Springer Texts in Statistics

Kenneth Lange

Optimization

Second Edition

Springer

For further volumes:
http://www.springer.com/series/417

Kenneth Lange

Optimization

Second Edition

Springer

Kenneth Lange
Biomathematics, Human Genetics,
Statistics
University of California
Los Angeles, CA, USA

ISSN 1431-875X
ISBN 978-1-4614-5837-1 ISBN 978-1-4614-5838-8 (eBook)
DOI 10.1007/978-1-4614-5838-8
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012948598

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my daughters, Jane and Maggie

Preface to the Second Edition

If this book were my child, I would say that it has grown up after passing through a painful adolescence. It has rebelled at times over my indecision. While it has aspired to greatness, I have aspired to closure. Although the second edition of *Optimization* occasionally demanded more exertion than I could muster, writing it has broadened my intellectual horizons. I hope my own struggles to reach clarity will translate into an easier path for readers.

The book's stress on mathematical fundamentals continues. Indeed, I was tempted to re-title it *Mathematical Analysis and Optimization*. I resisted this temptation because my ultimate goal is still to teach optimization theory. Nonetheless, there is a new chapter on the gauge integral and expanded treatments of differentiation and convexity. The focus remains on finite-dimensional optimization. The sole exception to this rule occurs in the new chapter on the calculus of variations. In my view functional analysis is just too high a rung on the ladder of mathematical abstraction.

Covering all of optimization theory is simply out of the question. Even though the second edition is more than double the length of the first, many important topics are omitted. The most grievous omissions are the simplex algorithm of linear programming and modern interior point methods. Fortunately, there are many admirable books devoted to these subjects. My development of adaptive barrier methods and exact penalty methods also partially compensates.

In addition to the two chapters on integration and the calculus of variations, four new chapters treat block relaxation (block descent and block ascent) and various advanced topics in the convex calculus, including the

Fenchel conjugate, subdifferentials, duality, feasibility, alternating projections, projected gradient methods, exact penalty methods, and Bregman iteration. My own interests in data mining and biological applications have dictated the nature of these chapters. High-dimensional problems are driving the discipline of optimization. These are qualitatively different from traditional problems, and standard algorithms such as Newton's method are often impractical. Penalization, model sparsity, and the MM algorithm now assume dominant roles. Fortunately, many of the challenging modern problems can also be phrased as convex programs.

In the first edition I eschewed the convention of setting vectors and matrices in boldface type. In the second edition I embrace it. Although this decision improves readability, it carries with it some residual ambiguity. The main difficulty lies in distinguishing constant vectors and matrices from vector and matrix-valued functions. In general, I have elected to set functions in ordinary type even when they are vector or matrix valued. The exceptions occur in the calculus of variations, where functions are considered vectors in infinite-dimensional spaces. Thus, a function appears in ordinary type when its argument is displayed and in boldface type when its argument is omitted.

Many people have helped me prepare this second edition. Hua Zhou and Tongtong Wu, my former postdoctoral fellows, and Eric Chi, my current postdoctoral fellow, deserve special credit. Without their assistance, the book would have been intellectually duller and graphically drearier. I would also like to thank my former doctoral students David Alexander, David Hunter, Mary Sehl, and Jinjin Zhou for proofreading and critiquing the new material. The students in my optimization class checked most of the exercises. I am indebted to Forrest Crawford, Gabriela Cybis, Gary Evans, Mitchell Johnson, Wesley Kerr, Kevin Keys, Omid Kohannim, Lewis Lee, Matthew Levinson, Lae Un Kim, John Ranola, and Moses Wilkes for their help.

Finally, let me report on my daughters Maggie and Jane, to whom this book is dedicated. Maggie is now embarked on a postdoctoral fellowship in medical ethics at Macquarie University in Sydney, Australia. Jane is completing her dissertation in biostatistics at the University of Washington. Assimilating their scholarship will keep me young for many years to come.

Preface to the First Edition

This foreword, like many forewords, was written afterwards. That is just as well because the plot of the book changed during its creation. It is painful to recall how many times classroom realities forced me to shred sections and start anew. Perhaps such adjustments are inevitable. Certainly I gained a better perspective on the subject over time. I also set out to teach optimization theory and wound up teaching mathematical analysis. The students in my classes are no less bright and eager to learn about optimization than they were a generation ago, but they tend to be less prepared mathematically. So what you see before you is a compromise between a broad survey of optimization theory and a textbook of analysis. In retrospect, this compromise is not so bad. It compelled me to revisit the foundations of analysis, particularly differentiation, and to get right to the point in optimization theory.

The content of courses on optimization theory varies tremendously. Some courses are devoted to linear programming, some to nonlinear programming, some to algorithms, some to computational statistics, and some to mathematical topics such as convexity. In contrast to their gaps in mathematics, most students now come well trained in computing. For this reason, there is less need to emphasize the translation of algorithms into computer code. This does not diminish the importance of algorithms, but it does suggest putting more stress on their motivation and theoretical properties. Fortunately, the dichotomy between linear and nonlinear programming is fading. It makes better sense pedagogically to view linear programming as a special case of nonlinear programming. This is the attitude taken in

the current book, which makes little mention of the simplex method and develops interior point methods instead. The real bridge between linear and nonlinear programming is convexity. I stress not only the theoretical side of convexity but also its applications in the design of algorithms for problems with either large numbers of parameters or nonlinear constraints.

This graduate-level textbook presupposes knowledge of calculus and linear algebra. I develop quite a bit of mathematical analysis from scratch and feature a variety of examples from linear algebra, differential equations, and convexity theory. Of course, the greater the prior exposure of students to this background material, the more quickly the beginning chapters can be covered. If the need arises, I recommend the texts [82, 134, 135, 188, 222, 223] for supplementary reading. There is ample material here for a fast-paced, semester-long course. Instructors should exercise their own discretion in skipping sections or chapters. For example, Chap. 10 on the EM algorithm primarily serves the needs of students in biostatistics and statistics. Overall, my intended audience includes graduate students in applied mathematics, biostatistics, computational biology, computer science, economics, physics, and statistics. To this list I would like to add upper-division majors in mathematics who want to see some rigorous mathematics with real applications. My own background in computational biology and statistics has obviously dictated many of the examples in the book.

Chapter 1 starts with a review of exact methods for solving optimization problems. These are methods that many students will have seen in calculus, but repeating classical techniques with fresh examples tends simultaneously to entertain, instruct, and persuade. Some of the exact solutions also appear later in the book as parts of more complicated algorithms.

Chapters 2 through 4 review undergraduate mathematical analysis. Although much of this material is standard, the examples may keep the interest of even the best students. Instructors should note that Carathéodory's definition rather than Fréchet's definition of differentiability is adopted. This choice eases the proof of many results. The gauge integral, another good addition to the calculus curriculum, is mentioned briefly.

Chapter 5 gets down to the serious business of optimization theory. McShane's clever proof of the necessity of the Karush–Kuhn–Tucker conditions avoids the complicated machinery of manifold theory and convex cones. It makes immediate use of the Mangasarian–Fromovitz constraint qualification. To derive sufficient conditions for optimality, I introduce second differentials by extending Carathéodory's definition of first differentials. To my knowledge, this approach to second differentials is new. Because it melds so effectively with second-order Taylor expansions, it renders critical proofs more transparent.

Chapter 6 treats convex sets, convex functions, and the relationship between convexity and the multiplier rule. The chapter concludes with the derivation of some of the classical inequalities of probability theory. Prior exposure to probability theory will obviously be an asset for readers here.

Chapters 8 and 9 introduce the MM and EM algorithms. These exploit convexity and the notion of majorization in transferring minimization of the objective function to a surrogate function. Minimizing the surrogate function drives the objective function downhill. The EM algorithm, which is a special case of the MM algorithm, arose in statistics. It is a slight misnomer to call these algorithms. They are really prescriptions for constructing algorithms. It takes experience and skill to wield these tools effectively, so careful attention to the examples is imperative.

Chapter 10 covers Newton's method and its statistical variants, scoring and the Gauss–Newton algorithm. To make this material less dependent on statistical knowledge, I have tried to motivate several algorithms from the perspective of positive definite approximation of the second differential of the objective function. Chapter 11 covers the conjugate gradient algorithm, quasi-Newton algorithms, and the method of trust regions. These classical subjects are in danger of being dropped from the curriculum of nonlinear programming. In my view, this would be a mistake.

Chapter 12 is devoted to convergence questions, both local and global. This material beautifully illustrates the virtues of soft analysis. Instructors wanting to emphasize practical matters may be tempted to sacrifice Chap. 12, but the constant interplay between theory and practice in designing new algorithms argues for its inclusion.

Chapter 13 on convex programming ends the book where more advanced treatises would start. I discuss adaptive barrier methods as a novel application of the MM algorithm, Dykstra's algorithm for finding feasible points in convex programming, and the rudiments of duality theory. These topics belong to the promised land. All you get here is a glimpse from the mountaintop looking out across the river.

Let me add a few words about notation. Lower-division undergraduate texts carefully distinguish between scalars and vectors by setting vectors in boldface type. This convention is considered cumbersome in higher mathematics and is dropped. However, mathematical analysis is plagued by a proliferation of superscripts and subscripts. I prefer to avoid superscripts because of the possible confusion with powers. This decision makes it difficult to distinguish an element of a vector sequence from a component of a vector. My compromise is to represent the mth entry of a vector sequence as $x_{(m)}$ and the nth component of that sequence element as x_{mn}. Similar conventions hold for matrices. Thus, M_{jkl} is the entry in row k and column l of the jth matrix $M_{(j)}$ of a sequence of matrices. Elements of scalar sequences are subscripted in the usual fashion without the enclosing parentheses.

I would like to thank my UCLA students for their help and patience in debugging this text. If it is readable, it is because their questions cut through the confusion. In retrospect, there were more contributing students than I can credit. Let me single out Jason Aten, Lara Bauman, Brian Dolan,

Wei-Hsun Liao, Andrew Nevai-Tucker, Robert Rovetti, and Andy Yip. Paul Maranian kindly prepared the index and proofread my last draft. Finally, I thank my ever helpful and considerate editor, John Kimmel.

I dedicate this book to my daughters, Jane and Maggie. It has been a privilege to be your father. Now that you are adults, I hope you can find the same pleasure in pursuing ideas that I have found in my professional life.

Contents

1
Elementary Optimization

1.1 Introduction

As one of the oldest branches of mathematics, optimization theory served as a catalyst for the development of geometry and differential calculus [258]. Today it finds applications in a myriad of scientific and engineering disciplines. The current chapter briefly surveys material that most students encounter in a good calculus course. This review is intended to showcase the variety of methods used to find the exact solutions of elementary problems. We will return to some of these methods later from a more rigorous perspective. One of the recurring themes in optimization theory is its close connection to inequalities. This chapter introduces a few classical inequalities; more will appear in succeeding chapters.

1.2 Univariate Optimization

The first optimization problems students encounter are univariate. Solution techniques for these simple problems are hardly limited to differential calculus. Our first two examples illustrate how plane geometry and algebra can play a role.

Example 1.2.1 *Heron's Problem*

The ancient mathematician Heron of Alexandria posed one of the earliest optimization problems. Consider the two points A and B and the line

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_1,

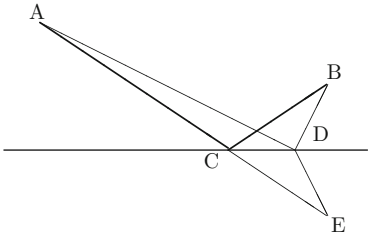

FIGURE 1.1. Diagram for Heron's problem

containing the points C and D drawn in Fig. 1.1. Heron's problem is to find the position of C on the line that minimizes the sum of the distances $|AC|$ and $|BC|$. The correct choice of C is determined by reflecting B across the given line to give E. From E we draw the line to A and note its intersection C with the original line. To demonstrate that C minimizes the total distance $|AC| + |BC|$, consider any other point D on the original horizontal line. By symmetry, $|AC| + |BC| = |AC| + |CE|$. Similarly, by symmetry, $|AD| + |BD| = |AD| + |DE|$. Because the sum of the lengths of two sides of a triangle exceeds the length of the third side, it follows immediately that $|AC| + |CE| \leq |AD| + |DE|$. Thus, C solves Heron's problem.

This example also has an optical interpretation. If we imagine that the horizontal line containing C lies on a mirror, then light travels along the quickest path between A and B via the mirror. This extremal principle can be explained by considering the wave nature of light, but we omit the long digression. It is interesting that the geometric argument automatically implies that the angle of incidence of the light ray equals the angle of reflection. ■

Example 1.2.2 *Simple Arithmetic-Geometric Mean Inequality*

If x and y are two nonnegative numbers, then $\sqrt{xy} \leq (x+y)/2$. This can be proved by noting that

$$\begin{aligned} 0 &\leq (\sqrt{x} - \sqrt{y})^2 \\ &= x - 2\sqrt{xy} + y. \end{aligned}$$

Evidently, equality holds if and only if $x = y$. As an application consider maximization of the function $f(x) = x(1-x)$. The inequality just derived shows that

$$f(x) \leq \left(\frac{x+1-x}{2}\right)^2 = \frac{1}{4},$$

with equality when $x = 1/2$. Thus, the maximum of $f(x)$ occurs at the point $x = 1/2$. One can interpret $f(x)$ as the area of a rectangle of fixed

perimeter 2 with sides of length x and $1-x$. The rectangle with the largest area is a square. The function $2f(x)$ is interpreted in population genetics as the fraction of a population that is heterozygous at a genetic locus with two alleles having frequencies x and $1-x$. Heterozygosity is maximized when the two alleles are equally frequent. ■

With the advent of differential calculus, it became possible to solve optimization problems more systematically. Before discussing concrete examples, it is helpful to review some of the standard theory. We restrict attention to real-valued functions defined on intervals. The intervals in question can be finite or infinite in extent and open or closed at either end. According to a celebrated theorem of Weierstrass, a continuous function $f(x)$ defined on a closed finite interval $[a, b]$ attains its minimum and maximum values on the interval. These extremal values are necessarily finite. The extremal points can occur at the endpoints a or b or at an interior point c. In the latter case, when $f(x)$ is differentiable, an even older principle of Fermat requires that $f'(c) = 0$. The stationarity condition $f'(c) = 0$ is no guarantee that c is optimal. It is possible for c to be a local rather than a global minimum or maximum or even to be a saddle point. However, it usually is a simple matter to check the endpoints a and b and any stationary points c. Collectively, these points are known as critical points.

If the domain of $f(x)$ is not a closed finite interval $[a, b]$, then the minimum or maximum of $f(x)$ may not exist. One can usually rule out such behavior by examining the limit of $f(x)$ as x approaches an open boundary. For example on the interval $[a, \infty)$, if $\lim_{x\to\infty} f(x) = \infty$, then we can be sure that $f(x)$ possesses a minimum on the interval, and we can find it by comparing the values of $f(x)$ at a and any stationary points c. On a half open interval such as $(a, b]$, we can likewise find a minimum whenever $\lim_{x\to a} f(x) = \infty$. Similar considerations apply to finding a maximum.

The nature of a stationary point c can be determined by testing the second derivative $f''(c)$. If $f''(c) > 0$, then c at least qualifies as a local minimum. Similarly, if $f''(c) < 0$, then c at least qualifies as a local maximum. The indeterminate case $f''(c) = 0$ is consistent with c being a local minimum, maximum, or saddle point. For example, $f(x) = x^4$ attains its minimum at 0 while $f(x) = x^3$ has a saddle point there. In both cases, $f''(0) = 0$. Higher-order derivatives or other qualitative features of $f(x)$ must be invoked to discriminate among these possibilities. If $f''(x) \geq 0$ for all x, then $f(x)$ is said to be convex. Any stationary point of a convex function is a minimum. If $f''(x) > 0$ for all x, then $f(x)$ is strictly convex, and there is at most one stationary point. Whenever it exists, the stationary point furnishes the global minimum. A concave function satisfies $f''(x) \leq 0$ for all x. Concavity bears the same relation to maxima as convexity does to minima.

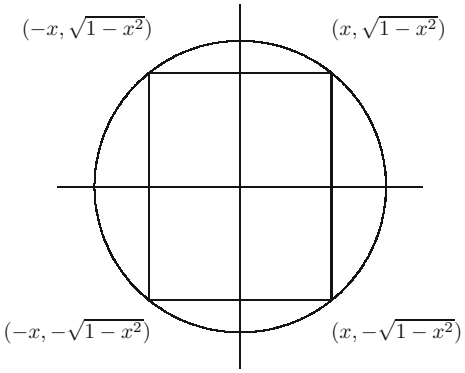

FIGURE 1.2. A *rectangle* inscribed in a *circle*

Example 1.2.3 (Kepler) *Largest Rectangle Inscribed in a Circle*

Figure 1.2 depicts a rectangle inscribed in a circle of radius 1 centered at the origin. If we suppose the vertical sides of the rectangle cross the horizontal axis at the point $(-x, 0)$ and $(x, 0)$, then Pythagoras's theorem gives the coordinates of the corners as noted in the figure. Here x is restricted to the interval $[0, 1]$. From these coordinates, it follows that the rectangle has area

$$f(x) \quad = \quad 4x\sqrt{1-x^2}.$$

Because $f(0) = f(1) = 0$, the maximum of $f(x)$ occurs somewhere in the open interval $(0, 1)$. Straightforward differentiation shows that

$$f'(x) \quad = \quad 4\sqrt{1-x^2} - \frac{4x^2}{\sqrt{1-x^2}}.$$

Setting $f'(x)$ equal to 0 and solving for x gives the critical point $x = 1/\sqrt{2}$ and the critical value $f(1/\sqrt{2}) = 2$. Since there is only one critical point on $(0, 1)$, it must be the maximum point. The largest inscribed rectangle is a square as expected. ■

Example 1.2.4 *Snell's Law*

Snell's law refers to an optical experiment involving two different media, say air and water. The less dense the medium, the faster light travels. Since light takes the path of least time, it bends at an interface such as that indicated by the horizontal axis in Fig. 1.3. Here we ask for the point $(x, 0)$ on the interface intersecting the light path. If we assume the speed of light above the interface is s_1 and below the interface is s_2, then the total travel time is given by

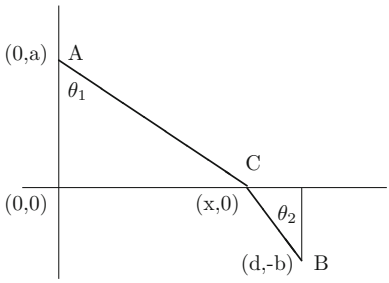

FIGURE 1.3. Diagram for Snell's law

$$f(x) = \frac{\sqrt{a^2+x^2}}{s_1} + \frac{\sqrt{b^2+(d-x)^2}}{s_2}. \tag{1.1}$$

