functions, how are we to compare the overhead for finding the stepsize and for updating the x vector with the work needed to compute the cost function value and its gradient? The informational complexity approach, which is discussed in detail in the book [NeY83] (see also [TrW80]), bypasses this difficulty by estimating the number of function (and possibly gradient) evaluations needed to find an exact or approximately optimal solution (as opposed to the number of necessary computational operations). In other respects, the informational and computational complexity approaches are similar.

(c) *Local analysis*: In this approach we focus on the local behavior of the method in a neighborhood of an optimal solution. Local analysis can describe quite accurately the behavior of a method near the solution by using series approximations, but ignores entirely the behavior of the method when far from the solution.

The main potential advantage of the computational and informational complexity approaches is that they provide information about the method's progress when far from the eventual limit. Unfortunately, however, this information is usually pessimistic as it accounts for the worst possible problem instance within the class considered. This has resulted in some striking discrepancies between the theoretical model predictions and practical real-world observations. For example, the most widely used linear programming method, the simplex method, is categorized as a "bad" method by worst-case complexity analysis, because it performs very poorly on some specially constructed examples, which, however, are highly unlikely in practice. On the other hand, the ellipsoid method of Khachiyan [Kha79]† is categorized as much better than the simplex method by worst-case complexity analysis, even though it performs very poorly on most practical linear programs.

The computational complexity approach has received considerable attention in the context of interior point methods. These methods, discussed in Section 5.1, were primarily motivated by Karmarkar's development of a linear programming algorithm with a polynomial complexity bound that was more favorable than the one of the ellipsoid method [Kar84]. It turned out, however, that the worst-case predictions for the required number of iterations of these methods were off by many orders of magnitude from the practically observed number of iterations. Furthermore, the interior point methods that perform best in practice have poor worst-case complexity, while the ones with the best complexity bounds are very slow in practice.

The local analysis approach, which will be adopted almost exclusively in this text, has enjoyed considerable success in predicting the behavior of various methods near nonsingular local minima where the cost function can be well approximated by a quadratic. Moreover it is often more intuitive than the computational complexity approach, and lends itself better to geometrical interpretations. However, the local analysis approach also has some drawbacks, the most important of which is that it does not account for the rate of progress in the initial iterations. Nonetheless, in many practical situations this is not a serious omission because progress is fast in the initial iterations and slows down only in the limit (the reasons for this seem hard to understand; they are problem-dependent). Furthermore, often in practice, starting points that are near a solution are easily obtainable by a combination of heuristics and experience from problems with similar data, in which case local analysis becomes more meaningful.

Local analysis has not been very successful for problems which either involve singular local minima or which are difficult in the sense that the principal methods take many iterations to get near their solution where local analysis applies. Theoretical guidance to help a practitioner who is

† The ellipsoid method was chronologically the first linear programming algorithm with a polynomial complexity bound; see [BGT81] or [BeT97] for a survey and discussion of this method.

faced with such problems is an important subject, which is still under active development.

1.3.1 The Local Analysis Approach

We now formalize the basic ingredients of our local rate of convergence analysis approach. These are:

(a) We restrict attention to sequences $\{x^k\}$ that converge to a unique limit point x^*.

(b) Rate of convergence is evaluated using an *error function* $e : \Re^n \mapsto \Re$ satisfying $e(x) \geq 0$ for all $x \in \Re^n$ and $e(x^*) = 0$. Typical choices are the Euclidean distance

$$e(x) = \|x - x^*\|$$

and the cost difference

$$e(x) = \big|f(x) - f(x^*)\big|.$$

(c) Our analysis is asymptotic; that is, we look at the rate of convergence of the tail of the error sequence $\big\{e(x^k)\big\}$.

(d) The generated error sequence $\big\{e(x^k)\big\}$ is compared with some "standard" sequences. In our case, we compare $\big\{e(x^k)\big\}$ with the geometric progression

$$\beta^k, \qquad k = 0, 1, \ldots,$$

where $\beta \in (0, 1)$ is some scalar. In particular, we say that $\big\{e(x^k)\big\}$ *converges linearly or geometrically*, if there exist $q > 0$ and $\beta \in (0, 1)$ such that for all k

$$e(x^k) \leq q\beta^k.$$

It is possible to show that linear convergence is obtained if for some $\beta \in (0, 1)$ we have

$$\limsup_{k\to\infty} \frac{e(x^{k+1})}{e(x^k)} \leq \beta;$$

that is, asymptotically, the error is decreasing by a factor of at least β at each iteration (see Exercise 1.3.6, which gives several additional convergence rate characterizations). If for every $\beta \in (0, 1)$, there exists q such that the condition $e(x^k) \leq q\beta^k$ holds for all k, we say that $\big\{e(x^k)\big\}$ converges *superlinearly*. This is true in particular, if

$$\lim_{k\to\infty} \frac{e(x^{k+1})}{e(x^k)} = 0.$$

To quantify further the notion of superlinear convergence, we may compare $\{e(x^k)\}$ with the sequence

$$(\beta)^{p^k}, \qquad k = 0, 1, \ldots,$$

(this is β raised to the power p raised to the power k) where $\beta \in (0,1)$, and $p > 1$ are some scalars. This sequence converges much faster than a geometric progression. We say that $\{e(x^k)\}$ *converges at least superlinearly with order* p, if there exist $q > 0$, $\beta \in (0,1)$, and $p > 1$ such that for all k

$$e(x^k) \leq q\,(\beta)^{p^k}.$$

The case where $p = 2$ is referred to as *quadratic convergence*. It is possible to show that superlinear convergence with order p is obtained if

$$\limsup_{k\to\infty} \frac{e(x^{k+1})}{e(x^k)^p} < \infty,$$

or equivalently, $e(x^{k+1}) = O\big(e(x^k)^p\big)$; see Exercise 1.3.7.

Most optimization algorithms that are of interest in practice produce sequences converging either linearly or superlinearly, at least when they converge to nonsingular local minima. Linear convergence is a fairly satisfactory rate of convergence for nonlinear programming algorithms, provided the factor β of the associated geometric progression is not too close to unity. Several nonlinear programming algorithms converge superlinearly for particular classes of problems. Newton's method is an important example, as we will see in the present section and also in Section 1.4. For convergence to singular local minima, slower than linear convergence rate is expected for most cases. One may then compare $\{e(x^k)\}$ with some standard sequences that converge sublinearly, such as $\{qk^{-p}\}$, where $q > 0$ and $p \geq 1$.

1.3.2 The Role of the Condition Number

Many of the important convergence rate characteristics of gradient methods reveal themselves when the cost function is quadratic. To see why, assume that a gradient method is applied to minimization of a twice continuously differentiable function $f : \Re^n \mapsto \Re$, and it generates a sequence $\{x^k\}$ converging to a nonsingular local minimum x^*. We have

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

Therefore, since $\nabla^2 f(x^*)$ is positive definite, f can be accurately approximated near x^* by the quadratic function

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

We thus expect that asymptotic convergence rate results obtained for the quadratic cost case have direct analogs for the general case. This conjecture has been substantiated by extensive numerical experimentation, and is one of the most reliable analytical guidelines in nonlinear programming.

For this reason, we take the positive definite quadratic case as our point of departure. In Exercise 1.3.3 we extend our analysis for the case where f has Lipschitz continuous gradient and is strongly convex (and hence also has positive definite Hessian when it is twice differentiable). We also discuss later in this section what happens when $\nabla^2 f(x^*)$ is not positive definite, in which case an analysis based on a quadratic model is inadequate.

Convergence Rate of Steepest Descent for Quadratic Functions

Suppose that the cost function f is quadratic with positive definite Hessian Q. We may assume without loss of generality that f is minimized at $x^* = 0$ and that $f(x^*) = 0$ [otherwise we can use the change of variables $y = x - x^*$ and subtract the constant $f(x^*)$ from $f(x)$]. Thus we have

$$f(x) = \tfrac{1}{2} x'Qx, \qquad \nabla f(x) = Qx, \qquad \nabla^2 f(x) = Q.$$

The steepest descent method takes the form

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = (I - \alpha^k Q)x^k.$$

Therefore, we have

$$\|x^{k+1}\|^2 = x^{k\prime}(I - \alpha^k Q)^2 x^k.$$

Since by Prop. A.18(b) of Appendix A, we have for all $x \in \Re^n$

$$x'(I - \alpha^k Q)^2 x \leq \big(\text{maximum eigenvalue of } (I - \alpha^k Q)^2\big)\|x\|^2,$$

we obtain

$$\|x^{k+1}\|^2 \leq \big(\text{maximum eigenvalue of } (I - \alpha^k Q)^2\big)\|x^k\|^2.$$

Using Prop. A.13 of Appendix A, it can be seen that the eigenvalues of $(I - \alpha^k Q)^2$ are equal to $(1 - \alpha^k \lambda_i)^2$, where λ_i are the eigenvalues of Q. Therefore, we have

$$\text{maximum eigenvalue of } (I - \alpha^k Q)^2 = \max\big\{(1 - \alpha^k m)^2, (1 - \alpha^k M)^2\big\},$$

where

$$m: \text{ smallest eigenvalue of } Q,$$

$$M: \text{ largest eigenvalue of } Q.$$

It follows that for $x^k \neq 0$, we have

$$\frac{\|x^{k+1}\|}{\|x^k\|} \leq \max\bigl\{|1-\alpha^k m|, |1-\alpha^k M|\bigr\}. \tag{1.38}$$

It can be seen that this inequality holds as an equation if x^k is proportional to an eigenvector corresponding to m and if $|1-\alpha^k m| \geq |1-\alpha^k M|$. Otherwise, if $|1-\alpha^k m| < |1-\alpha^k M|$, the inequality holds as an equation if x^k is proportional to an eigenvector corresponding to M.

The relation (1.38) is a fundamental convergence rate bound for the steepest descent method with a constant stepsize, which admits an extension to the case of a strongly convex cost function with Lipschitz continuous gradient (see Exercise 1.3.3). Figure 1.3.1 illustrates the bound of Eq. (1.38) as a function of the stepsize α^k. It can be seen that the value of α^k that minimizes the bound is

$$\alpha^* = \frac{2}{M+m},$$

in which case

$$\frac{\|x^{k+1}\|}{\|x^k\|} \leq \frac{M-m}{M+m}. \tag{1.39}$$

This is the best convergence rate bound for steepest descent with constant stepsize.

There is another interesting convergence rate result, which holds when α^k is chosen by the line minimization rule. This result quantifies the rate at which the cost decreases and has the form

$$\frac{f(x^{k+1})}{f(x^k)} \leq \left(\frac{M-m}{M+m}\right)^2. \tag{1.40}$$

The above inequality is verified in Prop. 1.3.1, given in the next subsection, where we collect and prove the more formal results of this section. It can be shown that the inequality is sharp in the sense that given any Q, there is a starting point x^0 such that this inequality holds as an equation for all k (see Fig. 1.3.2).

The ratio M/m is called the *condition number* of Q, and problems where M/m is large are referred to as *ill-conditioned*. Such problems are characterized by very elongated elliptical level sets. The steepest descent method converges slowly for these problems as indicated by the convergence rate bounds of Eqs. (1.38) and (1.40), and as illustrated in Fig. 1.3.2.

Scaling and Steepest Descent

Consider now the more general method

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k), \tag{1.41}$$

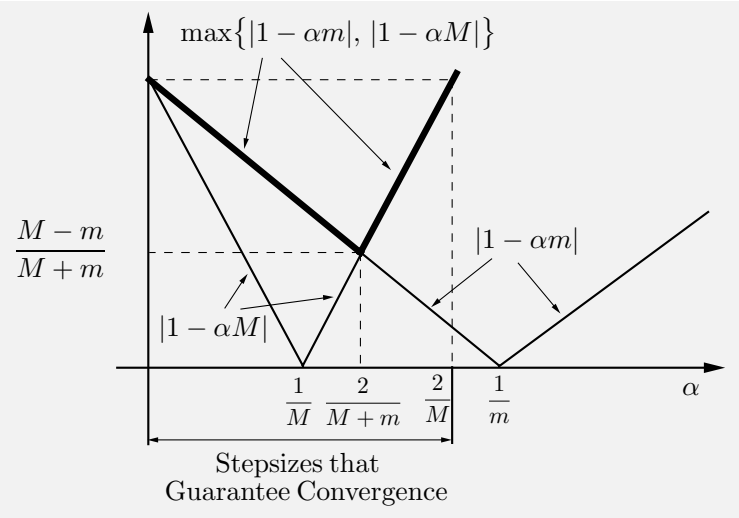

Figure 1.3.1. Illustration of the convergence rate bound

$$\frac{\|x^{k+1}\|}{\|x^k\|} \leq \max\big\{|1-\alpha m|, |1-\alpha M|\big\}$$

for steepest descent. The bound is minimized when α is such that $1 - \alpha m = \alpha M - 1$, i.e., for $\alpha = 2/(M+m)$.

where D^k is positive definite and symmetric; most of the gradient methods of interest have this form as discussed in Section 1.2. It turns out that we may view this iteration as a *scaled version of steepest descent*. In particular, this iteration is just steepest descent applied in a different coordinate system, which depends on D^k.

Indeed, let

$$S = (D^k)^{1/2}$$

denote the positive definite square root of D^k (cf. Prop. A.21 in Appendix A), and consider a transformation of variables defined by

$$x = Sy.$$

Then, in the space of y, the problem is written as

$$\begin{aligned} &\text{minimize} \quad h(y) \equiv f(Sy) \\ &\text{subject to} \quad y \in \Re^n. \end{aligned} \tag{1.42}$$

The steepest descent method for this problem takes the form

$$y^{k+1} = y^k - \alpha^k \nabla h(y^k). \tag{1.43}$$

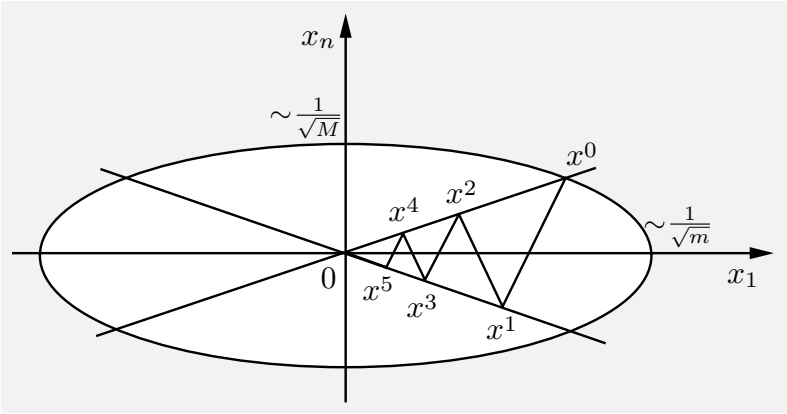

Figure 1.3.2. Example showing that the convergence rate bound

$$\frac{f(x^{k+1})}{f(x^k)} \leq \left(\frac{M-m}{M+m}\right)^2$$

is sharp for the steepest descent method with the line minimization rule. Consider the quadratic function

$$f(x) = \tfrac{1}{2}\sum_{i=1}^n \lambda_i x_i^2,$$

where $0 < m = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = M$. Any positive definite quadratic function can be put into this form by transformation of variables. Consider the starting point

$$x^0 = \left(m^{-1}, 0, \ldots, 0, M^{-1}\right)'$$

and apply the steepest descent method $x^{k+1} = x^k - \alpha^k \nabla f(x^k)$ with α^k chosen by the line minimization rule. We have $\nabla f(x^0) = (1, 0, \ldots, 0, 1)'$ and it can be verified that the minimizing stepsize is $\alpha^0 = 2/(M+m)$. Thus we obtain $x_1^1 = 1/m - 2/(M+m)$, $x_n^1 = 1/M - 2/(M+m)$, $x_i^1 = 0$ for $i = 2, \ldots, n-1$. Therefore,

$$x^1 = \left(\frac{M-m}{M+m}\right)\left(m^{-1}, 0, \ldots, 0, -M^{-1}\right)'$$

and, we can verify by induction that for all k,

$$x^{2k} = \left(\frac{M-m}{M+m}\right)^{2k} x^0, \qquad x^{2k+1} = \left(\frac{M-m}{M+m}\right)^{2k} x^1.$$

Thus, there exist starting points on the plane of points x of the form $x = (\xi_1, 0, \ldots, 0, \xi_n)'$, $\xi_1 \in \Re$, $\xi_n \in \Re$, in fact two lines shown in the figure, for which steepest descent converges in a way that the inequality

$$\frac{f(x^{k+1})}{f(x^k)} \leq \left(\frac{M-m}{M+m}\right)^2$$

is satisfied as an equation at each iteration.

Multiplying with S, we obtain

$$Sy^{k+1} = Sy^k - \alpha^k S\nabla h(y^k).$$

By passing back to the space of x, using the relations

$$x^k = Sy^k, \qquad \nabla h(y^k) = S\nabla f(x^k), \qquad S^2 = D^k, \tag{1.44}$$

we obtain

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k).$$

Thus the above gradient iteration is nothing but the steepest descent method (1.43) in the space of y.

We now apply the convergence rate results for steepest descent to the scaled iteration $y^{k+1} = y^k - \alpha^k \nabla h(y^k)$, obtaining

$$\frac{\|y^{k+1}\|}{\|y^k\|} \leq \max\bigl\{|1-\alpha^k m^k|, |1-\alpha^k M^k|\bigr\}$$

and

$$\frac{h(y^{k+1})}{h(y^k)} \leq \left(\frac{M^k - m^k}{M^k + m^k}\right)^2,$$

[cf. the convergence rate bounds (1.38) and (1.40), respectively], where m^k and M^k are the smallest and largest eigenvalues of the Hessian $\nabla^2 h(y)$, which is equal to $S\nabla^2 f(x)S = (D^k)^{1/2}Q(D^k)^{1/2}$. Using the equations

$$y^k = (D^k)^{-1/2}x^k, \qquad y^{k+1} = (D^k)^{-1/2}x^{k+1}$$

to pass back to the space of x, we obtain the convergence rate bounds

$$\frac{x^{k+1\prime}(D^k)^{-1}x^{k+1}}{x^{k\prime}(D^k)^{-1}x^k} \leq \max\bigl\{(1-\alpha^k m^k)^2, (1-\alpha^k M^k)^2\bigr\} \tag{1.45}$$

and

$$\frac{f(x^{k+1})}{f(x^k)} \leq \left(\frac{M^k - m^k}{M^k + m^k}\right)^2, \tag{1.46}$$

where

$$m^k: \text{ smallest eigenvalue of } (D^k)^{1/2}Q(D^k)^{1/2},$$
$$M^k: \text{ largest eigenvalue of } (D^k)^{1/2}Q(D^k)^{1/2}.$$

The stepsize that minimizes the right-hand side bound of Eq. (1.45) is

$$\frac{2}{M^k + m^k}. \tag{1.47}$$

The important point is that if M^k/m^k is much larger than unity, the convergence rate can be very slow, even if an optimal stepsize is used. Furthermore, we see that it is desirable to choose D^k as close as possible to Q^{-1}, so that $(D^k)^{1/2}$ is close to $Q^{-1/2}$ (cf. Prop. A.21 in Appendix A) and $M^k \approx m^k \approx 1$. Note that if D^k is so chosen, Eq. (1.47) shows that the stepsize $\alpha = 1$ is near optimal.

Diagonal Scaling

Many practical problems are ill-conditioned because of poor relative scaling of the optimization variables. By this we mean that the units in which the variables are expressed are incongruent in the sense that single unit changes of different variables have disproportionate effects on the cost.

As an example, consider a financial problem with two variables, *investment* denoted x_1 and expressed in dollars, and *interest rate* denoted x_2 and expressed in percentage points. If the effect on the cost function f due to a million dollar increment of investment is comparable to the effect due to a percentage point increment of interest rate, then the condition number will be of the order of 10^{12}!! [This rough calculation is based on estimating the condition number by the ratio

$$\frac{\partial^2 f(x_1, x_2)}{(\partial x_2^2)} \Bigg/ \frac{\partial^2 f(x_1, x_2)}{(\partial x_1)^2},$$

approximating the second partial derivatives by the finite difference formulas

$$\frac{\partial^2 f(x_1, x_2)}{(\partial x_1)^2} \approx \frac{f(x_1 + h_1, x_2) + f(x_1 - h_1, x_2) - 2f(x_1, x_2)}{h_1^2},$$

$$\frac{\partial^2 f(x_1, x_2)}{(\partial x_2)^2} \approx \frac{f(x_1, x_2 + h_2) + f(x_1, x_2 - h_2) - 2f(x_1, x_2)}{h_2^2},$$

and using the relations $f(x_1 + h_1, x_2) \approx f(x_1, x_2 + h_2)$, $f(x_1 - h_1, x_2) \approx f(x_1, x_2 - h_2)$, and $h_1 = 10^6$, $h_2 = 1$, which express the comparability of the effects of a million dollar investment increment and an interest rate percentage point increment.]

The ill-conditioning in such problems can be significantly alleviated by changing the units in which the optimization variables are expressed, which amounts to diagonal scaling of the variables. By this, we mean working in a new coordinate system of a vector y related to x by a transformation,

$$x = Sy,$$

where S is a diagonal matrix. In the absence of further information, a reasonable choice of S is one that makes all the diagonal elements of the Hessian of the cost

$$S\nabla^2 f(x)S$$

in the y-coordinate system approximately equal to unity. For this, we must have

$$s_i \approx \left(\frac{\partial^2 f(x)}{(\partial x_i)^2}\right)^{-1/2},$$

where s_i is the ith diagonal element of S. As discussed earlier, we may express any gradient algorithm in the space of variables y as a gradient

algorithm in the space of variables x. In particular, steepest descent in the y-coordinate system, when translated in the x-coordinate system, yields the *diagonally scaled steepest descent method*

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

where

$$D^k = \begin{pmatrix} d_1^k & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & d_2^k & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & 0 & d_{n-1}^k & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & d_n^k \end{pmatrix},$$

and

$$d_i^k \approx \left(\frac{\partial^2 f(x^k)}{(\partial x_i)^2} \right)^{-1}.$$

This method is also valid for nonquadratic problems as long as d_i^k are chosen to be positive. It is not guaranteed to improve the convergence rate of steepest descent, but it is simple and often surprisingly effective in practice. In particular, it tends to automatically correct mismatches of the units in which the various optimization variables are expressed.

Nonquadratic Cost Functions

It is possible to show that our main conclusions on rate of convergence carry over to the nonquadratic case for sequences converging to nonsingular local minima. Conceptually, this makes sense because in the neighborhood of a nonsingular local minimum a twice continuously differentiable cost function is very close to a positive definite quadratic (up to second order). The technical details of the proofs of such results are straightforward, but tend to be uninsightful and tedious, and for the most part will be omitted.

More specifically, let f be twice continuously differentiable and consider the gradient method

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k), \tag{1.48}$$

where D^k is positive definite and symmetric. Consider a generated sequence $\{x^k\}$, and assume that

$$x^k \to x^*, \qquad \nabla f(x^*) = 0, \qquad \nabla^2 f(x^*) : \text{ positive definite}, \tag{1.49}$$

and that $x^k \neq x^*$ for all k. Then, denoting

$$m^k : \text{ smallest eigenvalue of } (D^k)^{1/2} \nabla^2 f(x^k) (D^k)^{1/2},$$

$$M^k : \text{ largest eigenvalue of } (D^k)^{1/2} \nabla^2 f(x^k) (D^k)^{1/2},$$

it is possible to show the following:

(a) There holds

$$\limsup_{k\to\infty}\frac{(x^{k+1}-x^*)'(D^k)^{-1}(x^{k+1}-x^*)}{(x^k-x^*)'(D^k)^{-1}(x^k-x^*)} = \limsup_{k\to\infty}\max\bigl\{|1-\alpha^k m^k|^2,|1-\alpha^k M^k|^2\bigr\}.$$

(b) If α^k is chosen by the line minimization rule, there holds

$$\limsup_{k\to\infty}\frac{f(x^{k+1})-f(x^*)}{f(x^k)-f(x^*)}\le\limsup_{k\to\infty}\left(\frac{M^k-m^k}{M^k+m^k}\right)^2. \tag{1.50}$$

An alternative result for the case of the Armijo rule is given in Exercise 1.3.10 (with solution posted online).

From Eq. (1.50), we see that if D^k converges to some positive definite matrix as $x^k\to x^*$, the sequence $\{f(x^k)\}$ converges to $f(x^*)$ linearly. When

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

we have $\lim_{k\to\infty}M^k=\lim_{k\to\infty}m^k=1$ and Eq. (1.50) shows that the convergence rate of $\{f(x^k)\}$ is superlinear. A somewhat more general version of this result for the case of the Armijo rule is given in Prop. 1.3.2 in the next subsection. In particular, it is shown that if the direction

$$d^k=-D^k\nabla f(x^k)$$

approaches asymptotically the Newton direction $-\bigl(\nabla^2 f(x^k)\bigr)^{-1}\nabla f(x^k)$ and the Armijo rule is used with initial stepsize equal to one, the rate of convergence is superlinear.

There is a consistent theme that emerges from our analysis, namely that to achieve asymptotically fast convergence of the gradient method

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

one should try to choose the matrices D^k as close as possible to $\bigl(\nabla^2 f(x^*)\bigr)^{-1}$ so that the maximum and minimum eigenvalues of $(D^k)^{1/2}\nabla^2 f(x^*)(D^k)^{1/2}$ satisfy $M^k\approx 1$ and $m^k\approx 1$. Furthermore, when D^k is so chosen, the initial stepsize $s=1$ is a good choice for the Armijo rule and other related rules, or as a starting point for one-dimensional minimization procedures used in minimization stepsize rules. This finding has been supported by extensive numerical experience and is one of the most reliable guidelines for selecting and designing optimization algorithms for unconstrained problems. Note, however, that this guideline is valid only for problems where the cost function is twice differentiable and has positive definite Hessian near the points of interest. We discuss next problems where this condition is not satisfied.

Singular and Difficult Problems

Let us consider problems where the Hessian matrix either does not exist or is not positive definite at or near local minima of interest. Expressed mathematically, there are local minima x^* and directions d such that the slope of f along d, which is $\nabla f(x^* + \alpha d)'d$, changes very slowly or very rapidly with α, i.e., either

$$\lim_{\alpha \to 0} \frac{\nabla f(x^* + \alpha d)'d - \nabla f(x^*)'d}{\alpha} = 0, \tag{1.51}$$

or

$$\lim_{\alpha \to 0} \frac{\nabla f(x^* + \alpha d)'d - \nabla f(x^*)'d}{\alpha} = \infty. \tag{1.52}$$

The case of Eq. (1.51) is characterized by flatness of the cost along the direction d; large excursions from x^* along d produce small changes in cost. In the case of Eq. (1.52) the reverse is true; the cost rises steeply along d. An example is the function

$$f(x_1, x_2) = |x_1|^4 + |x_2|^{3/2},$$

where for the minimum $x^* = (0, 0)$, Eq. (1.51) holds along the direction $d = (1, 0)$ and Eq. (1.52) holds along the direction $d = (0, 1)$. Gradient methods that use directions that are comparable in size to the gradient may require very large stepsizes in the case of Eq. (1.51) and very small stepsizes in the case of Eq. (1.52). This suggests potential difficulties in the implementation of a good stepsize rule; certainly a constant stepsize does not look like an attractive possibility. Furthermore, in the Armijo rule, the initial stepsize should not be taken constant; it should be adjusted according to a suitable scheme, although designing such a scheme may not be easy.

One may view the cases of Eqs. (1.51) and (1.52) as corresponding to an "infinite condition number," thereby suggesting slower than linear convergence rate for the method of steepest descent. Proposition 1.3.3 of the next subsection quantifies the rate of convergence of gradient methods for the case of a convex function whose gradient satisfies the Lipschitz condition

$$\big\|\nabla f(x) - \nabla f(y)\big\| \leq L\|x - y\|, \tag{1.53}$$

for some L, and all x and y in a neighborhood of x^* [this assumption is consistent with the "flat" cost case of Eq. (1.51), but not with the "steep" cost case of Eq. (1.52)]. It is shown in particular that for a gradient method with several types of stepsize rules, we have

$$f(x^k) - f(x^*) = o(1/k).$$

This type of estimate suggests that for many practical singular problems one may be unable to obtain a highly accurate approximation of an optimal

solution. In the "steep" cost case where Eq. (1.52) holds for some directions d, computational examples suggest that the rate of convergence can be slower than linear for the method of steepest descent, although a formal analysis of this conjecture does not seem to have been published.

It should be noted that problems with singular local minima are not the only ones for which gradient methods may converge slowly. There are problems where a given method may have excellent asymptotic rate of convergence, but its progress when far from the eventual limit can be very slow. A prominent example is when the cost function is continuously differentiable but its Hessian matrix is discontinuous and possibly singular in some regions that are outside a small neighborhood of the solution; such functions arise for example in augmented Lagrangian methods for inequality constrained problems (see Section 5.2). Then the powerful Newton-like methods may require a very large number of iterations to get to the small neighborhood of the eventual limit where their convergence rate is favorable. What happens here is that these methods use second derivative information in sophisticated ways, but this information may be misleading due to the Hessian discontinuities.

Generally, there is a tendency to think that difficult problems should be addressed with sophisticated methods, such as Newton-like methods. This is often true, particularly for problems with nonsingular local minima that are poorly conditioned. However, it is important to realize that *often the reverse is true*, namely that for problems with "difficult" cost functions and singular local minima, it is best to use simple methods such as (perhaps diagonally scaled) steepest descent with simple stepsize rules such as a constant or a diminishing stepsize. The reason is that methods that use sophisticated descent directions and stepsize rules often rely on assumptions that are likely to be violated in difficult problems. We also note that for difficult problems, it may be helpful to supplement the steepest descent method with features that allow it to deal better with multiple local minima and peculiarities of the cost function. An often useful modification is to introduce extrapolation based on the preceding two iterates, which we discuss next.

Steepest Descent with Extrapolation

A variant of the steepest descent method, known as *gradient method with momentum*, involves extrapolation along the direction of the difference of the preceding two iterates:

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) + \beta^k (x^k - x^{k-1}), \tag{1.54}$$

where β^k is a scalar in $[0, 1)$, and we define $x_{-1} = x_0$. When α^k and β^k are chosen to be constant scalars α and β, respectively, the method is known as the *heavy ball method* [Pol64]; see Fig. 1.3.3. This is a sound method with

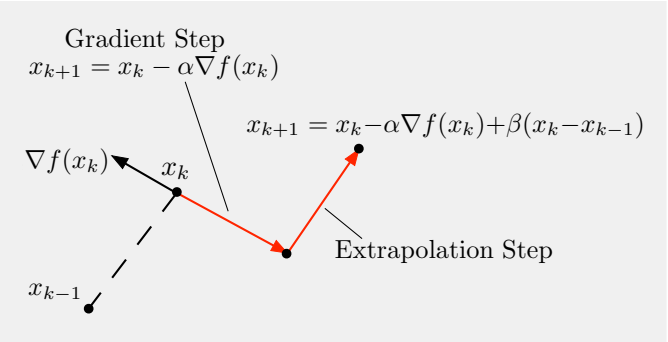

Figure 1.3.3. Illustration of the heavy ball method (1.54), where $\alpha^k \equiv \alpha$ and $\beta^k \equiv \beta$.

guaranteed convergence under a Lipschitz continuity assumption on ∇f. It can be shown to have faster convergence rate than the corresponding gradient method where $\alpha^k \equiv \alpha$ and $\beta^k \equiv 0$. In particular, for a positive definite quadratic problem with minimum at x^*, and with optimal choices of the constants α and β, the convergence rate of the heavy ball method is linear, and is governed by the formula

$$\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} \leq \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}; \tag{1.55}$$

(see Exercise 1.3.8, whose solution is posted online; also see [GFJ15] for a related convergence analysis). This formula has the same form as the one for the steepest descent method, but with M/m replaced by $\sqrt{M/m}$, which is a substantial improvement. Simple examples also suggest that with extrapolation, the steepest descent method is less prone to getting trapped at "shallow" local minima, and deals better with cost functions that are alternately very flat and very steep along the path of the algorithm.