The derivative of $f(x)$ is

$$f'(x) = \frac{x}{s_1\sqrt{a^2+x^2}} - \frac{d-x}{s_2\sqrt{b^2+(d-x)^2}}.$$

The minimum exists because $\lim_{|x|\to\infty} f(x) = \infty$. Although finding a stationary point is difficult, it is clear from the monotonicity of the functions $x/\left(s_1\sqrt{a^2+x^2}\right)$ and $(d-x)/\left(s_2\sqrt{b^2+(d-x)^2}\right)$ that it is unique. In trigonometric terms, Snell's law can be expressed as

$$\frac{\sin\theta_1}{s_1} = \frac{\sin\theta_2}{s_2}$$

using the angles at the minimum point as noted in Fig. 1.3. ■

Example 1.2.5 *The Functions* $f_n(x) = x^n e^x$

The functions $f_n(x) = x^n e^x$ for $n \geq 1$ exhibit interesting behavior. Figure 1.4 plots $f_n(x)$ for n between 1 and 3. It is clear that $\lim_{x\to-\infty} f_n(x) = 0$ and $\lim_{x\to\infty} f_n(x) = \infty$. These limits do not rule out the possibility of local maxima and minima. To find these we need

$$\begin{aligned} f_n'(x) &= (x^n + nx^{n-1})e^x \\ f_n''(x) &= [x^n + 2nx^{n-1} + n(n-1)x^{n-2}]e^x. \end{aligned}$$

Setting $f_n'(x) = 0$ produces the critical point $x = -n$, and when $n > 1$, the critical point $x = 0$. A brief calculation shows that $f_n''(-n) = (-n)^{n-1}e^{-n}$. Thus, $-n$ is a local minimum for n odd and a local maximum for n even. At 0 we have $f_2''(0) = 2$ and $f_n''(0) = 0$ for $n > 2$. Thus, the second derivative test fails for $n > 2$. However, it is clear from the variation of the sign of $f_n(x)$ to the right and left of 0 that 0 is a minimum of $f_n(x)$ for n even and

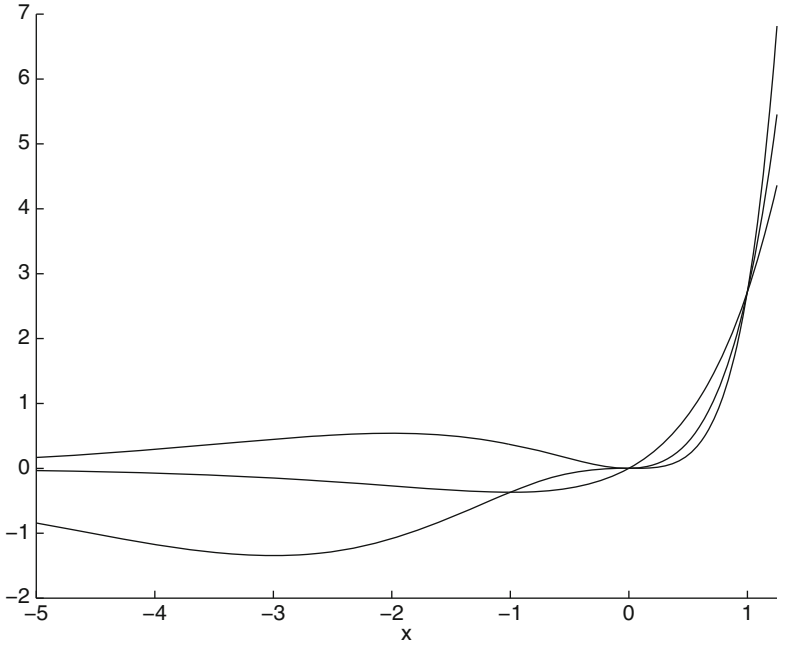

FIGURE 1.4. Plots of xe^x, x^2e^x, and x^3e^x

a saddle point of $f_n(x)$ for $n > 1$ and odd. One strength of modern graphing programs such as MATLAB is that they quickly suggest such conjectures. ■

Example 1.2.6 *Fenchel Conjugate of* $f_p(x) = |x|^p/p$ *for* $p > 1$

The Fenchel conjugate $f^\star(y)$ of a convex function $f(x)$ is defined by

$$f^\star(y) \quad = \quad \sup_x [yx - f(x)]. \tag{1.2}$$

Remarkably, $f^\star(y)$ is also convex. As a particular case of this result, we consider the Fenchel conjugate of $f_p(x)$. It turns out that $f_p^\star(y) = f_q(y)$, where

$$\frac{1}{p} + \frac{1}{q} \quad = \quad 1.$$

Here neither p nor q need be integers. According to the second derivative test, the function $f_p(x) = |x|^p/p$ is convex on the real line whenever $p > 1$. The possible failure of $f_p''(x)$ to exist at $x = 0$ does not invalidate this conclusion. To calculate $f_p^\star(y)$, we observe that $f_p'(x) = |x|^{p-1}\,\mathrm{sgn}(x)$. This clearly implies that $x = |y|^{1/(p-1)}\,\mathrm{sgn}(y)$ maximizes the concave function $g(x) = yx - f_p(x)$. At the maximum point

$$\begin{aligned} f_p^\star(y) &= yx - \frac{|x|^p}{p} \\ &= |y|^{1+1/(p-1)} - \frac{|y|^{p/(p-1)}}{p} \\ &= \frac{|y|^q}{q}, \end{aligned}$$

proving our claim.

Inserting the calculated value of $f_p^\star(y)$ in the definition (1.2) leads to Young's inequality

$$xy \leq \frac{|x|^p}{p} + \frac{|y|^q}{q}. \tag{1.3}$$

The double-dual identity $f_p^{\star\star}(x) = f_p(x)$ is a special case of a general result proved later in Proposition 14.3.2. Historically, the Fenchel conjugate was introduced by Legendre for smooth functions and later generalized by Fenchel to arbitrary functions. ■

1.3 Multivariate Optimization

Although multivariate optimization is more subtle, it typically parallels univariate optimization [125, 212, 247]. The most fundamental differences arise because of constraints. In unconstrained optimization, the right definitions and notation ease the generalization. Before discussing these issues of calculus, we look at two classical inequalities that can be established by purely algebraic techniques.

Example 1.3.1 *Cauchy-Schwarz Inequality*

Suppose $\boldsymbol{x}$ and $\boldsymbol{y}$ are any two points in R^n. The Cauchy-Schwarz inequality says

$$\left|\sum_{i=1}^n x_i y_i\right| \leq \left(\sum_{i=1}^n x_i^2\right)^{1/2} \left(\sum_{i=1}^n y_i^2\right)^{1/2}.$$

If we define the inner product

$$\boldsymbol{x}^* \boldsymbol{y} = \sum_{i=1}^n x_i y_i$$

using the transpose operator * and the Euclidean norm

$$\|\boldsymbol{x}\| = \left(\sum_{i=1}^n x_i^2\right)^{1/2},$$

then the inequality can be restated as $|\boldsymbol{x}^*\boldsymbol{y}| \leq \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|$. Equality occurs in the Cauchy-Schwarz inequality if and only if $\boldsymbol{y}$ is a multiple of $\boldsymbol{x}$ or vice versa.

In proving the inequality, we can immediately eliminate the case $\boldsymbol{x} = \boldsymbol{0}$ where all components of $\boldsymbol{x}$ are 0. Given that $\boldsymbol{x} \neq \boldsymbol{0}$, we introduce a scalar λ and consider the quadratic

$$\begin{aligned} 0 &\leq \|\lambda\boldsymbol{x} + \boldsymbol{y}\|^2 \\ &= \|\boldsymbol{x}\|^2\lambda^2 + 2\boldsymbol{x}^*\boldsymbol{y}\lambda + \|\boldsymbol{y}\|^2 \\ &= \frac{1}{a}(a\lambda + b)^2 + c - \frac{b^2}{a} \end{aligned}$$

with $a = \|\boldsymbol{x}\|^2$, $b = \boldsymbol{x}^*\boldsymbol{y}$, and $c = \|\boldsymbol{y}\|^2$. In order for this quadratic to be nonnegative for all λ, it is necessary and sufficient that $c - b^2/a \geq 0$, which is just an abbreviation for the Cauchy-Schwarz inequality. For the quadratic to attain the value 0, the condition $c - b^2/a = 0$ must hold. When the quadratic vanishes, $\boldsymbol{y} = -\lambda\boldsymbol{x}$. ■

Example 1.3.2 *Arithmetic-Geometric Mean Inequality*

One generalization of the simple arithmetic-geometric mean inequality of Example 1.2.2 takes the form

$$\sqrt[n]{x_1 \cdots x_n} \leq \frac{x_1 + \cdots + x_n}{n}, \tag{1.4}$$

where $x_1, \ldots, x_n$ are any n nonnegative numbers. For a purely algebraic proof of this fact, we first note that it is obvious if any $x_i = 0$. If all $x_i > 0$, then divide both sides of the inequality by $\sqrt[n]{x_1 \cdots x_n}$. This replaces x_i by $y_i = x_i / \sqrt[n]{x_1 \cdots x_n}$ and leads to the equality $\sqrt[n]{y_1 \cdots y_n} = 1$. It now suffices to prove that $y_1 + \cdots + y_n \geq n$, which is trivially valid when $n = 1$. For $n > 1$ we argue by induction. Clearly the assumption $\sqrt[n]{y_1 \cdots y_n} = 1$ implies that there are two numbers, say y_1 and y_2, with $y_1 \geq 1$ and $y_2 \leq 1$. If this is true, then $(y_1 - 1)(y_2 - 1) \leq 0$, or equivalently $y_1y_2 + 1 \leq y_1 + y_2$. Invoking the induction hypothesis, we now reason that

$$\begin{aligned} y_1 + \cdots + y_n &\geq 1 + y_1y_2 + y_3 + \cdots + y_n \\ &\geq 1 + (n - 1). \end{aligned}$$

■

As a prelude to discussing further examples, it is helpful to briefly summarize the theory to be developed later and often taken for granted in multidimensional calculus courses. The standard vocabulary and symbolism adopted here stress the minor adjustments necessary in going from one dimension to multiple dimensions.

For a real-valued function $f(\boldsymbol{x})$ defined on R^n, the differential $df(\boldsymbol{x})$ is the generalization of the derivative $f'(x)$. For our purposes, $df(\boldsymbol{x})$ is the row vector of partial derivatives; its transpose is the gradient vector $\nabla f(\boldsymbol{x})$. The symmetric matrix of second partial derivatives constitutes the second differential $d^2 f(\boldsymbol{x})$ or Hessian matrix. A stationary point $\boldsymbol{x}$ satisfies $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$. Fermat's principle says that all local maxima and minima on the interior of the domain of $f(\boldsymbol{x})$ are stationary points.

If $d^2 f(\boldsymbol{x})$ is positive definite at a stationary point $\boldsymbol{y}$, then $\boldsymbol{y}$ furnishes a local minimum. If $d^2 f(\boldsymbol{y})$ is negative definite, then $\boldsymbol{y}$ furnishes a local maximum. The function $f(\boldsymbol{x})$ is said to be convex if $d^2 f(\boldsymbol{x})$ is positive semidefinite for all $\boldsymbol{x}$; it is strictly convex if $d^2 f(\boldsymbol{x})$ is positive definite for all $\boldsymbol{x}$. Every stationary point of a convex function represents a global minimum. At most one stationary point exists per strictly convex function. Similar considerations apply to concave functions and global maxima, provided we substitute "negative" for "positive" throughout these definitions. These facts are rigorously proved in Chaps. 4, 5, and 6.

Example 1.3.3 *Least Squares Estimation*

Statisticians often estimate parameters by the method of least squares. To review the situation, consider n independent experiments with outcomes $y_1, \ldots, y_n$. We wish to predict y_i from p covariates (predictors) $x_{i1}, \ldots, x_{ip}$ known in advance. For instance, y_i might be the height of the ith child in a classroom of n children. Relevant predictors might be the heights x_{i1} and x_{i2} of i's mother and father and the sex of i coded as $x_{i3} = 1$ for a girl and $x_{i4} = 1$ for a boy. Here we take $p = 4$ and force $x_{i3}x_{i4} = 0$ so that only one sex is possible. If we use a linear predictor $\sum_{j=1}^{p} x_{ij}\theta_j$ of y_i, it is natural to estimate the regression coefficients θ_j by minimizing the sum of squares

$$f(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} x_{ij}\theta_j \Big)^2.$$

Differentiating $f(\boldsymbol{\theta})$ with respect to θ_j and setting the result equal to 0 produce

$$\sum_{i=1}^{n} x_{ij} y_i \;=\; \sum_{i=1}^{n} \sum_{k=1}^{p} x_{ij} x_{ik} \theta_k.$$

If we let $\boldsymbol{y}$ denote the column vector with entries y_i and $\boldsymbol{X}$ denote the matrix with entry x_{ij} in row i and column j, then these p normal equations can be written in vector form as

$$\boldsymbol{X}^* \boldsymbol{y} \;=\; \boldsymbol{X}^* \boldsymbol{X} \theta$$

and solved as

$$\hat{\boldsymbol{\theta}} \;=\; (\boldsymbol{X}^*\boldsymbol{X})^{-1}\boldsymbol{X}^*\boldsymbol{y}.$$

In order for the indicated matrix inverse $(\boldsymbol{X}^*\boldsymbol{X})^{-1}$ to exist, $n \geq p$ should hold and the matrix $\boldsymbol{X}$ must be of full rank. See Problem 22.

To check that our proposed solution $\hat{\boldsymbol{\theta}}$ represents the global minimum, we calculate the Hessian matrix $d^2 f(\boldsymbol{\theta})$. Its entries

$$\frac{\partial^2}{\partial\theta_j\partial\theta_k} f(\boldsymbol{\theta}) \;=\; 2\sum_{i=1}^{n} x_{ij}x_{ik}$$

permit us to identify $d^2 f(\boldsymbol{\theta})$ with the matrix $2\boldsymbol{X}^*\boldsymbol{X}$. Owing to the full rank assumption, the symmetric matrix $\boldsymbol{X}^*\boldsymbol{X}$ is positive definite. Hence, $f(\boldsymbol{\theta})$ is strictly convex, and $\hat{\boldsymbol{\theta}}$ is the global minimum. ■

1.4 Constrained Optimization

The subject of Lagrange multipliers has a strong geometric flavor. It deals with tangent vectors and directions of steepest ascent and descent. The classical theory, which is all we consider here, is limited to equality constraints. Inequality constraints were not introduced until later in the game.

The gradient direction $\nabla f(\boldsymbol{x}) = df(\boldsymbol{x})^*$ is the direction of steepest ascent of $f(\boldsymbol{x})$ near the point $\boldsymbol{x}$. We can motivate this fact by considering the linear approximation

$$f(\boldsymbol{x} + t\boldsymbol{u}) \;=\; f(\boldsymbol{x}) + tdf(\boldsymbol{x})\boldsymbol{u} + o(t)$$

for a unit vector $\boldsymbol{u}$ and a scalar t. The error term $o(t)$ becomes negligible compared to t as t decreases to 0. The inner product $df(\boldsymbol{x})\boldsymbol{u}$ in this approximation is greatest for the unit vector $\boldsymbol{u} = \nabla f(\boldsymbol{x})/\|\nabla f(\boldsymbol{x})\|$. Thus, $\nabla f(\boldsymbol{x})$ points locally in the direction of steepest ascent of $f(\boldsymbol{x})$. Similarly, $-\nabla f(\boldsymbol{x})$ points locally in the direction of steepest descent.

Now consider minimizing or maximizing $f(\boldsymbol{x})$ subject to the equality constraints $g_i(\boldsymbol{x}) = 0$ for $i = 1, \ldots, m$. A tangent direction $\boldsymbol{w}$ at the point $\boldsymbol{x}$ on the constraint surface satisfies $dg_i(\boldsymbol{x})\boldsymbol{w} = 0$ for all i. Of course, if the constraint surface is curved, we must interpret the tangent directions as specifying directions of infinitesimal movement. From the perpendicularity relation $dg_i(\boldsymbol{x})\boldsymbol{w} = 0$, it follows that the set of tangent directions is the orthogonal complement $S^{\perp}(\boldsymbol{x})$ of the vector subspace $S(\boldsymbol{x})$ spanned by the $\nabla g_i(\boldsymbol{x})$. To avoid degeneracies, the vectors $\nabla g_i(\boldsymbol{x})$ must be linearly independent. Figure 1.5 depicts level curves $g(\boldsymbol{x}) = c$ and gradients $\nabla g(\boldsymbol{x})$ for the function $\sin(x)\cos(y)$ over the square $[0, \pi] \times [-\frac{\pi}{2}, \frac{\pi}{2}]$. Tangent vectors are parallel to the level curves (contours) and perpendicular to the gradients (arrows).

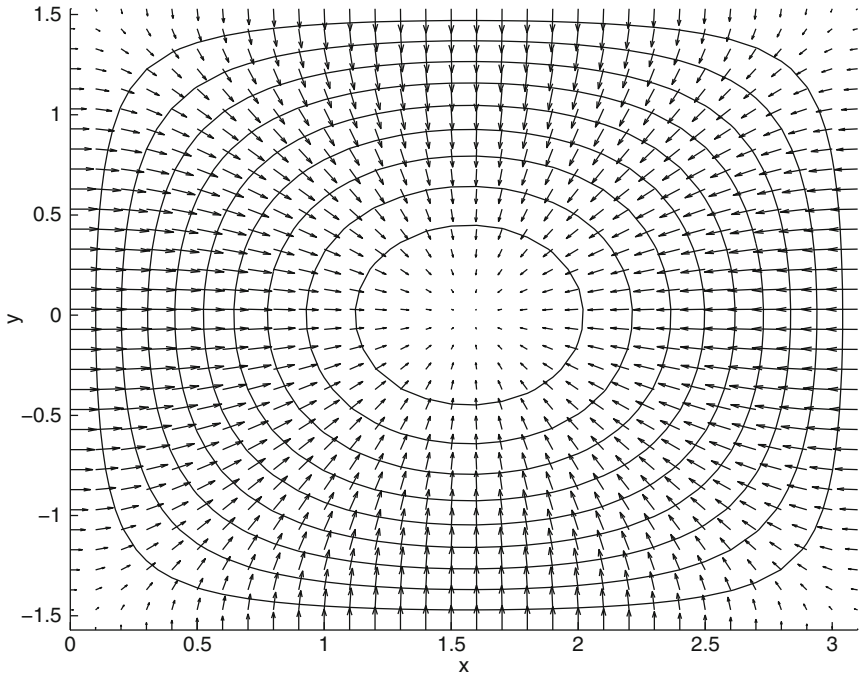

FIGURE 1.5. Level curves and steepest ascent directions for $\sin(x)\cos(y)$

At an optimal (or extremal) point $\boldsymbol{y}$, we have $df(\boldsymbol{y})\boldsymbol{w} = 0$ for every tangent direction $\boldsymbol{w} \in S^{\perp}(\boldsymbol{y})$; otherwise, we could move infinitesimally away from $\boldsymbol{y}$ in the tangent directions $\boldsymbol{w}$ and $-\boldsymbol{w}$ and both increase and decrease $f(\boldsymbol{x})$. In other words, $\nabla f(\boldsymbol{y})$ is a member of the double orthogonal complement $S^{\perp\perp}(\boldsymbol{y}) = S(\boldsymbol{y})$. This enables us to write

$$\nabla f(\boldsymbol{y}) \quad = \quad -\sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{y})$$

for properly chosen constants $\lambda_1, \ldots, \lambda_m$. Alternatively, the Lagrangian function

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\omega}) \quad = \quad f(\boldsymbol{x}) + \sum_{i=1}^{m} \omega_i g_i(\boldsymbol{x})$$

has a stationary point at $(\boldsymbol{y}, \boldsymbol{\lambda})$. In this regard, note that

$$\frac{\partial}{\partial \omega_i} \mathcal{L}(\boldsymbol{y}, \boldsymbol{\lambda}) \quad = \quad 0$$

owing to the constraint $g_i(\boldsymbol{y}) = 0$. The essence of the Lagrange multiplier rule consists in finding the stationary points of the Lagrangian. Although

our intuitive arguments need logical tightening in many places, they offer the basic geometric insights.

Example 1.4.1 *Projection onto a Hyperplane*

A hyperplane in R^n is the set of points $H = \{\boldsymbol{x} \in \mathsf{R}^n : \boldsymbol{z}^*\boldsymbol{x} = c\}$ for some vector $\boldsymbol{z} \in \mathsf{R}^n$ and scalar c. There is no loss in generality in assuming that $\boldsymbol{z}$ is a unit vector. If we seek the closest point on H to a point $\boldsymbol{y}$, then we must minimize $\|\boldsymbol{y} - \boldsymbol{x}\|^2$ subject to $\boldsymbol{x} \in H$. We accordingly form the Lagrangian

$$\mathcal{L}(\boldsymbol{x}, \lambda) \;=\; \|\boldsymbol{y} - \boldsymbol{x}\|^2 + \lambda(\boldsymbol{z}^*\boldsymbol{x} - c).$$

Setting the partial derivative with respect to x_i equal to 0 gives

$$-2(y_i - x_i) + \lambda z_i \;=\; 0.$$

This equality entails $\boldsymbol{x} = \boldsymbol{y} - \frac{1}{2}\lambda\boldsymbol{z}$ in vector notation. It follows that

$$c \;=\; \boldsymbol{z}^*\boldsymbol{x} \;=\; \boldsymbol{z}^*\boldsymbol{y} - \frac{1}{2}\lambda\|\boldsymbol{z}\|^2.$$

In view of the assumption $\|\boldsymbol{z}\| = 1$, we find that

$$\lambda \;=\; -2(c - \boldsymbol{z}^*\boldsymbol{y})$$

and consequently that

$$\boldsymbol{x} \;=\; \boldsymbol{y} + (c - \boldsymbol{z}^*\boldsymbol{y})\boldsymbol{z}.$$

If $\boldsymbol{y} \in H$ to begin with, then $\boldsymbol{x} = \boldsymbol{y}$. ■

Example 1.4.2 *Estimation of Multinomial Proportions*

As another statistical example, consider a multinomial experiment with m trials and observed successes $m_1, \ldots, m_n$ over n categories. The maximum likelihood estimate of the probability p_i of category i is $\hat{p}_i = m_i/m$, where $m = m_1 + \cdots + m_n$. To demonstrate this fact, let

$$L(\boldsymbol{p}) \;=\; \binom{m}{m_1, \ldots, m_n} \prod_{i=1}^{n} p_i^{m_i}$$

denote the likelihood. If $m_i = 0$ for some i, then we interpret $p_i^{m_i}$ as 1 even when $p_i = 0$. This convention makes it clear that we can increase $L(\boldsymbol{p})$ by replacing p_i by 0 and p_j by $p_j/(1 - p_i)$ for $j \neq i$. Thus, for purposes of maximum likelihood estimation, we can assume that all $m_i > 0$. Given this assumption, $L(\boldsymbol{p})$ tends to 0 when any p_i tends to 0. It follows that we can further restrict our attention to the interior region where all $p_i > 0$ and maximize the loglikelihood $\ln L(\boldsymbol{p})$ subject to the equality constraint

$\sum_{i=1}^n p_i = 1$. To find the maximum of $\ln L(\boldsymbol{p})$, we look for a stationary point of the Lagrangian

$$\mathcal{L}(\boldsymbol{p}, \lambda) \;=\; \ln \binom{m}{m_1, \ldots, m_n} + \sum_{i=1}^n m_i \ln p_i + \lambda\Big(\sum_{i=1}^n p_i - 1\Big).$$

Setting the partial derivative of $\mathcal{L}(\boldsymbol{p}, \lambda)$ with respect to p_i equal to 0 gives the equation

$$-\frac{m_i}{p_i} \;=\; \lambda.$$

These n equations are satisfied subject to the constraint by taking $\lambda = -m$ and $\hat{p}_i = m_i/m$. Thus, the necessary condition for a maximum holds at $\hat{p}$. One can show that $\hat{\boldsymbol{p}}$ furnishes the global maximum by exploiting the strict concavity of $L(\boldsymbol{p})$. Although we will omit the details of this argument, it is fair to point out that strict concavity follows from

$$\frac{\partial^2}{\partial p_i \partial p_j} \ln L(\boldsymbol{p}) \;=\; \begin{cases} -\frac{m_i}{p_i^2} & i = j \\ 0 & i \neq j \,. \end{cases}$$

In statistical applications, the negative second differential $-d^2 \ln L(\boldsymbol{p})$ is called the observed information matrix. ■

Example 1.4.3 *Eigenvalues of a Symmetric Matrix*

Let $\boldsymbol{M} = (m_{ij})$ be an $n \times n$ symmetric matrix. Recall that $\boldsymbol{M}$ has n real eigenvalues and n corresponding orthogonal eigenvectors. To find the minimum or maximum eigenvalue of $\boldsymbol{M}$, consider optimizing the quadratic form $\boldsymbol{x}^* \boldsymbol{M} \boldsymbol{x}$ subject to the constraint $\|\boldsymbol{x}\|^2 = 1$. To handle this nonlinear constraint, we introduce the Lagrangian

$$\mathcal{L}(\boldsymbol{x}, \lambda) \;=\; \boldsymbol{x}^* \boldsymbol{M} \boldsymbol{x} + \lambda(1 - \|\boldsymbol{x}\|^2).$$

Setting the partial derivative of $\mathcal{L}(\boldsymbol{x}, \lambda)$ with respect to x_i equal to 0 yields

$$2 \sum_{j=1}^n m_{ij} x_j - 2\lambda x_i \;=\; 0.$$

In matrix notation, this reduces to $\boldsymbol{M}\boldsymbol{x} = \lambda \boldsymbol{x}$. It follows that

$$\boldsymbol{x}^* \boldsymbol{M} \boldsymbol{x} \;=\; \lambda \boldsymbol{x}^* \boldsymbol{x} \;=\; \lambda.$$

Thus, the stationary points of the Lagrangian are eigenvectors of $\boldsymbol{M}$. The Lagrange multipliers are the corresponding stationary values or eigenvalues. The maximum and minimum eigenvalues occur among these stationary values. This result does not directly invoke the spectral decomposition theorem.

To prove that every symmetric matrix possesses a spectral decomposition, one can argue by induction. The scalar case is trivial, so suppose that the claim is true for every $(n-1) \times (n-1)$ symmetric matrix. If $\boldsymbol{M}$ is an $n \times n$ symmetric matrix, then as just noted, $\boldsymbol{M}$ has a maximum eigenvalue λ and corresponding unit eigenvector $\boldsymbol{x}$. Consider the deflated matrix $\boldsymbol{N} = \boldsymbol{M} - \lambda \boldsymbol{x}\boldsymbol{x}^*$ and the subspace $S = \{c\boldsymbol{x} : c \in \mathsf{R}\}$. It is clear that $\boldsymbol{N}$ maps S into $\boldsymbol{0}$. Furthermore, $\boldsymbol{N}$ also maps the perpendicular complement $S^\perp$ of S into itself. Indeed, if $\boldsymbol{y} \in S^\perp$, then

$$\boldsymbol{x}^*\boldsymbol{N}\boldsymbol{y} \;=\; \boldsymbol{x}^*(\boldsymbol{M}\boldsymbol{y} - \lambda \boldsymbol{x}\boldsymbol{x}^*\boldsymbol{y}) \;=\; \lambda \boldsymbol{x}^*\boldsymbol{y} - \lambda \boldsymbol{x}^*\boldsymbol{y} \;=\; 0.$$

By the induction hypothesis, the symmetric linear transformation induced by $\boldsymbol{N}$ on $S^\perp$ possesses a spectral decomposition. The $n-1$ eigenvectors of this decomposition, which are automatically perpendicular to $\boldsymbol{x}$, also serve as eigenvectors of $\boldsymbol{M}$. ■

Example 1.4.4 *A Minimization Problem with Two Constraints*

In three dimensions the plane $x_1 + x_2 + x_3 = 1$ intersects the cylinder $x_1^2 + x_2^2 = 1$ in an ellipse. To find the closest point on the ellipse to the origin, we construct the Lagrangian

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \;=\; x_1^2 + x_2^2 + x_3^2 + \lambda_1(x_1 + x_2 + x_3 - 1) + \lambda_2(x_1^2 + x_2^2 - 1).$$

The stationary points of $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ satisfy the system of equations

$$\begin{aligned} 2x_1 + \lambda_1 + 2\lambda_2 x_1 &= 0 \\ 2x_2 + \lambda_1 + 2\lambda_2 x_2 &= 0 \\ 2x_3 + \lambda_1 &= 0 \end{aligned} \tag{1.5}$$

in addition to the constraints. Since $\lambda_1 = -2x_3$, the first two equations of the system (1.5) can be recast as

$$\begin{aligned} (1 + \lambda_2)x_1 &= x_3 \\ (1 + \lambda_2)x_2 &= x_3. \end{aligned}$$

If $\lambda_2 = -1$, then $x_3 = 0$. In this case it is geometrically obvious that $(1, 0, 0)$ and $(0, 1, 0)$ are the only two points that satisfy the constraints $x_1 + x_2 = 1$ and $x_1^2 + x_2^2 = 1$. On the other hand, if $\lambda_2 \neq -1$, then $x_1 = x_2$. In this case, the constraints $2x_1 + x_3 = 1$ and $2x_1^2 = 1$ dictate the solutions $(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 1 - \sqrt{2})$ and $(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}, 1 + \sqrt{2})$. Of the four candidate points, it is easy to check that $(1, 0, 0)$ and $(0, 1, 0)$ are closest to the origin. ■

Example 1.4.5 *A Population Genetics Problem*

The multiplier rule is sometimes hard to apply, and ad hoc methods can lead to better results. In the setting of the multinomial distribution, consider the problem of maximizing the sum

$$f(\boldsymbol{p}) \;=\; \sum_{i<j}(2p_ip_j)^2\frac{1}{2}$$

subject to the constraints $\sum_{i=1}^n p_i = 1$ and all $p_i \geq 0$. This problem has a genetics interpretation involving a locus with n codominant alleles labeled $1,\ldots,n$. (See Sect. 8.4 for some genetics terminology.) At the locus, most genotype combinations of a mother, father, and child make it possible to infer which allele the mother contributes to the child and which allele the father contributes to the child. It turns out that the only ambiguous case occurs when all three family members share the same heterozygous genotype i/j, where $i \neq j$. The probability of this configuration is $(2p_ip_j)^2\frac{1}{2}$ if p_i and p_j are the population frequencies (proportions) of alleles i and j. Here $2p_ip_j$ is the frequency of an i/j mother or an i/j father and $\frac{1}{2}$ is the probability that one of them transmits an i allele and the other transmits a j allele. Thus, $f(\boldsymbol{p})$ represents the probability that the trio's genotypes do not permit inference of the child's maternal and paternal alleles.

The case $n = 2$ is particularly simple because the function $f(\boldsymbol{p})$ then reduces to $2(p_1p_2)^2$. In view of Example 1.2.2, the maximum of $\frac{1}{8}$ is attained when $p_1 = p_2 = \frac{1}{2}$. This suggests that the maximum for general n occurs when all $p_i = \frac{1}{n}$. Because there are $\binom{n}{2}$ heterozygous genotypes,

$$f\Big(\frac{1}{n}\mathbf{1}\Big) \;=\; \binom{n}{2}\Big(\frac{2}{n^2}\Big)^2\frac{1}{2} \;=\; \frac{n-1}{n^3},$$

which is strictly less than $\frac{1}{8}$ for $n \geq 3$. Our first guess is wrong, and we now conjecture that the maximum occurs on a boundary where all but two of the $p_i = 0$. If we permute the components of a maximum point, then symmetry dictates that the result will also be a maximum point. We therefore order the parameters so that $0 < p_1 \leq p_2 \leq \cdots \leq p_n$, avoiding for the moment the lower-dimensional case where $p_1 = 0$.

We now argue that we can increase $f(\boldsymbol{p})$ by increasing p_2 by $q \in [0, p_1]$ at the expense of decreasing p_1 by q. Consider the function

$$g(q) \;=\; 2(p_1-q)^2\sum_{i=3}^n p_i^2 + 2(p_2+q)^2\sum_{i=3}^n p_i^2 + 2(p_1-q)^2(p_2+q)^2$$

which equals the original objective function except for an additive constant independent of q. For $n \geq 3$, straightforward differentiation gives

$$\begin{aligned} g'(q) \;=\; & -4(p_1-q)\sum_{i=3}^n p_i^2 + 4(p_2+q)\sum_{i=3}^n p_i^2 \\ & - 4(p_1-q)(p_2+q)^2 + 4(p_1-q)^2(p_2+q) \end{aligned}$$

$$\begin{aligned} &= 4(p_2 - p_1 + 2q)\Big[\sum_{i=3}^{n} p_i^2 - (p_1 - q)(p_2 + q)\Big] \\ &\geq 4(p_2 - p_1 + 2q)\Big[\sum_{i=3}^{n} p_i^2 - (p_2 - q)(p_2 + q)\Big] \\ &= 4(p_2 - p_1 + 2q)\Big[\sum_{i=4}^{n} p_i^2 + p_3^2 - p_2^2 + q^2\Big] \\ &\geq 0 \end{aligned}$$

for $q \in [0, p_1]$. Thus, we should reduce p_1 to 0 and increase p_2 to $p_2 + p_1$. Furthermore, we should keep discarding the lowest positive p_j until all but two of the p_i equal 0. Finally, we set the remaining two p_i equal to $\frac{1}{2}$. This verifies our second conjecture. ■

Example 1.4.6 *Polygon of Greatest Area Inscribed in a Circle*

As a generalization of Example 1.2.3, consider a polygon inscribed in a circle. For a given number of vertices, the polygon with the greatest area is regular. Here is a proof using elementary plane geometry [213]. Let i, j, and k be three successive vertices as depicted in Fig. 1.6. If we fix the positions of i and k, then we can ask for the optimal placement of vertex j. The only part of the polygon in play is the triangle ijk. The area of the triangle equals half its base times its height. The base distance from i to k is fixed, and the height is clearly a maximum when j is moved to the position j' on the perpendicular bisector of the base. When $j = j'$, the sides ij and jk have equal length. Repeating this argument for all successive vertex triples demonstrates that all sides must have equal length and that the polygon is regular. This is a constrained optimization problem because all vertices are forced to lie on the given circle. Presumably one could reach the conclusion that the polygon is regular by invoking Lagrange multipliers, but the necessary machinery would obscure the underlying simplicity of the problem. ■

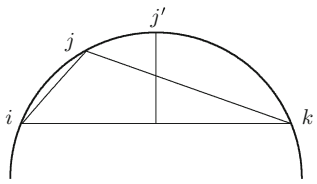

FIGURE 1.6. The *triangle* defined by three adjacent vertices

1.5 Problems

1. Given a point C in the interior of an acute angle, find the points A and B on the sides of the angle such that the perimeter of the triangle ABC is as short as possible. (Hint: Reflect C perpendicularly across the two sides of the angle to points C_1 and C_2, respectively. Let A and B be the intersections of the line segment connecting C_1 and C_2 with the two sides.)

2. Find the point in a triangle that minimizes the sum of the squared distances from the vertices. Show that this point is the intersection of the medians of the triangle.

3. Given an angle in the plane and a point in its interior, find the line that passes through the point and cuts off from the angle a triangle of minimal area. This triangle is determined by the vertex of the angle and the two points where the constructed line intersects the sides of the angle.

4. Find the minima of the functions

$$\begin{aligned} f(x) &= x \ln x \\ g(x) &= x - \ln x \\ h(x) &= x + \frac{1}{x} \end{aligned}$$

   on $(0, \infty)$. Demonstrate rigorously that your solutions are indeed the minima.

5. For $t > 0$ prove that $e^x > x^t$ for all $x > 0$ if and only if $t < e$ [69].

6. Consider the function

$$f(x) = 2(x+2)(x \ln x - x + 1) - 3(x-1)^2$$

   defined on the interval $(0, \infty)$. Show that $f(x) \geq f(1) = 0$ for all x. (Hint: Expand $f(x)$ in a third-order Taylor series around $x = 1$.)

7. Demonstrate that Euler's function $f(x) = x^2 - 1/\ln x$ possesses no local or global minima on either domain $(0, 1)$ or $(1, \infty)$.

8. Prove the harmonic-geometric mean inequality

$$\frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \cdots + \frac{1}{x_n}\right)} \leq \sqrt[n]{x_1 \cdots x_n}$$

   for n positive numbers $x_1, \ldots, x_n$.

9. Prove the arithmetic-quadratic mean inequality

$$\frac{1}{n}\sum_{i=1}^{n} x_i \quad \leq \quad \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)^{1/2}$$

for any nonnegative numbers $x_1, \ldots, x_n$.

10. Heron's classical formula for the area of a triangle with sides of length a, b, and c is $\sqrt{s(s-a)(s-b)(s-c)}$, where $s = (a+b+c)/2$ is the semiperimeter. Show that the triangle of fixed perimeter with greatest area is equilateral.

11. Consider the sequences

$$e_n \quad = \quad \left(1+\frac{1}{n}\right)^n, \qquad f_n \quad = \quad \left(1-\frac{1}{n}\right)^n.$$

It is well known that $\lim_{n\to\infty} e_n = e$ and $\lim_{n\to\infty} f_n = e^{-1}$. Use the arithmetic-geometric mean inequality to prove that $e_n \leq e_{n+1}$ and $f_n \leq f_{n+1}$. In addition, prove that $e_n \leq f_n^{-1}$, that e_n and f_n^{-1} have finite limits, and that these limits are equal [34]. (Hint: Write e_n as the product of 1 and n copies of $1 + \frac{1}{n}$.)