A method with similar structure as (1.54), proposed in [Nes83] and often called *Nesterov's method*, has received a lot of attention because it has theoretically interesting complexity properties. The iteration of this method is commonly described in two steps: first an extrapolation step, to compute

$$y^k = x^k + \beta^k (x^k - x^{k-1}) \tag{1.56}$$

with β^k chosen in a special way so that $\beta^k \to 1$, and then a gradient step with constant stepsize α, and gradient calculated at y^k,

$$x^{k+1} = y^k - \alpha \nabla f(y^k). \tag{1.57}$$

Compared to the method (1.54), it reverses the order of gradient calculation and extrapolation, and uses $\nabla f(y^k)$ in place of $\nabla f(x^k)$ (in addition to

using time-varying parameters β^k). Assuming only convexity of f and Lipschitz continuity of ∇f, the convergence rate of the method (1.56)-(1.57) is sublinear, but of better order than the one of the method (1.54). For a positive definite quadratic cost function, the convergence rate of both methods is linear and roughly comparable (with optimal choices of parameters). We refer to [Nes04], Section 2.2.1, for an analysis and discussion of the method (1.56)-(1.57); also [Ber15a], Section 6.2, and the references quoted there, including the paper [Tse08], which describes some extensions.

We finally note that extrapolation can also be used in the context of two implementations of the conjugate gradient method, which have superlinear convergence rate. These implementations are described in Section 2.1; see Exercises 2.1.5 and 2.1.6 (with solutions posted online).

1.3.3 Convergence Rate Results

We first derive the convergence rate of steepest descent with the minimization stepsize rule when the cost is quadratic.

Proposition 1.3.1: Consider the quadratic function

$$f(x) = \tfrac{1}{2} x' Q x, \tag{1.58}$$

where Q is positive definite and symmetric, and the method of steepest descent

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k), \tag{1.59}$$

where the stepsize α^k is chosen according to the minimization rule

$$f\big(x^k - \alpha^k \nabla f(x^k)\big) = \min_{\alpha \geq 0} f\big(x^k - \alpha \nabla f(x^k)\big).$$

Then, for all k,

$$f(x^{k+1}) \leq \left(\frac{M-m}{M+m}\right)^2 f(x^k),$$

where M and m are the largest and smallest eigenvalues of Q, respectively.

Proof: Let us denote

$$g^k = \nabla f(x^k) = Q x^k. \tag{1.60}$$

The result clearly holds if $g^k = 0$, so we assume $g^k \neq 0$. We first compute the minimizing stepsize α^k. We have

$$\frac{d}{d\alpha} f(x^k - \alpha g^k) = -g^{k\prime} Q (x^k - \alpha g^k) = -g^{k\prime} g^k + \alpha g^{k\prime} Q g^k.$$

By setting this derivative equal to zero, we obtain

$$\alpha^k = \frac{g^{k\prime} g^k}{g^{k\prime} Q g^k}. \tag{1.61}$$

We have, using Eqs. (1.58)-(1.60),

$$\begin{aligned} f(x^{k+1}) &= \tfrac{1}{2}(x^k - \alpha^k g^k)' Q (x^k - \alpha^k g^k) \\ &= \tfrac{1}{2}\left(x^{k\prime} Q x^k - 2\alpha^k g^{k\prime} Q x^k + (\alpha^k)^2 g^{k\prime} Q g^k\right) \\ &= \tfrac{1}{2}\left(x^{k\prime} Q x^k - 2\alpha^k g^{k\prime} g^k + (\alpha^k)^2 g^{k\prime} Q g^k\right) \end{aligned}$$

and using Eq. (1.61),

$$f(x^{k+1}) = \tfrac{1}{2}\left(x^{k\prime} Q x^k - \tfrac{(g^{k\prime} g^k)^2}{g^{k\prime} Q g^k}\right).$$

Thus, using the fact $f(x^k) = \frac{1}{2} x^{k\prime} Q x^k = \frac{1}{2} g^{k\prime} Q^{-1} g^k$, we obtain

$$f(x^{k+1}) = \left(1 - \frac{(g^{k\prime} g^k)^2}{(g^{k\prime} Q g^k)(g^{k\prime} Q^{-1} g^k)}\right) f(x^k). \tag{1.62}$$

At this point we need the following lemma.

Lemma 3.1: (Kantorovich Inequality) Let Q be a positive definite and symmetric $n \times n$ matrix. Then for any vector $y \in \Re^n, y \neq 0$, there holds

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

where M and m are the largest and smallest eigenvalues of Q, respectively.

Proof: Let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of Q and assume that

$$0 < m = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = M.$$

Let S be the matrix consisting of the n orthogonal eigenvectors of Q, normalized so that they have unit norm (cf. Prop. A.17 in Appendix A). Then, it can be seen that $S'QS$ is diagonal with diagonal elements $\lambda_1, \ldots, \lambda_n$. By using if necessary a transformation of the coordinate system that replaces y by Sx, we may assume that Q is diagonal and that its diagonal elements are $\lambda_1, \ldots, \lambda_n$. We have for $y = (y_1, \ldots, y_n)' \neq 0$

$$\frac{(y'y)^2}{(y'Qy)(y'Q^{-1}y)} = \frac{\left(\sum_{i=1}^n y_i^2\right)^2}{\left(\sum_{i=1}^n \lambda_i y_i^2\right)\left(\sum_{i=1}^n \frac{y_i^2}{\lambda_i}\right)}.$$

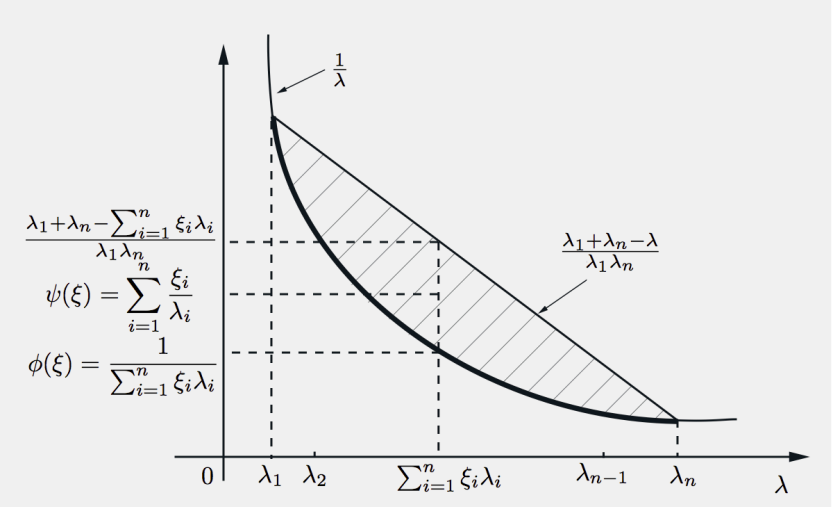

Figure 1.3.4. Proof of the Kantorovich inequality. Consider the function $1/\lambda$. The scalar $\sum_{i=1}^n \xi_i\lambda_i$ represents, for any $\xi = (\xi_1, \ldots, \xi_n)$ with $\xi_i \geq 0$, $\sum_{i=1}^n \xi_i = 1$, a point in the line segment $[\lambda_1, \lambda_n]$. Thus, the values $\phi(\xi) = 1/\sum_{i=1}^n \xi_i\lambda_i$ correspond to the thick part of the curve $1/\lambda$. On the other hand, the value $\psi(\xi) = \sum_{i=1}^n (\xi_i/\lambda_i)$ is a convex combination of $1/\lambda_1, \ldots, 1/\lambda_n$ and hence corresponds to a point in the shaded area in the figure. For the same vector ξ, both $\phi(\xi)$ and $\psi(\xi)$ are represented by points on the same vertical line. Hence,

$$\frac{\phi(\xi)}{\psi(\xi)} \geq \min_{\lambda_1 \leq \lambda \leq \lambda_n} \frac{\frac{1}{\lambda}}{\frac{\lambda_1+\lambda_n-\lambda}{\lambda_1\lambda_n}}.$$

The minimum is attained for $\lambda = (\lambda_1 + \lambda_n)/2$ and we obtain

$$\frac{\phi(\xi)}{\psi(\xi)} \geq \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2},$$

which is used to show the result.

By letting

$$\xi_j = \frac{y_j^2}{\sum_{i=1}^n y_i^2}$$

and by defining

$$\phi(\xi) = \frac{1}{\sum_{i=1}^n \xi_i\lambda_i}, \qquad \psi(\xi) = \sum_{i=1}^n \frac{\xi_i}{\lambda_i},$$

we obtain

$$\frac{(y'y)^2}{(y'Qy)(y'Q^{-1}y)} = \frac{\phi(\xi)}{\psi(\xi)}.$$

Figure 1.3.4 shows that we have

$$\frac{\phi(\xi)}{\psi(\xi)} \geq \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2},$$

which proves the desired inequality. **Q.E.D.**

Returning to the proof of Prop. 1.3.1, we have by using the Kantorovich inequality in Eq. (1.62)

$$f(x^{k+1}) \leq \left(1 - \frac{4Mm}{(M+m)^2}\right) f(x^k) = \left(\frac{M-m}{M+m}\right)^2 f(x^k).$$

Q.E.D.

The following proposition shows superlinear convergence for methods where d^k approaches the Newton direction $-\big(\nabla^2 f(x^*)\big)^{-1}\nabla f(x^k)$ and the Armijo rule is used.

Proposition 1.3.2: (Superlinear Convergence of Newton-Like Methods) Let f be twice continuously differentiable. Consider a sequence $\{x^k\}$ generated by the gradient method $x^{k+1} = x^k + \alpha^k d^k$ and suppose that

$$x^k \to x^*, \qquad \nabla f(x^*) = 0, \qquad \nabla^2 f(x^*) : \text{ positive definite.} \quad (1.63)$$

Assume further that $\nabla f(x^k) \neq 0$ for all k and

$$\lim_{k\to\infty} \frac{\big\|d^k + \big(\nabla^2 f(x^*)\big)^{-1}\nabla f(x^k)\big\|}{\big\|\nabla f(x^k)\big\|} = 0. \quad (1.64)$$

Then, if α^k is chosen by means of the Armijo rule with initial stepsize $s = 1$ and $\sigma < 1/2$, we have

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0. \quad (1.65)$$

Furthermore, there exists an integer $\bar{k} \geq 0$ such that $\alpha^k = 1$ for all $k \geq \bar{k}$ (i.e., eventually no reduction of the initial stepsize will be taking place).

Proof: We first prove that there exists a $\bar{k} \geq 0$ such that for all $k \geq \bar{k}$,

$$f(x^k + d^k) - f(x^k) \leq \sigma \nabla f(x^k)' d^k,$$

i.e., the unity initial stepsize passes the test of the Armijo rule. By the mean value theorem, we have

$$f(x^k + d^k) - f(x^k) = \nabla f(x^k)' d^k + \tfrac{1}{2} d^{k\prime} \nabla^2 f(\bar{x}^k) d^k,$$

where $\bar{x}^k$ is a point on the line segment joining x^k and $x^k + d^k$. Thus, it will be sufficient to show that for k sufficiently large, we have

$$\nabla f(x^k)' d^k + \tfrac{1}{2} d^k \nabla^2 f(\bar{x}^k) d^k \leq \sigma \nabla f(x^k)' d^k,$$

or equivalently,

$$(1 - \sigma) p^{k\prime} q^k + \tfrac{1}{2} q^{k\prime} \nabla^2 f(\bar{x}^k) q^k \leq 0, \tag{1.66}$$

where

$$p^k = \frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}, \qquad q^k = \frac{d^k}{\|\nabla f(x^k)\|}.$$

Indeed, from the assumption (1.64), we have

$$q^k + \big(\nabla^2 f(x^*)\big)^{-1} p^k \to 0.$$

Since $\nabla^2 f(x^*)$ is positive definite and $\|p^k\| = 1$, it follows that $\{q^k\}$ is a bounded sequence, and in view of $q^k = d^k / \|\nabla f(x^k)\|$ and $\nabla f(x^k) \to 0$, we obtain $d^k \to 0$. Hence, $x^k + d^k \to x^*$, and it follows that $\bar{x}^k \to x^*$ and $\nabla^2 f(\bar{x}^k) \to \nabla^2 f(x^*)$. We now write Eq. (1.64) as

$$q^k = -\big(\nabla^2 f(x^*)\big)^{-1} p^k + \beta^k,$$

where $\{\beta^k\}$ denotes a vector sequence with $\beta^k \to 0$. By using the above relation and the fact $\nabla^2 f(\bar{x}^k) \to \nabla^2 f(x^*)$, we may write Eq. (1.66) as

$$(1 - \sigma) p^{k\prime} \big(\nabla^2 f(x^*)\big)^{-1} p^k - \tfrac{1}{2} p^{k\prime} \big(\nabla^2 f(x^*)\big)^{-1} p^k \geq \gamma^k,$$

where $\{\gamma^k\}$ is some scalar sequence with $\gamma^k \to 0$. Thus Eq. (1.66) is equivalent to

$$\big(\tfrac{1}{2} - \sigma\big)\, p^{k\prime} \big(\nabla^2 f(x^*)\big)^{-1} p^k \geq \gamma^k.$$

Since $1/2 > \sigma$, $\|p^k\| = 1$, and $\nabla^2 f(x^*)$ is positive definite, the above relation holds for sufficiently large k. Thus, Eq. (1.66) holds, and it follows that the unity initial stepsize is acceptable for sufficiently large k.

To complete the proof, we note that from Eq. (1.64), we have

$$d^k + \big(\nabla^2 f(x^*)\big)^{-1} \nabla f(x^k) = \|\nabla f(x^k)\| \delta^k, \tag{1.67}$$

where δ^k is some vector sequence with $\delta^k \to 0$. We have

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

from which

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

$$\big\|\nabla f(x^k)\big\| = O\big(\|x^k - x^*\|\big).$$

Using the above two relations in Eq. (1.67), we obtain

$$d^k + x^k - x^* = o\big(\|x^k - x^*\|\big). \tag{1.68}$$

Since for sufficiently large k we have $d^k + x^k = x^{k+1}$, Eq. (1.68) yields

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

from which

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = \lim_{k\to\infty} \frac{o\big(\|x^k - x^*\|\big)}{\|x^k - x^*\|} = 0.$$

Q.E.D.

Note that the equation

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

[cf. Eq. (1.65)] implies that $\big\{\|x^k - x^*\|\big\}$ converges superlinearly (see Exercise 1.3.6). In particular, we see that Newton's method, combined with the Armijo rule with unity initial stepsize, has the property that when it converges to a local minimum x^* such that $\nabla^2 f(x^*)$ is positive definite, its rate of convergence is superlinear. The capture theorem (Prop. 1.2.3) together with the preceding proposition suggest that Newton-like methods with the Armijo rule and a unity initial stepsize converge to a local minimum x^* such that $\nabla^2 f(x^*)$ is positive definite, whenever they are started sufficiently close to such a local minimum. The proof of this is left as Exercise 1.3.4 for the reader (solution posted online).

We finally consider the convergence rate of gradient methods for singular problems with a convex cost function, and a stepsize that is either constant within an appropriate range (cf. Prop. 1.2.2), or is obtained by line minimization.

Proposition 1.3.3: (Convergence Rate of Gradient Methods for Singular Problems) Suppose that the cost function f is convex and its gradient satisfies for some L the Lipschitz condition

$$\big\|\nabla f(x) - \nabla f(y)\big\| \le L\|x - y\|, \qquad \forall\ x, y \in \Re^n. \tag{1.69}$$

Assume further that the set of global minima of f is nonempty and bounded. Consider a gradient method $x^{k+1} = x^k + \alpha^k d^k$ where for some $c > 0$ and all k we have

$$\nabla f(x^k)' d^k \leq -c\big\|\nabla f(x^k)\big\|^2, \qquad d^k \neq 0, \tag{1.70}$$

while α^k either satisfies for some $\epsilon \in (0, 1]$ and all k

$$\epsilon \leq \alpha^k \leq (2 - \epsilon)\bar{\alpha}^k, \tag{1.71}$$

where

$$\bar{\alpha}^k = \frac{\big|\nabla f(x^k)' d^k\big|}{L\|d^k\|^2},$$

or else satisfies $f(x^k + \alpha^k d^k) \leq f(x^k + \bar{\alpha}^k d^k)$. Then all limit points of $\{x^k\}$ are optimal and there exists at least one limit point. Moreover

$$f(x^k) - f^* = o(1/k),$$

where $f^* = \min_{x \in \Re^n} f(x)$ is the optimal value.

Proof: We assume that $\nabla f(x^k) \neq 0$; otherwise the method terminates finitely at a global minimum and the result holds trivially. Assume first that α^k is chosen by the rule (1.71). Then from the proof of Prop. 1.2.2, we have

$$f(x^k + \alpha^k d^k) - f(x^k) \leq -\tfrac{1}{2}\epsilon^2\big|\nabla f(x^k)' d^k\big|.$$

Combining this relation with Eq. (1.70), we obtain

$$f(x^k + \alpha^k d^k) \leq f(x^k) - \frac{c\epsilon^2}{2}\big\|\nabla f(x^k)\big\|^2. \tag{1.72}$$

The above relation holds also for $\alpha^k = \bar{\alpha}^k$, so that

$$f(x^k + \bar{\alpha}^k d^k) \leq f(x^k) - \frac{c\epsilon^2}{2}\big\|\nabla f(x^k)\big\|^2.$$

Thus for any of the possible stepsizes chosen by the algorithm, we have

$$f(x^{k+1}) \leq f(x^k) - \frac{c\epsilon^2}{2}\big\|\nabla f(x^k)\big\|^2. \tag{1.73}$$

Let X^* be the set of global minima of f. Since X^* is nonempty and compact, all the level sets of f are compact (Prop. B.10 in Appendix B). This together with the monotone decrease of $\big\{f(x^k)\big\}$, shows that $\{x^k\}$ is

bounded. Hence, by Prop. 1.2.1, all limit points of $\{x^k\}$ belong to X^*, and the distance of x^k from X^*, defined by

$$d(x^k, X^*) = \min_{x^* \in X^*} \|x^k - x^*\|,$$

converges to 0 and $e^k \to 0$. Using the convexity of f, we also have for every global minimum x^*

$$f(x^k) - f(x^*) \leq \nabla f(x^k)'(x^k - x^*) \leq \left\|\nabla f(x^k)\right\| \cdot \|x^k - x^*\|,$$

from which, by minimizing over $x^* \in X^*$,

$$f(x^k) - f^* \leq \left\|\nabla f(x^k)\right\| d(x^k, X^*). \tag{1.74}$$

Let us denote for all k

$$e^k = f(x^k) - f^*.$$

Combining Eqs. (1.73) and (1.74), we obtain

$$e^{k+1} \leq e^k - \frac{c\epsilon^2 (e^k)^2}{2d(x^k, X^*)^2}, \qquad \forall\ k, \tag{1.75}$$

where we assume without loss of generality that $d(x^k, X^*) \neq 0$.

We will show that Eq. (1.75) implies that $e^k = o(1/k)$. Indeed we have

$$0 < e^{k+1} \leq e^k \left(1 - \frac{c\epsilon^2 e^k}{2d(x^k, X^*)^2}\right),$$

$$0 < 1 - \frac{c\epsilon^2 e^k}{2d(x^k, X^*)^2},$$

from which

$$\begin{aligned}(e^{k+1})^{-1} &\geq (e^k)^{-1} \left(1 - \frac{c\epsilon^2 e^k}{2d(x^k, X^*)^2}\right)^{-1} \geq (e^k)^{-1} \left(1 + \frac{c\epsilon^2 e^k}{2d(x^k, X^*)^2}\right) \\ &= (e^k)^{-1} + \frac{c\epsilon^2}{2d(x^k, X^*)^2}.\end{aligned}$$

Summing this inequality over all k, we obtain

$$e^k \leq \left((e^0)^{-1} + \frac{c\epsilon^2}{2} \sum_{i=0}^{k-1} d(x^i, X^*)^{-2}\right)^{-1},$$

or

$$ke^k \leq \left(\frac{1}{ke^0} + \frac{c\epsilon^2}{2k} \sum_{i=0}^{k-1} d(x^i, X^*)^{-2}\right)^{-1}. \tag{1.76}$$

Since $d(x^i, X^*) \to 0$, we have $d(x^i, X^*)^{-2} \to \infty$ and

$$\frac{c\epsilon^2}{2k} \sum_{i=0}^{k-1} d(x^i, X^*)^{-2} \to \infty.$$

Therefore the right-hand side of Eq. (1.76) tends to 0, implying that $e^k = o(1/k)$. **Q.E.D.**

The key step in the preceding proof is that the stepsize rule is such that Eq. (1.73) holds. Indeed the proof goes through for any stepsize rule for which we have

$$f(x^{k+1}) \leq f(x^k) - \beta \left\| \nabla f(x^k) \right\|^2$$

for some constant $\beta > 0$ and all k. The proof can also be modified for the case where the Lipschitz condition (1.69) holds within the level set $\left\{ x \mid f(x) \leq f(x^0) \right\}$. Moreover, with additional assumptions on the structure of f, some more precise convergence rate results can be obtained. In particular, if f is convex, has a unique minimum x^*, and satisfies the following growth condition

$$f(x) - f(x^*) \geq q \|x - x^*\|^\beta, \qquad \forall\ x \text{ such that } f(x) \leq f(x^0),$$

for some scalars $q > 0$ and $\beta > 2$, it can be shown (see [Dun81]) that for the method of steepest descent with the Armijo rule we have

$$f(x^k) - f(x^*) = O\left(\frac{1}{k^{\frac{\beta}{\beta-2}}} \right).$$

E X E R C I S E S

1.3.1

Estimate the rate of convergence of steepest descent with the line minimization rule when applied to the function of two variables $f(x, y) = x^2 + 1.999xy + y^2$. Find a starting point for which this estimate is sharp (cf. Fig. 1.3.2).

1.3.2

Consider a positive definite quadratic problem with Hessian matrix Q. Suppose we use scaling with the diagonal matrix whose ith diagonal element is q_{ii}^{-1}, where q_{ii} is the ith diagonal element of Q. Show that if Q is 2×2, this diagonal scaling improves the condition number of the problem and the convergence rate of steepest descent. (*Note*: This need not be true for dimensions higher than 2.)

1.3.3 (Linear Convergence under Strong Convexity)

Let f be differentiable and satisfy the Lipschitz condition

$$\big\|\nabla f(x) - \nabla f(y)\big\| \leq L\|x-y\|, \qquad \forall\, x,y \in \Re^n.$$

Assume further that f is strongly convex, i.e., for some $\sigma > 0$, we have

$$\big(\nabla f(x) - \nabla f(y)\big)'(x-y) \geq \sigma\|x-y\|^2, \qquad \forall\, x,y \in \Re^n,$$

and let x^* be the unique minimum of f.

(a) Show that the mapping $G_\alpha(x) = x - \alpha\nabla f(x)$ of the steepest descent iteration with constant stepsize α satisfies

$$\big\|G_\alpha(x) - G_\alpha(y)\big\| \leq \max\big\{|1-\alpha L|,\, |1-\alpha\sigma|\big\}\,\|x-y\|, \qquad \forall\, x,y, \in \Re^n,$$

and is a contraction for all $\alpha \in (0, 2/L)$. *Abbreviated Proof*: For all $x, y \in \Re^n$, we have

$$\big\|G_\alpha(x) - G_\alpha(y)\big\|^2 = \big\|\big(x-\alpha\nabla f(x)\big) - \big(y - \alpha\nabla f(y)\big)\big\|^2.$$

Expanding the quadratic on the right-hand side, and using Prop. B.5(a), the Lipschitz condition, and the strong convexity condition, we obtain

$$\begin{aligned}
\big\|G_\alpha(x)-G_\alpha(y)\big\|^2 &\leq \|x-y\|^2 - 2\alpha\big(\nabla f(x) - \nabla f(y)\big)'(x-y) \\
&\qquad + \alpha^2\big\|\nabla f(x) - \nabla f(y)\big\|^2 \\
&\leq \|x-y\|^2 - \frac{2\alpha\sigma L}{\sigma+L}\|x-y\|^2 - \frac{2\alpha}{\sigma+L}\big\|\nabla f(x) - \nabla f(y)\big\|^2 \\
&\qquad + \alpha^2\big\|\nabla f(x) - \nabla f(y)\big\|^2 \\
&= \left(1 - \frac{2\alpha\sigma L}{\sigma+L}\right)\|x-y\|^2 \\
&\quad + \alpha\left(\alpha - \frac{2}{\sigma+L}\right)\big\|\nabla f(x) - \nabla f(y)\big\|^2 \\
&\leq \left(1 - \frac{2\alpha\sigma L}{\sigma+L}\right)\|x-y\|^2 \\
&\quad + \alpha\max\left\{L^2\left(\alpha - \frac{2}{\sigma+L}\right), \sigma^2\left(\alpha - \frac{2}{\sigma+L}\right)\right\}\|x-y\|^2 \\
&= \max\big\{(1-\alpha L)^2, (1-\alpha\sigma)^2\big\}\|x-y\|^2,
\end{aligned}$$

from which the desired inequality follows.

(b) Use part (a) to show that for the steepest descent method $x^{k+1} = G_\alpha(x^k)$ we have

$$\|x^{k+1} - x^*\| \leq \max\big\{|1-\alpha\sigma|, |1-\alpha L|\big\}\|x^k - x^*\|,$$

and that this relation generalizes the estimate of Eq. (1.38). From this relation, argue that the ratio L/σ plays the role of the condition number, which we have defined for twice differentiable f.

1.3.4 (Superlinear Convergence) www

Let f be twice continuously differentiable. Consider a sequence $\{x^k\}$ generated by the gradient method $x^{k+1} = x^k + \alpha^k d^k$ and suppose that x^* is a nonsingular local minimum. Assume that, for all k, $\nabla f(x^k) \neq 0$ and $d^k = d(x^k)$, where $d(\cdot)$ is a continuous function of x with

$$\lim_{x \to x^*,\, \nabla f(x) \neq 0} \frac{\big\| d(x) + \big(\nabla^2 f(x)\big)^{-1} \nabla f(x) \big\|}{\big\| \nabla f(x) \big\|} = 0.$$

Furthermore, α^k is chosen by means of the Armijo rule with initial stepsize $s = 1$ and $\sigma < 1/2$. Show that there exists an $\epsilon > 0$ such that if $\|x^0 - x^*\| < \epsilon$, then:

(a) $\{x^k\}$ converges to x^*.

(b) $\alpha^k = 1$ for all k.

(c) $\lim_{k \to \infty} \big(\|x^{k+1} - x^*\| / \|x^k - x^*\| \big) = 0$.

Hint: Use the line of argument of Prop. 1.3.2 together with the capture theorem (Prop. 1.2.3). Alternatively, instead of using the capture theorem, consult the proof of the subsequent Prop. 1.4.1.

1.3.5 (Steepest Descent with Errors)

Consider the steepest descent method

$$x^{k+1} = x^k - \alpha \big(\nabla f(x^k) + e^k \big),$$

where α is a constant stepsize, e^k is an error satisfying $\|e^k\| \leq \delta$ for all k, and f is the positive definite quadratic function

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

Let

$$q = \max\big\{ |1 - \alpha m|, |1 - \alpha M| \big\},$$

where

$$m: \text{ smallest eigenvalue of } Q, \qquad M: \text{ largest eigenvalue of } Q,$$

and assume that $q < 1$. Show that for all k, we have

$$\|x^k - x^*\| \leq \frac{\alpha \delta}{1 - q} + q^k \|x^0 - x^*\|.$$

1.3.6 (Convergence Rate Characterizations [Ber82a], p. 14)

Consider a scalar sequence $\{e^k\}$ with $e^k \geq 0$ for all k, and $e^k \to 0$. We say that $\{e^k\}$ converges *faster than linearly with convergence ratio* β, where $0 < \beta < 1$, if for every $\bar{\beta} \in (\beta, 1)$ and $q > 0$, there exists $\bar{k}$ such that

$$e^k \leq q\bar{\beta}^k, \qquad \forall\ k \geq \bar{k}.$$

We say that $\{e^k\}$ converges *slower than linearly with convergence ratio* β, where $0 < \beta < 1$, if for every $\bar{\beta} \in (\beta, 1)$ and $q > 0$, there exists $\bar{k}$ such that

$$q\bar{\beta}^k \leq e^k, \qquad \forall\ k \geq \bar{k}.$$

We say that $\{e^k\}$ converges *linearly with convergence ratio* β if it converges both faster and slower than linearly with convergence ratio β. Show that:

(a) $\{e^k\}$ converges faster than linearly with convergence ratio β if and only if

$$\limsup_{k\to\infty} (e^k)^{1/k} \leq \beta.$$

$\{e^k\}$ converges slower than linearly with convergence ratio β if and only if

$$\liminf_{k\to\infty} (e^k)^{1/k} \geq \beta.$$

$\{e^k\}$ converges linearly with convergence ratio β if and only if

$$\lim_{k\to\infty} (e^k)^{1/k} = \beta.$$

(b) Assume that $e^k \neq 0$ for all k, and denote

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

Show that if $0 < \beta_1 \leq \beta_2 < 1$, then $\{e^k\}$ converges faster than linearly with convergence ratio β_2 and slower than linearly with convergence ratio β_1. Furthermore, if $\beta_1 = \beta_2 = 0$, then $\{e^k\}$ converges superlinearly.

1.3.7

Consider a scalar sequence $\{e^k\}$ with $e^k > 0$ for all k, and $e^k \to 0$. Show that $\{e^k\}$ converges superlinearly with order p if

$$\limsup_{k\to\infty} \frac{e^{k+1}}{(e^k)^p} < \infty.$$

1.3.8 (The Heavy Ball Method [Pol64]) (www)

Consider the following variant of the steepest descent method:

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

where α is a constant positive stepsize and β is a scalar with $0 < \beta < 1$.

(a) Let f be the quadratic function $f(x) = (1/2)x'Qx + c'x$, where Q is positive definite and symmetric, and let m and M be the minimum and the maximum eigenvalues of Q, respectively. Show that the method converges linearly to the unique solution if $0 < \alpha < 2(1 + \beta)/M$. Show that with optimal choices of α and β, the ratio of linear convergence for the sequence $\{\|x^k - x^*\|\}$ is

$$\frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}};$$

cf. Eq. (1.55). [This is faster than the corresponding ratio of the steepest descent method where $\beta = 0$ and α is chosen optimally; cf. Eq. (1.39).] *Hint*: Write the iteration as

$$\begin{pmatrix} x^{k+1} \\ x^k \end{pmatrix} = \begin{pmatrix} (1+\beta)I - \alpha Q & -\beta I \\ I & 0 \end{pmatrix} \begin{pmatrix} x^k \\ x^{k-1} \end{pmatrix}$$

and show that v is an eigenvalue of the matrix in the above equation if and only if $v + \beta/v$ is equal to $1 + \beta - \alpha\lambda$ where λ is an eigenvalue of Q. (This is a challenging exercise.)

(b) It is generally conjectured that in comparison to steepest descent, the method is less prone to getting trapped at "shallow" local minima, and tends to behave better for difficult problems where the cost function is alternatively very flat and very steep. Argue for or against this conjecture.