12. Demonstrate that the function

$$f(\boldsymbol{x}) \quad = \quad 4x_1 + \frac{x_1}{x_2^2} + \frac{4x_2}{x_1}$$

on R^2 has the minimum value 8 for x_1 and x_2 positive. At what point $\boldsymbol{x}$ is the minimum attained? (Hint: Write

$$f(\boldsymbol{x}) \quad = \quad 4\left(\frac{4x_1 + \frac{x_1}{x_2^2} + \frac{2x_2}{x_1} + \frac{2x_2}{x_1}}{4}\right)$$

and apply the arithmetic-geometric mean inequality [34]. Attacking this problem by calculus is harder.)

13. Let $H_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$. Verify the inequality $n\sqrt[n]{n+1} \leq n + H_n$ for any positive integer n (Putnam Competition, 1975).

14. Consider an n-gon circumscribing the unit circle in R^2. Demonstrate that the n-gon has minimum area if and only if all of its n sides have equal length. (Hint: Let θ_m be the circular angle between the two points of tangency of sides m and $m+1$ [69]. Show that the area of the quadrilateral defined by the center of the circle, the two points of tangency, and the intersection of the two sides is given by $\tan\frac{\theta_m}{2}$.)

15. Find the minimum of the function

$$g(\boldsymbol{x}) \quad = \quad 2x_1^2 + x_2^2 + \frac{1}{2x_1^2 + x_2^2}$$

on R^2. (Hint: Consider $f(x) = x + 1/x$ on $(0, \infty)$.)

16. In forensic applications of genetics, the sum

$$s(\boldsymbol{p}) \quad = \quad 1 - 2\Big(\sum_{i=1}^n p_i^2\Big)^2 + \sum_{i=1}^n p_i^4$$

occurs [165]. Here the p_i are nonnegative and sum to 1. Prove rigorously that $s(\boldsymbol{p})$ attains its maximum $s_{\max} = 1 - \frac{2}{n^2} + \frac{1}{n^3}$ when all $p_i = \frac{1}{n}$. (Hint: To prove the claim about $s_{\max}$, note that without loss of generality one can assume $p_1 \leq p_2 \leq \cdots \leq p_n$. If $p_i < p_{i+1}$, then $s(\boldsymbol{p})$ can be increased by replacing p_i and p_{i+1} by $p_i + x$ and $p_{i+1} - x$ for x positive and sufficiently small.)

17. Suppose that a and b are real numbers satisfying $0 < a < b$. Prove that the origin locally minimizes $f(\boldsymbol{x}) = (x_2 - ax_1^2)(x_2 - bx_1^2)$ along every line $x_1 = ht$ and $x_2 = kt$ through the origin. Also show that $f(t, ct^2) < 0$ for $a < c < b$ and $t \neq 0$. The origin therefore affords a local minimum along each line through the origin but not a local minimum in the wider sense. If $c < a$ or $c > b$, then $f(t, ct^2) > 0$ for $t \neq 0$, and the paradox disappears.

18. Demonstrate that the function $x_1^2 + x_2^2(1 - x_1)^3$ has a unique stationary point in R^2, which is a local minimum but not a global minimum. Can this occur for a continuously differentiable function with domain R?

19. Find all of the stationary points of the function

$$f(\boldsymbol{x}) \quad = \quad x_1^2 x_2 e^{-x_1^2 - x_2^2}$$

in R^2. Classify each point as either a local minimum, a local maximum, or a saddle point.

20. Rosenbrock's function $f(\boldsymbol{x}) = 100(x_1^2 - x_2)^2 + (x_1 - 1)^2$ achieves its global minimum at $\boldsymbol{x} = \mathbf{1}$. Prove that $\nabla f(\mathbf{1}) = \mathbf{0}$ and $d^2 f(\mathbf{1})$ is positive definite.

21. Consider the polynomial

$$p(\boldsymbol{x}) \quad = \quad \Big[x_1^2 + (1 - x_2)^2\Big]\Big[x_2^2 + (1 - x_1)^2\Big]$$

in two variables. Show that $p(\boldsymbol{x})$ is symmetric in the sense that $p(x_1, x_2) = p(x_2, x_1)$, that $\lim_{\|\boldsymbol{x}\| \to \infty} p(\boldsymbol{x}) = \infty$, and that $p(\boldsymbol{x})$ does not attain its minimum along the diagonal $x_1 = x_2$.

22. Suppose that the $m \times n$ matrix $\boldsymbol{X}$ has full rank and that $m \geq n$. Show that the $n \times n$ matrix $\boldsymbol{X}^*\boldsymbol{X}$ is invertible and positive definite.

23. Consider two sets of positive numbers $x_1, \ldots, x_n$ and $\alpha_1, \ldots, \alpha_n$ such that $\sum_{i=1}^n \alpha_i = 1$. Prove the generalized arithmetic-geometric mean inequality

$$\prod_{i=1}^n x_i^{\alpha_i} \quad \leq \quad \sum_{i=1}^n \alpha_i x_i$$

by minimizing $\sum_{i=1}^n \alpha_i x_i$ subject to the constraint $\prod_{i=1}^n x_i^{\alpha_i} = c$.

24. Suppose the components of a vector $\boldsymbol{x} \in \mathsf{R}^n$ are positive and have product $\prod_{k=1}^n x_k = 1$. Prove that

$$\prod_{k=1}^n (1 + x_k) \quad \geq \quad 2^n.$$

25. Find the rectangular box in R^3 of greatest volume having a fixed surface area.

26. Let $S(\mathbf{0}, r) = \{\boldsymbol{x} \in \mathsf{R}^n : \|\boldsymbol{x}\| = r\}$ be the sphere of radius r centered at the origin. For $\boldsymbol{y} \in \mathsf{R}^n$, find the point of $S(\mathbf{0}, r)$ closest to $\boldsymbol{y}$.

27. Find the parallelepiped of maximum volume that can be inscribed in the ellipsoid

$$\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} + \frac{x_3^2}{c^2} = 1.$$

Assume that the parallelepiped is centered at the origin and has edges parallel to the coordinate axes.

28. A twice continuously differentiable function $f(\boldsymbol{x})$ on R^2 satisfies

$$\frac{\partial^2}{\partial x_1^2} f(\boldsymbol{x}) + \frac{\partial^2}{\partial x_2^2} f(\boldsymbol{x}) \quad > \quad 0$$

for all $\boldsymbol{x}$. Prove that $f(\boldsymbol{x})$ has no local maxima [69]. An example of such a function is $f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 = x_1^2 + x_2^2$.

29. Use the Cauchy-Schwarz inequality to verify the inequalities

$$\begin{aligned} \sum_{m=0}^n a_m x^m \quad &\leq \quad \frac{1}{\sqrt{1-x^2}} \left(\sum_{m=0}^n a_m^2 \right)^{\frac{1}{2}} \qquad 0 \leq x < 1 \\ \sum_{m=1}^n \frac{a_m}{m} \quad &\leq \quad \sqrt{\frac{\pi^2}{6}} \left(\sum_{m=1}^n a_m^2 \right)^{\frac{1}{2}} \end{aligned}$$

$$\begin{aligned}\sum_{m=1}^{n} \frac{a_m}{\sqrt{m+n}} &\le \sqrt{\ln 2}\left(\sum_{m=1}^{n} a_m^2\right)^{\frac{1}{2}} \\ \sum_{m=0}^{n} \binom{n}{m} a_m &\le \binom{2n}{n}^{\frac{1}{2}}\left(\sum_{m=1}^{n} a_m^2\right)^{\frac{1}{2}}.\end{aligned}$$

The upper bound n can be finite or infinite in the first two cases [243].

30. Verify Lagrange's identity

$$\left(\sum_{i=1}^{n} x_i y_i\right)^2 = \sum_{i=1}^{n} x_i^2 \sum_{j=1}^{n} y_j^2 - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i y_j - x_j y_i\right)^2.$$

How does this lead to a proof of the Cauchy-Schwarz inequality and its stated condition for equality?

31. Demonstrate the bound

$$\left|\sum_{m=1}^{n} a_m\right|^2 + \left|\sum_{m=1}^{n} (-1)^m a_m\right|^2 \le (n+2)\sum_{m=1}^{n} a_m^2.$$

This is better than the obvious bound $2n\sum_{m=1}^{n} a_m^2$ given by the Cauchy-Schwarz inequality [243]. (Hint: Let e_n and o_n be the sum of the a_m with m even and odd, respectively.)

32. Consider the function $f(x) = \sum_{m=1}^{n} p_m \cos(\alpha_m x)$ for a discrete probability distribution $p_1, \ldots, p_n$. Given that $g(x) = \cos \alpha x$ satisfies the identity $g(x)^2 = \frac{1}{2}[1 + g(2x)]$, show that $f(x)$ satisfies the inequality $f(x)^2 \le \frac{1}{2}[1 + f(2x)]$. This is the Harker-Kasper inequality from X-ray crystallography [243].

33. For positive numbers $b_1, \ldots, b_n$ and $h_1, \ldots, h_n$, show that

$$\min_{1 \le m \le n} \frac{h_m}{b_m} \le \frac{h_1 + \cdots + h_n}{b_1 + \cdots + b_n} \le \max_{1 \le m \le n} \frac{h_m}{b_m}.$$

This inequality of Cauchy has the baseball interpretation that a batting average of a team is never worse than that of its worst player and never better than that of its best player [243]. (Hint: Consider the case $n = 2$ and use induction.)

2
The Seven C's of Analysis

2.1 Introduction

The current chapter explains key concepts of mathematical analysis summarized by the six adjectives convergent, complete, closed, compact, continuous, and connected. Chapter 6 will add to these six c's the seventh c, convex. At first blush these concepts seem remote from practical problems of optimization. However, painful experience and exotic counterexamples have taught mathematicians to pay attention to details. Fortunately, we can benefit from the struggles of earlier generations and bypass many of the intellectual traps.

2.2 Vector and Matrix Norms

In multidimensional calculus, vector and matrix norms quantify notions of topology and convergence [48, 105, 117, 207]. Norms are also helpful in estimating rates of convergence of iterative methods for solving linear and nonlinear equations and optimizing functions. Functional analysis, which deals with infinite-dimensional vector spaces, uses norms on functions.

We have already met the Euclidean vector norm $\|\boldsymbol{x}\|$ on R^n. For most purposes, this norm suffices. It shares with other norms the four properties:

(a) $\|\boldsymbol{x}\| \geq 0$,

(b) $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = \mathbf{0}$,

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_2,

(c) $\|c\boldsymbol{x}\| = |c| \cdot \|\boldsymbol{x}\|$ for every real number c,

(d) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$.

Property (d) is known as the triangle inequality. To prove it for the Euclidean norm, we note that the Cauchy-Schwarz inequality implies

$$\begin{aligned} \|\boldsymbol{x} + \boldsymbol{y}\|^2 &= \|\boldsymbol{x}\|^2 + 2\boldsymbol{x}^*\boldsymbol{y} + \|\boldsymbol{y}\|^2 \\ &\leq \|\boldsymbol{x}\|^2 + 2\|\boldsymbol{x}\|\|\boldsymbol{y}\| + \|\boldsymbol{y}\|^2 \\ &= \left(\|\boldsymbol{x}\| + \|\boldsymbol{y}\|\right)^2. \end{aligned}$$

One immediate consequence of the triangle inequality is the further inequality

$$\big|\, \|\boldsymbol{x}\| - \|\boldsymbol{y}\| \,\big| \quad \leq \quad \|\boldsymbol{x} - \boldsymbol{y}\|.$$

Two other simple but helpful norms are the ℓ_1 and ℓ_∞ norms

$$\begin{aligned} \|\boldsymbol{x}\|_1 &= \sum_{i=1}^{n} |x_i| \\ \|\boldsymbol{x}\|_\infty &= \max_{1 \leq i \leq n} |x_i|. \end{aligned}$$

Some of the properties of these norms are explored in the problems. In the mathematical literature, the three norms are often referred to as the ℓ_2, ℓ_1, and ℓ_∞ norms.

An $m \times n$ matrix $\boldsymbol{A} = (a_{ij})$ can be viewed as a vector in R^{mn}. Accordingly, we define its Frobenius norm

$$\|\boldsymbol{A}\|_F \quad = \quad \left(\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \right)^{1/2} \quad = \quad \sqrt{\mathrm{tr}(\boldsymbol{A}\boldsymbol{A}^*)} \quad = \quad \sqrt{\mathrm{tr}(\boldsymbol{A}^*\boldsymbol{A})},$$

where $\mathrm{tr}(\cdot)$ is the matrix trace function. Our reasons for writing $\|\boldsymbol{A}\|_F$ rather than $\|\boldsymbol{A}\|$ will soon be apparent. In the meanwhile, the Frobenius matrix norm satisfies the additional condition

(e) $\|\boldsymbol{A}\boldsymbol{B}\| \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|$

for any two compatible matrices $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{B} = (b_{ij})$. Property (e) is verified by invoking the Cauchy-Schwarz inequality in

$$\begin{aligned} \|\boldsymbol{A}\boldsymbol{B}\|_F^2 &= \sum_{i,j} \Big| \sum_k a_{ik} b_{kj} \Big|^2 \\ &\leq \sum_{i,j} \Big(\sum_k a_{ik}^2 \Big) \Big(\sum_l b_{lj}^2 \Big) \\ &= \Big(\sum_{i,k} a_{ik}^2 \Big) \Big(\sum_{l,j} b_{lj}^2 \Big) \qquad (2.1) \\ &= \|\boldsymbol{A}\|_F^2 \|\boldsymbol{B}\|_F^2. \end{aligned}$$

The Frobenius norm does not satisfy the natural condition

(f) $\|\boldsymbol{I}\| = 1$

for an identity matrix $\boldsymbol{I}$. Indeed, an easy calculation shows that $\|\boldsymbol{I}\|_F = \sqrt{n}$ when $\boldsymbol{I}$ is $n \times n$.

To meet all of the conditions (a) through (f), we need to turn to induced matrix norms. Let $\|\cdot\|$ denote both the Euclidean norm on R^m and the Euclidean norm on R^n. The induced Euclidean norm on $m \times n$ matrices is defined by

$$\|\boldsymbol{A}\| \;=\; \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \;=\; \sup_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\|. \tag{2.2}$$

For reasons explained in Proposition 2.2.1, the induced norm (2.2) is called the spectral norm. The question of whether the indicated supremum exists definition (2.2) is settled by the inequalities

$$\|\boldsymbol{A}\boldsymbol{x}\| \;\leq\; \sum_{i=1}^{n} |x_i| \cdot \|\boldsymbol{A}\boldsymbol{e}_i\| \;\leq\; \left(\sum_{i=1}^{n} \|\boldsymbol{A}\boldsymbol{e}_i\|\right) \|\boldsymbol{x}\|,$$

where $\boldsymbol{x} = \sum_{i=1}^{n} x_i \boldsymbol{e}_i$ and $\boldsymbol{e}_i$ is the unit vector whose entries are all 0 except for $e_{ii} = 1$. More exotic induced matrix norms can be concocted by substituting non-Euclidean norms in the numerator and denominator of definition (2.2). For square matrices, the two norms ordinarily coincide. All of the defining properties of a matrix norm are trivial to check for an induced matrix norm. For instance, property (e) follows from

$$\begin{aligned} \|\boldsymbol{A}\boldsymbol{B}\| &= \sup_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{B}\boldsymbol{x}\| \\ &\leq \|\boldsymbol{A}\| \sup_{\|\boldsymbol{x}\|=1} \|\boldsymbol{B}\boldsymbol{x}\| \\ &= \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|. \end{aligned}$$

Definition (2.2) also clearly entails the equality $\|\boldsymbol{I}\| = 1$ when $m = n$.

The next proposition determines the value of the Euclidean norm $\|\boldsymbol{A}\|$. In the proposition, $\rho(\boldsymbol{M})$ denotes the absolute value of the dominant eigenvalue of the square matrix $\boldsymbol{M}$. This quantity is called the spectral radius of $\boldsymbol{M}$.

Proposition 2.2.1 *If $\boldsymbol{A} = (a_{ij})$ is an $m \times n$ matrix, then*

$$\|\boldsymbol{A}\| \;=\; \sqrt{\rho(\boldsymbol{A}^*\boldsymbol{A})} \;=\; \sqrt{\rho(\boldsymbol{A}\boldsymbol{A}^*)} \;=\; \|\boldsymbol{A}^*\|.$$

When $\boldsymbol{A}$ is symmetric, $\|\boldsymbol{A}\|$ reduces to $\rho(\boldsymbol{A})$. The norms $\|\boldsymbol{A}\|$ and $\|\boldsymbol{A}\|_F$ satisfy

$$\|\boldsymbol{A}\| \;\leq\; \|\boldsymbol{A}\|_F \;\leq\; \sqrt{n}\|\boldsymbol{A}\|. \tag{2.3}$$

Finally, when $\boldsymbol{A}$ is a row or column vector, the Euclidean matrix and vector norms of $\boldsymbol{A}$ coincide.

Proof: Choose an orthonormal basis of eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ for the symmetric matrix $\boldsymbol{A}^*\boldsymbol{A}$ with corresponding eigenvalues arranged so that $0 \leq \lambda_1 \leq \cdots \leq \lambda_n$. If $\boldsymbol{x} = \sum_{i=1}^n c_i \boldsymbol{u}_i$ is a unit vector, then $\sum_{i=1}^n c_i^2 = 1$, and

$$\begin{aligned} \|\boldsymbol{A}\|^2 &= \sup_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^*\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{x} \\ &= \sup_{\|\boldsymbol{x}\|=1} \sum_{i=1}^n \lambda_i c_i^2 \\ &\leq \lambda_n. \end{aligned}$$

Equality is achieved when $c_n = \pm 1$ and all other $c_i = 0$. If $\boldsymbol{A}$ is symmetric with eigenvalues μ_i arranged so that $|\mu_1| \leq \cdots \leq |\mu_n|$, then the $\boldsymbol{u}_i$ can be chosen to be the corresponding eigenvectors. In this case, clearly $\lambda_i = \mu_i^2$.

To prove that $\rho(\boldsymbol{A}^*\boldsymbol{A}) = \rho(\boldsymbol{A}\boldsymbol{A}^*)$, choose an eigenvalue $\lambda \neq 0$ of $\boldsymbol{A}^*\boldsymbol{A}$ with corresponding eigenvector $\boldsymbol{v}$. Multiplying the equation $\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$ on the left by $\boldsymbol{A}$ produces $(\boldsymbol{A}\boldsymbol{A}^*)\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{A}\boldsymbol{v}$. Because $\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$, the vector $\boldsymbol{A}\boldsymbol{v} \neq \boldsymbol{0}$. Thus, λ is an eigenvalue of $\boldsymbol{A}\boldsymbol{A}^*$ with eigenvector $\boldsymbol{A}\boldsymbol{v}$. Likewise, any eigenvalue $\omega \neq 0$ of $\boldsymbol{A}\boldsymbol{A}^*$ is also an eigenvalue of $\boldsymbol{A}^*\boldsymbol{A}$.

To verify the left bound of the pair of bounds (2.3), apply inequality (2.1) with $\boldsymbol{B} = \boldsymbol{x}$ in the definition of $\|\boldsymbol{A}\|$. The right bound follows from

$$\sum_{i=1}^m a_{ij}^2 = \|\boldsymbol{A}\boldsymbol{e}_j\|^2 \leq \|\boldsymbol{A}\|^2$$

by summing on j. Finally, suppose that $\boldsymbol{A}$ is a column vector. The two bounds (2.3) with $n = 1$ show that $\|\boldsymbol{A}\| = \|\boldsymbol{A}\|_F$. If $\boldsymbol{A}$ is a row vector, the same reasoning applied to $\boldsymbol{A}^*$ gives $\|\boldsymbol{A}\| = \|\boldsymbol{A}^*\| = \|\boldsymbol{A}^*\|_F = \|\boldsymbol{A}\|_F$. ■

2.3 Convergence and Completeness

A sequence $\boldsymbol{x}_m \in \mathsf{R}^n$ converges to $\boldsymbol{x}$, written $\lim_{m\to\infty} \boldsymbol{x}_m = \boldsymbol{x}$, provided $\lim_{m\to\infty} \|\boldsymbol{x}_m - \boldsymbol{x}\| = 0$ in the standard Euclidean norm. For convergence of $\boldsymbol{x}_m$ to $\boldsymbol{x}$ to occur, it is necessary and sufficient that each component sequence x_{mi} converge to x_i. Convergence of a sequence of matrices is defined similarly using either the Frobenius norm $\|\boldsymbol{A}\|_F$ or the induced matrix norm $\|\boldsymbol{A}\|$. The pair of bounds (2.3) shows that the two norms are equivalent in testing convergence.

Convergent sequences of vectors or matrices enjoy many useful properties. Some of these are mentioned in the next proposition.

Proposition 2.3.1 *In the following list, once a limit is assumed to exist for an item, it is assumed to exist for all subsequent items. With this proviso, we have:*

(a) *If* $\lim_{m\to\infty} \boldsymbol{x}_m = \boldsymbol{x}$, *then* $\lim_{m\to\infty} \|\boldsymbol{x}_m\| = \|\boldsymbol{x}\|$.

(b) *If* $\lim_{m\to\infty} \boldsymbol{y}_m = \boldsymbol{y}$, *then*

$$\lim_{m\to\infty} \boldsymbol{x}_m^* \boldsymbol{y}_m = \boldsymbol{x}^* \boldsymbol{y}.$$

(c) *If* a *and* b *are real scalars, then*

$$\lim_{m\to\infty} [a\boldsymbol{x}_m + b\boldsymbol{y}_m] = a\boldsymbol{x} + b\boldsymbol{y}.$$

(d) *If* $\lim_{m\to\infty} \boldsymbol{M}_m = \boldsymbol{M}$ *for a sequence of matrices compatible with* $\boldsymbol{x}$, *then*

$$\lim_{m\to\infty} \boldsymbol{M}_m \boldsymbol{x}_m = \boldsymbol{M}\boldsymbol{x}.$$

(e) *If* $\boldsymbol{M}$ *is square and invertible, then* $\boldsymbol{M}_m^{-1}$ *exists for large* m *and*

$$\lim_{m\to\infty} \boldsymbol{M}_m^{-1} = \boldsymbol{M}^{-1}.$$

(f) *Finally, if* $\lim_{m\to\infty} \boldsymbol{N}_m = \boldsymbol{N}$ *for a sequence of matrices compatible with* $\boldsymbol{M}$, *then*

$$\lim_{m\to\infty} \boldsymbol{M}_m \boldsymbol{N}_m = \boldsymbol{M}\boldsymbol{N}.$$

Proof: As a sample proof, part (d) follows from the inequalities

$$\begin{aligned} \|\boldsymbol{M}_m\boldsymbol{x}_m - \boldsymbol{M}\boldsymbol{x}\| &\leq \|\boldsymbol{M}_m\boldsymbol{x}_m - \boldsymbol{M}_m\boldsymbol{x}\| + \|\boldsymbol{M}_m\boldsymbol{x} - \boldsymbol{M}\boldsymbol{x}\| \\ &\leq \|\boldsymbol{M}_m\| \cdot \|\boldsymbol{x}_m - \boldsymbol{x}\| + \|\boldsymbol{M}_m - \boldsymbol{M}\| \cdot \|\boldsymbol{x}\| \\ \|\boldsymbol{M}_m\| &\leq \|\boldsymbol{M}_m - \boldsymbol{M}\| + \|\boldsymbol{M}\|. \end{aligned}$$

Part (e) will be proved after Example 2.3.3. ■

In some situations, we know that the members of a sequence become progressively closer together. A Cauchy sequence $\boldsymbol{x}_m$ exhibits a strong form of this phenomenon; namely, for every $\epsilon > 0$, there is an m such that $\|\boldsymbol{x}_p - \boldsymbol{x}_q\| \leq \epsilon$ for all $p, q \geq m$. The real line R is complete in the sense that every Cauchy sequence possesses a limit. The rational numbers are incomplete by contrast because a sequence of rationals can converge to an irrational. The completeness of R carries over to R^n. Indeed, if $\boldsymbol{x}_m$ is a Cauchy sequence, then under the Euclidean norm we have

$$|x_{pi} - x_{qi}| \leq \|\boldsymbol{x}_p - \boldsymbol{x}_q\|.$$

This shows that each component sequence is Cauchy and consequently possesses a limit x_i. The vector $\boldsymbol{x}$ with components x_i then furnishes a limit for the vector sequence $\boldsymbol{x}_m$.

Example 2.3.1 *Existence of Suprema and Infima*

The completeness of the real line is equivalent to the existence of least upper bounds or suprema. Consider a nonempty set $S \subset \mathsf{R}$ that is bounded above. If the set is finite, then its least upper bound is just its largest element. If the set is infinite, we choose a and b such that the interval $[a, b]$ contains an element of S and b is an upper bound of S. We can generate $\sup S$ by a bisection strategy. Bisect $[a, b]$ into the two subintervals $[a, (a + b)/2]$ and $[(a + b)/2, b]$. Let $[a_1, b_1]$ denote the left subinterval if $(a + b)/2$ provides an upper bound. Otherwise, let $[a_1, b_1]$ denote the right subinterval. In either case, $[a_1, b_1]$ contains an element of S. Now bisect $[a_1, b_1]$ and generate a subinterval $[a_2, b_2]$ by the same criterion. If we continue bisecting and choosing a left or right subinterval ad infinitum, then we generate two Cauchy sequences a_i and b_i with common limit c. By the definition of the sequence b_i, c furnishes an upper bound of S. By the definition of the sequence a_i, no bound of S is smaller than c. Establishing the existence of the greatest lower bound $\inf S$ for S bounded below proceeds similarly. If S is unbounded above, then $\sup S = \infty$, and if it is unbounded below, then $\inf S = -\infty$. ■

Example 2.3.2 *Limit Superior and Limit Inferior*

For a real sequence x_n. we define the limit superior and limit inferior by

$$\begin{aligned} \limsup_{n\to\infty} x_n &= \inf_m \sup_{n\geq m} x_n = \lim_{m\to\infty} \sup_{n\geq m} x_n \\ \liminf_{n\to\infty} x_n &= \sup_m \inf_{n\geq m} x_n = \lim_{m\to\infty} \inf_{n\geq m} x_n. \end{aligned}$$

If $\sup_n x_n = \infty$, then $\limsup_{n\to\infty} x_n = \infty$, and if $\lim_{n\to\infty} x_n = -\infty$, then $\limsup_{n\to\infty} x_n = -\infty$. From these definitions, one can also deduce that

$$\limsup_{n\to\infty} -x_n = -\liminf_{n\to\infty} x_n \tag{2.4}$$

and that

$$\liminf_{n\to\infty} x_n \leq \limsup_{n\to\infty} x_n. \tag{2.5}$$

The sequence x_n has a limit if and only if equality prevails in inequality (2.5). In this situation, the common value of the limit superior and inferior furnishes the limit of x_n. ■

Example 2.3.3 *Series Expansion for a Matrix Inverse*

If a square matrix $\boldsymbol{M}$ has norm $\|\boldsymbol{M}\| < 1$, then we can write

$$(\boldsymbol{I} - \boldsymbol{M})^{-1} = \sum_{i=0}^{\infty} \boldsymbol{M}^i.$$

To verify this claim, we first prove that the partial sums $\boldsymbol{S}_j = \sum_{i=0}^{j} \boldsymbol{M}^i$ form a Cauchy sequence. This fact is a consequence of the inequalities

$$\begin{aligned} \|\boldsymbol{S}_k - \boldsymbol{S}_j\| &= \Big\| \sum_{i=j+1}^{k} \boldsymbol{M}^i \Big\| \\ &\leq \sum_{i=j+1}^{k} \|\boldsymbol{M}^i\| \\ &\leq \sum_{i=j+1}^{k} \|\boldsymbol{M}\|^i \end{aligned}$$

for $k \geq j$ and the assumption $\|\boldsymbol{M}\| < 1$. If we let $\boldsymbol{S}$ represent the limit of the $\boldsymbol{S}_j$, then part (f) of Proposition 2.3.1 implies that $(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{S}_j$ converges to $(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{S}$. But $(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{S}_j = \boldsymbol{I} - \boldsymbol{M}^{j+1}$ also converges to $\boldsymbol{I}$. Hence, $(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{S} = \boldsymbol{I}$, and this verifies the claim $\boldsymbol{S} = (\boldsymbol{I}-\boldsymbol{M})^{-1}$. ■

With this result under our belts, we now demonstrate part (e) of Proposition 2.3.1. Because $\|\boldsymbol{M}^{-1}(\boldsymbol{M}-\boldsymbol{M}_m)\| \leq \|\boldsymbol{M}^{-1}\| \cdot \|\boldsymbol{M}-\boldsymbol{M}_m\|$, the matrix inverse $[\boldsymbol{I}-\boldsymbol{M}^{-1}(\boldsymbol{M}-\boldsymbol{M}_m)]^{-1}$ exists for large m. Therefore, we can write the inverse of

$$\begin{aligned} \boldsymbol{M}_m &= \boldsymbol{M} - (\boldsymbol{M}-\boldsymbol{M}_m) \\ &= \boldsymbol{M}[\boldsymbol{I} - \boldsymbol{M}^{-1}(\boldsymbol{M}-\boldsymbol{M}_m)] \end{aligned}$$

as

$$\boldsymbol{M}_m^{-1} = [\boldsymbol{I} - \boldsymbol{M}^{-1}(\boldsymbol{M}-\boldsymbol{M}_m)]^{-1}\boldsymbol{M}^{-1}.$$

The proof of convergence is completed by noting the bound

$$\begin{aligned} \|\boldsymbol{M}_m^{-1} - \boldsymbol{M}^{-1}\| &= \Big\| \sum_{i=1}^{\infty} [\boldsymbol{M}^{-1}(\boldsymbol{M}-\boldsymbol{M}_m)]^i \boldsymbol{M}^{-1} \Big\| \\ &\leq \sum_{i=1}^{\infty} \|\boldsymbol{M}^{-1}\|^i \|\boldsymbol{M}-\boldsymbol{M}_m\|^i \|\boldsymbol{M}^{-1}\| \\ &= \frac{\|\boldsymbol{M}^{-1}\|^2 \|\boldsymbol{M}-\boldsymbol{M}_m\|}{1 - \|\boldsymbol{M}^{-1}\| \cdot \|\boldsymbol{M}-\boldsymbol{M}_m\|}, \end{aligned}$$

applying in the process the matrix analog of part (a) of Proposition 2.3.1.

Example 2.3.4 *Matrix Exponential Function*

The exponential of a square matrix $\boldsymbol{M}$ is given by the series expansion

$$e^{\boldsymbol{M}} = \sum_{i=0}^{\infty} \frac{1}{i!} \boldsymbol{M}^i.$$

To prove the convergence of the series, it again suffices to show that the partial sums $\boldsymbol{S}_j = \sum_{i=0}^{j} \frac{1}{i!}\boldsymbol{M}^i$ form a Cauchy sequence. The bound

$$\|\boldsymbol{S}_k - \boldsymbol{S}_j\| \quad = \quad \Big\| \sum_{i=j+1}^{k} \frac{1}{i!}\boldsymbol{M}^i \Big\| \quad \leq \quad \sum_{i=j+1}^{k} \frac{1}{i!}\|\boldsymbol{M}\|^i$$

for $k \geq j$ is just what we need.

The matrix exponential function has many interesting properties. For example, the function $N(t) = e^{t\boldsymbol{M}}$ solves the differential equation

$$N'(t) \quad = \quad \boldsymbol{M}N(t)$$

subject to the initial condition $N(0) = \boldsymbol{I}$. Here t is a real parameter, and we differentiate the matrix $N(t)$ entry by entry. In Example 4.2.2 of Chap. 4, we will prove that $N(t) = e^{t\boldsymbol{M}}$ is the one and only solution. The law of exponents $e^{\boldsymbol{A}+\boldsymbol{B}} = e^{\boldsymbol{A}}e^{\boldsymbol{B}}$ for commuting matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is another interesting property of the matrix exponential function. One way of proving the law of exponents is to observe that $e^{t(\boldsymbol{A}+\boldsymbol{B})}$ and $e^{t\boldsymbol{A}}e^{t\boldsymbol{B}}$ both solve the differential equation

$$N'(t) \quad = \quad (\boldsymbol{A}+\boldsymbol{B})N(t)$$

subject to the initial condition $N(0) = \boldsymbol{I}$. Since the solution to such an initial value problem is unique, the two solutions must coincide at $t = 1$. ■

2.4 The Topology of R^n

Mathematics involves a constant interplay between the abstract and the concrete. We now consider some qualitative features of sets in R^n that generalize to more abstract spaces. For instance, there is the matter of boundedness. A set $S \subset \mathsf{R}^n$ is said to be bounded if it is contained in some ball $B(\mathbf{0}, r) = \{\boldsymbol{x} \in \mathsf{R}^n : \|\boldsymbol{x}\| < r\}$ of radius r centered at the origin $\mathbf{0}$. As we shall see in our discussion of compactness, boundedness takes on added importance when it is combined with the notion of closedness. A closed set is closed under the formation of limits. Thus, $S \subset \mathsf{R}^n$ is closed if for every convergent sequence $\boldsymbol{x}_m$ taken from S, we have $\lim_{m\to\infty} \boldsymbol{x}_m \in S$ as well.

It takes time and effort to appreciate the ramifications of these ideas. A few of the most pertinent ones for closedness are noted in the next proposition.

Proposition 2.4.1 *The collection of closed sets satisfy the following:*

(a) *The whole space* R^n *is closed.*

(b) *The empty set* $\emptyset$ *is closed.*

(c) *The intersection $S = \cap_\alpha S_\alpha$ of an arbitrary number of closed sets S_α is closed.*

(d) *The union $S = \cup_\alpha S_\alpha$ of a finite number of closed sets S_α is closed.*

Proof: All of these are easy. For part (d), observe that for any convergent sequence $\boldsymbol{x}_m$ taken from S, one of the sets S_α must contain an infinite subsequence $\boldsymbol{x}_{m_k}$. The limit of this subsequence exists and falls in S_α. ■

Some examples of closed sets are closed intervals $(-\infty, a]$, $[a, b]$, and $[b, \infty)$; closed balls $\{\boldsymbol{y} \in \mathsf{R}^n : \|\boldsymbol{y}-\boldsymbol{x}\| \leq r\}$; spheres $\{\boldsymbol{y} \in \mathsf{R}^n : \|\boldsymbol{y}-\boldsymbol{x}\| = r\}$; hyperplanes $\{\boldsymbol{x} \in \mathsf{R}^n : \boldsymbol{z}^*\boldsymbol{x} = c\}$; and closed halfspaces $\{\boldsymbol{x} \in \mathsf{R}^n : \boldsymbol{z}^*\boldsymbol{x} \leq c\}$. A closed set S of R^n is complete in the sense that all Cauchy sequences from S possess limits in S.

Example 2.4.1 *Finitely Generated Convex Cones*

A set C is a convex cone provided $\alpha\boldsymbol{u} + \beta\boldsymbol{v}$ is in C whenever the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are in C and the scalars α and β are nonnegative. A finitely generated convex cone can be written as

$$C \;=\; \Big\{\sum_{i=1}^m \alpha_i \boldsymbol{v}_i : \alpha_i \geq 0,\ i = 1, \ldots, m\Big\}.$$

Demonstrating that C is a closed set is rather subtle. Consider a sequence $\boldsymbol{u}_j = \sum_{i=1}^m \alpha_{ji}\boldsymbol{v}_i$ in C converging to a point $\boldsymbol{u}$. If the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ are linearly independent, then the coefficients α_{ji} are the unique coordinates of $\boldsymbol{u}_j$ in the finite-dimensional subspace spanned by the $\boldsymbol{v}_i$. To recover the α_{ji}, we introduce the matrix $\boldsymbol{V}$ with columns $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ and rewrite the original equation as $\boldsymbol{u}_j = \boldsymbol{V}\boldsymbol{\alpha}_j$. Multiplying this equation by first $\boldsymbol{V}^*$ and then by $(\boldsymbol{V}^*\boldsymbol{V})^{-1}$ on the left gives $\boldsymbol{\alpha}_j = (\boldsymbol{V}^*\boldsymbol{V})^{-1}\boldsymbol{V}^*\boldsymbol{u}_j$. This representations allows us to conclude that $\boldsymbol{\alpha}_j$ possesses a limit $\boldsymbol{\alpha}$ with nonnegative entries. Therefore, the limit $\boldsymbol{u} = \boldsymbol{V}\boldsymbol{\alpha}$ lies in C.

If we relax the assumption that the vectors are linearly independent, we must resort to an inductive argument to prove that C is closed. The case $m = 1$ is true because a single vector $\boldsymbol{v}_1$ is linearly independent. Assume that the claim holds for $m - 1$ vectors. If the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ are linearly independent, then we are done. If the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ are linearly dependent, then there exist scalars $\beta_1, \ldots, \beta_m$, not all 0, such that $\sum_{i=1}^m \beta_i \boldsymbol{v}_i = \boldsymbol{0}$. Without loss of generality, we can assume that $\beta_i < 0$ for at least one index i. We can express any point $\boldsymbol{u} \in C$ as

$$\boldsymbol{u} \;=\; \sum_{i=1}^m \alpha_i \boldsymbol{v}_i \;=\; \sum_{i=1}^m (\alpha_i + t\beta_i)\boldsymbol{v}_i$$

for an arbitrary scalar t. If we increase t gradually from 0, then there is a first value at which $\alpha_j + t\beta_j = 0$ for some index j. This shows that C can

be decomposed as the union

$$C = \bigcup_{j=1}^{m} \Big\{ \sum_{i \neq j} \gamma_i \boldsymbol{v}_i : \gamma_i \geq 0,\ i \neq j \Big\}.$$

Each of the convex cones $\{\sum_{i \neq j} \gamma_i \boldsymbol{v}_i : \gamma_i \geq 0,\ i \neq j\}$ is closed by the induction hypothesis. Since a finite union of closed sets is closed, C itself is closed. A straightforward extension of this argument establishes the stronger claim that every point in the cone can be represented as a positive combination of a linearly independent subset of $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$. ■

The complement $S^c = \mathsf{R}^n \setminus S$ of a closed set S is called an open set. Every $\boldsymbol{x} \in S^c$ is surrounded by a ball $B(\boldsymbol{x}, r)$ completely contained in S^c. If this were not the case, then we could construct a sequence of points $\boldsymbol{x}_m$ from S converging to $\boldsymbol{x}$, contradicting the closedness of S. This fact is the first of several mentioned in the next proposition.