(c) In support of your answer in (b), write a computer program to test the method with $\beta = 0$ and $\beta > 0$ with one-dimensional cost functions of the form

$$f(x) = \tfrac{1}{2}x^2\big(1 + \gamma\cos(x)\big),$$

where $\gamma \in (0, 1)$, and

$$f(x) = \tfrac{1}{2}\sum_{i=1}^m \big|z_i - \tanh(xy_i)\big|^2,$$

where z_i and y_i are given scalars.

1.3.9 (www)

Suppose that a vector sequence $\{e^k\}$ satisfies

$$\|e^{k+1} - e^k\| \leq \beta\|e^k - e^{k-1}\|, \qquad \forall\ k \geq \bar{k},$$

where $\bar{k}$ is a positive integer and $\beta \in (0, 1)$ is a scalar. Show that $\{e^k\}$ converges to some vector e^* linearly, and in fact we have

$$\|e^k - e^*\| \leq q\beta^k$$

for some scalar q and all k. *Hint*: Show that $\{e^k\}$ is a Cauchy sequence.

1.3.10 (Convergence Rate of Steepest Descent with the Armijo Rule) www

Let $f : \Re^n \mapsto \Re$ be a twice continuously differentiable function that satisfies

$$m\|y\|^2 \leq y'\nabla^2 f(x)y \leq M\|y\|^2, \qquad \forall\ x,\ y \in \Re^n,$$

where m and M are some positive scalars. Consider the steepest descent method $x^{k+1} = x^k - \alpha^k \nabla f(x^k)$ with α^k determined by the Armijo rule. Let x^* be the unique unconstrained minimum of f and let

$$r = 1 - \frac{4m\beta\sigma(1-\sigma)}{M}.$$

Show that for all k, we have

$$f(x^{k+1}) - f(x^*) \leq r\big(f(x^k) - f(x^*)\big),$$

and

$$\|x^k - x^*\|^2 \leq qr^k,$$

where q is some constant.

1.4 NEWTON'S METHOD AND VARIATIONS

In the last two sections we emphasized a basic tradeoff in gradient methods: implementation simplicity versus fast convergence. We have already discussed steepest descent, one of the simplest but also one of the slowest methods. We now consider its opposite extreme, Newton's method, which is arguably the most complex and also the fastest of the gradient methods (under appropriate conditions).

Newton's method consists of the iteration

$$x^{k+1} = x^k - \alpha^k \big(\nabla^2 f(x^k)\big)^{-1} \nabla f(x^k), \tag{1.77}$$

assuming that the Newton direction

$$d^k = -\big(\nabla^2 f(x^k)\big)^{-1} \nabla f(x^k) \tag{1.78}$$

is defined [i.e., $\nabla^2 f(x^k)$ is invertible] and is a direction of descent [i.e., $d^{k\prime}\nabla f(x^k) < 0$]. As explained in the preceding section, one may view this iteration as a scaled version of steepest descent where the "optimal" scaling matrix $\big(\nabla^2 f(x^k)\big)^{-1}$ is used. It is worth mentioning in this connection that *Newton's method is "scale-free,"* in the sense that it cannot be affected by a change in coordinate system as is true for steepest descent (see Exercise 1.4.1).

When the Armijo rule is used with initial stepsize $s = 1$, then no reduction of the stepsize will be necessary near a nonsingular minimum (positive definite Hessian), as shown in Prop. 1.3.2. Thus, near convergence the method takes the form

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k), \tag{1.79}$$

which will be referred to as the *pure form of Newton's method*. On the other hand, far from such a local minimum, the Hessian matrix may be singular or the Newton direction of Eq. (1.78) may not be a direction of descent because the Hessian $\nabla^2 f(x^k)$ is not positive definite. Thus the analysis of Newton's method has two principal aspects:

(a) Local convergence, dealing with the behavior of the pure form of the method near a nonsingular local minimum.

(b) Global convergence, addressing the modifications that are necessary to ensure that the method is valid and is likely to converge to a local minimum when started far from all local minima.

We consider these issues in this section and we also discuss some variations of Newton's method, which are aimed at reducing the overhead for computing the Newton direction.

Local Convergence

It can be shown that the pure form of Newton's method converges superlinearly when started close enough to a nonsingular local minimum. This is suggested by the local convergence result for gradient methods (the capture theorem of Prop. 1.2.4) together with the superlinear convergence result for Newton-like methods (Prop. 1.3.2). Results of this type hold for a more general form of Newton's method, which can be used to solve the system of n equations with n unknowns

$$g(x) = 0, \tag{1.80}$$

where $g : \Re^n \mapsto \Re^n$ is a continuously differentiable function. This method has the form

$$x^{k+1} = x^k - \left(\nabla g(x^k)'\right)^{-1} g(x^k), \tag{1.81}$$

and for the special case where $g(x)$ is equal to the gradient $\nabla f(x)$, it yields the pure form of Eq. (1.79). Note here that a continuously differentiable function $g : \Re^n \mapsto \Re^n$ need not be equal to the gradient of some function. In particular, $g(x) = \nabla f(x)$ for some $f : \Re^n \mapsto \Re$, if and only if the $n \times n$ matrix $\nabla g(x)$ is symmetric for all x (see [OrR70], p. 95). Thus, the equation version of Newton's method [cf. Eq. (1.81)] is more broadly applicable than the optimization version [cf. Eq. (1.79)].

Here is a simple argument that shows the fast convergence of Newton's method (1.81). Suppose that the method generates a sequence $\{x^k\}$ that converges to a vector x^* such that $g(x^*) = 0$ and $\nabla g(x^*)$ is invertible. Let us use a first order expansion around x^k to write

$$0 = g(x^*) = g(x^k) + \nabla g(x^k)'(x^* - x^k) + o\big(\|x^k - x^*\|\big).$$

By multiplying this relation with $\big(\nabla g(x^k)'\big)^{-1}$ we have

$$x^k - x^* - \big(\nabla g(x^k)'\big)^{-1} g(x^k) = o\big(\|x^k - x^*\|\big),$$

so for the pure Newton iteration, $x^{k+1} = x^k - \big(\nabla g(x^k)'\big)^{-1} g(x^k)$, we obtain

$$x^{k+1} - x^* = o\big(\|x^k - x^*\|\big).$$

Thus, for $x^k \neq x^*$,

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = \lim_{k\to\infty} \frac{o\big(\|x^k - x^*\|\big)}{\|x^k - x^*\|} = 0,$$

implying superlinear convergence. This argument can also be used to show convergence to x^* if the initial vector x^0 is sufficiently close to x^*. The following proposition proves a more detailed result.

Proposition 1.4.1: Consider a function $g : \Re^n \mapsto \Re^n$, and a vector x^* such that $g(x^*) = 0$. For $\delta > 0$, let S_δ denote the sphere $\{x \mid \|x - x^*\| \leq \delta\}$. Assume that g is continuously differentiable within some sphere $S_{\bar{\delta}}$ and that $\nabla g(x^*)$ is invertible.

(a) There exists $\delta > 0$ such that if $x^0 \in S_\delta$, the sequence $\{x^k\}$ generated by the iteration

$$x^{k+1} = x^k - \big(\nabla g(x^k)'\big)^{-1} g(x^k)$$

is defined, belongs to S_δ, and converges to x^*. Furthermore, $\big\{\|x^k - x^*\|\big\}$ converges superlinearly.

(b) Assume that for some $L > 0$, $M > 0$, $\delta \in (0, \bar{\delta}]$, and for all x and y in S_δ,

$$\big\|\nabla g(x) - \nabla g(y)\big\| \leq L\|x - y\|, \qquad \Big\|\big(\nabla g(x)'\big)^{-1}\Big\| \leq M. \quad (1.82)$$

Then, if $x^0 \in S_\delta$, we have

$$\|x^{k+1} - x^*\| \leq \frac{LM}{2}\|x^k - x^*\|^2, \qquad \forall\ k = 0, 1, ...,$$

so if $LM\delta/2 < 1$ and $x^0 \in S_\delta$, $\{\|x^k - x^*\|\}$ converges superlinearly with order at least two.

Proof: (a) Choose $\delta > 0$ so that $\big(\nabla g(x)'\big)^{-1}$ exists for $x \in S_\delta$, and let

$$M = \sup_{x \in S_\delta} \left\| \big(\nabla g(x)'\big)^{-1} \right\|.$$

Assuming that $x^0 \in S_\delta$, and using the relation

$$g(x^k) = \int_0^1 \nabla g\big(x^* + t(x^k - x^*)\big)' dt (x^k - x^*),$$

we estimate $\|x^{k+1} - x^*\|$ as

$$\begin{aligned}
\|x^{k+1} - x^*\| &= \left\| x^k - x^* - \big(\nabla g(x^k)'\big)^{-1} g(x^k) \right\| \\
&= \left\| \big(\nabla g(x^k)'\big)^{-1} \big(\nabla g(x^k)'(x^k - x^*) - g(x^k)\big) \right\| \\
&= \left\| \big(\nabla g(x^k)'\big)^{-1} \left(\nabla g(x^k)' - \int_0^1 \nabla g\big(x^* + t(x^k - x^*)\big)' dt \right) (x^k - x^*) \right\| \\
&= \left\| \big(\nabla g(x^k)'\big)^{-1} \left(\int_0^1 \Big[\nabla g(x^k)' - \nabla g\big(x^* + t(x^k - x^*)\big)' \Big] dt \right) (x^k - x^*) \right\| \\
&\leq M \left(\int_0^1 \big\| \nabla g(x^k) - \nabla g\big(x^* + t(x^k - x^*)\big) \big\| \, dt \right) \|x^k - x^*\|.
\end{aligned}$$

By continuity of ∇g, we can take δ sufficiently small to ensure that $\{\|x^k - x^*\|\}$ is monotonically decreasing and that the term under the integral sign is arbitrarily small for all k. The convergence of x^k to x^* and the superlinear convergence of $\|x^k - x^*\|$ follow.

(b) If the condition (1.82) holds, the preceding relation yields

$$\|x^{k+1} - x^*\| \leq M \left(\int_0^1 Lt \|x^k - x^*\| dt \right) \|x^k - x^*\| = \frac{LM}{2}\|x^k - x^*\|^2.$$

Q.E.D.

A related result is the following. Its proof requires a simple modification of the proof of Prop. 1.4.1(a), and is left as Exercise 1.4.2 for the reader.

Proposition 1.4.2: Under the assumptions of Prop. 1.4.1(a), given any $r > 0$, there exists a $\delta > 0$ such that if $\|x^k - x^*\| < \delta$, then

$$\|x^{k+1} - x^*\| \leq r\|x^k - x^*\|, \qquad \big\|g(x^{k+1})\big\| \leq r\big\|g(x^k)\big\|.$$

Thus, the pure form of Newton's method converges extremely fast once it gets "near" a solution x^* where $\nabla g(x^*)$ is invertible, typically taking a handful of iterations to achieve very high solution accuracy; see Fig. 1.4.1. Unfortunately, it is usually difficult to predict whether a given starting point is sufficiently near to a solution for the fast convergence rate of Newton's method to take hold right away. Thus, in practice one can only expect that *eventually* the fast convergence rate of Newton's method will take hold. Figure 1.4.2 illustrates how the method can fail to converge when started far from a solution.

Global Convergence

Newton's method in its pure form for unconstrained minimization of f has several serious drawbacks.

(a) The inverse $\big(\nabla^2 f(x^k)\big)^{-1}$ may fail to exist, in which case the method breaks down. This will happen, for example, in regions where f is linear ($\nabla^2 f = 0$).

(b) The pure form is not a descent method; i.e., possibly $f(x^{k+1}) > f(x^k)$.

(c) The pure form tends to be attracted by local maxima just as much as it is attracted by local minima. It just tries to solve the system of equations $\nabla f(x) = 0$.

For these reasons, it is necessary to modify the pure form of Newton's method to turn it into a reliable minimization algorithm. There are several schemes that accomplish this by converting the pure form into a gradient method with a gradient related direction sequence. Simultaneously the modifications are such that, near a nonsingular local minimum, the algorithm assumes the pure form of Newton's method (1.79) and achieves the attendant fast convergence rate.

A simple possibility is to replace the Newton direction by the steepest descent direction (possibly after diagonal scaling), whenever the Newton direction is either not defined or is not a descent direction.† With proper

† Interestingly, this motivated the development of steepest descent by M. Augustin Cauchy. In his original paper [Cau47], Cauchy states as motivation for the steepest descent method its capability to obtain a close approximation to the

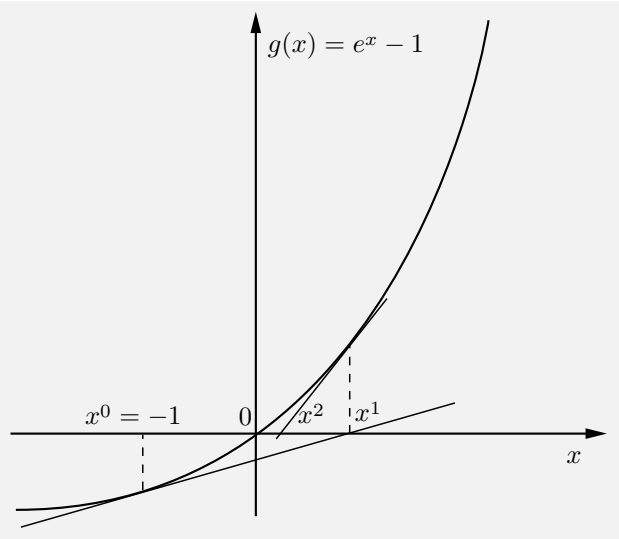

Figure 1.4.1. Fast convergence of Newton's method for solving the equation $e^x - 1 = 0$.

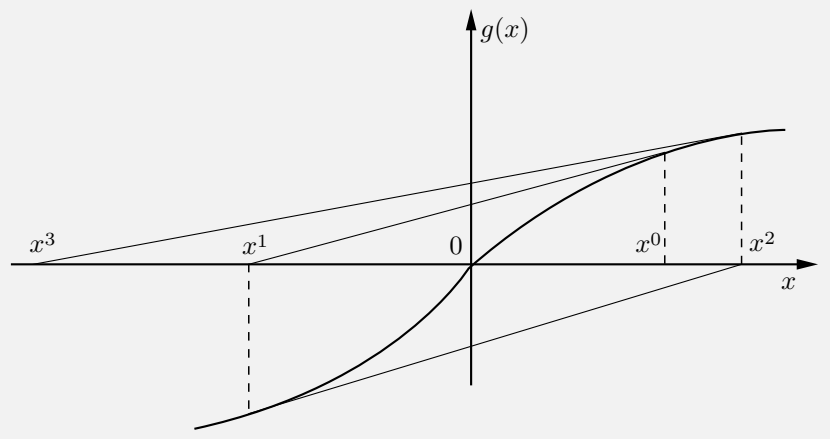

Figure 1.4.2. Divergence of Newton's method for solving an equation $g(x) = 0$ of a single variable x, when the starting point is far from the solution. This phenomenon typically occurs when $\big\|\nabla g(x)\big\|$ tends to decrease as $\|x\| \to \infty$.

safeguards, such a method has appropriate convergence and asymptotic

---

solution, in which case "... one can obtain new approximations very rapidly with the aid of the linear or Newton's method ..." (Note the attribution to Newton by Cauchy.)

rate of convergence properties (see Exercise 1.4.3, and for a related method, see Exercise 1.4.4). However, its performance at the early iterations may be quite slow, whether the Newton direction or the steepest descent direction is used in these iterations.

Generally, no modified version of Newton's method can be guaranteed to converge fast in the early iterations, but there are schemes that can use second derivative information effectively, even when the Hessian is not positive definite. These schemes are based on making diagonal modifications to the Hessian; that is, they obtain the direction d^k by solving a system of the form

$$\big(\nabla^2 f(x^k) + \Delta^k\big)d^k = -\nabla f(x^k),$$

whenever the Newton direction does not exist or is not a descent direction. Here Δ^k is a diagonal matrix such that

$$\nabla^2 f(x^k) + \Delta^k : \text{ positive definite.}$$

We outline some possibilities in the next two subsections.

1.4.1 Modified Cholesky Factorization

It can be shown that every positive definite matrix Q has a unique factorization of the form

$$Q = LL',$$

where L is a lower triangular matrix; this is known as the *Cholesky factorization of* Q (see Appendix D). Systems of equations of the form $Qx = b$ can be solved by first solving for y the triangular system $Ly = b$, and then by solving for x the triangular system $L'x = y$. These triangular systems can be solved easily [in $O(n^2)$ operations, as opposed to general systems, which require $O(n^3)$ operations; see Appendix D]. Since calculation of the Newton direction involves solution of the system

$$\nabla^2 f(x^k)d^k = -\nabla f(x^k),$$

it is natural to compute d^k by attempting to form the Cholesky factorization of $\nabla^2 f(x^k)$. During this process, one can detect whether $\nabla^2 f(x^k)$ is either nonpositive definite or nearly singular, in which case some of the diagonal elements of $\nabla^2 f(x^k)$ are suitably increased to ensure that the resulting matrix is positive definite. This is done sequentially during the factorization process, so in the end we obtain

$$L^k L^{k\prime} = \nabla^2 f(x^k) + \Delta^k,$$

where L^k is lower triangular and nonsingular, and Δ^k is diagonal.

As an illustration, consider the 2-dimensional case (for the general case, see Appendix D). Let

$$\nabla^2 f(x^k) = \begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix}$$

and let the desired factorization be of the form

$$LL' = \begin{pmatrix} \alpha & 0 \\ \gamma & \beta \end{pmatrix} \cdot \begin{pmatrix} \alpha & \gamma \\ 0 & \beta \end{pmatrix}.$$

We choose α, β, and γ, so that $\nabla^2 f(x^k) = LL'$ if $\nabla^2 f(x^k)$ is positive definite, and we appropriately modify h_{11} and h_{22} otherwise. This determines the first diagonal element α according to the relation

$$\alpha = \begin{cases} \sqrt{h_{11}} & \text{if } h_{11} > 0 \\ \sqrt{h_{11} + \delta_1} & \text{otherwise} \end{cases}$$

where δ_1 is such that $h_{11} + \delta_1 > 0$. Given α, we can calculate γ by equating the corresponding elements of $\nabla^2 f(x^k)$ and LL'. We obtain $\gamma\alpha = h_{12}$ or

$$\gamma = \frac{h_{12}}{\alpha}.$$

We can now calculate the second diagonal element β by equating the corresponding elements of $\nabla^2 f(x^k)$ and LL', after appropriately modifying h_{22} if necessary,

$$\beta = \begin{cases} \sqrt{h_{22} - \gamma^2} & \text{if } h_{22} > \gamma^2, \\ \sqrt{h_{22} - \gamma^2 + \delta_2} & \text{otherwise,} \end{cases}$$

where δ_2 is such that $h_{22} - \gamma^2 + \delta_2 > 0$. The method for choosing the increments δ_1 and δ_2 is largely heuristic. One possibility is discussed in Appendix D, which also describes more sophisticated versions of the above procedure where a positive increment is added to the diagonal elements of the Hessian even when the corresponding diagonal elements of the factorization are positive but very close to zero.

Given the $L^k L^{k\prime}$ factorization, the direction d^k is obtained by solving the system

$$L^k L^{k\prime} d^k = -\nabla f(x^k).$$

The next iterate is

$$x^{k+1} = x^k + \alpha^k d^k,$$

where α^k is chosen according to the Armijo rule or one of the other stepsize rules we have discussed.

To guarantee convergence, the increments added to the diagonal elements of the Hessian can be chosen so that $\{d^k\}$ is gradient related (cf. Prop. 1.2.1). Also, these increments can be chosen to be zero near a nonsingular local minimum. In particular, with proper safeguards, near such a point, the method becomes identical to the pure form of Newton's method and achieves the corresponding superlinear convergence rate (see Appendix D).

1.4.2 Trust Region Methods

As explained in Section 1.2, the pure Newton step is obtained by minimizing over d the second order approximation of f around x^k, given by

$$f^k(d) = f(x^k) + \nabla f(x^k)'d + \tfrac{1}{2}d'\nabla^2 f(x^k)d.$$

We know that $f^k(d)$ is a good approximation of $f(x^k + d)$ when d is in a small neighborhood of zero, but the difficulty is that with unconstrained minimization of $f^k(d)$ one may obtain a step that lies outside this neighborhood. It therefore makes sense to consider a *restricted Newton step* d^k, which is obtained by minimizing $f^k(d)$ over a suitably small neighborhood of zero, called the *trust region*:

$$d^k \in \arg \min_{||d|| \le \gamma^k} f^k(d), \tag{1.83}$$

where γ^k is some positive scalar.† An approximate solution of the constrained minimization problem of Eq. (1.83) can be obtained quickly using the fact that it has only one constraint. We refer to the specialized literature, including [MoS83] and the book [CGT00], for an account of approximate solution methods.

An important observation here is that even if $\nabla^2 f(x^k)$ is not positive definite or, more generally, even if the pure Newton direction is not a descent direction, the restricted Newton step d^k improves the cost, provided $\nabla f(x^k) \neq 0$ and γ^k is sufficiently small. The reason is that, in view of Eq. (1.83), $f^k(d^k)$ is smaller than $f(x^k)$ [which is equal to $f^k(0)$], and $f(x^k+d^k)$ is very close to its second order expansion $f^k(d^k)$ when $\|d^k\|$ is small.

More specifically, we have for all d with $||d|| \le \gamma^k$

$$f(x^k + d) = f^k(d) + o\big((\gamma^k)^2\big),$$

so that

$$\begin{aligned} f(x^k + d^k) &= f^k(d^k) + o\big((\gamma^k)^2\big) \\ &= f(x^k) + \min_{||d|| \le \gamma^k} \big\{\nabla f(x^k)'d + \tfrac{1}{2}d'\nabla^2 f(x^k)d\big\} + o\big((\gamma^k)^2\big). \end{aligned}$$

Therefore, denoting

$$\tilde{d}^k = -\frac{\nabla f(x^k)}{\big\|\nabla f(x^k)\big\|}\gamma^k,$$

† It can be shown that the restricted Newton step d^k also solves a system of the form $\big(\nabla^2 f(x^k) + \delta^k I\big)d = -\nabla f(x^k)$, where I is the identity matrix and δ^k is a nonnegative scalar (a Lagrange multiplier in the terminology of Chapter 4), so the preceding method of determining d^k fits the general framework of using a correction of the Hessian matrix by a positive semidefinite matrix.

we have

$$\begin{aligned} f(x^k + d^k) \leq & f(x^k) + \nabla f(x^k)' \tilde{d}^k + \tfrac{1}{2} \tilde{d^k}' \nabla^2 f(x^k) \tilde{d}^k + o\big((\gamma^k)^2\big) \\ = & f(x^k) - \gamma^k \big\| |\nabla f(x^k)| \big\| + \frac{(\gamma^k)^2}{2 \big\| \nabla f(x^k) \big\|^2} \nabla f(x^k)' \nabla^2 f(x^k) \nabla f(x^k) \\ & + o\big((\gamma^k)^2\big). \end{aligned}$$

For γ^k sufficiently small, the negative term $-\gamma^k \big\| |\nabla f(x^k)| \big\|$ dominates the last two terms on the right-hand side above, showing that

$$f(x^k + d^k) < f(x^k).$$

It can be seen in fact from the preceding relations that a cost improvement is possible even when $\nabla f(x^k) = 0$, provided γ^k is sufficiently small and f has a direction of negative curvature at x^k, i.e., $\nabla^2 f(x^k)$ is not positive semidefinite. Thus the preceding procedure will fail to improve the cost only if $\nabla f(x^k) = 0$ and $\nabla^2 f(x^k)$ is positive semidefinite, i.e., x^k satisfies the first *and* the second order necessary conditions. In particular, one can make progress even if x^k is a stationary point that is not a local minimum.

We are thus motivated to consider a method of the form

$$x^{k+1} = x^k + d^k,$$

where d^k is the restricted Newton step corresponding to a suitably chosen scalar γ^k as per Eq. (1.83). Here, for a given x^k, γ^k should be small enough so that there is cost improvement; one possibility is to start from an initial trial γ^k and successively reduce γ^k by a certain factor as many times as necessary until a cost reduction occurs [$f(x^{k+1}) < f(x^k)$]. The choice of the initial trial value for γ^k is crucial here; if it is chosen too large, a large number of reductions may be necessary before a cost improvement occurs; if it is chosen too small the convergence rate may be poor. In particular, to maintain the superlinear convergence rate of Newton's method, as x^k approaches a nonsingular local minimum, one should select the initial trial value of γ^k sufficiently large so that the restricted Newton step and the pure Newton step coincide.

A reasonable way to adjust the initial trial value for γ^k is to increase this value when the method appears to be progressing well and to decrease this value otherwise. One can measure progress by using the ratio of actual over predicted cost improvement [based on the approximation $f^k(d)$]

$$r^k = \frac{f(x^k) - f(x^{k+1})}{f(x^k) - f^k(d^k)}.$$

In particular, it makes sense to increase the initial trial value for γ ($\gamma^{k+1} > \gamma^k$) if this ratio is close to or above unity, and decrease γ otherwise. The

following algorithm is a typical example of such a method. Given x^k and an initial trial value γ^k, it determines x^{k+1} and an initial trial value γ^{k+1} by using two thresholds σ_1, σ_2 with $0 < \sigma_1 \leq \sigma_2 \leq 1$ and two factors β_1, β_2 with $0 < \beta_1 < 1 < \beta_2$ (typical values are $\sigma_1 = 0.2$, $\sigma_2 = 0.8$, $\beta_1 = 0.25$, $\beta_2 = 2$).

Step 1: Find

$$d^k \in \arg \min_{||d|| \leq \gamma^k} f^k(d), \tag{1.84}$$

If $f^k(d^k) = f(x^k)$ stop (x^k satisfies the first and second order necessary conditions for a local minimum); else go to Step 2.

Step 2: If $f(x^k + d^k) < f(x^k)$ set

$$x^{k+1} = x^k + d^k \tag{1.85}$$

calculate

$$r^k = \frac{f(x^k) - f(x^{k+1})}{f(x^k) - f^k(d^k)} \tag{1.86}$$

and go to Step 3; else set $\gamma^k := \beta_1 ||d^k||$ and go to Step 1.

Step 3: Set

$$\gamma^{k+1} = \begin{cases} \beta_1 ||d^k|| & \text{if } r^k < \sigma_1, \\ \beta_2 \gamma^k & \text{if } \sigma_2 \leq r^k \text{ and } \|d^k\| = \gamma^k, \\ \gamma^k & \text{otherwise.} \end{cases} \tag{1.87}$$

Go to the next iteration.

Assuming that f is twice continuously differentiable, it is possible to show that the above algorithm is convergent in the sense that if $\{x^k\}$ is a bounded sequence, there exists a limit point of $\{x^k\}$ that satisfies the first and the second order necessary conditions for optimality. Furthermore, if $\{x^k\}$ converges to a nonsingular local minimum x^*, then asymptotically, the method is identical to the pure form of Newton's method, thereby attaining a superlinear convergence rate; see the references given at the end of the chapter for proofs of these and other related results for trust region methods.

1.4.3 Variants of Newton's Method

We will now briefly consider approximate implementations of Newton's method. The idea is to calculate the Newton direction approximately, with the aim of economizing on computational overhead with relatively small degradation of the convergence rate.

Newton's Method with Periodic Reevaluation of the Hessian

A variation of Newton's method is obtained if the Hessian matrix $\nabla^2 f$ is recomputed every $p > 1$ iterations rather than at every iteration. In particular, this method, in unmodified form, is given by

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

where

$$D^{ip+j} = \left(\nabla^2 f(x^{ip})\right)^{-1}, \qquad j = 0, 1, ..., p-1, \ i = 0, 1, ...$$

The idea here is to save the computation and the inversion (or factorization) of the Hessian for the iterations where $j \neq 0$. This reduction in overhead is achieved at the expense of what is usually a small degradation in speed of convergence.

Truncated Newton Methods

We have so far assumed that the Newton system $\nabla^2 f(x^k) d^k = -\nabla f(x^k)$ will be solved exactly for the direction d^k by Cholesky factorization or Gaussian elimination, which require a finite number of arithmetic operations [$O(n^3)$]. When the dimension n is large, the calculation required for exact solution may be prohibitive. An alternative is to use an approximate solution, which may be obtained with an iterative method. This approach is often useful for solving very large linear systems of equations, arising in the solution of partial differential equations, where an adequate approximation to the solution can often be obtained by iterative methods quite fast, while the computation to find the exact solution can be overwhelming.

Generally, solving for d any system of the form $H^k d = -\nabla f(x^k)$, where H^k is a positive definite symmetric $n \times n$ matrix, can be done by solving the quadratic optimization problem

$$\begin{aligned} &\text{minimize} \quad \tfrac{1}{2} d' H^k d + \nabla f(x^k)' d \\ &\text{subject to} \quad d \in \Re^n, \end{aligned} \tag{1.88}$$

whose cost function gradient is zero at d if and only if $H^k d = -\nabla f(x^k)$. Suppose that an iterative descent method is used for solution and the starting point is $d^0 = 0$. Since the quadratic cost is reduced at each iteration and its value at the starting point is zero, we obtain after each iteration a vector d^k satisfying $\frac{1}{2} d^{k\prime} H^k d^k + \nabla f(x^k)' d^k < 0$, from which, using the positive definiteness of H^k,

$$\nabla f(x^k)' d^k < 0.$$

Thus the approximate solution d^k of the system $H^k d = -\nabla f(x^k)$, obtained after any positive number of iterations, is a descent direction.

Possible iterative methods for solving the direction finding problem (1.88) include the conjugate gradient method to be presented in Section 2.1 and the coordinate descent method to be discussed in Section 2.3.1. For the conjugate gradient method to be practical, the calculation of matrix-vector products of the form $H^k d$ must be convenient, and for this, the presence of special structure of H^k may be important. An example of this type is network optimization problems, to be discussed in Section 3.8. Also, the idea of implementing Newton's method by using an iterative method applies more generally to constrained forms of the method, to be discussed in Chapter 3.

Conditions on the accuracy of the approximate solution d^k that ensure linear or superlinear rate of convergence are given in Exercise 1.4.5. Generally, the superlinear convergence rate property of the method to a nonsingular local minimum is maintained if the approximate Newton directions d^k satisfy

$$\lim_{k\to\infty} \frac{\left\|\nabla^2 f(x^k)d^k + \nabla f(x^k)\right\|}{\left\|\nabla f(x^k)\right\|} = 0,$$

(cf. Prop. 1.3.2). Thus, for superlinear convergence rate, the norm of the error in solving the Newton system must become negligible relative to the gradient norm in the limit.

1.4.4 Least Squares and the Gauss-Newton Method

We will now consider a specialized Newton-like method for solving least squares problems of the form

$$\begin{aligned} &\text{minimize} \quad f(x) = \tfrac{1}{2}\left\|g(x)\right\|^2 = \tfrac{1}{2}\sum_{i=1}^m \left\|g_i(x)\right\|^2 \\ &\text{subject to} \quad x \in \Re^n, \end{aligned} \tag{1.89}$$

where g is a continuously differentiable function with component functions $g_1, \ldots, g_m$, where $g_i : \Re^n \to \Re^{r_i}$. Usually $r_i = 1$, but it is sometimes convenient to consider the more general case.

Least squares problems are common in many practical contexts. An important case arises when g consists of n scalar-valued functions and we want to solve the system of n equations with n unknowns $g(x) = 0$. We can formulate this as the least squares optimization problem (1.89) [x^* solves the system $g(x) = 0$ if and only if it minimizes $\frac{1}{2}\left\|g(x)\right\|^2$ and the optimal value is zero]. Here are some other examples:

Example 1.4.1 (Model Construction – Curve Fitting)

Suppose that we want to estimate n parameters of a mathematical model so that it fits well a physical system, based on a set of input-output data. In particular, we hypothesize an approximate relation of the form

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

where h is a known function representing the model and

$x \in \Re^n$ is a vector of unknown parameters,
$z \in \Re^r$ is the model's output,
$y \in \Re^p$ is the model's input.

Given a set of m input-output data pairs $(y_1, z_1), \ldots, (y_m, z_m)$ from measurements of the physical system that we try to model, we want to find the vector of parameters x that matches best the data in the sense that it minimizes the sum of squared errors

$$\tfrac{1}{2} \sum_{i=1}^{m} \big\| z_i - h(x, y_i) \big\|^2.$$

For example, to fit the data pairs by a cubic polynomial approximation, we would choose

$$h(x, y) = x_3 y^3 + x_2 y^2 + x_1 y + x_0,$$

where $x = (x_0, x_1, x_2, x_3)$ is the vector of unknown coefficients of the cubic polynomial.

The next two examples are really special cases of the preceding one.

Example 1.4.2 (Dynamic System Identification)

A common model for a single input-single output dynamic system is to relate the input sequence $\{y_k\}$ to the output sequence $\{z_k\}$ by a linear equation of the form

$$\sum_{j=0}^{n} \alpha_j z_{k-j} = \sum_{j=0}^{n} \beta_j y_{k-j}.$$

Given a record of inputs and outputs $y_1, z_1, \ldots, y_m, z_m$ from the true system, we would like to find a set of parameters $\{\alpha_j, \beta_j \mid j = 0, 1, \ldots, n\}$ that matches this record best in the sense that it minimizes

$$\sum_{k=n}^{m} \left(\sum_{j=0}^{n} \alpha_j z_{k-j} - \sum_{j=0}^{n} \beta_j y_{k-j} \right)^2.$$

This is a least-squares problem.

Example 1.4.3 (Neural Networks)

A least squares modeling approach that has received a lot of attention is provided by *neural networks*. Here the model is specified by a multistage system, also called a *multilayer perceptron*. The kth stage consists of n_k *activation units*, each being a single input-single output mapping of a given form $\phi : \Re \mapsto \Re$ (examples will be given shortly). The output of the jth

activation unit of the $(k+1)$st stage is denoted by x_{k+1}^j and the input is a linear function of the output vector $x_k = (x_k^1, \ldots, x_k^{n_k})$ of the kth stage. Thus

$$x_{k+1}^j = \phi\left(u_k^{0j} + \sum_{s=1}^{n_k} x_k^s u_k^{sj}\right), \qquad j = 1, \ldots, n_{k+1}, \tag{1.90}$$

where the coefficients u_k^{sj} (also called *weights*) are to be determined. In variants of this approach there may be some constraints on the weights in order to induce desired connectivity structures between the stages, which are designed to produce a particular effect and/or exploit some known structure of the input. However, for simplicity we will not consider this possibility; the algorithmic ideas to be described generalize.

Suppose that the multilayer perceptron has N stages, and let u denote the vector of the weights of all the stages:

$$u = \{u_k^{sj} \mid k = 0, \ldots, N-1,\ s = 0, \ldots, n_k,\ j = 1, \ldots, n_{k+1}\}.$$

Then, for a given vector u of weights, an input vector x_0 to the first stage produces a unique output vector x_N from the Nth stage via Eq. (1.90). Thus, we may view the multilayer perceptron as a mapping h that is parameterized by u and transforms the input vector x_0 into an output vector of the form $x_N = h(u, x_0)$. Suppose that we have m sample input-output pairs $(y_1, z_1), \ldots, (y_m, z_m)$ from a physical system that we are trying to model. Then, by selecting u appropriately, we can try to match the mapping of the multilayer perceptron with the mapping of the physical system. A common way to do this is to minimize over u the sum of squared errors

$$\tfrac{1}{2}\textstyle\sum_{i=1}^m \big\|z_i - h(u, y_i)\big\|^2.$$

In neural network terminology, finding the optimal weights u is referred to as *training the network*. Incremental gradient methods, to be discussed in Section 2.4.1, are often used for this purpose (see e.g., [BeT96], [Hay11]).

Examples of activation units are functions such as

$$\phi(\xi) = \frac{1}{1+e^{-\xi}}, \qquad \text{(sigmoidal function)},$$

$$\phi(\xi) = \frac{e^{\xi} - e^{-\xi}}{e^{\xi} + e^{-\xi}}, \qquad \text{(hyperbolic tangent function)},$$

whose gradients are zero as the argument ξ approaches $-\infty$ and ∞. For these functions, it is possible to show that with a sufficient number of activation units and a number of stages $N \geq 2$, a multilayer perceptron can approximate arbitrarily closely very complex input-output maps; see [Cyb89]. In practice, a number N that is considerably larger than 2 is often considered, in combination with functions ϕ specially tailored to particular types of problems, giving rise to so called *deep neural networks*, which have attained considerable success in a variety of applications; see e.g., [HDY12], [SHM16].

Neural network training problems can be quite challenging. Their cost function is typically nonconvex and involves multiple local minima. For large

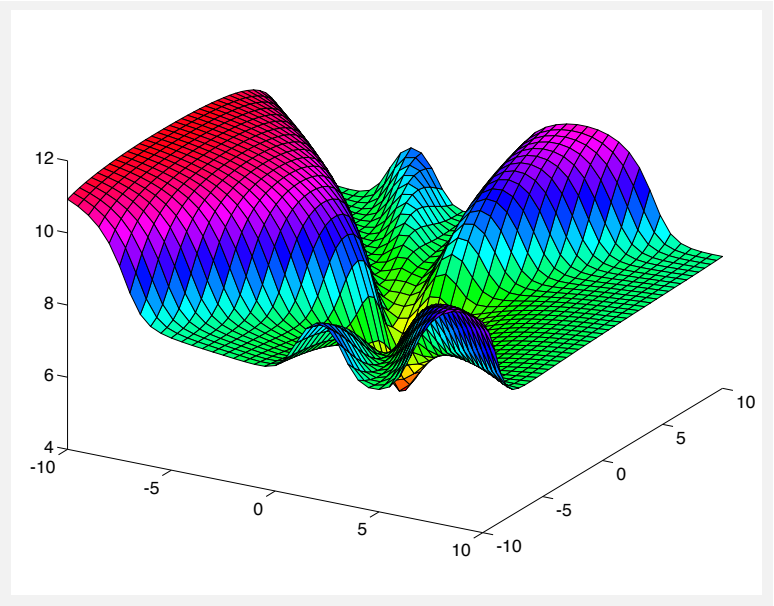

Figure 1.4.3. Three-dimensional plot of a least squares cost function

$$\tfrac{1}{2}\sum_{i=1}^{5}\big(z_i - \phi(u_1 y_i + u_0)\big)^2,$$

for a neural network training problem where there are only two weights u_0 and u_1, five data pairs, and ϕ is the hyperbolic tangent function. The data of the problem are given in Exercise 2.4.3. The cost function tends to a constant as u is changed along rays of the form $r\bar{u}$, where $r > 0$ and $\bar{u}$ is a fixed vector.

values of the weights u_k^{ij}, the cost becomes "flat." In fact, the cost function tends to a constant as u is changed along rays of the form $r\bar{u}$, where $r > 0$ and $\bar{u}$ is a fixed vector; see Fig. 1.4.3. For u near the origin, the cost function can be quite complicated, alternately involving flat and steep regions.

The next example deals with an important context where neural networks are often used.

Example 1.4.4 (Classification - Hypothesis Testing)

Let us consider a problem of classifying objects based on the values of their characteristics. Here we use the term "object" generically. In some contexts, the classification may relate to persons or situations. In other cases, an object may represent a hypothesis, and the problem is to decide which of the hypotheses is true, based on some data.

We assume that each object is presented to us with a vector y, and we wish to classify it in one of s categories $1, \ldots, s$. For example, the vector y may represent data, such as the results of a collection of tests on a medical

patient, and we may wish to classify the patient as being healthy or as having one of several types of illnesses.

A classical classification approach is to assume that for each category $j = 1, \ldots, s$, we know the conditional probability $p(j \mid y)$ that an object is of category j given data y. Then, given data y, we decide on the category $j^*(y)$ having maximum posterior probability, i.e.,

$$j^*(y) \in \arg \max_{j=1,\ldots,s} p(j \mid y). \tag{1.91}$$

This is called the Maximum a Posteriori rule (or MAP rule for short; see for example the book [BeT08] for a discussion).

Suppose now that the probabilities $p(j \mid y)$, viewed as functions of y, are unknown, but instead we have a sample consisting of data for m object-category pairs. Then we may try to estimate $p(j \mid y)$ based on the following simple fact: out of all functions $f_j(y)$ of y, $p(j \mid y)$ is the one that minimizes the expected value of $(z_j - f_j(y))^2$, where

$$z_j = \begin{cases} 1 & \text{if } y \text{ is of category } j, \\ 0 & \text{otherwise.} \end{cases}$$

To this end, we adopt a parametric approach. For each category $j = 1, \ldots, s$, we estimate the probability $p(j \mid y)$ with a function $h_j(x_j, y)$ that is parameterized by a vector x_j. The function h_j may be provided for example by a neural network (cf. Example 1.4.3). Then, denoting y_i the data vector of the ith object, we obtain x_j by minimizing the least squares function

$$\tfrac{1}{2} \textstyle\sum_{i=1}^m \big(z_j^i - h_j(x_j, y_i)\big)^2,$$

where

$$z_j^i = \begin{cases} 1 & \text{if } y_i \text{ is of category } j, \\ 0 & \text{otherwise.} \end{cases}$$

This minimization approximates the minimization of the expected value of $\big(z_j - f_j(y)\big)^2$. Once the optimal parameter vectors x_j^*, $j = 1, \ldots, s$, have been obtained, we may use them to classify a new object with data vector y according to the rule

$$\text{Estimated Object Category} \in \arg \max_{j=1,\ldots,s} h_j(x_j^*, y),$$

which approximates the MAP rule (1.91).

For the simpler case where there are just two categories, say A and B, a similar formulation is to hypothesize a relation of the following form between data vector y and category of an object:

$$\text{Object Category} = \begin{cases} A & \text{if } h(x, y) = 1, \\ B & \text{if } h(x, y) = -1, \end{cases}$$

where h is a given function and x is an unknown vector of parameters. Given a set of m data pairs $(z_1, y_1), \ldots, (z_m, y_m)$ of representative objects of known category, where y_i is the data vector of the ith object, and

$$z_i = \begin{cases} 1 & \text{if } y \text{ is of category } A, \\ -1 & \text{if } y \text{ is of category } B, \end{cases}$$

we obtain x by minimizing the least squares function

$$\tfrac{1}{2}\sum_{i=1}^{m}\big(z_i - h(x,y_i)\big)^2.$$

The optimal parameter vector x^* is used to classify a new object with data vector y according to the rule

$$\text{Estimated Object Category} = \begin{cases} A & \text{if } h(x^*,y) > 0, \\ B & \text{if } h(x^*,y) < 0. \end{cases}$$

There are several variations on the above theme, for which we refer to the specialized literature. Furthermore, there are several alternative optimization-based methods for classification (see Example 2.4.1, and the book [Ber15a], which also gives many references to the extensive literature on this subject).

The Gauss-Newton Method

We now consider the *Gauss-Newton method*, which is a specialized method for minimizing the least squares cost $(1/2)\big\|g(x)\big\|^2$. Given a point x^k, the pure form of the Gauss-Newton iteration is based on linearizing g to obtain

$$\tilde{g}(x,x^k) = g(x^k) + \nabla g(x^k)'(x - x^k)$$

and then minimizing the norm of the linearized function $\tilde{g}$:

$$\begin{aligned} x^{k+1} &\in \arg\min_{x\in\Re^n} \tfrac{1}{2}\big\|\tilde{g}(x,x^k)\big\|^2 \\ &= \arg\min_{x\in\Re^n} \tfrac{1}{2}\big\{\|g(x^k)\|^2 + 2(x-x^k)'\nabla g(x^k)g(x^k) \\ &\qquad\qquad + (x-x^k)'\nabla g(x^k)\nabla g(x^k)'(x-x^k)\big\}. \end{aligned}$$

Assuming that the $n\times n$ matrix $\nabla g(x^k)\nabla g(x^k)'$ is invertible, the above quadratic minimization yields

$$x^{k+1} = x^k - \big(\nabla g(x^k)\nabla g(x^k)'\big)^{-1}\nabla g(x^k)g(x^k). \tag{1.92}$$

Note that if g is already a linear function, we have $\big\|g(x)\big\|^2 = \big\|\tilde{g}(x,x^k)\big\|^2$, and the method converges in a single iteration. Note also that the direction

$$-\big(\nabla g(x^k)\nabla g(x^k)'\big)^{-1}\nabla g(x^k)g(x^k)$$

used in the above iteration is a descent direction since $\nabla g(x^k)g(x^k)$ is the gradient at x^k of the least squares cost function $(1/2)\big\|g(x)\big\|^2$ and $\big(\nabla g(x^k)\nabla g(x^k)'\big)^{-1}$ is a positive definite matrix.

To deal with the case where the matrix $\nabla g(x^k)\nabla g(x^k)'$ is singular (as well as enhance convergence when this matrix is nearly singular), the method is often implemented in the modified form

$$x^{k+1} = x^k - \alpha^k \big(\nabla g(x^k)\nabla g(x^k)' + \Delta^k\big)^{-1} \nabla g(x^k) g(x^k), \tag{1.93}$$

where α^k is a stepsize chosen by one of the stepsize rules that we have discussed, and Δ^k is a diagonal matrix such that

$$\nabla g(x^k)\nabla g(x^k)' + \Delta^k : \text{positive definite.}$$

For example, Δ^k may be chosen in accordance with the Cholesky factorization scheme outlined in Section 1.4.1. An early proposal, known as the *Levenberg-Marquardt method*, is to choose Δ^k to be a positive multiple of the identity matrix. With these choices of Δ^k, it can be seen that the directions used by the method are gradient related, and the convergence results of Section 1.2.2 apply.

Relation to Newton's Method

The Gauss-Newton method bears a close relation to Newton's method. In particular, assuming each g_i is a scalar function, the Hessian of the cost function $(1/2)\big\|g(x)\big\|^2$ is

$$\nabla g(x^k)\nabla g(x^k)' + \sum_{i=1}^m \nabla^2 g_i(x^k) g_i(x^k), \tag{1.94}$$

so it is seen that the Gauss-Newton iterations (1.92) and (1.93) are approximate versions of their Newton counterparts, where the second order term

$$\sum_{i=1}^m \nabla^2 g_i(x^k) g_i(x^k) \tag{1.95}$$

is neglected. Thus, in the Gauss-Newton method, we save the computation of this term at the expense of some deterioration in the convergence rate. If, however, the neglected term (1.95) is relatively small near a solution, the convergence rate of the Gauss-Newton method is satisfactory. This is often true in many applications such as for example when g is nearly linear, and also when the components $g_i(x)$ are small near the solution.

A case in point is when $m = n$ and the problem is to solve the system $g(x) = 0$. Then the neglected term is zero at a solution, and assuming $\nabla g(x^k)$ is invertible, we have

$$\big(\nabla g(x^k)\nabla g(x^k)'\big)^{-1} \nabla g(x^k) g(x^k) = \big(\nabla g(x^k)'\big)^{-1} g(x^k).$$

Thus the pure form of the Gauss-Newton method (1.92) takes the form

$$x^{k+1} = x^k - \big(\nabla g(x^k)'\big)^{-1} g(x^k),$$

and is identical to Newton's method for solving the system $g(x) = 0$ [rather than Newton's method for minimizing $\big\|g(x)\big\|^2$]. The convergence rate is typically superlinear in this case (cf. Prop. 1.4.1).

E X E R C I S E S

1.4.1 (Scale-Free Character of Newton's Method)

The purpose of this exercise is to show that Newton's method is unaffected by linear scaling of the variables. Consider a linear invertible transformation of variables $x = Sy$. Write the pure form of Newton's method in the space of the variables y and show that it generates the sequence $y^k = S^{-1}x^k$, where $\{x^k\}$ is the sequence generated by Newton's method in the space of the variables x.

1.4.2 (www)

Show Prop. 1.4.2. *Hint*: For the second relation, let

$$M(x) = \int_0^1 \nabla g\big(x^* + t(x - x^*)\big)' dt,$$

so that $g(x) = M(x)(x - x^*)$. Argue that for some $\delta > 0$ the eigenvalues of $M(x)'M(x)$ lie between some positive scalars γ and Γ for all x with $\|x - x^*\| \leq \delta$. Show that

$$\gamma\|x - x^*\|^2 \leq \big\|g(x)\big\|^2 \leq \Gamma\|x - x^*\|^2, \qquad \forall\ x \text{ with } \|x - x^*\| \leq \delta.$$

1.4.3 (Combination of Newton and Steepest Descent Methods)

Consider the iteration $x^{k+1} = x^k + \alpha^k d^k$ where α^k is chosen by the Armijo rule with initial stepsize $s = 1$, $\sigma \in (0, 1/2)$, and d^k is equal to

$$d_N^k = -\big(\nabla^2 f(x^k)\big)^{-1} \nabla f(x^k)$$

if $\nabla^2 f(x^k)$ is nonsingular and the following two inequalities hold:

$$c_1\big\|\nabla f(x^k)\big\|^{p_1} \leq -\nabla f(x^k)' d_N^k,$$

$$\|d_N^k\|^{p_2} \leq c_2\big\|\nabla f(x^k)\big\|;$$

otherwise

$$d^k = -D\nabla f(x^k),$$

where D is a fixed positive definite symmetric matrix. The scalars c_1, c_2, p_1, and p_2 satisfy $c_1 > 0$, $c_2 > 0$, $p_1 > 2$, $p_2 > 1$. Show that the sequence $\{d^k\}$ is gradient related. Furthermore, every limit point of $\{x^k\}$ is stationary, and if $\{x^k\}$ converges to a nonsingular local minimum x^*, the rate of convergence of $\big\{\|x^k - x^*\|\big\}$ is superlinear.

1.4.4 (Armijo Rule Along a Curved Path)

This exercise provides a globally convergent variant of Newton's method, which combines the Newton and steepest descent directions along a curved path. Let $f : \Re^n \mapsto \Re$ be twice continuously differentiable. At a point x^k, let $d_S^k = -D\nabla f(x^k)$ be a scaled steepest descent direction, where D is a fixed positive definite symmetric matrix. Let also d_N^k be the Newton direction $-\big(\nabla^2 f(x^k)\big)^{-1}\nabla f(x^k)$ if $\nabla^2 f(x^k)$ is nonsingular, and be equal to d_S^k otherwise. Consider the method

$$x^{k+1} = x^k + \alpha^k\big((1-\alpha^k)d_S^k + \alpha^k d_N^k\big),$$

where $\alpha^k = \beta^{m^k}$ and m^k is the first nonnegative integer m such that the following three inequalities hold:

$$f(x^k) - f\big(x^k + \beta^m\big((1-\beta^m)d_S^k + \beta^m d_N\big)\big) \geq -\sigma\beta^m\nabla f(x^k)'\big((1-\beta^m)d_S^k + \beta^m d_N\big),$$

$$c_1\min\big\{\nabla f(x^k)'D\nabla f(x^k),\, \|\nabla f(x^k)\|^3\big\} \leq -\nabla f(x^k)'\big((1-\beta^m)d_S^k + \beta^m d_N\big),$$

$$\big\|(1-\beta^m)d_S^k + \beta^m d_N\big\| \leq c_2\max\big\{\|D\nabla f(x^k)\|,\, \|\nabla f(x^k)\|^{1/2}\big\},$$

where β and σ are scalars satisfying $0 < \beta < 1$ and $0 < \sigma < 1/2$, and c_1 and c_2 are scalars satisfying $c_1 < 1$ and $c_2 > 1$. Show that the method is well defined in the sense that the stepsize α^k will be obtained after a finite number of trials. Furthermore, every limit point of $\{x^k\}$ is stationary and if $\{x^k\}$ converges to a nonsingular local minimum x^*, the rate of convergence of $\big\{\|x^k - x^*\|\big\}$ is superlinear. *Hint*: For a given x^k, each of the three inequalities is satisfied for m sufficiently large, so α^k is obtained after a finite number of trials. The directions $d^k = (1-\alpha^k)d_S^k + \alpha^k d_N^k$ are gradient related by construction (cf. the last two inequalities). Use the line of proof of Prop. 1.2.1 to show stationarity of the limit points of $\{x^k\}$. Use the line of proof of Prop. 1.3.2 to show that if $\{x^k\}$ converges to a nonsingular local minimum x^*, the convergence is superlinear (including the fact that $\alpha^k = 1$ for all sufficiently large k).

1.4.5 (www)

Consider a truncated Newton method with the stepsize chosen by the Armijo rule with initial stepsize $s = 1$ and $\sigma < 1/2$, and assume that $\{x^k\}$ converges to a nonsingular local minimum x^*. Assume that the matrices H^k and the directions d^k satisfy

$$\lim_{k\to\infty}\big\|H^k - \nabla^2 f(x^k)\big\| = 0, \qquad \lim_{k\to\infty}\frac{\big\|H^k d^k + \nabla f(x^k)\big\|}{\big\|\nabla f(x^k)\big\|} = 0.$$

Show that $\big\{\|x^k - x^*\|\big\}$ converges superlinearly.

1.4.6 (www)

Apply Newton's method with a constant stepsize to minimization of the function $f(x) = \|x\|^3$. Identify the range of stepsizes for which convergence is obtained, and show that it includes the unit stepsize. Show that for any stepsize within this range, the method converges linearly to $x^* = 0$. Explain this fact in light of Prop. 1.4.1.

1.4.7

Consider Newton's method with the given trust region implementation for the case of a positive definite quadratic cost function. Show that the method terminates in a finite number of iterations.

1.4.8

(a) Consider the pure form of Newton's method for the case of the cost function $f(x) = \|x\|^\beta$, where $\beta > 1$. For what starting points and values of β does the method converge to the optimal solution? What happens when $\beta \leq 1$?

(b) Repeat part (a) for the case where Newton's method with the Armijo rule is used.

1.4.9 (Necessary and Sufficient Conditions for Convergence of Iterative Methods for Linear Equations)

This exercise deals with the convergence of iterative algorithms for the system of linear equations $Ax = b$, where A is a given (possibly singular) $n \times n$ matrix and b is a vector in $\Re^n$. We assume that b lies in the range of A, so that the system has at least one solution. For a given $n \times n$ matrix D, we say that the iteration

$$x^{k+1} = x^k - \alpha D(Ax^k - b) \tag{1.96}$$

is *convergent* if there exists $\bar{\alpha} >$ such that for all $\alpha \in (0, \bar{\alpha}]$ and $x^0 \in \Re^n$ the sequence $\{x^k\}$ produced by the iteration converges to some solution of $Ax = b$. Show that the iteration is convergent if and only if the following conditions hold.

(i) Each eigenvalue of DA either has a positive real part or is equal to 0.

(ii) The dimension of the nullspace of DA is equal to the multiplicity of the 0 eigenvalue of DA.

(iii) The nullspace of A is equal to the nullspace of DA.

Note: The case where A is invertible is straightforward [then conditions (i)-(iii) are reduced to the condition that each eigenvalue of DA has positive real part]. The case where A is singular is challenging. To show that condition (ii) is necessary for a convergent iteration, use the fact that if it does not hold then there exists a vector v such that $DAv \neq 0$ and $(DA)^2 v = 0$. See [WaB13b] or [Ber12], Section 7.3.8, for a complete proof and related analysis.

1.4.10 (Iterative Solution of Nonlinear Equations)

This exercise deals with the iterative solution of the system of equations $g(x) = 0$, where $g : \Re^n \mapsto \Re^n$ is a continuously differentiable function. Consider the sequence $\{x^k\}$ generated by the iteration

$$x^{k+1} = x^k - D^k g(x^k), \qquad k = 0, 1, \ldots,$$

where $\{D^k\}$ is a sequence of $n \times n$ matrices.

(a) (*Linear Convergence*) Suppose that $\{x^k\}$ converges to some x^*. Denote

$$\bar{L} = \limsup_{k\to\infty} \left\| I - D^k \nabla g(x^*)' \right\|,$$

where $\|\cdot\|$ is the matrix norm induced by the standard Euclidean norm, and assume that $\bar{L} < 1$. Show that $g(x^*) = 0$. Moreover for every $L \in (\bar{L}, 1)$, there is an integer m such that

$$\|x^{k+1} - x^*\| \leq L\|x^k - x^*\|, \qquad \forall\ k \geq m.$$

(b) (*Local Convergence*) Let x^* be such that $g(x^*) = 0$ and $\nabla g(x^*)$ is nonsingular. Let also $D(x)$ be a matrix function such that

$$\limsup_{x\to x^*} \left\| I - D(x) \nabla g(x^*)' \right\| < 1.$$

Show that there is a neighborhood N of x^* such that if $x^0 \in N$, the sequence $\{x^k\}$ generated by the algorithm with $D^k = D(x^k)$ remains in N and converges to x^*.

Note: This is a challenging exercise; see [Hes80], Section 1.4, for a complete proof and related analysis.

1.5 NOTES AND SOURCES

Section 1.2: The steepest descent method dates to Cauchy [Cau47], who attributes to Newton the unity stepsize version of what we call Newton's method. The modern theory of gradient methods evolved in the 60s, when several practical convergent stepsize rules were proposed starting with the work of Goldstein [Gol62], [Gol64], Armijo [Arm66], and others. Shortly afterwards, general methods of convergence analysis were formulated starting with the work of Zangwill [Zan69], which was followed by the works of Ortega and Rheinboldt [OrR70], Daniel [Dan71], and Polak [Pol71]. The Armijo stepsize rule is the most popular of a broad variety of rules that enforce descent and provide convergence guarantees without requiring a full line minimization.

The capture theorem (Prop. 1.2.3) was formulated and proved by the author for the case of a nonsingular local minimum in [Ber82a], Prop. 1.12. It was extended to the form given here by Dunn [Dun93c].

Gradient methods with errors are discussed in Poljak [Pol87], and Bertsekas and Tsitsiklis [BeT96], [BeT00]. Parallel and asynchronous stochastic gradient methods converge under very weak conditions; see Tsitsiklis, Bertsekas, and Athans [TBA86], Bertsekas and Tsitsiklis [BeT89], Section 7.8. The convergence rate of these methods is discussed by Duchi, Chaturapruek, and Re [DCR15].

Section 1.3: For further discussion of various measures of rate of convergence, see Ortega and Rheinboldt [OrR70], Bertsekas [Ber82a], and Barzilai and Dempster [BaD93]. The convergence rate of steepest descent with line minimization was analyzed by Kantorovich [Kan45]. The case of a constant stepsize was analyzed by Goldstein [Gol64], and Levitin and Poljak [LeP65]. For analysis of the convergence rate of steepest descent for singular problems, see Dunn [Dun81], [Dun87], and Poljak [Pol87].

Section 1.4: The modern analysis of Newton's method is generally attributed to Kantorovich [Kan39], [Kan49], although the method has a long history, reviewed among others by Deuflhard [Deu12] and Ypma [Ypm95]. For an analysis of the case where the method converges to a singular point, see Decker and Kelley [DeK80], Decker, Keller, and Kelley [DKK83], and Hughes and Dunn [HuD84]. An alternative analysis, based on the notion of self-concordance, which is related to the interior point algorithms described in Chapter 5, is given by Nesterov and Nemirovskii [NeN94] (for a more accessible account, see Boyd and Vandenbergue [BoV04]).

The modification to extend the region of convergence of Newton's method by modifying the Cholesky factorization was given by Gill and Murray [GiM74]; see also Gill, Murray, and Wright [GMW81]. The use of a trust region has been discussed in the paper by Moré and Sorensen [MoS83], and in the book by Conn, Gould, and Toint [CGT00]. Extensive accounts of various aspects of Newton-like methods are given in the books by Goldstein [Gol67], Hestenes [Hes80], Gill, Murray, and Wright [GMW81], Dennis and Schnabel [DeS83], Luenberger Lue84], Nazareth [Naz94], Kelley [Kel99], Fletcher [Fle00], and Nocedal and Wright [NoW06]. The truncated Newton method is discussed in Dembo, Eisenstadt, and Steihaug [DES82], Nash [Nas85], and Nash and Sofer [NaS89].

There is a vast literature on least squares problems. They arise in many practical contexts, including statistical data analysis, where they are often referred to as *regression problems*. In addition to the Gauss-Newton method, they are also often solved with the incremental methods to be discussed in Section 2.4, particularly when m, the number of terms in the least squares sum, is very large.

References

[AFB06] Ahn, S., Fessler, J., Blatt, D., and Hero, A. O., 2006. "Convergent Incremental Optimization Transfer Algorithms: Application to Tomography," IEEE Transactions on Medical Imaging, Vol. 25, pp. 283-296.

[AHR93] Anstreicher, K. M., den Hertog, D., Roos, C., and Terlaky, T., 1993. "A Long Step Barrier Method for Convex Quadratic Programming," Algorithmica, Vol. 10, pp. 365-382.

[AHR97] Auslender, A., Cominetti, R., and Haddou, M., 1997. "Asymptotic Analysis for Penalty and Barrier Methods in Convex and Linear Programming," Math. Operations Res., Vol. 22, pp. 43-62.

[AHU58] Arrow, K. J., Hurwicz, L., and Uzawa, H., (Eds.), 1958. Studies in Linear and Nonlinear Programming, Stanford Univ. Press, Stanford, CA.

[AHU61] Arrow, K. J., Hurwicz, L., and Uzawa, H., 1961. "Constraint Qualifications in Maximization Problems," Naval Research Logistics Quarterly, Vol. 8, pp. 175-191.

[AMO91] Ahuja, R. K., Magnanti, T. L., and Orlin, J. B., 1991. "Some Recent Advances in Network Flows," SIAM Review, Vol. 33, pp. 175-219.

[AaL97] Aarts, E., and Lenstra, J. K., 1997. Local Search in Combinatorial Optimization, Wiley, N. Y.

[Aba67] Abadie, J., 1967. "On the Kuhn-Tucker Theorem," in Nonlinear Programming, Abadie, J., (Ed.), North Holland, Amsterdam.

[Ali92] Alizadeh, F., 1992. "Optimization over the Positive-Definite Cone: Interior Point Methods and Combinatorial Applications," in Pardalos, P., (Ed.), Advances in Optimization and Parallel Computing, North Holland, Amsterdam.

[Ali95] Alizadeh, F., 1995. "Interior-Point Methods in Semidefinite Programming with Applications in Combinatorial Applications," SIAM J. on Optimization, Vol. 5, pp. 13-51.

[AnH13] Andersen, M. S., and Hansen, P. C., 2013. "Generalized Row-Action Methods for Tomographic Imaging," Numerical Algorithms, Vol. 67, pp. 1-24.

[AnV94] Anstreicher, K. M., and Vial, J.-P., 1994. "On the Convergence of an Infeasible Primal-Dual Interior-Point Method for Convex Programming," Optimization Methods and Software, Vol. 3, pp. 273-283.

[Arm66] Armijo, L., 1966. "Minimization of Functions Having Continuous Partial Derivatives," Pacific J. Math., Vol. 16, pp. 1-3.

[Ash72] Ash, R. B., 1972. Real Analysis and Probability, Academic Press, N. Y.

[AtV95] Atkinson, D. S., and Vaidya, P. M., 1995. "A Cutting Plane Algorithm for Convex Programming that Uses Analytic Centers," Math. Programming, Vol. 69, pp. 1-44.

[AuC90] Auslender, A., and Cominetti, R., 1990. "First and Second Order Sensitivity Conditions," Optimization, Vol. 21, pp. 1-13.