Proposition 2.4.2 *The collection of open sets satisfy the following:*

(a) *Every open set is a union of balls, and every union of balls is an open set.*

(b) *The whole space* R^n *is open.*

(c) *The empty set* $\emptyset$ *is open.*

(d) *The union* $S = \cup_\alpha S_\alpha$ *of an arbitrary number of open sets* S_α *is open.*

(e) *The intersection* $S = \cap_\alpha S_\alpha$ *of a finite number of open sets* S_α *is open.*

Proof: Again these are easy. Parts (d) and (e) are consequences of the set identities

$$\begin{aligned} (\cap_\alpha S_\alpha)^c &= \cup_\alpha S_\alpha^c \\ (\cup_\alpha S_\alpha)^c &= \cap_\alpha S_\alpha^c \end{aligned}$$

and parts (c) and (d) of Proposition 2.4.1. ■

Some examples of open sets are open intervals $(-\infty, a)$, (a, b), and (b, ∞); balls $\{\boldsymbol{y} \in \mathsf{R}^n : \|\boldsymbol{y} - \boldsymbol{x}\| < r\}$, and open halfspaces $\{\boldsymbol{x} \in \mathsf{R}^n : \boldsymbol{z}^* \boldsymbol{x} < c\}$. Any open set surrounding a point is called a neighborhood of the point. Some examples of sets that are neither closed nor open are the unbalanced intervals $(a, b]$ and $[a, b)$, the discrete set $V = \{n^{-1} : n = 1, 2, \ldots\}$, and the rational numbers. If we append the limit 0 to the set V, then it becomes closed.

A boundary point $\boldsymbol{x}$ of a set S is the limit of a sequence of points from S and also the limit of a different sequence of points from S^c. Closed sets contain all of their boundary points, and open sets contain none of their

boundary points. The interior of S is the largest open set contained within S. The closure of S is the smallest closed set containing S. For instance, the boundary of the ball $B(\boldsymbol{x}, r) = \{\boldsymbol{y} \in \mathsf{R}^n : \|\boldsymbol{y} - \boldsymbol{x}\| < r\}$ is the sphere $S(\boldsymbol{x}, r) = \{\boldsymbol{y} \in \mathsf{R}^n : \|\boldsymbol{y} - \boldsymbol{x}\| = r\}$. The closure of $B(\boldsymbol{x}, r)$ is the closed ball $C(\boldsymbol{x}, r) = \{\boldsymbol{y} \in \mathsf{R}^n : \|\boldsymbol{y} - \boldsymbol{x}\| \leq r\}$, and the interior of $C(\boldsymbol{x}, r)$ is $B(\boldsymbol{x}, r)$.

A closed bounded set is said to be compact. Finite intervals $[a, b]$ are typical compact sets. Compact sets can be defined in several equivalent ways. The most important of these is the Bolzano-Weierstrass characterization. In preparation for this result, let us define a multidimensional interval $[\boldsymbol{a}, \boldsymbol{b}]$ in R^n to be the Cartesian product

$$[\boldsymbol{a}, \boldsymbol{b}] \quad = \quad [a_1, b_1] \times \cdots \times [a_n, b_n]$$

of n one-dimensional intervals. We will only consider closed intervals. The diameter of $[\boldsymbol{a}, \boldsymbol{b}]$ is the greatest separation between any two of its points; this clearly reduces to the distance $\|\boldsymbol{a} - \boldsymbol{b}\|$ between its extreme corners.

Proposition 2.4.3 (Bolzano-Weierstrass) *A set $S \subset \mathsf{R}^n$ is compact if and only if every sequence $\boldsymbol{x}_m$ in S has a convergent subsequence $\boldsymbol{x}_{m_i}$ with limit in S.*

Proof: Suppose every sequence $\boldsymbol{x}_m$ in S has a convergent subsequence $\boldsymbol{x}_{m_i}$ with limit in S. If S is unbounded, then we can define a sequence $\boldsymbol{x}_m$ with $\|\boldsymbol{x}_m\| \geq m$. Clearly, this sequence has no convergent subsequence. If S is not closed, then there is a convergent sequence $\boldsymbol{x}_m$ with limit $\boldsymbol{x}$ outside S. Clearly, no subsequence of $\boldsymbol{x}_m$ can converge to a limit in S. Thus, the subsequence property implies compactness.

For the converse, let $\boldsymbol{x}_m$ be a sequence in the compact set S. Because S is bounded, it is contained in a multidimensional interval $[\boldsymbol{a}, \boldsymbol{b}]$. If infinitely many of the $\boldsymbol{x}_m$ coincide, then these can be used to construct a constant subsequence that trivially converges to a point of S. Otherwise, let T_0 denote the infinite set $\cup_{m=1}^{\infty}\{\boldsymbol{x}_m\}$.

The rest of the proof adapts the bisection strategy of Example 2.3.1. The first stage of the bisection divides $[\boldsymbol{a}, \boldsymbol{b}]$ into 2^n subintervals of equal volume. Each of these subintervals can be written as $[\boldsymbol{a}_1, \boldsymbol{b}_1]$, where $a_{1j} = a_j$ and $b_{1j} = (a_j + b_j)/2$ or $a_{1j} = (a_j + b_j)/2$ and $b_{1j} = b_j$. There is no harm in the fact that these subintervals overlap along their boundaries. It is only vital to observe that one of the subintervals contains an infinite subset $T_1 \subset T_0$. Let us choose such a subinterval and label it using the generic notation $[\boldsymbol{a}_1, \boldsymbol{b}_1]$. We now inductively repeat the process. At stage $i + 1$ we divide the previously chosen subinterval $[\boldsymbol{a}_i, \boldsymbol{b}_i]$ into 2^n subintervals of equal volume. Each of these subintervals can be written as $[\boldsymbol{a}_{i+1}, \boldsymbol{b}_{i+1}]$, where either $a_{i+1,j} = a_{ij}$ and $b_{i+1,j} = (a_{ij} + b_{ij})/2$ or $a_{i+1,j} = (a_{ij} + b_{ij})/2$ and $b_{i+1,j} = b_{ij}$. One of these subintervals, which we label $[\boldsymbol{a}_{i+1}, \boldsymbol{b}_{i+1}]$ for convenience, contains an infinite subset $T_{i+1} \subset T_i$.

We continue this process ad infinitum. in the process choosing $\boldsymbol{x}_{m_i}$ from T_i and $m_i > m_{i-1}$. Because $T_i \subset [\boldsymbol{a}_i, \boldsymbol{b}_i]$ and the diameter of $[\boldsymbol{a}_i, \boldsymbol{b}_i]$ tends

to 0, the subsequence $\boldsymbol{x}_{m_i}$ is Cauchy. By virtue of the completeness of R^n, this subsequence converges to some point $\boldsymbol{x}$, which necessarily belongs to the closed set S. ■

In many instances it is natural to consider a subset S of R^n as a topological space in its own right. Notions of distance and convergence carry over immediately, but we must exercise some care in defining closed and open sets. In the relative topology, a subset $T \subset S$ is closed if and only if it can be represented as the intersection $T = S \cap C$ of S with a closed set C of R^n. If T is closed in S, then the obvious choice of C is the closure of T in R^n. Likewise, $T \subset S$ is open in the relative topology if and only if it can be represented as the intersection $T = S \cap O$ of S with an open set O of R^n. These two definitions are consistent with an open set being the relative complement of a closed set and vice versa. They are also consistent with the development of continuous functions sketched in the next section.

2.5 Continuous Functions

Continuous functions are the building blocks of mathematical analysis. Continuity is such an intuitive notion that ancient mathematicians did not even bother to define it. Proper recognition of continuity had to wait until differentiability was thoroughly explored. Our approach to continuity emphasizes convergent sequences. A function $f(\boldsymbol{x})$ from R^m to R^n is said to be continuous at $\boldsymbol{y}$ if $f(\boldsymbol{x}_i)$ converges to $f(\boldsymbol{y})$ for every sequence $\boldsymbol{x}_i$ that converges to $\boldsymbol{y}$. If the domain of $f(\boldsymbol{x})$ is a subset S of R^m, then the sequences $\boldsymbol{x}_i$ and the point $\boldsymbol{y}$ are confined to S. Finally, $f(\boldsymbol{x})$ is said to be continuous if it is continuous at every point $\boldsymbol{y}$ of its domain.

The definition of continuity through convergent sequences tends to be simpler to apply than the competing ϵ and δ approach of calculus. We leave it to the reader to show that the two definitions are fully equivalent. Either definition has powerful consequences. For instance, it is clear that a vector-valued function is continuous if and only if each of its component functions is continuous. Before enumerating other less obvious consequences, it is helpful to forge a few tools for recognizing and constructing continuous functions. Fortunately, the collection of continuous functions is closed under many standard algebraic operations. Here are a few examples.

Proposition 2.5.1 *Given that the vector-valued functions $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ and matrix-valued function $M(\boldsymbol{x})$ and $N(\boldsymbol{x})$ are continuous and compatible whenever necessary, the following algebraic combinations are continuous:*

(a) *The norm $\|f(\boldsymbol{x})\|$.*

(b) *The inner product $f(\boldsymbol{x})^* g(\boldsymbol{x})$.*

(c) *The linear combination $\alpha f(\boldsymbol{x}) + \beta g(\boldsymbol{x})$ for real scalars α and β.*

(d) *The matrix-vector product* $M(\boldsymbol{x})f(\boldsymbol{x})$.

(e) *The matrix inverse* $M^{-1}(\boldsymbol{x})$ *when* $M(\boldsymbol{x})$ *is square and invertible.*

(f) *The matrix product* $M(\boldsymbol{x})N(\boldsymbol{x})$.

(g) *The functional composition* $f \circ g(\boldsymbol{x}) = f[g(\boldsymbol{x})]$.

Proof: Parts (a) through (f) are all immediate by-products of Proposition 2.3.1 and the definition of continuity. For part (g), suppose $\boldsymbol{x}_i$ tends to $\boldsymbol{x}$. Then $f(\boldsymbol{x}_i)$ tends to $f(\boldsymbol{x})$, and so $f \circ g(\boldsymbol{x}_i)$ tends to $f \circ g(\boldsymbol{x})$. ■

Example 2.5.1 *Rational Functions*

Because the coordinate variables x_i of $\boldsymbol{x} \in \mathsf{R}^n$ are continuous, all polynomials in these variables are continuous as well. For example, the determinant of a square matrix is a continuous function of the entries of the matrix. A quotient of two polynomials (rational function) in the coordinate variables x_i of $\boldsymbol{x} \in \mathsf{R}^n$ is continuous where its denominator does not vanish. Finally, any linear transformation of one vector space into another is continuous. ■

Example 2.5.2 *Distance to a Set*

The distance $\operatorname{dist}(\boldsymbol{x}, S)$ from a point $\boldsymbol{x} \in \mathsf{R}^n$ to a set S is defined by

$$\operatorname{dist}(\boldsymbol{x}, S) \quad = \quad \inf_{\boldsymbol{z} \in S} \|\boldsymbol{z} - \boldsymbol{x}\|.$$

To prove that the function $\operatorname{dist}(\boldsymbol{x}, S)$ is continuous in $\boldsymbol{x}$, take the infimum over $\boldsymbol{z} \in S$ of both sides of the triangle inequality

$$\|\boldsymbol{z} - \boldsymbol{x}\| \quad \leq \|\boldsymbol{z} - \boldsymbol{y}\| + \|\boldsymbol{y} - \boldsymbol{x}\|.$$

This demonstrates that $\operatorname{dist}(\boldsymbol{x}, S) \leq \operatorname{dist}(\boldsymbol{y}, S) + \|\boldsymbol{y} - \boldsymbol{x}\|$. Reversing the roles of $\boldsymbol{x}$ and $\boldsymbol{y}$ then leads to the inequality

$$|\operatorname{dist}(\boldsymbol{x}, S) - \operatorname{dist}(\boldsymbol{y}, S)| \quad \leq \quad \|\boldsymbol{y} - \boldsymbol{x}\|,$$

establishing continuity. ■

In generalizing continuity to more abstract topological spaces, the characterizations in the next proposition are crucial.

Proposition 2.5.2 *The following conditions are equivalent for a function* $f(\boldsymbol{x})$ *from* $T \subset \mathsf{R}^m$ *to* R^n*:*

(a) $f(\boldsymbol{x})$ *is continuous.*

(b) *The inverse image* $f^{-1}(S)$ *of every closed set* S *is closed.*

(c) *The inverse image* $f^{-1}(S)$ *of every open set* S *is open.*

Proof: To prove that (a) implies (b), suppose $\boldsymbol{x}_i$ is a sequence in $f^{-1}(S)$ tending to $\boldsymbol{x} \in T$. Then the conclusion $\lim_{i\to\infty} f(\boldsymbol{x}_i) = f(\boldsymbol{x})$ identifies $f(\boldsymbol{x})$ as an element of the closed set S and therefore $\boldsymbol{x}$ as belonging to $f^{-1}(S)$. Conditions (b) and (c) are equivalent because of the relation

$$f^{-1}(S)^c \quad = \quad f^{-1}(S^c)$$

between inverse images and set complements. Finally, to prove that (c) entails (a), suppose that $\lim_{i\to\infty} \boldsymbol{x}_i = \boldsymbol{x}$. For any $\epsilon > 0$, the inverse image of the ball $B[f(\boldsymbol{x}), \epsilon]$ is open by assumption. Consequently, there exists a neighborhood $T \cap B(\boldsymbol{x}, \delta)$ mapped into $B[f(\boldsymbol{x}), \epsilon]$. In other words,

$$\|f(\boldsymbol{x}_i) - f(\boldsymbol{x})\| \quad < \quad \epsilon$$

whenever $\|\boldsymbol{x}_i - \boldsymbol{x}\| < \delta$, which is sufficient to validate continuity. ■

Example 2.5.3 *Continuity of* $\sqrt[m]{x}$

The root function $f(x) = \sqrt[m]{x}$ is the functional inverse of the power function $g(x) = x^m$. We have already noted that $g(x)$ is continuous. On the interval $(0, \infty)$, it is also strictly increasing and maps the open interval (a, b) onto the open interval (a^m, b^m). Put another way, $f^{-1}[(a, b)] = (a^m, b^m)$. (Here we implicitly invoke the intermediate value property proved in Proposition 2.7.1.) Because the inverse image of a union of open intervals is a union of open intervals, application of part (c) of Proposition 2.5.2 establishes the continuity of $f(x)$. ■

Example 2.5.4 *The Set of Positive Definite Matrices*

A symmetric $n \times n$ matrix $\boldsymbol{M} = (m_{ij})$ can be viewed as a point in R^m for $m = \binom{n}{2} + n$. To demonstrate that the subset S of positive definite matrices is open in R^m, we invoke the classical criterion of Sylvester. (See Problem 29 of Chap. 5 or [136].) This test for positive definiteness uses the determinants of the principal submatrices of $\boldsymbol{M}$. The kth of these submatrices $\boldsymbol{M}_k$ is the $k \times k$ upper left block of $\boldsymbol{M}$. If $\boldsymbol{M}$ is positive definite, then one can show that $\boldsymbol{M}_k$ is positive definite by taking a nontrivial $k \times 1$ vector $\boldsymbol{x}_k$ and padding it with zeros to construct a nontrivial $n \times 1$ vector $\boldsymbol{x}$. It is then clear that $\boldsymbol{x}_k^* \boldsymbol{M}_k \boldsymbol{x}_k = \boldsymbol{x}^* \boldsymbol{M} \boldsymbol{x} > 0$. Because $\boldsymbol{M}_k$ is positive definite, its determinant $\det \boldsymbol{M}_k > 0$. Conversely, if all of the $\det \boldsymbol{M}_k > 0$, then $\boldsymbol{M}$ itself is positive definite.

Given this background, we write

$$S \quad = \quad \bigcap_{k=1}^{n} \{\boldsymbol{M} : \det \boldsymbol{M}_k > 0\}.$$

Because the functions $\det \boldsymbol{M}_k$ are continuous in the entries of $\boldsymbol{M}$, the inverse images $\{\boldsymbol{M} : \det \boldsymbol{M}_k > 0\}$ of the open set $(0, \infty)$ are open. Since a finite intersection of open sets is open, S itself is an open set. ■

As opposed to inverse images, the image of a closed (open) set under a continuous function need not be closed (open). However, continuous functions do preserve compactness.

Proposition 2.5.3 *Suppose the continuous function $f(\boldsymbol{x})$ maps the compact set $S \subset \mathsf{R}^m$ into R^n. Then the image $f(S)$ is compact.*

Proof: The key is to apply Proposition 2.4.3. Let $f(\boldsymbol{x}_i)$ be a sequence in $f(S)$. Extract a convergent subsequence $\boldsymbol{x}_{i_j}$ of $\boldsymbol{x}_i$ with limit $\boldsymbol{y} \in S$. Then the continuity of $f(\boldsymbol{x})$ compels $f(\boldsymbol{x}_{i_j})$ to converge to $f(\boldsymbol{y})$. ■

We now come to one of the most important results in optimization theory.