[AuE76] Aubin, J. P., and Ekeland, I., 1976. "Estimates of the Duality Gap in Nonconvex Optimization," Math. Operations Res., Vol. 1, pp. 225-245.

[AuT03] Auslender, A., and Teboulle, M., 2003. Asymptotic Cones and Functions in Optimization and Variational Inequalities, Springer, N. Y.

[Aus76] Auslender, A., 1976. Optimization: Methodes Numeriques, Mason, Paris.

[Aus92] Auslender, A., 1992. "Asymptotic Properties of the Fenchel Dual Functional and Applications to Decomposition Properties," Vol. 73, pp. 427-449.

[Aus96] Auslender, A., 1996. "Non Coercive Optimization Problems," Math. of Operations Research, Vol. 21, pp. 769-782.

[Aus97] Auslender, A., 1997. "How to Deal with the Unbounded in Optimization: Theory and Algorithms," Math. Programing, Vol. 79, pp. 3-18.

[Avr76] Avriel, M., 1976. Nonlinear Programming: Analysis and Methods, Prentice-Hall, Englewood Cliffs, N. J.

[BCS99] Bonnans, J. F., Cominetti, R., and Shapiro, A., 1999. "Second Order Optimality Conditions Based on Parabolic Second Order Tangent Sets," SIAM J. on Optimization, Vol. 9, pp. 466-492.

[BGG84] Bertsekas, D. P., Gafni, E. M., and Gallager, R. G., 1984. "Second Derivative Algorithms for Minimum Delay Distributed Routing in Networks," IEEE Trans. on Communications, Vol. 32, pp. 911-919.

[BGI95] Burachik, R., Grana Drummond, L. M., Iusem, A. N., and Svaiter, B. F., 1995. "Full Convergence of the Steepest Descent Method with Inexact Line Searches," Optimization, Vol. 32, pp. 137-146.

[BGL06] Bonnans, J. F., Gilbert, J. C., Lemaréchal, C., and Sagastizábal, S. C., 2006. Numerical Optimization: Theoretical and Practical Aspects, Springer, N. Y.

[BGS72] Bazaraa, M. S., Goode, J. J., and Shetty, C. M., 1972. "Constraint Qualifications Revisited," Management Science, Vol. 18, pp. 567-573.

[BGT81] Bland, R. G., Goldfarb, D., and Todd, M. J., 1981. "The Ellipsoid Method: A Survey," Operations Research, Vol. 29, pp. 1039-91.

[BHG08] Blatt, D., Hero, A. O., Gauchman, H., 2008. "A Convergent Incremental Gradient Method with a Constant Step Size," SIAM J. Optimization, Vol. 18, pp. 29-51.

[BHT87] Bertsekas, D. P., Hossein, P., and Tseng, P., 1987. "Relaxation Methods for Network Flow Problems with Convex Arc Costs," SIAM J. on Control and Optimization, Vol. 25, pp. 1219-1243.

[BJS90] Bazaraa, M. S., Jarvis, J. J., and Sherali, H. D., 1990. Linear Programming and Network Flows, 2nd edition, Wiley, N. Y.

[BLY15] Bragin, M. A., Luh, P. B., Yan, J. H., Yu, N., and Stern, G. A., 2015. "Convergence of the Surrogate Lagrangian Relaxation Method," J. of Optimization Theory and Applications, Vol. 164, pp. 173-201.

[BMM95a] Ball, M. O., Magnanti, T. L., Monma, C. L., and Nemhauser, G. L., 1995. Network Models, Handbooks in OR and MS, Vol. 7, North-Holland, Amsterdam.

[BMM95b] Ball, M. O., Magnanti, T. L., Monma, C. L., and Nemhauser, G. L., 1995. Network Routing, Handbooks in OR and MS, Vol. 8, North-Holland, Amsterdam.

[BMR00] Birgin, E. G., Martinez, J. M., and Raydan, M., 2000. "Nonmonotone Spectral

Projected Gradient Methods on Convex Sets," SIAM J. on Optimization, Vol. 10, pp. 1196-1211.

[BMS99] Boltyanski, V., Martini, H., and Soltan, V., 1999. Geometric Methods and Optimization Problems, Kluwer, Boston.

[BMT90] Burke, J. V., Moré, J. J., and Toraldo, G., 1990. "Convergence Properties of Trust Region Methods for Linear and Convex Constraints," Math. Programming, Vol. 47, pp. 305-336.

[BNO03] Bertsekas, D. P., Nedić, A., and Ozdaglar, A. E., 2003. Convex Analysis and Optimization, Athena Scientific, Belmont, MA.

[BOT06] Bertsekas, D. P., Ozdaglar, A. E., and Tseng, P., 2006 "Enhanced Fritz John Optimality Conditions for Convex Programming," SIAM J. on Optimization, Vol. 16, pp. 766-797.

[BPC11] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J., 2011. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers, Now Publishers Inc, Boston, MA.

[BPT92] Bonnans, J. F., Panier, E. R., Tits, A. L., and Zhou, J. L., 1992. "Avoiding the Maratos Effect by Means of a Nonmonotone Line Search II. Inequality Constrained Problems – Feasible Iterates," SIAM J. Numer. Anal., Vol. 29, pp. 1187-1202.

[BPT97a] Bertsekas, D. P., Polymenakos, L. C., and Tseng, P., 1997. "An ϵ-Relaxation Method for Separable Convex Cost Network Flow Problems," SIAM J. on Optimization, Vol. 7, pp. 853-870.

[BPT97b] Bertsekas, D. P., Polymenakos, L. C., and Tseng, P., 1997. "Epsilon-Relaxation and Auction Methods for Separable Convex Cost Network Flow Problems," in Network Optimization, Pardalos, P. M., Hearn, D. W., and Hager, W. W., (Eds.), Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, N. Y., pp. 103-126.

[BSL14] Bergmann, R., Steidl, G., Laus, F., and Weinmann, A., 2014. "Second Order Differences of Cyclic Data and Applications in Variational Denoising," arXiv preprint arXiv:1405.5349.

[BSS93] Bazaraa, M. S., Sherali, H. D., and Shetty, C. M., 1993. Nonlinear Programming Theory and Algorithms, 2nd edition, Wiley, N. Y.

[BST14] Bolte, J., Sabach, S., and Teboulle, M., 2014. "Proximal Alternating Linearized Minimization for Nonconvex and Nonsmooth Problems," Math. Programming, Vol. 146, pp. 1-36.

[BTW82] Boggs, P. T., Tolle, J. W., and Wang, P., 1982. "On the Local Convergence of Quasi-Newton Methods for Constrained Optimization," SIAM J. on Control and Optimization, Vol. 20, pp. 161-171.

[BaB88] Barzilai, J., and Borwein, J. M., 1988. "Two-Point Step Size Gradient Methods," IMA J. Numerical Analysis, Vol. 8, pp. 141-148.

[BaC11] Bauschke, H. H., and Combettes, P. L., 2011. Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer, NY.

[BaD93] Barzilai, J., and Dempster, M. A. H., 1993. "Measuring Rates of Convergence of Numerical Algorithms," J. Opt. Theory and Appl., Vol. 78, pp. 109-125.

[BaL89] Bayer, D. A., and Lagarias, J. C., 1989. "The Nonlinear Geometry of Linear Programming. I. Affine and Projective Scaling Trajectories. II. Legendre Transform Coordinates and Central Trajectories. III. Projective Legendre Transform Coordinates and Hilbert Geometry," Trans. Amer. Math. Soc., Vol. 314, pp. 499-581.

[BaT85] Balas, E., and Toth, P., 1985. "Branch and Bound Methods," in The Traveling

Salesman Problem, Lawler, E., Lenstra, J. K., Rinnoy Kan, A. H. G., and Shmoys, D. B., (Eds.), Wiley, N. Y., pp. 361-401.

[BaW75] Balinski, M., and Wolfe, P., (Eds.), 1975. Nondifferentiable Optimization, Math. Programming Study 3, North-Holland, Amsterdam.

[Bac14] Bacak, M., 2014. "Computing Medians and Means in Hadamard Spaces," arXiv preprint arXiv:1210.2145v3.

[Bac16] Bacak, M., 2016. "A variational Approach to Stochastic Minimization of Convex Functionals," arXiv preprint arXiv:1605.03289.

[BeE88] Bertsekas, D. P., and Eckstein, J., 1988. "Dual Coordinate Step Methods for Linear Network Flow Problems," Math. Programming, Vol. 42, pp. 203-243.

[BeG82] Bertsekas, D. P., and Gafni, E., 1982. "Projection Methods for Variational Inequalities with Application to the Traffic Assignment Problem," Math. Programming Studies, Vol. 17, pp. 139-159.

[BeG83] Bertsekas, D. P., and Gafni, E., 1983. "Projected Newton Methods and Optimization of Multicommodity Flows," IEEE Trans. Automat. Control, Vol. AC-28, pp. 1090-1096.

[BeG92] Bertsekas, D. P., and Gallager, R. G., 1992. Data Networks, 2nd edition, Prentice-Hall, Englewood Cliffs, N. J.

[BeM71] Bertsekas, D. P, and Mitter, S. K., 1971. "Steepest Descent for Optimization Problems with Nondifferentiable Cost Functionals," Proc. 5th Annual Princeton Confer. Inform. Sci. Systems, Princeton, N. J., pp. 347-351.

[BeM73] Bertsekas, D. P., and Mitter, S. K., 1973. "A Descent Numerical Method for Optimization Problems with Nondifferentiable Cost Functionals," SIAM J. on Control, Vol. 11, pp. 637-652.

[BeN01] Ben-Tal, A., and Nemirovski, A., 2001. Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications, SIAM, Philadelphia.

[BeO02] Bertsekas, D. P., and Ozdaglar, A. E., 2002. "Pseudonormality and a Lagrange Multiplier Theory for Constrained Optimization," J. Opt. Th. and Appl., Vol. 114, pp. 287-343.

[BeS15] Beck, A., and Shtern, S., 2015. "Linearly Convergent Away-Step Conditional Gradient for Non-Strongly Convex Functions," arXiv preprint arXiv:1504.05002.

[BeT88] Bertsekas, D. P., and Tseng, P., 1988. "Relaxation Methods for Minimum Cost Ordinary and Generalized Network Flow Problems," Operations Research, Vol. 36, pp. 93-114.

[BeT89] Bertsekas, D. P., and Tsitsiklis, J. N., 1989. Parallel and Distributed Computation: Numerical Methods, Prentice-Hall, Englewood Cliffs, N. J; republished by Athena Scientific, Belmont, MA, 1997.

[BeT91] Bertsekas, D. P., and Tsitsiklis, J. N., 1991. "Some Aspects of Parallel and Distributed Iterative Algorithms - A Survey," Automatica, Vol. 27, pp. 3-21.

[BeT94] Bertsekas, D. P., and Tseng, P., 1994. "Partial Proximal Minimization Algorithms for Convex Programming," SIAM J. on Optimization, Vol. 4, pp. 551-572.

[BeT96] Bertsekas, D. P., and Tsitsiklis, J. N., 1996. Neuro-Dynamic Programming, Athena Scientific, Belmont, MA.

[BeT97] Bertsimas, D., and Tsitsiklis, J. N., 1997. Introduction to Linear Optimization, Athena Scientific, Belmont, MA.

[BeT00] Bertsekas, D. P., and Tsitsiklis, J. N., 2000. "Gradient Convergence of Gradient Methods with Errors," SIAM J. on Optimization, Vol. 36, pp. 627-642.

[BeT07] Bertsekas, D. P., and Tseng, P., 2007. "Set Intersection Theorems and Existence of Optimal Solutions," Mathematical Programming, Vol. 110, pp. 287-314.

[BeT08] Bertsekas, D. P., and Tsitsiklis, J. N., 2008. Introduction to Probability, 2nd Edition, Athena Scientific, Belmont, MA.

[BeT13] Beck, A., and Tetruashvili, L., 2013. "On the Convergence of Block Coordinate Descent Type Methods," SIAM J. on Optimization, Vol. 23, pp. 2037-2060.

[BeY09] Bertsekas, D. P., and Yu, H., 2009. "Projected Equation Methods for Approximate Solution of Large Linear Systems," J. of Computational and Applied Mathematics, Vol. 227, pp. 27-50.

[BeY10] Bertsekas, D. P., and Yu, H., 2010. "Asynchronous Distributed Policy Iteration in Dynamic Programming," Proc. of Allerton Conf. on Communication, Control and Computing, Allerton Park, Ill, pp. 1368-1374.

[BeY11] Bertsekas, D. P., and Yu, H., 2011. "A Unifying Polyhedral Approximation Framework for Convex Optimization," SIAM J. on Optimization, Vol. 21, pp. 333-360.

[BeZ82] Ben-Tal, A., and Zowe, J., 1982. "A Unified Theory of First and Second-Order Conditions for Extremum Problems in Topological Vector Spaces," Math. Programming Studies, Vol. 19, pp. 39-76.

[BeZ97] Ben-Tal, A., and Zibulevsky, M., 1997. "Penalty/Barrier Multiplier Methods for Convex Programming Problems," SIAM J. on Optimization, Vol. 7, pp. 347-366.

[Ben62] Benders, J. F., 1962. "Partitioning Procedures for Solving Mixed Variables Programming Problems," Numer. Math., Vol. 4, pp. 238-252.

[Ber71] Bertsekas, D. P., 1971. "Control of Uncertain Systems With a Set-Membership Description of the Uncertainty," Ph.D. Thesis, Dept. of EECS, MIT; may be downloaded from http://web.mit.edu/dimitrib/www/publ.html.

[Ber72] Bertsekas, D. P., 1972. "Stochastic Optimization Problems with Nondifferentiable Cost Functionals with an Application in Stochastic Programming," Proc. 1972 IEEE Conf. Decision and Control, pp. 555-559.

[Ber73] Bertsekas, D. P., 1973. "Stochastic Optimization Problems with Nondifferentiable Cost Functionals," J. of Optimization Theory and Applications, Vol. 12, pp. 218-231.

[Ber74] Bertsekas, D. P., 1974. "Partial Conjugate Gradient Methods for a Class of Optimal Control Problems," IEEE Trans. Automat. Control, Vol. 19, pp. 209-217.

[Ber75a] Bertsekas, D. P., 1975. "Necessary and Sufficient Conditions for a Penalty Method to be Exact," Math. Programming, Vol. 9, pp. 87-99.

[Ber75b] Bertsekas, D. P., 1975. "Combined Primal-Dual and Penalty Methods for Constrained Optimization, SIAM J. on Control, Vol. 13, pp. 521-544.

[Ber75c] Bertsekas, D. P., 1975. "Nondifferentiable Optimization Via Approximation," Math. Programming Study 3, Balinski, M., and Wolfe, P., (Eds.), North-Holland, Amsterdam, pp. 1-25.

[Ber75d] Bertsekas, D. P., 1975. "On the Method of Multipliers for Convex Programming," IEEE Transactions on Aut. Control, Vol. 20, pp. 385-388.

[Ber76a] Bertsekas, D. P., 1976. "On Penalty and Multiplier Methods for Constrained Optimization," SIAM J. on Control and Optimization, Vol. 14, pp. 216-235.

[Ber76b] Bertsekas, D. P., 1976. "Multiplier Methods: A Survey," Automatica, Vol. 12, pp. 133-145.

[Ber76c] Bertsekas, D. P., 1976. "On the Goldstein-Levitin-Poljak Gradient Projection Method," IEEE Trans. Automat. Control, Vol. 21, pp. 174-184.

[Ber77] Bertsekas, D. P., 1977. "Approximation Procedures Based on the Method of Multipliers," J. Opt. Th. and Appl., Vol. 23, pp. 487-510.

[Ber78] Bertsekas, D. P., 1978. "Local Convex Conjugacy and Fenchel Duality," Preprints of Triennial World Congress of IFAC, Helsinki, Vol. 2, pp. 1079-1084.

[Ber79a] Bertsekas, D. P., 1979. "Convexification Procedures and Decomposition Algorithms for Large-Scale Nonconvex Optimization Problems," J. Opt. Th. and Appl., Vol. 29, pp. 169-197.

[Ber79b] Bertsekas, D. P., 1979. "A Distributed Algorithm for the Assignment Problem," Lab. for Information and Decision Systems Working Paper, M.I.T.

[Ber80a] Bertsekas, D. P., 1980. "A Class of Optimal Routing Algorithms for Communication Networks," Proc. of the Fifth International Conference on Computer Communication, Atlanta, Ga., pp. 71-76.

[Ber80b] Bertsekas, D. P., 1980. "Variable Metric Methods for Constrained Optimization Based on Differentiable Exact Penalty Functions," Proc. Allerton Conference on Communication, Control, and Computation, Allerton Park, Ill., pp. 584-593.

[Ber81] Bertsekas, D. P., 1981. "A New Algorithm for the Assignment Problem," Math. Programming, Vol. 21, pp. 152-171.

[Ber82a] Bertsekas, D. P., 1982. Constrained Optimization and Lagrange Multiplier Methods, Academic Press, N. Y.; republished by Athena Scientific, Belmont, MA, 1997.

[Ber82b] Bertsekas, D. P., 1982. "Projected Newton Methods for Optimization Problems with Simple Constraints," SIAM J. on Control and Optimization, Vol. 20, pp. 221-246.

[Ber82c] Bertsekas, D. P., 1982. "Enlarging the Region of Convergence of Newton's Method for Constrained Optimization," J. Opt. Th. and Appl., Vol. 36, pp. 221-252.

[Ber82d] Bertsekas, D. P., 1982. "Distributed Dynamic Programming," IEEE Trans. Aut. Control, Vol. AC-27, pp. 610-616.

[Ber82e] Bertsekas, D. P., 1982. "Notes on Nonlinear Programming and Discrete-Time Optimal Control," Lab. for Information and Decision Systems Report LIDS-P-919, MIT.

[Ber83] Bertsekas, D. P., 1983. "Distributed Asynchronous Computation of Fixed Points," Math. Programming, Vol. 27, pp. 107-120.

[Ber85] Bertsekas, D. P., 1985. "A Unified Framework for Minimum Cost Network Flow Problems," Math. Programming, Vol. 32, pp. 125-145.

[Ber86] Bertsekas, D. P., 1986. "Distributed Relaxation Methods for Linear Network Flow Problems," Proceedings of 25th IEEE Conference on Decision and Control, pp. 2101-2106.

[Ber91] Bertsekas, D. P., 1991. Linear Network Optimization: Algorithms and Codes, M.I.T. Press, Cambridge, MA.

[Ber92] Bertsekas, D. P., 1992. "Auction Algorithms for Network Problems: A Tutorial Introduction," Computational Optimization and Applications, Vol. 1, pp. 7-66.

[Ber96a] D. P. Bertsekas, 1996. "Thevenin Decomposition and Network Optimization," J. Opt. Theory and Appl., Vol. 89, pp. 1-15.

[Ber96b] Bertsekas, D. P., 1996. "Incremental Least Squares Methods and the Extended Kalman Filter," SIAM J. on Optimization, Vol. 6, pp. 807-822.

[Ber97] Bertsekas, D. P., 1997. "A New Class of Incremental Gradient Methods for Least Squares Problems," SIAM J. on Optimization, Vol. 7, pp. 913-926.

[Ber98] Bertsekas, D. P., 1998. Network Optimization: Continuous and Discrete Models, Athena Scientific, Belmont, MA.

[Ber99] Bertsekas, D. P., 1999. "A Note on Error Bounds for Convex and Nonconvex Problems," Computational Optimization and Applications, Vol. 12, pp. 41-51.

[Ber05a] Bertsekas, D. P., 2005. Dynamic Programming and Optimal Control, 3rd Edition, Vol. I, Athena Scientific, Belmont, MA.

[Ber05b] Bertsekas, D. P., 2005. "Dynamic Programming and Suboptimal Control: A Survey from ADP to MPC," Fundamental Issues in Control, Special Issue for the CDC-ECC-05, European J. of Control, Vol. 11, Nos. 4-5.

[Ber05c] Bertsekas, D. P., 2005. "Lagrange Multipliers with Optimal Sensitivity Properties in Constrained Optimization," Lab. for Information and Decision Systems Report 2632, MIT; in Proc. of the 2004 Erice Workshop on Large Scale Nonlinear Optimization, Erice, Italy, Kluwer.

[Ber09] Bertsekas, D. P., 2009. Convex Optimization Theory, Athena Scientific, Belmont, MA.

[Ber10a] Bertsekas, D. P., 2010. "Extended Monotropic Programming and Duality," Lab. for Information and Decision Systems Report LIDS-P-2692, MIT, March 2006, corrected in Feb. 2010.

[Ber10b] Bertsekas, D. P., 2010. "Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey," Lab. for Information and Decision Systems Report LIDS-P-2848, MIT.

[Ber11a] Bertsekas, D. P., 2011. "Incremental Proximal Methods for Large Scale Convex Optimization," Math. Programming, Vol. 129, pp. 163-195.

[Ber11b] Bertsekas, D. P., 2011. "Centralized and Distributed Newton Methods for Network Optimization and Extensions," Lab. for Information and Decision Systems Report LIDS-P-2866, MIT.

[Ber12] Bertsekas, D. P., 2012. Dynamic Programming and Optimal Control: Approximate Dynamic Programming, 4th Edition, Vol. II, Athena Scientific, Belmont, MA.

[Ber13] Bertsekas, D. P., 2013. Abstract Dynamic Programming, Athena Scientific, Belmont, MA.

[Ber15a] Bertsekas, D. P., 2015. Convex Optimization Algorithms, Athena Scientific, Belmont, MA.

[Ber15b] Bertsekas, D. P., 2015. "Incremental Aggregated Proximal and Augmented Lagrangian Algorithms," Lab. for Information and Decision Systems Report LIDS-P-3176, MIT, September 2015.

[BiL97] Birge, J. R., and Louveaux, 1997. Introduction to Stochastic Programming, Springer-Verlag, New York, N. Y.

[Bia15] Bianchi, P., 2015. "Ergodic Convergence of a Stochastic Proximal Point Algorithm," arXiv preprint arXiv:1504.05400.

[Bis95] Bishop, C. M, 1995. Neural Networks for Pattern Recognition, Oxford University Press, N. Y.

[BoL00] Borwein, J. M., and Lewis, A. S., 2000. Convex Analysis and Nonlinear Optimization, Springer-Verlag, N. Y.

[BoL05] Bottou, L., and LeCun, Y., 2005. "On-Line Learning for Very Large Datasets," Applied Stochastic Models in Business and Industry, Vol. 21, pp. 137-151.

[BoS00] Bonnans, J. F., and Shapiro, A., 2000. Perturbation Analysis of Optimization Problems, Springer-Verlag, N. Y.

[BoT80] Boggs, P. T., and Tolle, J. W., 1980. "Augmented Lagrangians which are Quadratic in the Multiplier," J. Opt. Th. and Appl., Vol. 31, pp. 17-26.

[BoV04] Boyd, S., and Vandenbergue, L., 2004. Convex Optimization, Cambridge Univ. Press, Cambridge, U.K.

[Bog90] Bogart, K. P., 1990. Introductory Combinatorics, Harcourt Brace Jovanovich, Inc., New York, N. Y.

[Bon89a] Bonnans, J. F., 1989. "A Variant of a Projected Variable Metric Method for Bound Constrained Optimization Problems," Report, INRIA, France.

[Bon89b] Bonnans, J. F., 1989. "Asymptotic Admissibility of the Unit Stepsize in Exact Penalty Methods," SIAM J. on Control and Optimization, Vol. 27, pp. 631-641.

[Bon92] Bonnans, J. F., 1992. "Directional Derivatives of Optimal Solutions in Smooth Nonlinear Programming," J. Opt. Theory and Appl., Vol. 73, pp. 27-45.

[Bon94] Bonnans, J. F., 1994. "Local Analysis of Newton Type Methods for Variational Inequalities and Nonlinear Programming," J. Applied Math. Optimization, Vol. 29, pp. 161-186.

[Bor08] Borkar, V. S., 2008. Stochastic Approximation: A Dynamical Systems Viewpoint, Cambridge Univ. Press.

[Brä93] Brännlund, U., 1993. "On Relaxation Methods for Nonsmooth Convex Optimization," Doctoral Thesis, Royal Institute of Technology, Stockholm, Sweden.

[Bro70] Broyden, C. G., 1970. "The Convergence of a Class of Double Rank Minimization Algorithms," J. Inst. Math. Appl., Vol. 6, pp. 76-90.

[BuM88] Burke, J. V., and Moré, J. J., 1988. "On the Identification of Active Constraints," SIAM J. Numer. Anal., Vol. 25, pp. 1197-1211.

[BuQ98] Burke, J. V., and Qian, M., 1998. "A Variable Metric Proximal Point Algorithm for Monotone Operators," SIAM J. on Control and Optimization, Vol. 37, pp. 353-375.

[CCP70] Canon, M. D., Cullum, C. D., and Polak, E., 1970. Theory of Optimal Control and Mathematical Programming, McGraw-Hill, N. Y.

[CCP98] Cook, W., Cunningham, W., Pulleyblank, W., and Schrijver, A., 1998. Combinatorial Optimization, Wiley, N. Y.

[CFM75] Camerini, P. M., Fratta, L., and Maffioli, F., 1975. "On Improving Relaxation Methods by Modified Gradient Techniques," Math. Programming Studies, Vol. 3, pp. 26-34.

[CGT91] Conn, A. R., Gould, N. I. M., and Toint, P. L., 1991. "A Globally Convergent Augmented Lagrangian Algorithm for Optimization with General Constraints and Simple Bounds," SIAM J. Numer. Anal., Vol. 28, pp. 545-572.

[CGT92] Conn, A. R., Gould, N. I. M., and Toint, P. L., 1992. "LANCELOT: A FORTRAN Package for Large-Scale Nonlinear Optimization," Springer-Verlag, N. Y.

[CGT00] Conn, A. R., Gould, N. I. M., and Toint, P. L., 2000. Trust Region Methods, SIAM, Philadelphia, PA.

[CHY16] Chen, C., He, B., Ye, Y., and Yuan, X., 2016. "The Direct Extension of ADMM for Multi-Block Convex Minimization Problems is not Necessarily Convergent," Math. Programming, Series A, Vol. 155, pp. 57-79.

[CPS92] Cottle, R., Pang, J. S., and Stone, R. E., 1992. The Linear Complementarity Problem, Academic Press, Boston.

[CPS11] Choi, S. C. T., Paige, C. C., and Saunders, M. A., 2011. "MINRES-QLP: A Krylov Subspace Method for Indefinite or Singular Symmetric Systems," SIAM Journal on Scientific Computing, Vol. 33, pp. 1810-1836.

[CaC68] Canon, M. D., and Cullum, C. D., 1968. "A Tight Upper Bound on the Rate of Convergence of the Frank-Wolfe Algorithm," SIAM J. on Control, Vol. 6, pp. 509-516.

[CaF97] Caprara, A., and Fischetti, M., 1997. "Branch and Cut Algorithms," in Annotated Bibliographies in Combinatorial Optimization, Dell'Amico, M., Maffioli, F., and Martello, S., (Eds.), Wiley, Chisester, Chapter 4.

[CaG74] Cantor, D. G., Gerla, M., 1974. "Optimal Routing in Packet Switched Computer Networks," IEEE Trans. on Computing, Vol. C-23, pp. 1062-1068.

[CaM87] Calamai, P. H., and Moré, J. J., 1987. "Projected Gradient Methods for Linearly Constrained Problems," Math. Programming, Vol. 39, pp. 98-116.

[Cam94] Cameron, P. J., 1994. Combinatorics: Topics, Techniques, Algorithms, Cambridge Univ. Press.

[Car61] Carroll, C. W., 1961. "The Created Response Surface Technique for Optimizing Nonlinear Restrained Systems," Operations Research, Vol. 9, pp. 169-184.

[Cau47] Cauchy, M. A., 1847. "Analyse Mathématique–Méthode Générale Pour La Résolution des Systémes d' Eq́uations Simultanées," Comptes Vendus Acad. Sc., Paris.

[CeH87] Censor, Y., and Herman, G. T., 1987. "On Some Optimization Techniques in Image Reconstruction from Projections," Applied Numer. Math., Vol. 3, pp. 365-391.

[CeZ92] Censor, Y., and Zenios, S. A., 1992. "The Proximal Minimization Algorithm with D-Functions," J. Opt. Theory and Appl., Vol. 73, pp. 451-464.

[CeZ97] Censor, Y., and Zenios, S. A., 1997. Parallel Optimization: Theory, Algorithms, and Applications, Oxford University Press, N. Y.

[ChG59] Cheney, E. W., and Goldstein, A. A., 1959. "Newton's Method for Convex Programming and Tchebycheff Approximation," Numer. Math., Vol. I, pp. 253-268.

[ChT93] Chen, G., and Teboulle, M., 1993. "Convergence Analysis of a Proximal-Like Minimization Algorithm Using Bregman Functions," SIAM J. on Optimization, Vol. 3, pp. 538-543.

[ChT94] Chen, G., and Teboulle, M., 1994. "A Proximal-Based Decomposition Method for Convex Minimization Problems," Math. Programming, Vol. 64, pp. 81-101.

[Cla83] Clarke, F. H., 1983. Nonsmooth Analysis and Optimization, Wiley-Interscience, N. Y.

[CoC82a] Coleman, T. F., and Conn, A. R., 1982. "Nonlinear Programming Via an Exact Penalty Function: Asymptotic Analysis," Math. Progamming, Vol. 24, pp. 123-136.

[CoC82b] Coleman, T. F., and Conn, A. R., 1982. "Nonlinear Programming Via an Exact Penalty Function: Global Analysis," Math. Programming, Vol. 24, pp. 137-161.

[CoL94] Correa, R., and Lemarechal, C., 1994. "Convergence of Some Algorithms for Convex Minimization," Math. Programming, Vol. 62, pp. 261-276.

[CoT13] Couellan, N. P., and Trafalis, T. B., 2013. "On-line SVM Learning via an

Incremental Primal-Dual Technique," Optimization Methods and Software, Vol. 28, pp. 256-275.

[Coh80] Cohen, G., 1980. "Auxiliary Problem Principle and Decomposition of Optimization Problems," J. Opt. Theory and Appl., Vol. 32, pp. 277-305.

[Cro58] Croes, G. A., 1958. "A Method for Solving Traveling Salesman Problems," Operations Research, Vol. 6, pp. 791-812.

[Cry71] Cryer, C. W., 1971. "The Solution of a Quadratic Programming Problem Using Systematic Overrelaxation," SIAM J. on Control, Vol. 9, pp. 385-392.

[Cul71] Cullum, J., 1971. "An Explicit Procedure for Discretizing Continuous Optimal Control Problems," J. Opt. Theory and Appl., Vol. 8, pp. 15-34.

[Cyb89] Cybenko, 1989. "Approximation by Superpositions of a Sigmoidal Function," Math. of Control, Signals, and Systems, Vol. 2, pp. 303-314.

[DCD14] Defazio, A. J., Caetano, T. S., and Domke, J., 2014. "Finito: A Faster, Permutable Incremental Gradient Method for Big Data Problems," Proceedings of the 31st ICML, Beijing.

[DCR15] Duchi, J. C., Chaturapruek, S., and Re, R., 2015. "Asynchronous Stochastic Convex Optimization," arXiv preprint arXiv:1508.00882.

[DES82] Dembo, R. S., Eisenstadt, S. C., and Steihaug, T., 1982. "Inexact Newton Methods," SIAM J. Numer. Anal., Vol. 19, pp. 400-408.

[DFJ54] Dantzig, G. B., Fulkerson, D. R., and Johnson, S. M., 1954. "Solution of a Large-Scale Traveling-Salesman Problem," Operations Research, Vol. 2, pp. 393-410.