Proposition 2.5.4 (Weierstrass) *Let $f(\boldsymbol{x})$ be a continuous real-valued function defined on a set S of R^n. If the set $T = \{\boldsymbol{x} \in S : f(\boldsymbol{x}) \geq f(\boldsymbol{y})\}$ is compact for some $\boldsymbol{y} \in S$, then $f(\boldsymbol{x})$ attains its supremum on S. Similarly, if $T = \{\boldsymbol{x} \in S : f(\boldsymbol{x}) \leq f(\boldsymbol{y})\}$ is compact for some $\boldsymbol{y} \in S$, then $f(\boldsymbol{x})$ attains its infimum on S. Both conclusions apply when S itself is compact.*

Proof: Consider the question of whether the function $f(\boldsymbol{x})$ attains its supremum $u = \sup_{\boldsymbol{x} \in S} f(\boldsymbol{x})$. The set $f(T)$ is bounded by virtue of Proposition 2.5.3, and the supremum of $f(\boldsymbol{x})$ on T coincides with u. For every positive integer i choose a point $\boldsymbol{x}_i \in T$ such that $f(\boldsymbol{x}_i) \geq u - 1/i$. In view of the compactness of T, we can extract a convergent subsequence of $\boldsymbol{x}_i$ with limit $\boldsymbol{z} \in T$. The continuity of $f(\boldsymbol{x})$ along this subsequence then implies that $f(\boldsymbol{z}) = u$. ■

Example 2.5.5 *Closest Point in a Set*

To prove that the distance $\operatorname{dist}(\boldsymbol{x}, S)$ is achieved for some $\boldsymbol{z} \in S$, we must assume that S is closed. In finding the closest point to $\boldsymbol{x}$ in S, choose any point $\boldsymbol{y} \in S$. The set $T = S \cap \{\boldsymbol{z} : \|\boldsymbol{z} - \boldsymbol{x}\| \leq \|\boldsymbol{y} - \boldsymbol{x}\|\}$ is both closed and bounded and therefore compact. Proposition 2.5.4 now informs us that the continuous function $\boldsymbol{z} \mapsto \|\boldsymbol{z} - \boldsymbol{x}\|$ attains its infimum on S. ■

Example 2.5.6 *Equivalence of Norms*

Every norm $\|\boldsymbol{x}\|_\dagger$ on R^n is equivalent to the Euclidean norm $\|\boldsymbol{x}\|$ in the sense that there exist positive constants a and b such that the inequalities

$$a\|\boldsymbol{x}\| \quad \leq \quad \|\boldsymbol{x}\|_\dagger \quad \leq \quad b\|\boldsymbol{x}\| \tag{2.6}$$

hold for all $\boldsymbol{x}$. To prove the right inequality in (2.6), let $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ denote the standard basis. Then conditions (c) and (d) defining a norm indicate that $\boldsymbol{x} = \sum_{i=1}^n x_i \boldsymbol{e}_i$ satisfies

$$\|\boldsymbol{x}\|_\dagger \quad \leq \quad \sum_{i=1}^n |x_i| \cdot \|\boldsymbol{e}_i\|_\dagger \quad = \quad \|\boldsymbol{x}\| \sum_{i=1}^n \|\boldsymbol{e}_i\|_\dagger.$$

This proves the upper bound with $b = \sum_{i=1}^n \|\boldsymbol{e}_i\|_\dagger$.

To establish the lower bound, we note that property (c) of a norm allows us to restrict attention to the sphere $S = \{\boldsymbol{x} : \|\boldsymbol{x}\| = 1\}$. Now the function $\boldsymbol{x} \mapsto \|\boldsymbol{x}\|_\dagger$ is uniformly continuous on R^n because

$$\big|\, \|\boldsymbol{x}\|_\dagger - \|\boldsymbol{y}\|_\dagger \,\big| \quad \leq \quad \|\boldsymbol{x} - \boldsymbol{y}\|_\dagger \quad \leq \quad b\|\boldsymbol{x} - \boldsymbol{y}\|$$

follows from the upper bound just demonstrated. Since the sphere S is compact, the continuous function $\boldsymbol{x} \to \|\boldsymbol{x}\|_\dagger$ attains its lower bound a on S. In view of property (b) defining a norm, $a > 0$. ■

Example 2.5.7 *The Fundamental Theorem of Algebra*

Consider a polynomial $p(z) = c_n z^n + c_{n-1} z^{n-1} + \cdots + c_0$ in the complex variable z with $c_n \neq 0$. The fundamental theorem of algebra says that $p(z)$ has a root. d'Alembert suggested an interesting optimization proof of this fact [30]. We begin by observing that if we identify a complex number with an ordered pair of real numbers, then the domain of the real-valued function $|p(z)|$ is R^2. The identity

$$|p(z)| \quad = \quad |z|^n \left| c_n + \frac{c_{n-1}}{z} + \cdots + \frac{c_0}{z^n} \right|$$

shows that $|p(z)|$ tends to ∞ whenever $|z|$ tends to ∞. Therefore, the set $T = \{z : |p(z| \leq d\}$ is compact for any d, and Proposition 2.5.4 implies that $|p(z)|$ attains its minimum at some point y. Expanding $p(z)$ around y gives a polynomial

$$q(z) \quad = \quad p(z + y) \quad = \quad b_n z^n + b_{n-1} z^{n-1} + \cdots + b_0$$

with the same degree as $p(z)$. Furthermore, the minimum of $|q(z)|$ occurs at $z = 0$. Suppose $b_1 = \cdots = b_{k-1} = 0$ and $b_k \neq 0$. For some angle $\theta \in [0, 2\pi)$, the scaled complex exponential

$$u \quad = \quad \left| \frac{b_0}{b_k} \right|^{1/k} e^{i\theta/k}$$

is a root of the equation $b_k u^k + b_0 = 0$. The function $f(t) = |q(tu)|$ clearly satisfies $f(t) \geq |b_0|$ and

$$f(t) \quad = \quad |b_k t^k u^k + b_0| + o(t^k) \quad = \quad |b_0(1 - t^k)| + o(t^k)$$

for t small and positive. These two conditions are compatible only if $b_0 = 0$. Hence, the minimum of $|q(z)| = |p(z + y)|$ is 0. ■

Example 2.5.8 *Continuity of the Roots of a Polynomial*

As a followup to the previous example, let us prove that the roots of a polynomial depend continuously on its coefficients [261]. One has to exercise

caution in stating this result. First, we limit ourselves to monic polynomials $p(z) = z^n + c_{n-1}z^{n-1} + \cdots + c_0$. Second, we rely on the fact that a monic polynomial can be written in factored form as

$$p(z) \quad = \quad (z - r_1) \cdots (z - r_n) \tag{2.7}$$

based on the roots guaranteed by the fundamental theorem of algebra. Let r be a root of $p(z)$ of multiplicity m and $q(z) = z^n + d_{n-1}z^{n-1} + \cdots + d_0$ be a second monic polynomial of the same degree n as $p(z)$. We now interpret continuity to mean that for every $\epsilon > 0$, there exists a $\delta > 0$ such that $q(z)$ has at least m roots within ϵ of r whenever the coefficient vector $\boldsymbol{d}$ of $q(z)$ satisfies $\|\boldsymbol{d} - \boldsymbol{c}\| < \delta$. Here we use the Euclidean norm on R^{2n}. In proving this result, we need the simple bound

$$|r_j| \quad \leq \quad \max\Big\{1, \sum_{i=0}^{n-1} |c_i|\Big\}.$$

on the roots of a monic polynomial $p(z)$ in terms of its coefficients. The proof of the bound is an immediate consequence of the identity

$$r_j \quad = \quad -\sum_{i=0}^{n-1} c_i r_j^{i-n+1}.$$

We are now in a position to verify the asserted continuous dependence. Suppose it fails for the polynomial $p(z)$ and the specified root r. Then for some $\epsilon > 0$ there exists a sequence $q_k(z)$ of monic polynomials of degree n with fewer than m roots within ϵ of r but whose coefficients d_{ki} converge to the coefficients c_i. Since the coefficients of the $q_k(z)$ converge, by the above inequality, the roots s_{ki} of the $q_k(z)$ are bounded. We can therefore extract a subsequence $q_{k_l}(z)$ whose roots converge to the complex numbers t_i. At most $m - 1$ of the t_i equal r. The representation

$$p(z) \quad = \quad \lim_{l \to \infty} q_{k_l}(z) \quad = \quad (z - t_1) \cdots (z - t_n)$$

is at odds with the representation (2.7) of $p(z)$. Indeed, one has m roots equal to r, and the other has at most $m - 1$ roots equal to r. This contradiction proves the claimed continuity of the roots.

As an illustration consider the quadratic $p(z) = z^2 - 2z + 1 = (z - 1)^2$ with the root 1 of multiplicity 2. For $\delta > 0$ small the related polynomial $z^2 - 2z + 1 - \delta$ has the real roots $1 \pm \sqrt{\delta}$ while the polynomial $z^2 - 2z + 1 + \delta$ has the complex roots $1 \pm \sqrt{-\delta}$. A more important application concerns the continuity of the eigenvalues of a matrix. Suppose the sequence $\boldsymbol{M}_k$ of square matrices converges to the square matrix $\boldsymbol{M}$. Then the sequence of characteristic polynomials $\det(z\boldsymbol{I} - \boldsymbol{M}_k)$ converges to the characteristic polynomial $\det(z\boldsymbol{I} - \boldsymbol{M})$. It follows that the eigenvalues of $\boldsymbol{M}_k$ converge to the eigenvalues of $\boldsymbol{M}$ in the sense just explained. ■

A function $f(\boldsymbol{x})$ is said to be uniformly continuous on its domain S if for every $\epsilon > 0$ there exists a $\delta > 0$ such that $\|f(\boldsymbol{y}) - f(\boldsymbol{x})\| < \epsilon$ whenever $\|\boldsymbol{y} - \boldsymbol{x}\| < \delta$. This sounds like ordinary continuity, but the chosen δ does not depend on the pivotal point $\boldsymbol{x} \in S$. One of the virtues of a compact domain is that it forces uniform continuity.

Proposition 2.5.5 (Heine) *Every continuous function $f(\boldsymbol{x})$ from a compact set S of R^m into R^n is uniformly continuous.*

Proof: Suppose $f(\boldsymbol{x})$ fails to be uniformly continuous. Then for some $\epsilon > 0$, there exist sequences $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ from S such that $\lim_{i\to\infty} \|\boldsymbol{x}_i - \boldsymbol{y}_i\| = 0$ and $\|f(\boldsymbol{x}_i) - f(\boldsymbol{y}_i)\| \geq \epsilon$. Since S is compact, we can extract a subsequence of $\boldsymbol{x}_i$ that converges to a point $\boldsymbol{u} \in S$. Along the corresponding subsequence of $\boldsymbol{y}_i$ we can extract a subsubsequence that converges to a point $\boldsymbol{v} \in S$. Substituting the constructed subsubsequences for $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ if necessary, we may assume that $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ both converge to the same limit $\boldsymbol{u} = \boldsymbol{v}$. The condition $\|f(\boldsymbol{x}_i) - f(\boldsymbol{y}_i)\| \geq \epsilon$ now contradicts the continuity of $f(\boldsymbol{x})$ at $\boldsymbol{u}$. ■

Example 2.5.9 *Rigid Motions*

Uniform continuity certainly appears in the absence of compactness. One spectacular example is a rigid motion. By this we mean a function $f(\boldsymbol{x})$ of R^n into itself with the property $\|f(\boldsymbol{y}) - f(\boldsymbol{x})\| = \|\boldsymbol{y} - \boldsymbol{x}\|$ for every choice of $\boldsymbol{x}$ and $\boldsymbol{y}$. We can better understand the rigid motion $f(\boldsymbol{x})$ by investigating the translated rigid motion $g(\boldsymbol{x}) = f(\boldsymbol{x}) - f(\boldsymbol{0})$ that maps the origin $\boldsymbol{0}$ into itself. Because $g(\boldsymbol{x})$ preserves distances, it also preserves inner products. This fact is evident from the equalities

$$\begin{aligned}
\|\boldsymbol{y} - \boldsymbol{x}\|^2 &= \|\boldsymbol{y}\|^2 - 2\boldsymbol{y}^*\boldsymbol{x} + \|\boldsymbol{x}\|^2 \\
\|g(\boldsymbol{y}) - g(\boldsymbol{x})\|^2 &= \|g(\boldsymbol{y})\|^2 - 2g(\boldsymbol{y})^* g(\boldsymbol{x}) + \|g(\boldsymbol{x})\|^2 \\
\|g(\boldsymbol{y})\|^2 &= \|\boldsymbol{y}\|^2 \\
\|g(\boldsymbol{x})\|^2 &= \|\boldsymbol{x}\|^2.
\end{aligned}$$

The inner product identity

$$g(\boldsymbol{y})^* g(\boldsymbol{x}) = \boldsymbol{y}^*\boldsymbol{x}$$

is only possible if $g(\boldsymbol{y})$ is linear. To demonstrate this assertion, note that $g(\boldsymbol{x})$ maps the standard orthonormal basis $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ onto the orthonormal basis $g(\boldsymbol{e}_1), \ldots, g(\boldsymbol{e}_n)$. Because

$$\begin{aligned}
g(\alpha\boldsymbol{x} + \beta\boldsymbol{y})^* g(\boldsymbol{e}_i) &= (\alpha\boldsymbol{x} + \beta\boldsymbol{y})^* \boldsymbol{e}_i \\
&= \alpha\boldsymbol{x}^*\boldsymbol{e}_i + \beta\boldsymbol{y}^*\boldsymbol{e}_i \\
&= [\alpha g(\boldsymbol{x}) + \beta g(\boldsymbol{y})]^* g(\boldsymbol{e}_i)
\end{aligned}$$

holds for all i, it follows that $g(\alpha\boldsymbol{x}+\beta\boldsymbol{y}) = \alpha g(\boldsymbol{x}) + \beta g(\boldsymbol{y})$. In other words, $g(\boldsymbol{x})$ is linear. The linear transformations that preserve angles and distances are precisely the orthogonal transformations. Thus, the rigid motion $f(\boldsymbol{x})$ reduces to an orthogonal transformation $\boldsymbol{U}\boldsymbol{x}$ followed by the translation $f(\boldsymbol{0})$. Conversely, it is trivial to prove that every such transformation

$$f(\boldsymbol{x}) \quad = \quad \boldsymbol{U}\boldsymbol{x} + f(\boldsymbol{0})$$

is a rigid motion. ■

Example 2.5.10 *Multilinear Maps*

A k-linear map $\boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k]$ transforms points from the k-fold Cartesian product $\mathsf{R}^m \times \mathsf{R}^m \times \cdots \times \mathsf{R}^m$ into points in R^n and satisfies the rules

$$\begin{aligned} \boldsymbol{M}[\boldsymbol{u}_1, \ldots, c\boldsymbol{u}_j, \ldots, \boldsymbol{u}_k] &= c\boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_j, \ldots, \boldsymbol{u}_k] \\ \boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_j + \boldsymbol{v}_j, \ldots, \boldsymbol{u}_k] &= \boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_j, \ldots, \boldsymbol{u}_k] \\ &\quad + \boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{v}_j, \ldots, \boldsymbol{u}_k] \end{aligned}$$

for every scalar c, index j, vector $\boldsymbol{v}_j$, and combination of vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$. For example, matrix multiplication $\boldsymbol{u} \mapsto \boldsymbol{A}\boldsymbol{u}$ is 1-linear and the determinant map $\boldsymbol{U} \mapsto \det \boldsymbol{U}$ is k-linear on the columns $\boldsymbol{u}_j$ of a $k \times k$ matrix. A k-linear map into the real line ($n = 1$) is called a k-linear form. The k-linear form

$$[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k] \quad \mapsto \quad \prod_{j=1}^{k} \boldsymbol{v}_j^* \boldsymbol{u}_j \tag{2.8}$$

for any fixed combination $[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k]$ of vectors is often useful in applications. A k-linear map $\boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_j, \ldots, \boldsymbol{u}_k]$ is said to be symmetric if

$$\boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_i, \ldots, \boldsymbol{u}_j, \ldots, \boldsymbol{u}_k] \quad = \quad \boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_j, \ldots, \boldsymbol{u}_i, \ldots, \boldsymbol{u}_k]$$

for all pairs of indices i and j and antisymmetric if

$$\boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_i, \ldots, \boldsymbol{u}_j, \ldots, \boldsymbol{u}_k] \quad = \quad -\boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_j, \ldots, \boldsymbol{u}_i, \ldots, \boldsymbol{u}_k].$$

The determinant function is antisymmetric.

One can easily check that the collection $\mathsf{L}^k(\mathsf{R}^m, \mathsf{R}^n)$ of k-linear maps from $\mathsf{R}^m \times \mathsf{R}^m \times \cdots \times \mathsf{R}^m$ to R^n forms a vector space under pointwise addition and scalar multiplication. Its dimension is $m^k n$. Indeed, let $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m$ be a basis for R^m and $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$ be a basis for R^n. If $\boldsymbol{u}_i = \sum_{j=1}^{m} c_{ij}\boldsymbol{e}_j$, then the expansion

$$\boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k] \quad = \quad \sum_{j_1=1}^{m} \cdots \sum_{j_k=1}^{m} \Big(\prod_{i=1}^{k} c_{i,j_i} \Big) \boldsymbol{M}[\boldsymbol{e}_{j_1}, \ldots, \boldsymbol{e}_{j_k}] \tag{2.9}$$

correctly suggests that the k-linear maps with

$$\boldsymbol{M}_{j_1,\ldots,j_k,l}[\boldsymbol{e}_{i_1},\ldots,\boldsymbol{e}_{i_k}] \;=\; \begin{cases} \boldsymbol{f}_l & i_1 = j_1,\ldots,i_k = j_k \\ \boldsymbol{0} & \text{otherwise} \end{cases}$$

constitute a basis of $\mathsf{L}^k(\mathsf{R}^m,\mathsf{R}^n)$. For a linear form $\boldsymbol{M}[\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k]$, it is helpful to think of the numbers $\boldsymbol{M}[\boldsymbol{e}_{i_1},\ldots,\boldsymbol{e}_{i_k}]$ as coefficients that define the linear form just as the coefficients of a matrix define the corresponding linear transformation.

Equation (2.9) also implies that $\boldsymbol{M}[\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k]$ is a continuous function of its arguments. Therefore, the norm

$$\|\boldsymbol{M}\| \;=\; \sup_{\boldsymbol{u}_j \neq \boldsymbol{0}\,\forall j} \frac{\|\boldsymbol{M}[\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k]\|}{\|\boldsymbol{u}_1\|\cdots\|\boldsymbol{u}_k\|} \;=\; \sup_{\|\boldsymbol{u}_j\|=1\,\forall j} \|\boldsymbol{M}[\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k]\|$$

on $\mathsf{L}^k(\mathsf{R}^m,\mathsf{R}^n)$ induced by the Euclidean norms on R^m and R^n is finite. For example, the norm of the k-linear form (2.8) is $\prod_{j=1}^k \|\boldsymbol{v}_j\|$. This value is attained by choosing $\boldsymbol{u}_j = \|\boldsymbol{v}_j\|^{-1}\boldsymbol{v}_j$ and serves as an absolute upper bound on the k-linear form on unit vectors by virtue of the Cauchy-Schwarz inequality. The inequality

$$\|\boldsymbol{M}[\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k]\| \;\leq\; \|\boldsymbol{M}\|\|\boldsymbol{u}_1\|\cdots\|\boldsymbol{u}_k\| \tag{2.10}$$

is an immediate consequence of the definition of $\|\boldsymbol{M}\|$. Problem 33 asks the reader to verify that the map $(\boldsymbol{M},\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k) \mapsto \boldsymbol{M}[\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k]$ is jointly continuous in its $k+1$ variables. ■

2.6 Semicontinuity

For real-valued functions, the notions of lower and upper semicontinuity are often useful substitutes for continuity. A real-valued function $f(\boldsymbol{x})$ with domain $T \subset \mathsf{R}^m$ is lower semicontinuous if the set $\{\boldsymbol{x} \in T : f(\boldsymbol{x}) \leq c\}$ is closed in T for every constant c. Given the duality of closed and open sets, an equivalent condition is that $\{\boldsymbol{x} \in T : f(\boldsymbol{x}) > c\}$ is open in T for every constant c. A real-valued function $g(\boldsymbol{x})$ is said to be upper semicontinuous if and only if $f(\boldsymbol{x}) = -g(\boldsymbol{x})$ is lower semicontinuous. Owing to this simple relationship, we will confine our attention to lower semicontinuous functions. The next proposition gives two alternative definitions.