[DHS06] Dai, Y. H., Hager, W. W., Schittkowski, K., and Zhang, H., 2006. "The Cyclic Barzilai-Borwein Method for Unconstrained Optimization," IMA J. of Numerical Analysis, Vol. 26, pp. 604-627.

[DKK83] Decker, D. W., Keller, H. B., and Kelley, C. T., 1983. "Convergence Rates for Newton's Method at Singular Points," SIAM J. Numer. Anal., Vol. 20, pp. 296-314.

[DaW60] Dantzig, G. B., and Wolfe, P., 1960. "Decomposition Principle for Linear Programs," Operations Research, Vol. 8, pp. 101-111.

[DaY14a] Davis, D., and Yin, W., 2014. "Convergence Rate Analysis of Several Splitting Schemes," arXiv preprint arXiv:1406.4834.

[DaY14b] Davis, D., and Yin, W., 2014. "Convergence Rates of Relaxed Peaceman-Rachford and ADMM Under Regularity Assumptions, arXiv preprint arXiv:1407.5210.

[Dan67] Danskin, J. M., 1967. The Theory of Max-Min and its Application to Weapons Allocation Problems, Springer, NY.

[Dan71] Daniel, J. W., 1971. The Aproximate Minimization of Functionals, Prentice-Hall, Englewood Cliffs, N. J.

[Dav59] Davidon, W. C., 1959. "Variable Metric Method for Minimization," Argonne National Lab., Report ANL-5990 (Rev.), Argonne, Ill. Reprinted with a new preface in SIAM J. on Optimization, Vol. 1, 1991, pp. 1-17.

[Dav76] Davidon, W. C., 1976. "New Least Squares Algorithms," J. Opt. Theory and Appl., Vol. 18, pp. 187-197.

[DeK80] Decker, D. W., and Kelley, C. T., 1980. "Newton's Method at Singular Points, Parts I and II," SIAM J. Numer. Anal., Vol. 17, pp. 66-70, 465-471.

[DeM77] Dennis, J. E., and Moré, J. J., 1977. "Quasi-Newton Methods: Motivation and Theory," SIAM Review, Vol. 19, pp. 46-89.

[DeR70] Demjanov, V. F., and Rubinov, A. M., 1970. Approximate Methods in Optimization Problems, American Elsevier, N. Y.

[DeS83] Dennis, J. E., and Schnabel, R. E., 1983. Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Prentice-Hall, Englewood Cliffs, N. J.

[DeT83] Dembo, R. S., and Tulowitzki, U., 1983. "On the Minimization of Quadratic Functions Subject to Box Constraints," Working Paper Series B No. 71, School of Organization and Management, Yale Univ., New Haven, Conn.

[DeT91] Dennis, J. E., and Torczon, V., 1991. "Direct Search Methods on Parallel Machines," SIAM J. on Optimization, Vol. 1, pp. 448-474.

[DeT93] De Angelis, P. L., and Toraldo, G., 1993. "On the Identification Property of a Projected Gradient Method," SIAM J. Numer. Anal., Vol. 30, pp. 1483-1497.

[DeV85] Demjanov, V. F., and Vasilév, L. V., 1985. Nondifferentiable Optimization, Optimization Software, N. Y.

[Del12] Delfour, M. C., 2012. Introduction to Optimization and Semidifferential Calculus, SIAM, Phila.

[Deu12] Deuflhard, P., 2012. "A Short History of Newton's Method," Documenta Mathematica, Optimization Stories, pp. 25-30.

[DiG79] DiPillo, G., and Grippo, L., 1979. "A New Class of Augmented Lagrangians in Nonlinear Programming," SIAM J. on Control and Optimization, Vol. 17, pp. 618-628.

[DiG89] DiPillo, G., and Grippo, L., 1989. "Exact Penalty Functions in Constrained Optimization," SIAM J. on Control and Optimization, Vol. 27, pp. 1333-1360.

[Dix72a] Dixon, L. C. W., 1972. "Quasi-Newton Algorithms Generate Identical Points," Math. Programming, Vol. 2, pp. 383-387.

[Dix72b] Dixon, L. C. W., 1972. "Quasi-Newton Algorithms Generate Identical Points. II. The Proofs of Four New Theorems," Math. Programming, Vol. 3, pp. 345-358.

[DoJ86] Dontchev, A. L., and Jongen, H. Th., 1986. "On the Regularity of the Kuhn-Tucker Curve," SIAM J. on Control and Optimization, Vol. 24, pp. 169-176.

[DoR09] Dontchev, A. L., and Rockafellar, R. T., 2009. Implicit Functions and Solution Mappings," 2nd edition, Springer, N. Y.

[DuB89] Dunn, J. C., and Bertsekas, D. P., 1989. "Efficient Dynamic Programming Implementations of Newton's Method for Unconstrained Optimal Control Problems," J. Opt. Theory and Appl., Vol. 63, pp. 23-38.

[DuM65] Dubovitskii, M. D., and Milyutin, A. A., 1965. "Extremum Problems in the Presence of Restriction," USSR Comp. Math. and Math. Phys., Vol. 5, pp. 1-80.

[DuS83] Dunn, J. C., and Sachs, E., 1983. "The Effect of Perturbations on the Convergence Rates of Optimization Algorithms," Appl. Math. Optim., Vol. 10, pp. 143-157.

[DuZ89] Du, D.-Z., and Zhang, X.-S., 1989. "Global Convergence of Rosen's Gradient Projection Method," Math. Programming, Vol. 44, pp. 357-366.

[Dun79] Dunn, J. C., 1979. "Rates of Convergence for Conditional Gradient Algorithms Near Singular and Nonsingular Extremals," SIAM J. on Control and Optimization, Vol. 17, pp. 187-211.

[Dun80a] Dunn, J. C., 1980. "Convergence Rates for Conditional Gradient Sequences Generated by Implicit Step Length Rules," SIAM J. on Control and Optimization, Vol. 18, pp. 473-487.

[Dun80b] Dunn, J. C., 1980. "Newton's Method and the Goldstein Step Length Rule for

Constrained Minimization Problems," SIAM J. on Control and Optimization, Vol. 18, pp. 659-674.

[Dun81] Dunn, J. C., 1981. "Global and Asymptotic Convergence Rate Estimates for a Class of Projected Gradient Processes," SIAM J. on Control and Optimization, Vol. 19, pp. 368-400.

[Dun87] Dunn, J. C., 1987. "On the Convergence of Projected Gradient Processes to Singular Critical Points," J. Opt. Theory and Appl., Vol. 55, pp. 203-216.

[Dun88a] Dunn, J. C., 1988. "Gradient Projection Methods for Systems Optimization Problems," Control and Dynamic Systems, Vol. 29, pp. 135-195.

[Dun88b] Dunn, J. C., 1988. "A Projected Newton Method for Minimization Problems with Nonlinear Inequality Constraints," Numer. Math., Vol. 53, pp. 377-409.

[Dun91a] Dunn, J. C., 1991. "Scaled Gradient Projection Methods for Optimal Control Problems and Other Structured Nonlinear Programs," in New Trends in Systems Theory, Conte, G., et al, (Eds.), Birkhöuser, Boston, MA.

[Dun91b] Dunn, J. C., 1991. "A Subspace Decomposition Principle for Scaled Gradient Projection Methods: Global Theory," SIAM J. on Control and Optimization, Vol. 29, pp. 219-246.

[Dun93a] Dunn, J. C., 1993. "A Subspace Decomposition Principle for Scaled Gradient Projection Methods: Local Theory," SIAM J. on Control and Optimization, Vol. 31, pp. 219-246.

[Dun93b] Dunn, J. C., 1993. "Second-Order Multiplier Update Calculations for Optimal Control Problems and Related Large Scale Nonlinear Programs," SIAM J. on Optimization, Vol. 3, pp. 489-502.

[Dun93c] Dunn, J. C., 1993. Private Communication.

[Dun94] Dunn, J. C., 1994. "Gradient-Related Constrained Minimization Algorithms in Function Spaces: Convergence Properties and Computational Implications," in Large Scale Optimization: State of the Art, Hager, W. W., Hearn, D. W., and Pardalos, P. M., (Eds.), Kluwer, Boston.

[Eas58] Eastman, W. L., 1958. Linear Programming with Pattern Constraints, Ph.D. Thesis, Harvard University, Cambridge, MA.

[EcB90] Eckstein, J., and Bertsekas, D. P., 1990. "An Alternating Direction Method for Linear Programming," Report LIDS-P-1967, Lab. for Info. and Dec. Systems, M.I.T.

[EcB92] Eckstein, J., and Bertsekas, D. P., 1992. "On the Douglas-Rachford Splitting Method and the Proximal Point Algorithm for Maximal Monotone Operators," Math. Programming, Vol. 55, pp. 293-318.

[Eck94a] Eckstein, J., 1994. "Nonlinear Proximal Point Algorithms Using Bregman Functions, with Applications to Convex Programming," Math. Operations Res., Vol. 18, pp. 202-226.

[Eck94b] Eckstein, J., 1994. "Parallel Alternating Direction Multiplier Decomposition of Convex Programs," J. Opt. Theory and Appl., Vol. 80, pp. 39-62.

[EkT76] Ekeland, I., and Teman, R., 1976. Convex Analysis and Variational Problems, North-Holland Publ., Amsterdam.

[ElM75] Elzinga, J., and Moore, T. G., 1975. "A Central Cutting Plane Algorithm for the Convex Programming Problem," Math. Programming, Vol. 8, pp. 134-145.

[Eve63] Everett, H., 1963. "Generalized Lagrange Multiplier Method for Solving Problems of Optimal Allocation of Resources," Operations Research, Vol. 11, pp. 399-417.

[Evt85] Evtushenko, Y. G., 1985. Numerical Optimization Techniques, Optimization Software, N. Y.

[FAJ14] Feyzmahdavian, H. R., Aytekin, A., and Johansson, M., 2014. "A Delayed Proximal Gradient Method with Linear Convergence Rate," in Prop. of 2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1-6.

[FGK73] Fratta, L., Gerla, M., and Kleinrock, L., 1973. "The Flow Deviation Method: An Approach to Store-and-Forward Communication Network Design," Networks, Vol. 3, pp. 97-133.

[FGW02] Forsgren, A., Gill, P. E., and Wright, M. H., 2002. "Interior Methods for Nonlinear Optimization," SIAM Review, Vol. 44, pp. 525-597.

[FaF63] Fadeev, D. K., and Fadeeva, V. N., 1963. Computational Methods of Linear Algebra, Freeman, San Francisco, CA.

[FaP03] Facchinei, F., and Pang, J.-S., 2003. Finite-Dimensional Variational Inequalities and Complementarity Problems, Springer Verlag, N. Y.

[Fab73] Fabian, V., 1973. "Asymptotically Efficient Stochastic Approximation: The RM Case," Ann. Statist., Vol. 1, pp. 486-495.

[FeM91] Ferris, M. C., and Mangasarian, O. L., 1991. "Parallel Constraint Distribution," SIAM J. on Optimization, Vol. 1, pp. 487-500.

[Fen49] Fenchel, W., 1949. "On Conjugate Convex Functions," Canad. J. Math., Vol. 1, pp. 73-77.

[Fen51] Fenchel, W., 1951. "Convex Cones, Sets, and Functions," Mimeographed Notes, Princeton Univ.

[Fey16] Feyzmahdavian, H. R., 2016. Performance Analysis of Positive Systems and Optimization Algorithms with Time-Delays, Doctoral Thesis, KTH, Sweden.

[FiM68] Fiacco, A. V., and McCormick, G. P., 1968. Nonlinear Programming: Sequential Unconstrained Minimization Techniques, Wiley, N. Y.

[Fia83] Fiacco, A. V., 1983. Introduction to Sensitivity and Stability Analysis in Nonlinear Programming, Academic Press, N. Y.

[FlH95] Florian, M. S., and Hearn, D., 1995. "Network Equilibrium Models and Algorithms," Handbooks in OR and MS, Ball, M. O., Magnanti, T. L., Monma, C. L., and Nemhauser, G. L., (Eds.), Vol. 8, North-Holland, Amsterdam, pp. 485-550.

[FlP63] Fletcher, R., and Powell, M. J. D., 1963. "A Rapidly Convergent Descent Algorithm for Minimization," Comput. J., Vol. 6, pp. 163-168.

[FlP95] Floudas, C., and Pardalos, P. M., (Eds.), 1995. State of the Art in Global Optimization: Computational Methods and Applications, Kluwer, Boston.

[Fla92] Flam, S. D., 1992. "On Finite Convergence and Constraint Identification of Subgradient Projection Methods," Math. Programming, Vol. 57, pp 427-437.

[Fle70a] Fletcher, R., 1970. "A New Approach to Variable Metric Algorithms," Computer J., Vol. 13, pp. 317-322.

[Fle70b] Fletcher, R., 1970. "A Class of Methods for Nonlinear Programming with Termination and Convergence Properties," in Integer and Nonlinear Programming, Abadie, J., (Ed.), pp. 157-173, North-Holland Publ., Amsterdam.

[Fle00] Fletcher, R., 2000. Practical Methods of Optimization, 2nd edition, Wiley, NY.

[Flo95] Floudas, C. A., 1995. Nonlinear and Mixed-Integer Optimization: Fundamentals and Applications, Oxford University Press, N. Y.

[FoG83] Fortin, M., and Glowinski, R., (Eds.), 1983. Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems, North-Holland, Amsterdam.

[FoS11] Fong, D. C. L., and Saunders, M., 2011. "LSMR: An Iterative Algorithm for Sparse Least-Squares Problems," SIAM Journal on Scientific Computing, Vol. 33, pp. 2950-2971.

[FoS12] Fong, D. C. L., and Saunders, M., 2012. "CG Versus MINRES: An Empirical Comparison," SQU Journal for Science, Vol. 17, pp. 44-62.

[FrG16] Freund, R., and Grigas, P., 2016. "New Analysis and Results for the Frank-Wolfe Method," Math. Programming, Series A, Vol. 155, pp. 199-230.

[FrW56] Frank, M., and Wolfe, P., 1956. "An Algorithm for Quadratic Programming," Naval Research Logistics Quarterly, Vol. 3, pp. 95-110.

[Fra12] Frauendorfer, K., 2012. Stochastic Two-Stage Programming, Springer, N. Y.

[Fre91] Freund, R. M., 1991. "Theoretical Efficiency of a Shifted Barrier Function Algorithm in Linear Programming," Linear Algebra and Appl., Vol. 152, pp. 19-41.

[Fri56] Frisch, M. R., 1956. "La Resolution des Problemes de Programme Lineaire par la Methode du Potential Logarithmique," Cahiers du Seminaire D'Econometrie, Vol. 4, pp. 7-20.

[Fuk92] Fukushima, M., 1992. "Application of the Alternating Direction Method of Multipliers to Separable Convex Programming," Comp. Opt. and Appl., Vol. 1, pp. 93-111.

[GFJ15] Ghadimi, E., Feyzmahdavian, H. R., and Johansson, M., 2015. "Global Convergence of the Heavy-Ball Method for Convex Optimization," in European Control Conference (ECC), pp. 310-315.

[GHV92] Goffin, J. L., Haurie, A., and Vial, J. P., 1992. "Decomposition and Nondifferentiable Optimization with the Projective Algorithm," Management Science, Vol. 38, pp. 284-302.

[GKT51] Gale, D., Kuhn, H. W., and Tucker, A. W., 1951. "Linear Programming and the Theory of Games," in Activity Analysis of Production and Allocation, Koopmans, T. C., (Ed.), Wiley, N. Y.

[GKX10] Gupta, M. D., Kumar, S., and Xiao, J. 2010. "L1 Projections with Box Constraints," arXiv preprint arXiv:1010.0141.

[GLL91] Grippo, L., Lampariello, F., and Lucidi, S., 1991. "A Class of Nonmonotone Stabilization Methods in Unconstrained Minimization," Numer. Math., Vol. 59, pp. 779-805.

[GLY94] Goffin, J. L., Luo, Z.-Q., and Ye, Y., 1994. "On the Complexity of a Column Generation Algorithm for Convex or Quasiconvex Feasibility Problems," in Large Scale Optimization: State of the Art, Hager, W. W., Hearn, D. W., and Pardalos, P. M., (Eds.), Kluwer, Boston.

[GLY96] Goffin, J. L., Luo, Z.-Q., and Ye, Y., 1996. "Complexity Analysis of an Interior Cutting Plane Method for Convex Feasibility Problems," SIAM J. on Optimization, Vol. 6, pp. 638-652.

[GMW81] Gill, P. E., Murray, W., and Wright, M. H., 1981. Practical Optimization, Academic Press, N. Y.

[GMW91] Gill, P. E., Murray, W., and Wright, M. H., 1991. Numerical Linear Algebra and Optimization, Vol. I, Addison-Wesley, Redwood City, CA.

[GOP15a] Gurbuzbalaban, M., Ozdaglar, A., and Parrilo, P., 2015. "On the Convergence Rate of Incremental Aggregated Gradient Algorithms," arXiv preprint arXiv:1506.02081.

[GOP15b] Gurbuzbalaban, M., Ozdaglar, A., and Parrilo, P., 2015. "Convergence Rate of Incremental Gradient and Newton Methods," arXiv preprint arXiv:1510.08562.

[GOP15c] Gurbuzbalaban, M., Ozdaglar, A., and Parrilo, P., 2015. "Why Random Reshuffling Beats Stochastic Gradient Descent," arXiv preprint arXiv:1510.08560.

[GaB82] Gafni, E. M., and Bertsekas, D. P., 1982. "Convergence of a Gradient Projection Method," Report LIDS-P-1201, Lab. for Info. and Dec. Systems, M.I.T.

[GaB84] Gafni, E. M., and Bertsekas, D. P., 1984. "Two-Metric Projection Methods for Constrained Optimization," SIAM J. on Control and Optimization, Vol. 22, pp. 936-964.

[GaD88] Gawande, M., and Dunn, J. C., 1988. "Variable Metric Gradient Projection Processes in Convex Feasible Sets Defined by Nonlinear Inequalities," Appl. Math. Optim., Vol. 17, pp. 103-119.

[GaJ88] Gauvin, J., and Janin, R., 1988. "Directional Behavior of Optimal Solutions in Nonlinear Mathematical Programming," Math. of Operations Res., Vol. 13, pp. 629-649.

[GaM76] Gabay, D., and Mercier, B., 1976. "A Dual Algorithm for the Solution of Nonlinear Variational Problems via Finite-Element Approximations," Comp. Math. Appl., Vol. 2, pp. 17-40.

[GaM92] Gaudioso, M., and Monaco, M. F, 1992. "Variants to the Cutting Plane Approach for Convex Nondifferentiable Optimization," Optimization, Vol. 25, pp. 65-75.

[Gab79] Gabay, D., 1979. Methodes Numeriques pour l'Optimization Non Lineaire, These de Doctorat d'Etat et Sciences Mathematiques, Univ. Pierre at Marie Curie (Paris VI).

[Gab82] Gabay, D., 1982. "Reduced Quasi-Newton Methods with Feasibility Improvement for Nonlinearly Constrained Optimization," Math. Programming Studies, Vol. 16, pp. 18-44.

[Gab83] Gabay, D., 1983. "Applications of the Method of Multipliers to Variational Inequalities," in M. Fortin and R. Glowinski, eds., Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems, North-Holland, Amsterdam.

[Gai94] Gaivoronski, A. A., 1994. "Convergence Analysis of Parallel Backpropagation Algorithm for Neural Networks," Optimization Methods and Software, Vol. 4, pp. 117-134.

[Gal77] Gallager, R. G., 1977. "A Minimum Delay Routing Algorithm Using Distributed Computation," IEEE Trans. on Communications, Vol. 25, pp. 73-85.

[Gau77] Gauvin, J., 1977. "A Necessary and Sufficient Condition to Have Bounded Multipliers in Convex Programming," Math. Programming., Vol. 12, pp. 136-138.

[Geo70] Geoffrion, A. M., 1970. "Elements of Large-Scale Mathematical Programming, I, II," Management Science, Vol. 16, pp. 652-675, 676-691.

[Geo74] Geoffrion, A. M., 1974. "Lagrangian Relaxation for Integer Programming," Math. Programming Studies, Vol. 2, pp. 82-114.

[Geo77] Geoffrion, A. M., 1977. "Objective Function Approximations in Mathematical Programming," Math. Programming, Vol. 13, pp. 23-27.

[GiK95] Gilmore, P., and Kelley, C. T., 1995. "An Implicit Filtering Algorithm for Optimization of Functions with Many Local Minima," SIAM J. on Optimization, Vol. 5, pp. 269-285.

[GiM74] Gill, P. E., and Murray, W., (Eds.), 1974. Numerical Methods for Constrained Optimization, Academic Press, N. Y.

[GlL97] Glover, F., and Laguna, M., 1997. Tabu Search, Kluwer, Boston.

[GlM75] Glowinski, R. and Marrocco, A., 1975. "Sur l' Approximation par Elements Finis d' Ordre un et la Resolution par Penalisation-Dualite d'une Classe de Problemes de Dirichlet Non Lineaires" Revue Francaise d'Automatique Informatique Recherche Operationnelle, Analyse Numerique, R-2, pp. 41-76.

[GlP79] Glad, T., and Polak, E., 1979. "A Multiplier Method with Automatic Limitation of Penalty Growth," Math. Programming, Vol. 17, pp. 140-155.

[Gla79] Glad, T., 1979. "Properties of Updating Methods for the Multipliers in Augmented Lagrangians," J. Opt. Th. and Appl., Vol. 28, pp. 135-156.

[GoP79] Gonzaga, C., and Polak, E., 1979. "On Constraint Dropping Schemes and Optimality Functions for a Class of Outer Approximations Algorithms," SIAM J. on Control and Optimization, Vol. 17, pp.477-493.

[GoT71] Gould, F. J., and Tolle, J., 1971. "A Necessary and Sufficient Condition for Constrained Optimization," SIAM J. Applied Math., Vol. 20, pp. 164-172.

[GoT72] Gould, F. J., and Tolle, J., 1972. "Geometry of Optimality Conditions and Constraint Qualifications," Math. Programming, Vol. 2, pp. 1-18.

[GoV90] Goffin, J. L., and Vial, J. P., 1990. "Cutting Planes and Column Generation Techniques with the Projective Algorithm," J. Opt. Th. and Appl., Vol. 65, pp. 409-429.

[GoV99] Goffin, J. L., and Vial, J. P., 1999. "Convex Nondifferentiable Optimization: A Survey Focussed on the Analytic Center Cutting Plane Method," Logilab Technical Report, Department of Management Studies, University of Geneva, Switzerland; also GERAD Tech. Report G-99-17, McGill Univ., Montreal, Canada.

[Gof77] Goffin, J. L., 1977. "On Convergence Rates of Subgradient Optimization Methods," Math. Programming, Vol. 13, pp. 329-347.

[Gol62] Goldstein, A. A., 1962. "Cauchy's Method of Minimization," Numer. Math., Vol. 4, pp. 146-150.

[Gol64] Goldstein, A. A., 1964. "Convex Programming in Hibert Space," Bull. Amer. Math. Soc., Vol. 70, pp. 709-710.

[Gol67] Goldstein, A. A., 1967. Constructive Real Analysis, Harper and Row, N. Y.

[Gol70] Goldfarb, D., 1970. "A Family of Variable-Metric Methods Derived by Variational Means," Math. Comp., Vol. 24, pp. 23-26.

[Gol85] Golshtein, E. G., 1985. "A Decomposition Method for Linear and Convex Programming Problems," Matecon, Vol. 21, pp. 1077-1091.

[Gol89] Goldberg, D. E., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning, Addison Wesley, Reading, MA.

[Gom58] Gomory, R. E., 1958. "Outline of an Algorithm for Integer Solutions to Linear Programs," Bulletin of the American Mathematical Society, Vol. 64, pp. 275-278.

[Gon91] Gonzaga, C. C., 1991. "Large Step Path-Following Methods for Linear Programming, Part I: Barrier Function Method," SIAM J. on Optimization, Vol. 1, pp. 268-279.

[Gon92] Gonzaga, C. C., 1992. "Path Following Methods for Linear Programming," SIAM Review, Vol. 34, pp. 167-227.

[Gon00] Gonzaga, C. C., 2000. "Two Facts on the Convergence of the Cauchy Algorithm," J. of Optimization Theory and Applications, Vol. 107, pp. 591-600.

[GrS00] Grippo, L., and Sciandrone, M., 2000. "On the Convergence of the Block Nonlinear Gauss-Seidel Method Under Convex Constraints," Operations Research Letters, Vol. 26, pp. 127-136.

[GrW08] Griewank, A., and Walther, A., 2008. Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, SIAM.

[Gri94] Grippo, L., 1994. "A Class of Unconstrained Minimization Methods for Neural Network Training," Optimization Methods and Software, Vol. 4, pp. 135-150.

[GuM86] Guelat, J., and Marcotte, P., 1986. "Some Comments on Wolfe's 'Away Step'," Math. Programming, Vol. 35, pp. 110-119.

[Gui69] Guignard, M., 1969. "Generalized Kuhn-Tucker Conditions for Mathematical Programming Problems in a Banach Space," SIAM J. on Control, Vol. 7, pp. 232-241.

[Gul92] Guler, O., 1992. "New Proximal Point Algorithms for Convex Minimization," SIAM J. on Optimization, Vol. 2, pp. 649-664.

[Gul94] Guler, O., 1994. "Limiting Behavior of Weighted Central Paths in Linear Programming," Math. Programming, Vol. 65, pp. 347-363.

[HDY12] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior A., et al. 2012. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," Signal Processing Magazine, IEEE, Vol. 29, pp. 82-97.

[HKR95] den Hertog, D., Kaliski, J., Roos, C., and Terlaky, T., 1995. "A Path-Following Cutting Plane Method for Convex Programming," Annals of Operations Research, Vol. 58, pp. 69-98.

[HLV87] Hearn, D. W., Lawphongpanich, S., and Ventura, J. A., 1987. "Restricted Simplicial Decomposition: Computation and Extensions," Math. Programming Studies, Vol. 31, pp. 119-136.

[HPT00] Horst, R., Pardalos, P. M., and Thoai, N. V., 2000. Introduction to Global Optimization, 2nd Edition, Kluwer, Boston.

[HDR13] Hong, M., Wang, X., Razaviyayn, M., and Luo, Z. Q., 2013. "Iteration Complexity Analysis of Block Coordinate Descent Methods," arXiv preprint arXiv:1310.6957.

[HaB70] Haarhoff, P. C., and Buys, J. D, 1970. "A New Method for the Optimization of a Nonlinear Function Subject to Nonlinear Constraints," Computer J., Vol. 13, pp. 178-184.

[HaH93] Hager, W. W., and Hearn, D. W., 1993. "Application of the Dual Active Set Algorithm to Quadratic Network Optimization," Computational Optimization and Applications, Vol. 1, pp. 349-373.

[HaM79] Han, S. P., and Mangasarian, O. L., 1979. "Exact Penalty Functions in Nonlinear Programming," Math. Programming, Vol. 17, pp. 251-269.

[Hag99] Hager, W. W., 1999. "Stabilized Sequential Quadratic Programming," Computational Optimization and Applications, Vol. 12.

[Han77] Han, S. P., 1977. "A Globally Convergent Method for Nonlinear Programming," J. Opt. Th. and Appl., Vol. 22, pp. 297-309.

[Hay11] Haykin, S., 2011. Neural Networks and Learning Machines, (3rd Ed.), Pearson Education, Upper Saddle River, N. J.

[HeK70] Held, M., and Karp, R. M., 1970. "The Traveling Salesman Problem and Minimum Spanning Trees," Operations Research, Vol. 18, pp. 1138-1162.

[HeK71] Held, M., and Karp, R. M., 1971. "The Traveling Salesman Problem and Minimum Spanning Trees: Part II," Math. Programming, Vol. 1, pp. 6-25.

[HeL89] Hearn, D. W., and Lawphongpanich, S., 1989. "Lagrangian Dual Ascent by Generalized Linear Programming," Operations Res. Letters, Vol. 8, pp. 189-196.

[HeS52] Hestenes, M. R., and Stiefel, E. L., 1952. "Methods of Conjugate Gradients for Solving Linear Systems," J. Res. Nat. Bur. Standards Sect. B, Vol. 49, pp. 409-436.

[Hei96] Heinkenschloss, M., 1996. "Projected Sequential Quadratic Programming Methods," SIAM J. on Optimization, Vol. 6, pp. 373-417.

[Her94] den Hertog, D., 1994. Interior Point Approach to Linear, Quadratic, and Convex Programming, Kluwer, Dordrecht, The Netherlands.

[Hes69] Hestenes, M. R., 1969. "Multiplier and Gradient Methods," J. Opt. Th. and Appl., Vol. 4, pp. 303-320.

[Hes75] Hestenes, M. R., 1975. Optimization Theory: The Finite Dimensional Case, Wiley, N. Y.

[Hes80] Hestenes, M. R., 1980. Conjugate Direction Methods in Optimization, Springer-Verlag, Berlin and N. Y.

[HiL93] Hiriart-Urruty, J.-B., and Lemarechal, C., 1993. Convex Analysis and Minimization Algorithms, Vols. I and II, Springer-Verlag, Berlin and N. Y.

[Hil57] Hildreth, C., 1957. "A Quadratic Programming Procedure," Naval Res. Logist. Quart., Vol. 4, pp. 79-85. See also "Erratum," Naval Res. Logist. Quart., Vol. 4, p. 361.

[HoJ61] Hooke, R., and Jeeves, T. A., 1961. "Direct Search Solution of Numerical and Statistical Problems," J. Assoc. Comp. Mach., Vol. 8, pp. 212-221.

[HoK71] Hoffman, K., and Kunze, R., 1971. Linear Algebra, Prentice-Hall, Englewood Cliffs, N. J.

[HoL13] Hong, M., and Luo, Z. Q., 2013. "On the Linear Convergence of the Alternating Direction Method of Multipliers," arXiv preprint arXiv:1208.3922.

[Hoh77] Hohenbalken, B. von, 1977. "Simplicial Decomposition in Nonlinear Programming," Math. Programming, Vol. 13, pp. 49-68.

[Hol74] Holloway, C. A., 1974. "An Extension of the Frank and Wolfe Method of Feasible Directions," Math. Programming, Vol. 6, pp. 14-27.

[HuD84] Hughes, G. C., and Dunn, J. C., 1984. "Newton-Goldstein Convergence Rates for Convex Constrained Minimization Problems with Singular Solutions," Appl. Math. Optim., Vol. 12, pp. 203-230.

[HuL88] Huang, C., and Litzenberger, R. H., 1988. Foundations of Financial Economics, Prentice-Hall, Englewood Cliffs, N. J.

[IJT15a] Iusem, A., Jofre, A., and Thompson, P., 2015. "Approximate Projection Methods for Monotone Stochastic Variational Inequalities," Math. Programming, to appear.

[IJT15b] Iusem, A., Jofre, A., and Thompson, P., 2015. "Incremental Constraint Projection Methods for Monotone Stochastic Variational Inequalities," Math. Operations Res., to appear.

[IST94] Iusem, A. N., Svaiter, B., and Teboulle, M., 1994. "Entropy-Like Proximal Methods in Convex Programming," Math. Operations Res., Vol. 19, pp. 790-814.

[IbF96] Ibaraki, S., and Fukushima, M., 1996. "Partial Proximal Method of Multipliers

for Convex Programming Problems," J. of Operations Research Society of Japan, Vol. 39, pp. 213-229.

[IbK88] Ibaraki, T., and Katoh, N., 1988. Resource Allocation Problems: Algorithmic Approaches, M.I.T. Press, Cambridge, MA.

[Iof94] Ioffe, A., 1994. "On Sensitivity Analysis of Nonlinear Programs in Banach Spaces: The Approach via Composite Unconstrained Minimization," SIAM J. on Optimization, Vol. 4, pp. 1-43.

[IuT95] Iusem, A. N., and Teboulle, M., 1995. "Convergence Rate Analysis of Nonquadratic Proximal Methods for Convex and Linear Programming," Math. Operations Res., Vol. 20.

[Ius99] Iusem, A. N., 1999. "Augmented Lagrangian Methods and Proximal Point Methods for Convex Minimization," Investigacion Operativa, Vol. 8, pp. 11-49 .

[JRT95] Junger, M., Reinelt, G., and Rinaldi, G., 1995. "Practical Problem Solving with Cutting Plane Algorithms in Combinatorial Oprimization," in Combinatorial Optimization, Cook, W. J., Lovasz, L., and Seymour, P., (Eds.), DIMACS Series in Discrete Mathematics and Computer Science, AMS, pp. 11-152.

[JaS95] Jarre, F., and Saunders, M. A., 1995. "A Practical Interior-Point Method for Convex Programming," SIAM J. Optimization, Vol. 5, pp. 149-171.

[Jag13] Jaggi, M., 2013. "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization," Proc. of ICML 2013.

[JeW92] Jeyakumar, V., and Wolkowicz, H., 1992. "Generalizations of Slater's Constraint Qualification for Infinite Convex Programs," Math. Programming, Vol. 57, pp. 85-101.

[Joh48] John, F., 1948. "Extremum Problems with Inequalities as Subsidiary Conditions," in Studies and Essays: Courant Anniversary Volume, K. O. Friedrichs, Neugebauer, O. E., and Stoker, J. J., (Eds.), Wiley-Interscience, N. Y., pp. 187-204.

[KAK89] Korst, J., Aarts, E. H., and Korst, A., 1989. Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing, Wiley, N. Y.

[KLT03] Kolda, T. G., Lewis, R. M., and Torczon, V., 2003. "Optimization by Direct Search: New Perspectives on Some Classical and Modern Methods," SIAM Review, Vol. 45, pp. 385-482.

[KMN91] Kojima, M., Meggido, N., Noma, T., and Yoshise, A., 1991. A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems, Springer-Verlag, Berlin.

[KMY89] Kojima, M., Mizuno, S., and Yoshise, A., 1989. "A Primal-Dual Interior Point Algorithm for Linear Programming," in Progress in Mathematical Programming, Interior Point and Related Methods, Meggido, N., (Ed.), Springer-Verlag, N. Y., pp. 29-47.

[KaW94] Kall, P., and Wallace, S. W., 1994. Stochastic Programming, Wiley, Chichester, UK.

[Kan39] Kantorovich, L. V., 1939. "The Method of Successive Approximations for Functional Equations," Acta Math., Vol. 71, pp. 63-97.

[Kan45] Kantorovich, L. V., 1945. "On an Effective Method of Solution of Extremal Problems for a Quadratic Functional," Dokl. Akad. Nauk SSSR, Vol. 48, pp. 483-487.

[Kan49] Kantorovich, L. V., 1949. "On Newton's Method," Trudy Mat. Inst. Steklov, Vol. 28, pp. 104-144. Translated in Selected Articles in Numerical Analysis by C. D. Benster, 104.1-144.2.

[Kar39] Karush, W., 1939. "Minima of Functions of Several Variables with Inequalities as Side Conditions," M.S. Thesis, Department of Math., University of Chicago.

[Kar84] Karmarkar, N., 1984. "A New Polynomial-Time Algorithm for Linear Programming," Combinatorica, Vol. 4, pp. 373-395.

[KeG88] Keerthi, S. S., and Gilbert, E. G., 1988. "Optimal, Infinite Horizon Feedback Laws for a General Class of Constrained Discete Time Systems: Stability and Moving-Horizon Approximations," J. Optimization Theory Appl., Vo. 57, pp. 265-293.

[Kel60] Kelley, J. E., 1960. "The Cutting-Plane Method for Solving Convex Programs," J. Soc. Indust. Appl. Math., Vol. 8, pp. 703-712.

[Kel99] Kelley, C. T., 1999. Iterative Methods for Optimization, SIAM, Philadelphia.

[Kha79] Khachiyan, L. G., 1979. "A Polynomial Algorithm for Linear Programming," Soviet Math. Doklady, Vol. 20, pp. 191-194.

[KiN91] Kim, S., and Nazareth, J. L., 1991. "The Decomposition Principle and Algorithms for Linear Programming," Linear Algebra and its Applications, Vol. 152, pp. 119-133.

[KiR92] King, A., and Rockafellar, R. T., 1992. "Sensitivity Analysis for Nonsmooth Generalized Equations," Math. Programming, Vol. 55, pp. 193-212.

[KlH98] Klatte, D., and Henrion, R., 1998. "Regularity and Stability in Nonlinear Semi-Infinite Optimization," in Semi-Infinite Programming, Reemtsen, R., and Ruckman, J. J., (Eds.), Kluwer, Boston, pp. 69-102.

[KoB72] Kort, B. W., and Bertsekas, D. P., 1972. "A New Penalty Function Method for Constrained Minimization," Proc. 1972 IEEE Confer. Decision Control, New Orleans, LA, pp. 162-166.

[KoB76] Kort, B. W., and Bertsekas, D. P., 1976. "Combined Primal-Dual and Penalty Methods for Convex Programming," SIAM J. on Control and Optimization, Vol. 14, pp. 268-294.

[KoM98] Kontogiorgis, S., and Meyer, R. R., 1998. "A Variable-Penalty Alternating Directions Method for Convex Optimization," Math. Programming, Vol. 83, pp. 29-53.

[KoN93] Kortanek, K. O., and No, H., 1993. "A Central Cutting Plane Algorithm for Convex Semi-Infinite Programming Problems," SIAM J. on Optimization, Vol. 3, pp. 901-918.

[KoZ93] Kortanek, K. O., and Zhu, J., 1993. "A Polynomial Barrier Algorithm for Linearly Constrained Convex Programming Problems," Math. Operations Res., Vol. 18, pp. 116-127.

[KoZ95] Kortanek, K. O., and Zhu, J., 1995. "On Controlling the Parameter in the Logarithmic Barrier Term for Convex Programming Problems," J. Opt. Th. and Appl., Vol. 84, pp. 117-143.

[Koh74] Kohonen, T., 1974. "An Adaptive Associative Memory Principle," IEEE Trans. on Computers, Vol. C-23, pp. 444-445.

[Kor75] Kort, B. W., 1975. "Combined Primal-Dual and Penalty Function Algorithms for Nonlinear Programming," Ph.D. Thesis, Dept. of Engineering-Economic Systems, Stanford Univ., Stanford, Ca.

[Kor76] Korpelevich, G. M., 1976. "The Extragradient Method for Finding Saddle Points and Other Problems," Matecon, Vol. 12, pp. 747-756.

[KuC78] Kushner, H. J., and Clark, D. S., 1978. Stochastic Approximation Methods for Constrained and Unconstrained Systems, Springer-Verlag, N. Y.

[KuT51] Kuhn, H. W., and Tucker, A. W., 1951. "Nonlinear Programming," in Proc. of the Second Berkeley Symposium on Math. Statistics and Probability, Neyman, J., (Ed.), Univ. of California Press, Berkeley, CA, pp. 481-492.

[KuY97] Kushner, H. J., and Yin, G., 1997. Stochastic Approximation Methods, Springer-Verlag, N. Y.

[Kuh76] Kuhn, H. W., 1976. "Nonlinear Programming: A Historical View," in Nonlinear Programming, Cottle, R. W., and Lemke, C. E., (Eds.), SIAM-AMS Proc., Vol. IX, American Math. Soc., Providence, RI, pp. 1-26.

[LJS12] Lacoste-Julien, S., Jaggi, M., Schmidt, M., and Pletscher, P., 2012. "Block-Coordinate Frank-Wolfe Optimization for Structural SVMs," arXiv preprint arXiv:1207-.4747.

[LMS92] Lustig, I. J., Marsten, R. E., and Shanno, D. F., 1992. "On Implementing Mehrotra's Predictor-Corrector Interior-Point Method for Linear Programming," SIAM J. on Optimization, Vol. 2, pp. 435-449.

[LPW92] Ljung, L., Pflug, G., and Walk, H., 1992. Stochastic Approximation and Optimization of Random Systems, Birkhauser, Boston.

[LRW98] Lagarias, J. C., Reeds, J. A., Wright, M. H., and Wright, P. E., 1998. "Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions," SIAM J. on Optimization, Vol. 9, pp. 112-147.

[LST01] Lucidi, S., Sciandrone, M., and Tseng, P., 2001. "Objective-Derivative-Free Methods for Constrained Optimization," Math. Programming, Vol. 92, pp. 37-59.

[LVB98] Lobo, M. S., Vandenberghe, L., Boyd, S., and Lebret, H., 1998. "Applications of Second-Order Cone Programming," Linear Algebra and Applications, Vol. 284, pp. 193-228.

[LaD60] Land, A. H., and Doig, A. G., 1960. "An Automatic Method for Solving Discrete Programming Problems," Econometrica, Vol. 28, pp. 497-520.

[LaJ13] Lacoste-Julien, S., and Jaggi, M., 2013. "An Affine Invariant Linear Convergence Analysis for Frank-Wolfe Algorithms," arXiv preprint arXiv:1312.7864.

[LaT85] Lancaster, P., and Tismenetsky, M., 1985. The Theory of Matrices, Academic Press, N. Y.

[LaW78] Lasdon, L. S., and Waren, A. D., 1978. "Generalized Reduced Gradient Software for Linearly and Nonlinearly Constrained Problems," in Design and Implementation of Optimization Software, Greenberg, H. J., (Ed.), Sijthoff and Noordhoff, Holland, pp. 335-362.

[Lan15] Landi, G., 2015. "A Modified Newton Projection Method for ℓ_1-Regularized Least Squares Image Deblurring," J. of Mathematical Imaging and Vision, Vol. 51, pp. 195-208.

[Las70] Lasdon, L. S., 1970. Optimization Theory for Large Systems, Macmillian, N. Y.; republished by Dover Pubs, N. Y., 2002.

[Law76] Lawler, E., 1976. Combinatorial Optimization: Networks and Matroids, Holt, Reinhart, and Winston, N. Y.

[LeL10] Leventhal, D., and Lewis, A. S., 2010. "Randomized Methods for Linear Constraints: Convergence Rates and Conditioning," Math. of Operations Research, Vol. 35, pp. 641-654.

[LeO16] Lewis, A. S., and Overton, M. L., 2013. "Nonsmooth Optimization via Quasi-Newton Methods," Math. Programming, Vol. 141, pp. 135-163.

[LeP65] Levitin, E. S., and Poljak, B. T., 1965. "Constrained Minimization Methods," Ž. Vyčisl. Mat. i Mat. Fiz., Vol. 6, pp. 787-823.

[LeS93] Lemaréchal, C., and Sagastizábal, C., 1993. "An Approach to Variable Metric Bundle Methods," in Systems Modelling and Optimization, Proc. of the 16th IFIP-TC7 Conference, Compiègne, Henry, J., and Yvon, J.-P., (Eds.), Lecture Notes in Control and Information Sciences 197, pp. 144-162.

[Lem74] Lemarechal, C., 1974. "An Algorithm for Minimizing Convex Functions," in Information Processing '74, Rosenfeld, J. L., (Ed.), pp. 552-556, North-Holland, Amsterdam.

[LiM79] Lions, P. L., and Mercier, B., 1979. "Splitting Algorithms for the Sum of Two Nonlinear Operators," SIAM J. on Numerical Analysis, Vol. 16, pp. 964-979.

[LiP87] Lin, Y. Y., and Pang, J.-S., 1987. "Iterative Methods for Large Convex Quadratic Programs: A Survey," SIAM J. on Control and Optimization, Vol. 18, pp. 383-411.

[Lin07] Lin, C. J., 2007. "Projected Gradient Methods for Nonnegative Matrix Factorization," Neural Computation, Vol. 19, pp. 2756-2779.

[LuB96] Lucena, A., and Beasley, J. E., 1996. "Branch and Cut Algorithms," in Advances in Linear and Integer Programming, Beasley, J. E., (Ed.), Oxford University Press, N. Y., Chapter 5.

[LuT91] Luo, Z. Q., and Tseng, P., 1991. "On the Convergence of a Matrix-Splitting Algorithm for the Symmetric Monotone Linear Complementarity Problem," SIAM J. on Control and Optimization, Vol. 29, pp. 1037-1060.

[LuT92a] Luo, Z. Q., and Tseng, P., 1992. "On the Convergence of the Coordinate Descent Method for Convex Differentiable Minimization," J. Opt. Th. and Appl., Vol. 72, pp. 7-35.

[LuT92b] Luo, Z. Q., and Tseng, P., 1992. "On the Linear Convergence of Descent Methods for Convex Essentially Smooth Minimization," SIAM J. on Control and Optimization, Vol. 30, pp. 408-425.

[LuT93a] Luo, Z. Q., and Tseng, P., 1993. "Error Bounds and Convergence Analysis of Feasible Descent Methods: A General Approach," Annals of Operations Res., Vol. 46, pp. 157-178.

[LuT93b] Luo, Z. Q., and Tseng, P., 1993. "Error Bound and Reduced Gradient Projection Algorithms for Convex Minimization over a Polyhedral Set," SIAM J. on Optimization, Vol. 3, pp. 43-59.

[LuT93c] Luo, Z. Q., and Tseng, P., 1993. "On the Convergence Rate of Dual Ascent Methods for Linearly Constrained Convex Minimization," Math. of Operations Res., Vol. 18, pp. 846-867.

[LuT94a] Luo, Z. Q., and Tseng, P., 1994. "Analysis of an Approximate Gradient Projection Method with Applications to the Backpropagation Algorithm," Optimization Methods and Software, Vol. 4, pp. 85-101.

[LuT94b] Luo, Z. Q., and Tseng, P., 1994. "On the Rate of Convergence of a Distributed Asynchronous Routing Algorithm," IEEE Transactions on Automatic Control, Vol. 39, pp. 1123-1129.

[LuX15] Lu, Z., and Xiao, L., 2015. "On the Complexity Analysis of Randomized Block-Coordinate Descent Methods," Math. Programming, Vol. 152, pp. 615-642.

[LuY16] Luenberger, D. G., and Ye, Y., 1916. Introduction to Linear and Nonlinear Programming, 4th edition, Springer, N. Y.

[Lue69] Luenberger, D. G., 1969. Optimization by Vector Space Methods, Wiley, N. Y.

[Lue84] Luenberger, D. G., 1984. Introduction to Linear and Nonlinear Programming, 2nd edition, Addison-Wesley, Reading, MA.

[Lue98] Luenberger, D. G., 1998. Investment Science, Oxford University Press, N. Y.

[Luo91] Luo, Z. Q., 1991. "On the Convergence of the LMS Algorithm with Adaptive Learning Rate for Linear Feedforward Networks," Neural Computation, Vol. 3, pp. 226-245.

[Luq84] Luque, F.J., 1984. "Asymptotic Convergence Analysis of the Proximal Point Algorithm," SIAM J. on Control and Optimization, Vol. 22, pp. 277-293.

[MMS91] McShane, K. A., Monma, C. L., and Shanno, D., 1991. "An Implementation of a Primal-Dual Interior Point Method for Linear Programming," ORSA J. on Computing, Vol. 1, pp. 70-83.

[MMZ95] McKenna, M. P., Mesirov, J. P., and Zenios, S. A., 1995. "Data Parallel Quadratic Programming on Box-Constrained Problems," SIAM J. on Optimization, Vol. 5, pp. 570-589.

[MOT95] Mahey, P., Oualibouch, S., and Tao, P. D., 1995. "Proximal Decomposition on the Graph of a Maximal Monotone Operator," SIAM J. on Optimization, Vol. 5, pp. 454-466.

[MRR00] Mayne, D. Q., Rawlings, J. B., Rao, C. V., and Scokaert, P. O. M., 2000. "Constrained Model Predictive Control: Stability and Optimality," Automatica, Vol. 36, pp. 789-814.

[MSQ98] Mifflin, R., Sun, D., and Qi, L., 1998. "Quasi-Newton Bundle-Type Methods for Nondifferentiable Convex Optimization," SIAM J. on Optimization, Vol. 8, pp. 583-603.

[MTW93] Monteiro, R. D. C., Tsuchiya, T., and Wang, Y., 1993. "A Simplified Global Convergence Proof of the Affine Scaling Algorithm," Annals of Operations Res., Vol. 47, pp. 443-482.

[MYF03] Moriyama, H., Yamashita, N., and Fukushima, M., 2003. "The Incremental Gauss-Newton Algorithm with Adaptive Stepsize Rule," Computational Optimization and Applications, Vol. 26, pp. 107-141.

[MaD87] Mangasarian, O. L., and De Leone, R., 1987. "Parallel Successive Overelaxation Methods for Symmetric Linear Complementarity Problems and Linear Programs," J. of Optimization Th. and Appl., Vol. 54, pp. 437-446.

[MaD88] Mangasarian, O. L., and De Leone, R., 1988. "Parallel Gradient Projection Successive Overrelaxation for Symmetric Linear Complementarity Problems," Annals of Operations Res., Vol. 14, pp. 41-59.

[MaF67] Mangasarian, O. L., and Fromovitz, S., 1967. "The Fritz John Necessary Optimality Conditions in the Presence of Equality and Inequality Constraints," J. Math. Anal. and Appl., Vol. 17, pp. 37-47.

[MaP82] Mayne, D. Q., and Polak, E., 1982. "A Superlinearly Convergent Algorithm for Constrained Optimization Problems," Math. Programming Studies, Vol. 16, pp. 45-61.

[MaS94] Mangasarian, O. L., and Solodov, M. V., 1994. "Serial and Parallel Backpropagation Convergence Via Nonmonotone Perturbed Minimization," Optimization Methods and Software, Vol. 4, pp. 103-116.

[Mai13] Mairal, J., 2013. "Optimization with First-Order Surrogate Functions," arXiv preprint arXiv:1305.3120.

[Mai14] Mairal, J., 2014. "Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning," arXiv preprint arXiv:1402.4419.

[Man69] Mangasarian, O. L., 1969. Nonlinear Programming, Prentice-Hall, Englewood Cliffs, N. J.; also SIAM, Classics in Applied Mathematics 10, Phila., PA., 1994.

[Mar70] Martinet, B., 1970. "Regularisation d'Inequations Variationnelles par Approximations Successives," Rev. Francaise Inf. Rech. Oper., Vol. 4, pp. 154-158.

[Mar72] Martinet, B., 1972. "Determination Approchee d'un Point Fixe d'une Application Pseudo-Contractante," C. R. Acad. Sci. Paris, 274A, pp. 163-165.

[Mar78] Maratos, N., 1978. "Exact Penalty Function Algorithms for Finite Dimensional and Control Optimization Problems," Ph.D. Thesis, Imperial College Sci. Tech, Univ. of London.

[McK98] McKinnon, K. I. M., 1998. "Convergence of the Nelder-Mead Simplex Method to a Non-Stationary Point," SIAM J. on Optimization, Vol. 9, pp. 148-158.

[McL80] McLinden, L., 1980. "The Complementarity Problem for Maximal Monotone Multifunctions," in Variational Inequalities and Complementarity Problems, Cottle, R., Giannessi, F., and Lions, J.-L., (Eds.), Wiley, N. Y., pp. 251-270.

[McS73] McShane, E. J., 1973. "The Lagrange Multiplier Rule," American Mathematical Monthly, Vol. 80, pp. 922-925.

[Meg88] Megiddo, N., 1988. "Pathways to the Optimal Set in Linear Programming," in Progress in Mathematical Programming, Megiddo, N., (Ed.), Springer-Verlag, N. Y., pp. 131-158.

[Meh92] Mehrotra, S., 1992. "On the Implementation of a Primal-Dual Interior Point Method," SIAM J. on Optimization, Vol. 2, pp. 575-601.

[Mey79] Meyer, R. R., 1979. "Two-Segment Separable Programming," Management Science, Vol. 25, pp. 385-395.

[Mey07] Meyn, S., 2007. Control Techniques for Complex Networks, Cambridge Univ. Press, N. Y.

[Mif96] Mifflin, R., 1996. "A Quasi-Second-Order Proximal Bundle Algorithm," Math. Programming, Vol. 73, pp. 51-72.

[Mig94] Migdalas, A., 1994. "A Regularization of the Frank-Wolfe Method and Unification of Certain Nonlinear Programming Methods," Math. Programming, Vol. 65, pp. 331-345.

[Min60] Minty, G. J., 1960. "Monotone Networks," Proc. Roy. Soc. London, A, Vol. 257, pp. 194-212.

[Min86] Minoux, M., 1986. Mathematical Programming: Theory and Algorithms, Wiley, N. Y.

[Mit66] Mitter, S. K., 1966. "Successive Approximation Methods for the Solution of Optimal Control Problems," Automatica, Vol. 3, pp. 135-149.

[MoA89a] Monteiro, R. D. C., and Adler, I., 1989. "Interior Path Following Primal-Dual Algorithms, Part I: Linear Programming," Math. Programming, Vol. 44, pp. 27-41.

[MoA89b] Monteiro, R. D. C., and Adler, I., 1989. "Interior Path Following Primal-Dual Algorithms, Part II: Convex Quadratic Programming," Math. Programming, Vol. 44, pp. 43-66.

[MoL99] Morari, M., and Lee, J. H., 1999. "Model Predictive Control: Past, Present, and Future," Computers and Chemical Engineering, Vol. 23, pp. 667-682.

[MoS83] Moré, J. J., and Sorensen, D. C., 1983. "Computing a Trust Region Step," SIAM J. on Scientific and Statistical Computing, Vol. 4, pp. 553-572.

[MoT89] Moré, J. J., and Toraldo, G., 1989. "Algorithms for Bound Constrained Quadratic Programming Problems," Numer. Math., Vol. 55, pp. 377-400.

[MoW93] Moré, J. J., and Wright, S. J., 1993. Optimization Software Guide, SIAM, Frontiers in Applied Mathematics 14, Phila., PA.

[Mor88] Mordukhovich, B. S., 1988. Approximation Methods in Problems of Optimization and Control, Nauka, Moscow.

[Mor06] Mordukhovich, B. S., 2006. Variational Analysis and Generalized Differentiation I: Basic Theory, Springer, N. Y.

[MuP75] Mukai, H., and Polak, E., 1975. "A Quadratically Convergent Primal-Dual Algorithm with Global Convergence Properties for Solving Optimization Problems with Equality Constraints," Math. Programming, Vol. 9, pp. 336-349.

[MuR95] Mulvey, J. M., and Ruszcynski, A., 1995. "A New Scenario Decomposition Method for Large Scale Stochastic Optimization," Operations Research, Vol. 43, pp. 477-490.

[MuS87] Murtagh, B. A., and Saunders, M. A., 1987. "MINOS 5.1 User's Guide," Technical Report SOL-83-20R, Stanford Univ.

[Mur92] Murty, K. G., 1992. Network Programming, Prentice-Hall, Englewood Cliffs, N. J.

[NBB01] Nedić, A., Bertsekas, D. P., and Borkar, V. S., 2001. "Distributed Asynchronous Incremental Subgradient Methods," in Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, Butnariu, D., Censor, Y., and Reich, S., (Eds.), Elsevier Science, Amsterdam, Netherlands.

[NSL15] Nutini, J., Schmidt, M., Laradji, I. H., Friedlander, M., and Koepke, H., 2015. "Coordinate Descent Converges Faster with the Gauss-Southwell Rule Than Random Selection," arXiv preprint arXiv:1506.00552.

[NaQ96] Nazareth, J. L., and Qi, L., 1996. "Globalization of Newton's Method for Solving Nonlinear Equations," Numerical Linear Algebra with Applications, Vol. 3, pp. 239-249.

[NaS89] Nash, S. G., and Sofer, 1989. "Block Truncated-Newton Methods for Parallel Optimization," Math. Programming, Vol. 45, pp. 529-546.

[NaT02] Nazareth, L., and Tseng, P., 2002. "Gilding the Lily: A Variant of the Nelder-Mead Algorithm Based on Golden-Section Search," Computational Optimization and Applications, Vol. 22, pp. 133-144.

[Nag93] Nagurney, A., 1993. Network Economics: A Variational Inequality Approach, Kluwer, Dordrecht, The Netherlands.

[Nas85] Nash, S. G., 1985. "Preconditioning of Truncated-Newton Methods," SIAM J. on Scientific and Statistical Computing, Vol. 6, pp. 599-616.

[Naz94] Nazareth, J. L., 1994. The Newton-Cauchy Framework: A Unified Approach to Unconstrained Nonlinear Minimization, Lecture Notes in Computer Science No. 769, Springer-Verlag, Berlin and New York.

[Naz96] Nazareth, J. L., 1996. "Lagrangian Globalization: Solving Nonlinear Equations via Constrained Optimization," in Mathematics of Numerical Analysis, Renegar, J., Shub, M., and Smale, S., (Eds.), Lectures in Applied Mathematics, Vol. 32, The American Mathematical Society, Providence, RI, pp. 533-542.

[NeB00] Nedić, A., and Bertsekas, D. P., 2000. "Convergence Rate of Incremental Subgradient Algorithms," Stochastic Optimization: Algorithms and Applications," S. Uryasev and P. M. Pardalos, Eds., Kluwer, pp. 263-304.

[NeB01] Nedić, A., and Bertsekas, D. P., 2001. "Incremental Subgradient Methods for Nondifferentiable Optimization," SIAM J. on Optimization, Vol. 12, pp. 109-138.

[NeB10] Nedić, A., and Bertsekas, D. P., 2010. "The Effect of Deterministic Noise in Subgradient Methods," Math. Programming, Ser. A, Vol. 125, pp. 75-99.

[NeN94] Nesterov, Y., and Nemirovskii, A., 1994. Interior Point Polynomial Algorithms in Convex Programming, SIAM, Studies in Applied Mathematics 13, Phila., PA.

[NeW88] Nemhauser, G. L., and Wolsey, L. A., 1988. Integer and Combinatorial Optimization, Wiley, N. Y.

[NeY83] Nemirovsky, A., and Yudin, D. B., 1983. Problem Complexity and Method Efficiency, Wiley, N. Y.

[Ned11] Nedić, A., 2011. "Random Algorithms for Convex Minimization Problems," Math. Programming, Vol. 129, pp. 225?253.

[Nes83] Nesterov, Y., 1983. "A Method for Unconstrained Convex Minimization Problem with the Rate of Convergence $O(1/k^2)$," Doklady AN SSSR, Vol. 269, pp. 543-547; translated as Soviet Math. Dokl.

[Nes95] Nesterov, Y., 1995. "Complexity Estimates of Some Cutting Plane Methods Based on Analytic Barrier," Math. Programming, Vol. 69, pp. 149-176.

[Nes04] Nesterov, Y., 2004. Introductory Lectures on Convex Optimization, Kluwer Academic Publisher, Dordrecht, The Netherlands.

[Nes12] Nesterov, Y., 2012. "Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems," SIAM J. on Optimization, Vol. 22, pp. 341-362.

[Neu28] Neumann, J. von, 1928. "Zur Theorie der Gesellschaftsspiele," Math. Ann., Vol. 100, pp. 295-320.

[NgS79] Nguyen, V. H., and Strodiot, J. J., 1979. "On the Convergence Rate of a Penalty Function Method of Exponential Type," J. Opt. Th. and Appl., Vol. 27, pp. 495-508.

[NoW06] Nocedal, J., and Wright, S. J., 2006. Numerical Optimization, 2nd Edition, Springer, NY.

[Noc80] Nocedal, J., 1980. "Updating Quasi-Newton Matrices with Limited Storage," Math. of Computation, Vol. 35, pp. 773-782.

[OLR85] O'hEigeartaigh, M., Lenstra, S. K., and Rinnoy Kan, A. H. G., (Eds.), 1985. Combinatorial Optimization: Annotated Bibliographies, Wiley, N. Y.

[OrL74] Oren, S. S., and Luenberger, D. G., 1974. "Self-Scaling Variable Metric Algorithm, Part I," Management Science, Vol. 20, pp. 845-862.

[OrR70] Ortega, J. M., and Rheinboldt, W. C., 1970. Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, N. Y.

[Ore73] Oren, S. S., 1973. "Self-Scaling Variable Metric Algorithm, Part II," Management Science, Vol. 20, pp. 863-874.

[OzB03] Ozdaglar, A. E., and Bertsekas, D. P., 2003. "Routing and Wavelength Assignment in Optical Networks," IEEE Trans. on Networking, pp. 259-272.

[OzB04] Ozdaglar, A. E., and Bertsekas, D. P., 2004. "The Relation Between Pseudonormality and Quasiregularity in Constrained Optimization," Optimization Methods and Software, Vol. 19, pp. 493–506.

[PCC93] Pulleyblank, W., Cook, W., Cunningham, W., and Schrijver, A., 1993. An Introduction to Combinatorial Optimization, Wiley, N. Y.

[PaM89] Pantoja, J. F. A. D., and Mayne, D. Q., 1989. "Sequential Quadratic Pro-

gramming Algorithm for Discrete Optimal Control Problems with Control Inequality Constraints," Intern. J. on Control, Vol. 53, pp. 823-836.

[PaR87] Pardalos, P. M., and Rosen, J. B., 1987. Constrained Global Optimization: Algorithms and Applications, Springer-Verlag, N. Y.

[PaS82] Papadimitriou, C. H., and Steiglitz, K., 1982. Combinatorial Optimization: Algorithms and Complexity, Prentice-Hall, Englewood Cliffs, N. J.

[PaT91] Panier, E. R., and Tits, A. L., 1991. "Avoiding the Maratos Effect by Means of a Nonmonotone Line Search. I.," SIAM J. on Numer. Anal., Vol. 28, pp. 1183-1195.

[Pan84] Pang, J.-S., 1984. "On the Convergence of Dual Ascent Methods for Large-Scale Linearly Constrained Optimization Problems," Unpublished Manuscript, School of Management, Univ. of Texas, Dallas, Texas.

[Pap81] Papavassilopoulos, G., 1981. "Algorithms for a Class of Nondifferentiable Problems," J. Opt. Th. and Appl., Vol. 34, pp. 41-82.

[Pap82] Pappas, T. N., 1982. "Solution of Nonlinear Equations by Davidon's Least Squares Method," M.S. Thesis, Dept. of Electrical Engineering and Computer Science, M.I.T., Cambridge, MA.

[Pat93] Patriksson, M., 1993. "Partial Linearization Methods for Nonlinear Programming," J. Opt. Th. and Appl., Vol. 78, pp. 227-246.

[Pat98] Patriksson, M., 1998. Nonlinear Programming and Variational Inequalities: A Unified Approach, Kluwer, Dordtrecht, The Netherlands.

[Per78] Perry, A., 1978. "A Modified Conjugate Gradient Algorithm," Operations Research, Vol. 26, pp. 1073-1078.

[Pfl96] Pflug, G. C., 1996. Optimization of Stochastic Models, Kluwer, Boston.

[PiP73] Pironneau, O., and Polak, E., 1973. "Rate of Convergence of a Class of Methods of Feasible Directions," SIAM J. Numer. Anal., Vol. 10, pp. 161-173.

[PiZ92] Pinar, M. C., and Zenios, S. A., 1992. "Parallel Decomposition of Multicommodity Network Flows Using a Linear-Quadratic Penalty Algorithm," ORSA J. on Computing, Vol. 4, pp. 235-249.

[PiZ94] Pinar, M. C., and Zenios, S. A., 1994. "On Smoothing Exact Penalty Functions for Convex Constrained Problems," SIAM J. on Optimization, Vol. 4, pp. 486-511.

[PoH91] Polak, E., and He, L., 1991. "Finite-Termination Schemes for Solving Semi-infinite Satisfycing Problems," J. Opt. Theory and Appl., Vol. 70, pp. 429-442.

[PoR69] Polak, E., and Ribiere, G., 1969. "Note sur la Convergence de Methodes de Directions Conjugees," Rev. Fr. Inform. Rech. Oper., Vol. 16-R1, pp. 35-43.

[PoT73a] Poljak, B. T., and Tsypkin, Y. Z., 1973. "Pseudogradient Adaptation and Training Algorithms," Automation and Remote Control, pp. 45-68.

[PoT73b] Poljak, B. T., and Tretjakov, N. V., 1973. "The Method of Penalty Estimates for Conditional Extremum Problems," Z. VyČisl. Mat. i Mat. Fiz., Vol. 13, pp. 34-46.

[PoT80a] Polak, E., and Tits, A. L., 1980. "A Globally Convergent, Implementable Multiplier Method with Automatic Penalty Limitation," Applied Math. and Optimization, Vol. 6, pp. 335-360.

[PoT80b] Poljak, B. T., and Tsypkin, Y. Z., 1980. "Adaptive Estimation Algorithms (Convergence, Optimality, Stability)," Automation and Remote Control, Vol. 40, pp. 378-389.

[PoT81] Poljak, B. T., and Tsypkin, Y. Z., 1981. "Optimal Pseudogradient Adaptation Algorithms," Automation and Remote Control, Vol. 41, pp. 1101-1110.

[PoT97] Polyak, R., and Teboulle, M., 1997. "Nonlinear Rescaling and Proximal-Like Methods in Convex Optimization," Math. Programming, Vol. 76, pp. 265-284.

[Pol64] Poljak, B. T., 1964. "Some Methods of Speeding up the Convergence of Iteration Methods," Z. VyČisl. Mat. i Mat. Fiz., Vol. 4, pp. 1-17.

[Pol69a] Poljak, B. T., 1969. "The Conjugate Gradient Method in Extremal Problems," Z. Vyčisl. Mat. i Mat. Fiz., Vol. 9, pp. 94-112.

[Pol69b] Poljak, B. T., 1969. "Minimization of Unsmooth Functionals," Z. Vyčisl. Mat. i Mat. Fiz., Vol. 9, pp. 14-29.

[Pol70] Poljak, B. T., 1970. "Iterative Methods Using Lagrange Multipliers for Solving Extremal Problems with Constraints of the Equation Type," Z. VyČisl. Mat. i Mat. Fiz., Vol. 10, pp. 1098-1106.

[Pol71] Polak, E., 1971. Computational Methods in Optimization: A Unified Approach, Academic Press, N. Y.

[Pol73] Polak, E., 1973. "A Historical Survey of Computational Methods in Optimal Control," SIAM Review, Vol. 15, pp. 553-584.

[Pol79] Poljak, B. T., 1979. "On Bertsekas' Method for Minimization of Composite Functions," Internat. Symp. Systems Opt. Analysis, Benoussan, A., and Lions, J. L., (Eds.), pp. 179-186, Springer-Verlag, Berlin and N. Y.

[Pol87] Poljak, B. T., 1987. Introduction to Optimization, Optimization Software Inc., N. Y.

[Pol92] Polyak, R., 1992. "Modified Barrier Functions (Theory and Methods)," Math. Programming, Vol. 54, pp. 177-222.

[Pol97] Polak, E., 1997. Optimization: Algorithms and Consistent Approximations, Springer-Verlag, N. Y.

[Pot94] Potra, F. A., 1994. "A Quadratically Convergent Predictor-Corrector Method for Solving Linear Programs from Infeasible Starting Points," Math. Programming, Vol. 67, pp. 383–406.

[Pow64] Powell, M. J. D., 1964. "An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Derivatives," The Computer Journal, Vol. VII, pp. 155-162.

[Pow69] Powell, M. J. D., 1969. "A Method for Nonlinear Constraints in Minimizing Problems," in Optimization, Fletcher, R., (Ed.), Academic Press, N. Y, pp. 283-298.

[Pow73] Powell, M. J. D., 1973. "On Search Directions for Minimization Algorithms," Math. Programming, Vol. 4, pp. 193-201.

[Pre95] Prekopa, A., 1995. Stochastic Programming, Kluwer, Boston.

[PsD75] Pschenichny, B. N., and Danilin, Y. M., 1975. "Numerical Methods in Extremal Problems," MIR, Moscow, (Engl. trans., 1978).

[Psc70] Pschenichny, B. N., 1970. "Algorithms for the General Problem of Mathematical Proramming," Kibernetika (Kiev), Vol. 6, pp. 120-125.

[Pyt98] Pytlak, R., 1998. "An Efficient Algorithm for Large-Scale Nonlinear Programming Problems with Simple Bounds on the Variables," SIAM J. on Optimization, Vol. 8, pp. 532-560.

[RGV14] Richard, E., Gaiffas, S., and Vayatis, N., 2014. "Link Prediction in Graphs with Autoregressive Features," J. of Machine Learning Research, Vol. 15, pp. 565-593.

[RHL13] Razaviyayn, M., Hong, M., and Luo, Z. Q., 2013. "A Unified Convergence Analysis of Block Successive Minimization Methods for Nonsmooth Optimization," SIAM J. on Optimization, Vol. 23, pp. 1126-1153.

[RSP16] Reddi, S. J., Sra, S., Poczos, B., and Smola, A., 2016. "Fast Incremental Method for Nonconvex Optimization," arXiv preprint arXiv:1603.06159.

[Ray93] Raydan, M., 1993. "On the Barzilai and Borwein Choice of Steplength for the Gradient Method," IMA J. Num. Anal., Vol. 13, pp. 321-326.

[Ray97] Raydan, M., 1997. "The Barzilai and Borwein Gradient Method for the Large Scale Unconstrained Minimization Problem," SIAM J. on Optimization, Vol. 7, pp. 26-33.

[ReR98] Reemtsen, R., and Ruckman, J. J., (Eds.), 1998. Semi-Infinite Programming, Kluwer, Boston.

[Ren01] Renegar, J., 2001. A Mathematical View of Interior-Point Methods in Convex Optimization, SIAM, Phila.

[RiT14] Richtarik, P., and Takac, M., 2014. "Iteration Complexity of Randomized Block-Coordinate Descent Methods for Minimizing a Composite Function," Math. Programming, Vol. 144, pp. 1-38.

[RoW91] Rockafellar, R. T., and Wets, R. J.-B., 1991. "Scenarios and Policy Aggregation in Optimization under Uncertainty," Math. of Operations Res., Vol. 16, pp. 119-147.

[RoW98] Rockafellar, R. T., and Wets, R. J.-B., 1998. Variational Analysis, Springer-Verlag, Berlin.

[Rob74] Robinson, S. M., 1974. "Perturbed Kuhn-Tucker Points and Rates of Convergence for a Class of Nonlinear Programming Algorithms," Math. Programming, Vol. 7, pp. 1-16.

[Rob87] Robinson, S. M., 1987. "Local Structure of Feasible Sets in Nonlinear Programming, Part III. Stability and Sensitivity," Math. Programming Studies, Vol. 30, pp. 45-66.

[Roc67] Rockafellar, R. T., 1967. "Convex Programming and Systems of Elementary Monotonic Relations," J. of Math. Analysis and Applications, Vol. 19, pp. 543-564.

[Roc70] Rockafellar, R. T., 1970. Convex Analysis, Princeton Univ. Press, Princeton, N. J.

[Roc73a] Rockafellar, R. T., 1973. "A Dual Approach to Solving Nonlinear Programming Problems by Unconstrained Optimization," Math. Programming, pp. 354-373.

[Roc73b] Rockafellar, R. T., 1973. "The Multiplier Method of Hestenes and Powell Applied to Convex Programming," J. Opt. Th. and Appl., Vol. 12, pp. 555-562.

[Roc74] Rockafellar, R. T., 1974. "Augmented Lagrange Multiplier Functions and Duality in Nonconvex Programming," SIAM J. on Control, Vol. 12, pp. 268-285.

[Roc76a] Rockafellar, R. T., 1976. "Monotone Operators and the Proximal Point Algorithm," SIAM J. on Control and Optimization, Vol. 14, pp. 877-898.

[Roc76b] Rockafellar, R. T., 1976. "Augmented Lagrangians and Applications of the Proximal Point Algorithm in Convex Programming," Math. Operations Res., Vol. 1, pp. 97-116.

[Roc76c] Rockafellar, R. T., 1976. "Solving a Nonlinear Programming Problem by Way of a Dual Problem," Symp. Matematica, Vol. 27, pp. 135-160.

[Roc81] Rockafellar, R. T., 1981. "Monotropic Programming: Descent Algorithms and Duality," in Nonlinear Programming 4, by Mangasarian, O. L., Meyer, R. R., and Robinson, S. M., (Eds.), Academic Press, N. Y., pp. 327-366.

[Roc84] Rockafellar, R. T., 1984. Network Flows and Monotropic Optimization, Wiley, N. Y.; republished by Athena Scientific, Belmont, MA, 1998.

[Roc90] Rockafellar, R. T., 1990. "Computational Schemes for Solving Large-Scale Problems in Extended Linear-Quadratic Programming," Math. Programming, Vol. 48, pp. 447-474.

[Roc93] Rockafellar, R. T., 1993. "Lagrange Multipliers and Optimality," SIAM Review, Vol. 35, pp. 183-238.

[Ros60a] Rosenbrock, H. H., 1960. "An Automatic Method for Finding the Greatest or Least Value of a Function," Computer J., Vol. 3, pp. 175-184.

[Ros60b] Rosen, J. B., 1960. "The Gradient Projection Method for Nonlinear Programming, Part I, Linear Constraints," SIAM J. Applied Math., Vol. 8, pp. 514-553.

[RuK04] Rubinstein, R. Y., and Kroese, D. P., 2004. The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Springer, N. Y.

[Rud76] Rudin, W., 1976. Real Analysis, McGraw-Hill, N. Y.

[Rus86] Ruszczynski, A., 1986. "A Regularized Decomposition Method for Minimizing a Sum of Polyhedral Functions," Math. Programming, Vol. 35, pp. 309-333.

[Rus89] Ruszczynski, A., 1989. "An Augmented Lagrangian Decomposition Method for Block Diagonal Linear Programming Problems," Operations Res. Letters, Vol. 8, pp. 287-294.

[Rus95] Ruszczynski, A., 1995. "On Convergence of an Augmented Lagrangian Decomposition Method for Sparse Convex Optimization," Math. of Operations Res., Vol. 20, pp. 634-656.

[Rus97] Ruszczynski, A., 1997. "Decomposition Methods in Stochastic Programming," Math. Programming, Vol. 79, pp. 333-353.

[Rus06] Ruszczynski, A., 2006. Nonlinear Optimization, Princeton Univ. Press, Princeton, N. J.

[SBC93] Saarinen, S., Bramley, R., and Cybenko, G., 1993. "Ill-Conditioning in Neural Network Training Problems," SIAM J. Sci. Comput., Vol. 14, pp. 693-714.

[SBK64] Shah, B., Buehler, R., and Kempthorne, O., 1964. "Some Algorithms for Minimizing a Function of Several Variables," J. Soc. Indust. Appl. Math., Vol. 12, pp. 74-92.

[SFR09] Schmidt, M., Fung, G., and Rosales, R., 2009. "Optimization Methods for ℓ_1-Regularization," Univ. of British Columbia, Technical Report TR-2009-19.

[SHH62] Spendley, W. G., Hext, G. R., and Himsworth, F. R., 1962. "Sequential Application of Simplex Designs in Optimisation and Evolutionary Operation," Technometrics, Vol. 4, pp. 441-461.

[SHM16] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser J., et al. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search," Nature, Vol. 529, pp. 484-489.

[SKS12] Schmidt, M., Kim, D., and Sra, S., 2012. "Projected Newton-Type Methods in Machine Learning," in Optimization for Machine Learning, by Sra, S., Nowozin, S., and Wright, S. J., (eds.), MIT Press, Cambridge, MA, pp. 305-329.

[SLB13] Schmidt, M., Le Roux, N., and Bach, F., 2013. "Minimizing Finite Sums with the Stochastic Average Gradient," arXiv preprint arXiv:1309.2388.

[SaS86] Saad, Y., and Schultz, M. H., 1986. "GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems," SIAM J. Sci. Statist. Comput., Vol. 7, pp. 856-869.

[Sah96] Sahinidis, N. V., 1996. "BARON: A General Purpose Global Optimization Software Package," Journal of Global Optimization, Vol. 8, pp. 201-205.

[Sah04] Sahinidis, N. V., 2004. "Optimization Under Uncertainty: State-of-the-Art and Opportunities," Computers and Chemical Engineering, Vol. 28, pp. 971-983.

[Sak66] Sakrison, D. T., 1966. "Stochastic Approximation: A Recursive Method for Solving Regression Problems," in Advances in Communication Theory and Applications, 2, A. V. Balakrishnan, ed., Academic Press, NY, pp. 51-106.

[ScF14] Schmidt, M., and Friedlander, M. P., 2014. "Coordinate Descent Converges Faster with the Gauss-Southwell Rule than Random Selection," Advances in Neural Information Processing Systems 27 (NIPS 2014).

[Sch82] Schnabel, R. B., 1982. "Determining Feasibility of a Set of Nonlinear Inequality Constraints," Math. Programming Studies, Vol. 16, pp. 137-148.

[Sch86] Schrijver, A., 1986. Theory of Linear and Integer Programming, Wiley, N. Y.

[Sch93] Schrijver, A., 1993. Combinatorial Optimization: Polyhedra and Efficiency, Springer, N. Y.

[Sch10] Schmidt, M., 2010. "Graphical Model Structure Learning with L1-Regularization," PhD Thesis, Univ. of British Columbia.

[Sch12] Schittkowski, K., 2012. Nonlinear Programming Codes: Information, Tests, Performance, Springer Science and Business Media.

[SeS86] Sen, S., and Sherali, H. D., 1986. "A Class of Convergent Primal-Dual Subgradient Algorithms for Decomposable Convex Programs," Math. Programming, Vol. 35, pp. 279-297.

[ShA99] Sherali, H. D., and Adams, W. P., 1999. A Reformulation-Linearization Technique for Solving Discrete and Continuous Nonconvex Problems, Kluwer, Boston.

[ShD14] Shapiro, A., and Dentcheva D., 2014. Lectures on Stochastic Programming: Modeling and Theory, SIAM, Phila.

[Sha70] Shanno, D. F., 1970. "Conditioning of Quasi-Newton Methods for Function Minimization," Math. Comput., Vol. 27, pp. 647-656.

[Sha78] Shanno, D. F., 1978. "Conjugate Gradient Methods with Inexact Line Searches," Math. of Operations Res., Vol. 3, pp. 244-256.

[Sha79] Shapiro, J. E., 1979. Mathematical Programming Structures and Algorithms, Wiley, N. Y.

[Sha88] Shapiro, A., 1988. "Sensitivity Analysis of Nonlinear Programs and Differentiability Properties of Metric Projections," SIAM J. on Control and Optimization, Vol. 26, pp. 628-645.

[Sho85] Shor, N. Z., 1985. Minimization Methods for Nondifferentiable Functions, Springer-Verlag, Berlin.

[Sho98] Shor, N. Z., 1998. Nondifferentiable Optimization and Polynomial Problems, Kluwer, Dordrecht, the Netherlands.

[Sla50] Slater, M., 1950. "Lagrange Multipliers Revisited: A Contribution to Non-Linear Programming," Cowles Commission Discussion Paper, Math. 403.

[Sol98] Solodov, M. V., 1998. "Incremental Gradient Algorithms with Stepsizes Bounded Away from Zero," Computational Optimization and Applications, Vol. 11, pp. 23-35.

[Son86] Sonnevend, G., 1986. "An "Analytical Centre" for Polyhedrons and New Classes of Global Algorithms for Linear (Smooth, Convex) Programming," Lecture Notes in Control and Information Sciences, Vol. 84, pp. 866-878.

[Spa03] Spall, J. C., 2003. Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control, J. Wiley, Hoboken, N. J.

[Spa12] Spall, J. C., 2012. "Cyclic Seesaw Process for Optimization and Identification," J. of Optimization Theory and Applications, Vol. 154, pp. 187-208.

[Spi85] Spingarn, J. E., 1985. "Applications of the Method of Partial Inverses to Convex Programming: Decomposition," Math. Programming, Vol. 32, pp. 199-223.

[StW75] Stephanopoulos, G., and Westerberg, A. W., 1975. "The Use of Hestenes' Method of Multipliers to Resolve Dual Gaps in Engineering System Optimization," J. Opt. Th. and Applications, Vol. 15, pp. 285-309.

[Str76] Strang, G., 1976. Linear Algebra and Its Applications, Academic Press, N. Y.

[TBA86] Tsitsiklis, J. N., Bertsekas, D. P., and Athans, M., 1986. "Distributed Asynchronous Deterministic and Stochastic Gradient Optimization Algorithms," IEEE Trans. on Aut. Control, Vol. AC-31, pp. 803-812.

[TBT90] Tseng, P., Bertsekas, D. P., and Tsitsiklis, J. N., 1990. "Partially Asynchronous Algorithms for Network Flow and Other Problems," SIAM J. on Control and Optimization, Vol. 28, pp. 678-710.

[TTA15] Toulis, P., Tran, D., and Airoldi, E. M., 2015. "Stability and Optimality in Stochastic Gradient Descent," arXiv preprint arXiv:1505.02417.

[TZY95] Tapia, R. A., Zhang, Y., and Ye, Y., 1995. "On the Convergence of the Iteration Sequence in Primal-Dual Interior-Point Methods," Math. Programming, Vol. 68, pp. 141-154.

[TaM85] Tanikawa, A., and Mukai, H., 1985. "A New Technique for Nonconvex Primal-Dual Decomposition," IEEE Trans. on Aut. Control, Vol. AC-30, pp. 133-143.

[TaP13] Talischi, C., and Paulino, G. H., 2013. "A Consistent Operator Splitting Algorithm and a Two-Metric Variant: Application to Topology Optimization," arXiv preprint arXiv:1307.5100.

[Tap77] Tapia, R. A., 1977. "Diagonalized Multiplier Methods and Quasi-Newton Methods for Constrained Minimization," J. Opt. Th. and Applications, Vol. 22, pp. 135-194.

[Teb92] Teboulle, M., 1992. "Entropic Proximal Mappings with Applications to Nonlinear Programming," Math. Operations Res., Vol. 17, pp. 1-21.

[ToT90] Toint, P. L., and Tuyttens, D., 1990. "On Large Scale Nonlinear Network Optimization," Math. Programming, Vol. 48, pp. 125-159.

[ToV67] Topkis, D. M., and Veinott, A. F., 1967. "On the Convergence of Some Feasible Directions Algorithms for Nonlinear Programming," SIAM J. on Control, Vol. 5, pp. 268-279.

[Tor91] Torczon, V., 1991. "On the Convergence of the Multidimensional Search Algorithm," SIAM J. on Optimization, Vol. 1, pp. 123-145.

[TrW80] Traub, J. F., and Wozniakowski, H., 1980. A General Theory of Optimal Algorithms, Academic Press, N. Y.

[TsB86] Tsitsiklis, J. N., and Bertsekas, D. P., 1986. "Distributed Asynchronous Optimal Routing in Data Networks," IEEE Trans. on Automatic Control, Vol. 31, pp. 325-331.

[TsB87] Tseng, P., and Bertsekas, D. P., 1987. "Relaxation Methods for Problems with Strictly Convex Separable Costs and Linear Constraints," Math. Programming, Vol. 38, pp. 303-321.

[TsB90] Tseng, P., and Bertsekas, D. P., 1990. "Relaxation Methods for Monotropic Programs," Math. Programming, Vol. 46, pp. 127-151.

[TsB91] Tseng, P., and Bertsekas, D. P., 1991. "Relaxation Methods for Problems with Strictly Convex Costs and Linear Constraints," Math. Operations Res., Vol. 16, pp. 462-481.

[TsB93] Tseng, P., and Bertsekas, D. P., 1993. "On the Convergence of the Exponential Multiplier Method for Convex Programming," Math. Programming, Vol. 60, pp. 1-19.

[TsB00] Tseng, P., and Bertsekas, D. P., 2000. "An Epsilon-Relaxation Method for Separable Convex Cost Generalized Network Flow Problems," Math. Programming, Vol. 88, pp. 85-104.

[Tse89] Tseng, P., 1989. "A Simple Complexity Proof for a Polynomial-Time Linear Programming Algorithm," Operations Res. Letters, Vol. 8, pp. 155-159.

[Tse90] Tseng, P., 1990. "Dual Ascent Methods for Problems with Strictly Convex Costs and Linear Constraints: A Unified Approach," SIAM J. on Control and Optimization, Vol. 28, pp. 214-242.

[Tse91a] Tseng, P., 1991. "On the Rate of Convergence of a Partially Asynchronous Gradient Projection Algorithm," SIAM J. on Optimization, Vol. 4, pp. 603-619.

[Tse91b] Tseng, P., 1991. "Relaxation Method for Large Scale Linear Programming using Decomposition," Math. of Operations Res., Vol. 17, pp. 859-880.

[Tse91c] Tseng, P., 1991. "Decomposition Algorithm for Convex Differentiable Minimization," J. Opt. Theory and Appl., Vol. 70, pp. 109-135.

[Tse92] Tseng, P., 1992. "Complexity Analysis of a Linear Complementarity Algorithm Based on a Lyapunov Function," Math. Programming, Vol. 53, pp. 297-306.

[Tse93] Tseng, P., 1993. "Dual Coordinate Ascent Methods for Non-Strictly Convex Minimization," Math. Programming, Vol. 59, pp. 231-247.

[Tse95a] Tseng, P., 1995. "Fortified-Descent Simplicial Search Method," Report, Dept. of Math., University of Washington, Seattle, Wash.; also in SIAM J. on Optimization, Vol. 10, 2000, pp. 269-288.

[Tse95b] Tseng, P., 1995. "Simplified Analysis of an $O(nL)$-Iteration Infeasible Predictor-Corrector Path Following Method for Monotone LCP,"in Recent Trends in Optimization Theory and Applications, Agarwal, R. P., (Ed.), World Scientific, pp. 423-434.

[Tse98] Tseng, P., 1998. "Incremental Gradient(-Projection) Method with Momentum Term and Adaptive Stepsize Rule," SIAM J. on Optimization, Vol. 8, pp. 506-531.

[Tse00] Tseng, P., 2000. "A Modified Forward-Backward Splitting Method for Maximal Monotone Mappings," SIAM J. on Control and Optimization, Vol. 38, pp. 431-446.

[Tse01a] Tseng, P., 2001. "Convergence of Block Coordinate Descent Methods for Non-differentiable Minimization," J. Optim. Theory Appl., Vol. 109, pp. 475-494.

[Tse01b] Tseng, P., 2001. "An Epsilon Out-of-Kilter Method for Monotropic Programming," Math. of Operations Research, Vol. 26, pp. 221-233.

[Tse04] Tseng, P., 2004. "An Analysis of the EM Algorithm and Entropy-Like Proximal Point Methods," Math. Operations Research, Vol. 29, pp. 27-44.

[Tse08] Tseng, P., 2008. "On Accelerated Proximal Gradient Methods for Convex-Concave Optimization," Report, Math. Dept., Univ. of Washington.

[VaB95] Vandenberghe, L., and Boyd, S., 1995. "A Primal-Dual Potential Reduction Method for Problems Involving Matrix Inequalities," Math. Programming, Vol. 69, pp. 205-236.

[VaW89] Varaiya, P., and Wets, R. J-B., 1989. "Stochastic Dynamic Optimization Approaches and Computation," Mathematical Programming: State of the Art, M. Iri and K. Tanabe (eds.), Kluwer, Boston, pp. 309-332.

[VeH93] Ventura, J. A., and Hearn, D. W., 1993. "Restricted Simplicial Decomposition for Convex Constrained Problems," Math. Programming, Vol. 59, pp. 71-85.

[Ven67] Venter, J. H., 1967. "An Extension of the Robbins-Monro Procedure," Ann. Math. Statist., Vol. 38, pp. 181-190.

[WDS13] Weinmann, A., Demaret, L., and Storath, M., 2013. "Total Variation Regularization for Manifold-Valued Data," arXiv preprint arXiv:1312.7710.

[WHM13] Wang, X., Hong, M., Ma, S., Luo, Z. Q., 2013. "Solving Multiple-Block Separable Convex Minimization Problems Using Two-Block Alternating Direction Method of Multipliers," arXiv preprint arXiv:1308.5294.

[WQB98] Wei, Z., Qi, L., and Birge, J. R., 1998. "New Method for Nonsmooth Convex Optimization," J. of Inequalities and Applications, Vol. 2, pp. 157-179.

[WSK14] Wytock, M., Sra, S., and Kolter, J. K., 2014. "Fast Newton Methods for the Group Fused Lasso," Proc. of 2014 Conf. on Uncertainty in Artificial Intelligence.

[WaB13a] Wang, M., and Bertsekas, D. P., 2013. "Incremental Constraint Projection-Proximal Methods for Nonsmooth Convex Optimization," Lab. for Information and Decision Systems Report LIDS-P-2907, MIT; SIAM Journal on Optimization, Vol. 26, 2016, pp. 681-717.

[WaB13b] Wang, M., and Bertsekas, D. P., 2013. "Convergence of Iterative Simulation-Based Methods for Singular Linear Systems," Stochastic Systems, Vol. 3, pp. 38-95.

[WaB15] Wang, M., and Bertsekas, D. P., 2015. "Incremental Constraint Projection Methods for Variational Inequalities." Math. Programming," Vol. 150.2, pp. 321-363.

[WaB16] Wang, M., and Bertsekas, D. P., 2016. "Stochastic First-Order Methods with Random Constraint Projection." SIAM J. on Optimization, Vol. 26, pp. 681-717.

[WeO13] Wei, E., and Ozdaglar, A., 2013. "On the $O(1/k)$ Convergence of Asynchronous Distributed Alternating Direction Method of Multipliers," arXiv preprint arXiv:1307.8254.

[Web29] Weber, A., 1929. Theory of Location of Industries, (Engl. Transl. by C. J. Friedrich), Univ. of Chicago Press, Chicago, Ill.

[WiH60] Widrow, B., and Hoff, M. E., 1960. "Adaptive Switching Circuits," Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, Part 4, pp. 96-104.

[Wil63] Wilson, R. B., 1963. "A Simplicial Algorithm for Concave Programming," Ph.D. Thesis, Grad. Sch. Business Admin., Harvard Univ., Cambridge, MA.

[Wol98] Wolsey, L. A., 1998. Integer Programming, Wiley, N. Y.

[Wri92] Wright, S. J., 1992. "An Interior Point Algorithm for Linearly Constrained Optimization," SIAM J. on Optimization, Vol. 2, pp. 450-473.

[Wri93a] Wright, S. J., 1993. "Identifiable Surfaces in Constrained Optimization," SIAM J. on Control and Optimization, Vol. 31, pp. 1063-1079.

[Wri93b] Wright, S. J., 1993. "Interior Point Methods for Optimal Control of Discrete Time Systems," J. Opt. Theory and Appl., Vol. 77, pp. 161-187.

[Wri93c] Wright, S. J., 1993. "A Path-Following Infeasible-Interior-Point Algorithm for Linear Complementarity Problems," Optimization Methods and Software, Vol. 2, pp. 79-106.

[Wri94] Wright, S. J., 1994. "An Infeasible-Interior-Point Algorithm for Linear Complementarity Problems," Math. Programming, Vol. 67, pp. 29-52.

[Wri96] Wright, S. J., 1996. "A Path-Following Interior-Point Algorithm for Linear and Quadratic Problems," Annals of Operations Res., Vol. 62, pp. 103-130.

[Wri97a] Wright, S. J., 1997. Primal-Dual Interior Point Methods, SIAM, Phila., PA.

[Wri97b] Wright, S. J., 1997. "Applying New Optimization Algorithms to Model Predictive Control," Chemical Process Control-V, CACHE, AIChE Symposium Series No. 316, Vol. 93, pp. 147-155.

[Wri98] Wright, S. J., 1998. "Superlinear Convergence of a Stabilized SQP Method to a Degenerate Solution," Computational Optimization and Applications, Vol. 11, pp. 253-275.

[WuB01] Wu, C., and Bertsekas, D. P., 2001. "Distributed Power Control Algorithms for Wireless Networks," IEEE Trans. on Vehicular Technology, Vol. 50, pp. 504-514.

[YSQ14] You, K., Song, S., and Qiu, L., 2014. "Randomized Incremental Least Squares for Distributed Estimation Over Sensor Networks," Preprints of the 19th World Congress The International Federation of Automatic Control Cape Town, South Africa.

[YVG10] Yu, J., Vishwanathan, S. V. N., Gunter, S., and Schraudolph, N. N., 2010. "A Quasi-Newton Approach to Nonsmooth Convex Optimization Problems in Machine Learning," J. of Machine Learning Research, Vol. 11, pp. 1145-1200.

[Ye92] Ye, Y., 1992. "A Potential Reduction Algorithm Allowing Column Generation," SIAM J. on Optimization, Vol. 2, pp. 7-20.

[Ye97] Ye, Y., 1997. Interior Point Algorithms: Theory and Analysis, Wiley Interscience, N. Y.

[Ypm95] Ypma, T. J., 1995. "Historical Development of the Newton-Raphson Method," SIAM Review, Vol. 37, pp. 531-551.

[You15] Yousefpour, R., 2015. "Combination of Steepest Descent and BFGS Methods for Nonconvex Nonsmooth Optimization," Numerical Algorithms, pp. 1-34.

[ZLW99] Zhao, X., Luh, P. B., and Wang, J., 1999. "Surrogate Gradient Algorithm for Lagrangian Relaxation," J. Opt. Theory and Appl., Vol. 100, pp. 699-712.

[ZTP93] Zhang, Y., Tapia, R. A., and Potra, F., 1993. "On the Superlinear Convergence of Interior-Point Algorithms for a General Class of Problems," SIAM J. on Optimization, Vol. 3 pp. 413-422.

[Zal02] Zalinescu, C., 2002. Convex Analysis in General Vector Spaces, World Scientific, Singapore.

[Zan67a] Zangwill, W. I., 1967. "Minimizing a Function Without Calculating Derivatives," The Computer Journal, Vol. X, pp. 293-296.

[Zan67b] Zangwill, W. I., 1967. "Nonlinear Programming via Penalty Functions," Management Science, Vol. 13, pp. 344-358.

[Zan69] Zangwill, W. I., 1969. Nonlinear Programming, Prentice-Hall, Englewood Cliffs, N. J.

[ZhT92] Zhang, Y., and Tapia, R. A., 1992. "Superlinear and Quadratic Convergence of Primal-Dual Interior-Point Algorithms for Linear Programming Revisited," J. Opt. Theory and Appl., Vol. 73, pp. 229-242.

[ZhT93] Zhang, Y., and Tapia, R. A., 1993. "A Superlinearly Convergent Polynomial Primal-Dual Interior-Point Algorithm for Linear Programming," SIAM J. on Optimization, Vol. 3, pp. 118-133.

[Zho93] Zhou, L., 1993. "A Simple Proof of the Shapley-Folkman Theorem," Economic Theory, Vol. 3, pp. 371-372.

[Zhu95] Zhu, C., 1995. "On the Primal-Dual Steepest Descent Algorithm for Extended Linear-Quadratic Programming," SIAM J. on Optimization, Vol. 5, pp. 114-128.

[Zou60] Zoutendijk, G., 1960. Methods of Feasible Directions, Elsevier Publ. Co., Amsterdam.

[Zou76] Zoutendijk, G., 1976. Mathematical Programming Methods, North Holland, Amsterdam.