Proposition 2.6.1 *A necessary and sufficient condition for $f(\boldsymbol{x})$ to be lower semicontinuous is that*

$$f(\boldsymbol{x}) \;\leq\; \liminf_{n\to\infty} f(\boldsymbol{x}_n) \tag{2.11}$$

whenever $\lim_{n\to\infty} \boldsymbol{x}_n = \boldsymbol{x}$ in T. Another necessary and sufficient condition is that the epigraph $\{(\boldsymbol{x},y) \in T \times \mathsf{R} : f(\boldsymbol{x}) \leq y\}$ is a closed set.

Proof: Suppose $f(\boldsymbol{x})$ is lower semicontinuous and $\lim_{n\to\infty} \boldsymbol{x}_n = \boldsymbol{x}$. For any $\epsilon > 0$, the point $\boldsymbol{x}$ lies in the open set $\{\boldsymbol{y} \in T : f(\boldsymbol{y}) > f(\boldsymbol{x}) - \epsilon\}$. Hence, $f(\boldsymbol{x}_n) > f(\boldsymbol{x}) - \epsilon$ for all sufficient large n. But this implies inequality (2.11). Similarly, if $\boldsymbol{x}_n$ converges to $\boldsymbol{x}$ and $y_n \geq f(\boldsymbol{x}_n)$ converges to y, then the inequality $y \geq \liminf_{n\to\infty} f(\boldsymbol{x}_n) \geq f(\boldsymbol{x})$ follows, and the epigraph is closed. Thus, both stated conditions are necessary.

For sufficiency, suppose inequality (2.11) holds. Consider a sequence $\boldsymbol{x}_n$ in the set $\{\boldsymbol{y} \in T : f(\boldsymbol{y}) \leq c\}$ with limit $\boldsymbol{x}$. It is clear that $f(\boldsymbol{x}) \leq c$ as well. Thus, $\{\boldsymbol{y} \in T : f(\boldsymbol{y}) \leq c\}$ is closed. To deal with the second sufficient condition, suppose the epigraph is closed, but $f(\boldsymbol{x})$ is not lower semicontinuous. Then there exists a sequence $\boldsymbol{x}_n$ converging to $\boldsymbol{x}$ in T and an $\epsilon > 0$ such that $f(\boldsymbol{x}) - \epsilon > \liminf_{n\to\infty} f(\boldsymbol{x}_n)$. It follows that the pair $(\boldsymbol{x}_n, f(\boldsymbol{x}) - \epsilon)$ is in the epigraph for infinitely many n. Because the epigraph is closed, this forces the contradiction that $(\boldsymbol{x}, f(\boldsymbol{x}) - \epsilon)$ belongs to the epigraph. ■

Part of the motivation for defining semicontinuity is to generalize Proposition 2.5.4. The result stated there for global maxima holds for upper semicontinuous functions, and the result for global minima holds for lower semicontinuous functions. The proof carries over almost word for word. It is also obvious that any continuous function is lower semicontinuous, and any function that is both lower and upper semicontinuous is continuous. Fortunately, the closure properties of lower semicontinuous functions are quite flexible.

Proposition 2.6.2 *The collection of lower semicontinuous functions with common domain $T \subset \mathsf{R}^m$ satisfies the following rules:*

(a) *If $f_k(\boldsymbol{x})$ is a family of lower semicontinuous functions, then $\sup_k f_k(\boldsymbol{x})$ is lower semicontinuous.*

(b) *If $f_k(\boldsymbol{x})$ is a finite family of lower semicontinuous functions, then $\min_k f_k(\boldsymbol{x})$ is lower semicontinuous.*

(c) *If $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are lower semicontinuous, then $f(\boldsymbol{x}) + g(\boldsymbol{x})$ is lower semicontinuous.*

(d) *If $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are both positive and lower semicontinuous, then $f(\boldsymbol{x})g(\boldsymbol{x})$ is lower semicontinuous.*

(e) *If $f(\boldsymbol{x})$ is lower semicontinuous and $g(\boldsymbol{x})$ is continuous with range U contained in T, then $f \circ g(\boldsymbol{x})$ is lower semicontinuous.*

Proof: These rules follow from the set identities

$$\begin{aligned} \{\boldsymbol{x} : \sup_k f_k(\boldsymbol{x}) > c\} &= \cup_k \{\boldsymbol{x} : f_k(\boldsymbol{x}) > c\} \\ \{\boldsymbol{x} : \min_k f_k(\boldsymbol{x}) > c\} &= \cap_k \{\boldsymbol{x} : f_k(\boldsymbol{x}) > c\} \end{aligned}$$

$$\begin{aligned}\{\boldsymbol{x} : f(\boldsymbol{x}) + g(\boldsymbol{x}) > c\} &= \cup_d(\{\boldsymbol{x} : f(\boldsymbol{x}) > c - d\} \cap \{\boldsymbol{x} : g(\boldsymbol{x}) > d\}) \\ \{\boldsymbol{x} : f(\boldsymbol{x})g(\boldsymbol{x}) > c\} &= \cup_{d>0}(\{\boldsymbol{x} : f(\boldsymbol{x}) > d^{-1}c\} \cap \{\boldsymbol{x} : g(\boldsymbol{x}) > d\}) \\ \{\boldsymbol{y} : f \circ g(\boldsymbol{y}) > c\} &= g^{-1}[\{\boldsymbol{x} : f(\boldsymbol{x}) > c\}]\end{aligned}$$

and the properties of open sets and continuous functions summarized in Propositions 2.4.2 and 2.5.2. ■

Example 2.6.1 *Row and Column Rank*

Every $m \times n$ matrix $\boldsymbol{A}$ has a well defined nullity and rank. Although these are not continuous functions of $\boldsymbol{A}$, the former function is upper semicontinuous, and the latter function is lower semicontinuous. In view of the dimension identity $\text{nullity}(\boldsymbol{A}) = n - \text{rank}(\boldsymbol{A})$, to validate both claims it suffices to show that $\text{rank}(\boldsymbol{A})$ is lower semicontinuous. Consider an arbitrary constant c and an arbitrary matrix $\boldsymbol{A} = (a_{ij})$ with $\text{rank}(\boldsymbol{A}) > c$. If we abbreviate $\text{rank}(\boldsymbol{A}) = r$, then there exist row indices $1 \leq i_1 < \cdots < i_r \leq m$ and column indices $1 \leq j_1 < \cdots < j_r \leq n$ such that the submatrix

$$\begin{pmatrix} a_{i_1 j_1} & \cdots & a_{i_1 j_r} \\ \vdots & \ddots & \vdots \\ a_{i_r j_1} & \cdots & a_{i_r j_r} \end{pmatrix}$$

has nonzero determinant. Because the determinant function is continuous, the same submatrix has nonvanishing determinant for all $m \times n$ matrices $\boldsymbol{B}$ close to $\boldsymbol{A}$. It follows that $\{\boldsymbol{A} : \text{rank}(\boldsymbol{A}) > c\}$ is an open set and therefore that $\text{rank}(\boldsymbol{A})$ is lower semicontinuous. ■

2.7 Connectedness

Roughly speaking, a set is disconnected if it can be split into two pieces sharing no boundary. A set is connected if it is not disconnected. One way of making this vague distinction precise is to consider a set S disconnected if there exists a real-valued continuous function $\phi(\boldsymbol{x})$ defined on S and having range $\{0, 1\}$. The nonempty subsets $A = \phi^{-1}(0)$ and $B = \phi^{-1}(1)$ then constitute the two disconnected pieces of S. According to part (b) of Proposition 2.5.2, both A and B are closed. Because one is the complement of the other, both are also open.

Arcwise connectedness is a variation on the theme of connectedness. A set is said to be arcwise connected if for any pair of points $\boldsymbol{x}$ and $\boldsymbol{y}$ of the set there is a continuous function $f(t)$ from the interval $[0, 1]$ into the set satisfying $f(0) = \boldsymbol{x}$ and $f(1) = \boldsymbol{y}$. We will see shortly that arcwise connectedness implies connectedness. On open sets, the two notions coincide.

Can we identify the connected subsets of the real line? Intuition suggests that the only connected subsets are intervals. Here a single point x is viewed

as the interval $[x, x]$. Suppose S is a connected subset of R, and let a and b be two points of S. In order for S to be an interval, every point $c \in (a, b)$ should be in S. If S fails to contain an intermediate point c, then we can define a continuous function $\phi(x)$ disconnecting S by taking $\phi(x) = 0$ for $x < c$ and $\phi(x) = 1$ for $x > c$. Thus, every connected subset must be an interval.

To prove the converse, suppose a disconnecting function $\phi(x)$ lives on an interval. Select points a and b of the interval with $\phi(a) = 0$ and $\phi(b) = 1$. Without loss of generality we can take $a < b$. On $[a, b]$ we now carry out the bisection strategy of Example 2.3.1, selecting the right or left subinterval at each stage so that the values of $\phi(x)$ at the endpoints of the selected subinterval disagree. Eventually, bisection leads to a subinterval contradicting the uniform continuity of $\phi(x)$ on $[a, b]$. Indeed, there is a number δ such that $|\phi(y) - \phi(x)| < 1$ whenever $|y - x| < \delta$; at some stage, the length of the subinterval containing points with both values of $\phi(x)$ falls below δ.

This result is the first of four characterizing connected sets.

Proposition 2.7.1 *Connected subsets of* R^n *have the following properties:*

(a) *A subset of the real line is connected if and only if it is an interval.*

(b) *The image of a connected set under a continuous function is connected.*

(c) *The union* $S = \cup_\alpha S_\alpha$ *of an arbitrary collection of connected subsets is connected if one of the sets* S_β *has a nonempty intersection* $S_\beta \cap S_\alpha$ *with every other set* S_α*.*

(d) *Every arcwise connected set* S *is connected.*

Proof: To prove part (b) let $f(\boldsymbol{x})$ be a continuous map from a connected set $S \subset \mathsf{R}^m$ into R^n. If the image $f(S)$ is disconnected, then there is a continuous function $\phi(\boldsymbol{x})$ disconnecting it. The composition $\phi \circ f(\boldsymbol{x})$ is continuous by part (g) of Proposition 2.5.1 and serves to disconnect S, contradicting the connectedness of S. To prove (c) suppose that the continuous function $\phi(\boldsymbol{x})$ disconnects the union S. Then there exists $\boldsymbol{y} \in S_{\alpha_1}$ and $\boldsymbol{z} \in S_{\alpha_2}$ with $\phi(\boldsymbol{y}) = 0$ and $\phi(\boldsymbol{z}) = 1$. Choose $\boldsymbol{u} \in S_\beta \cap S_{\alpha_1}$ and $\boldsymbol{v} \in S_\beta \cap S_{\alpha_2}$. If $\phi(\boldsymbol{u}) \neq \phi(\boldsymbol{v})$, then $\phi(\boldsymbol{x})$ disconnects S_β. If $\phi(\boldsymbol{u}) = \phi(\boldsymbol{v})$, then $\phi(\boldsymbol{y}) \neq \phi(\boldsymbol{u})$ or $\phi(\boldsymbol{z}) \neq \phi(\boldsymbol{v})$. In the former case $\phi(\boldsymbol{x})$ disconnects S_{α_1}, and in the latter case $\phi(\boldsymbol{x})$ disconnects S_{α_2}. Finally, to prove part (d), suppose the arcwise connected set S fails to be connected. Then there exists a continuous disconnecting function $\phi(\boldsymbol{x})$ with $\phi(\boldsymbol{y}) = 0$ and $\phi(\boldsymbol{z}) = 1$. Let $f(t)$ be an arc in S connecting $\boldsymbol{y}$ and $\boldsymbol{z}$. The continuous function $\phi \circ f(t)$ then serves to disconnect $[0, 1]$. ■

Example 2.7.1 *The Intermediate Value Property*

Consider a continuous function $f(x)$ from an interval $[a, b]$ to the real line. The intermediate value theorem asserts that the image $f([a, b])$ coincides with the interval $[\min f(x), \max f(x)]$. This theorem, which is a

consequence of properties (a) and (b) of Proposition 2.7.1, has many applications. For example, suppose $g(x)$ is a continuous function from $[0,1]$ into $[0,1]$. If $f(x) = g(x) - x$, then it is obvious that $f(0) \geq 0$ and $f(1) \leq 0$. It follows that $f(x) = 0$ for some x. In other words, $g(x)$ has a fixed point satisfying $g(x) = x$. ■

Example 2.7.2 *Connectedness of Spheres*

The set $S(\boldsymbol{x}, r)$ in R^n is the image of the continuous map $\boldsymbol{y} \mapsto \boldsymbol{x} + r\boldsymbol{y}/\|\boldsymbol{y}\|$ of the domain $T = \mathsf{R}^n \setminus \mathbf{0}$. Hence, to prove connectedness when $n > 1$, it suffices to prove that T is connected. To achieve this, we argue that T is arcwise connected. Consider two points $\boldsymbol{u}$ and $\boldsymbol{v}$ in T. If $\mathbf{0}$ does not lie on the line segment between $\boldsymbol{u}$ and $\boldsymbol{v}$, then we can use the function $f(t) = \boldsymbol{u} + t(\boldsymbol{v} - \boldsymbol{u})$ to connect $\boldsymbol{u}$ and $\boldsymbol{v}$. If $\mathbf{0}$ lies on the line segment, choose any $\boldsymbol{w}$ not on the line determined by $\boldsymbol{u}$ and $\boldsymbol{v}$. Now the continuous function

$$f(t) \quad = \quad \begin{cases} \boldsymbol{u} + 2t(\boldsymbol{w} - \boldsymbol{u}) & t \in [0, \frac{1}{2}] \\ \boldsymbol{w} + (2t-1)(\boldsymbol{v} - \boldsymbol{w}) & t \in [\frac{1}{2}, 1] \end{cases}$$

connects $\boldsymbol{u}$ and $\boldsymbol{v}$. The sphere $S(x, r)$ in R reduces to the two points $x - r$ and $x + r$ and is disconnected. ■

2.8 Uniform Convergence

Many delicate issues of analysis revolve around the question of whether a given property of a sequence of functions $f_m(\boldsymbol{x})$ is preserved under a passage to a limit. As a simple example, consider the sequence $f_m(x) = x^m$ of continuous functions defined on the unit interval $[0,1]$. It is clear that $f_m(x)$ converges pointwise to the discontinuous function

$$f(x) \quad = \quad \begin{cases} 0 & 0 \leq x < 1 \\ 1 & x = 1\,. \end{cases}$$

The failure of $f(x)$ to be continuous suggests that an additional hypothesis must be imposed. The key hypothesis is uniform convergence. This requires for each $\epsilon > 0$ that there exists an integer k such that $|f_m(\boldsymbol{x}) - f(\boldsymbol{x})| < \epsilon$ for all $m \geq k$ and all $\boldsymbol{x}$. Here the adjective "uniform" refers to the assumption that the same k works for all $\boldsymbol{x}$. Of course, k is allowed to depend on ϵ.

Proposition 2.8.1 *Suppose the sequence of continuous functions $f_m(\boldsymbol{x})$ maps a domain $D \subset \mathsf{R}^p$ into R^q. If $f_m(\boldsymbol{x})$ converges uniformly to $f(\boldsymbol{x})$ on D, then $f(\boldsymbol{x})$ is also continuous.*

Proof: Choose $\boldsymbol{y} \in D$ and $\epsilon > 0$, and take k so that $\|f_m(\boldsymbol{x}) - f(\boldsymbol{x})\| < \frac{\epsilon}{3}$ for all $m \geq k$ and $\boldsymbol{x}$. By virtue of the continuity of $f_k(\boldsymbol{x})$, there is a $\delta > 0$

such that $\|f_k(\boldsymbol{x}) - f_k(\boldsymbol{y})\| < \frac{\epsilon}{3}$ whenever $\|\boldsymbol{x} - \boldsymbol{y}\| < \delta$. Assuming that $\boldsymbol{y}$ is fixed and $\|\boldsymbol{x} - \boldsymbol{y}\| < \delta$, we have

$$\begin{aligned} \|f(\boldsymbol{x}) - f(\boldsymbol{y})\| &\leq \|f(\boldsymbol{x}) - f_k(\boldsymbol{x})\| + \|f_k(\boldsymbol{x}) - f_k(\boldsymbol{y})\| + \|f_k(\boldsymbol{y}) - f(\boldsymbol{y})\| \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \\ &= \epsilon. \end{aligned}$$

This shows that $f(\boldsymbol{x})$ is continuous at $\boldsymbol{y}$. ■

Example 2.8.1 *Weierstrass M-Test*

Suppose the entries $g_k(\boldsymbol{x})$ of a sequence of continuous functions satisfy $\|g_k(\boldsymbol{x})\| \leq M_k$, where $\sum_{k=1}^{\infty} M_k < \infty$. Then Cauchy's criterion and Proposition 2.8.1 together imply that the partial sums $f_l(\boldsymbol{x}) = \sum_{k=1}^{l} g_k(\boldsymbol{x})$ converge uniformly to the continuous function $f(\boldsymbol{x}) = \sum_{k=1}^{\infty} g_k(\boldsymbol{x})$. ■

2.9 Problems

1. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ be points in R^n. State and prove a necessary and sufficient condition under which the Euclidean norm equality

$$\|\boldsymbol{x}_1 + \cdots + \boldsymbol{x}_m\| = \|\boldsymbol{x}_1\| + \cdots + \|\boldsymbol{x}_m\|$$

holds. (Hints: Square and expand both sides. Use the necessary and sufficient conditions of the Cauchy-Schwarz inequality term by term.)

2. Show that it is possible to choose $n+1$ points $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in R^n such that $\|\boldsymbol{x}_i\| = 1$ for all i and $\|\boldsymbol{x}_i - \boldsymbol{x}_j\| = \|\boldsymbol{x}_k - \boldsymbol{x}_l\|$ for all pairs $i \neq j$ and $k \neq l$. These points define a regular simplex with vertices on the unit sphere. (Hint: One possibility is to take $\boldsymbol{x}_0 = n^{-1/2}\mathbf{1}$ and $\boldsymbol{x}_i = a\mathbf{1} + b\boldsymbol{e}_i$ for $i \geq 1$, where

$$a = -\frac{1+\sqrt{n+1}}{n^{3/2}}, \quad b = \sqrt{\frac{n+1}{n}}.$$

Any rotated version of these points also works.)

3. Show that

$$\|\boldsymbol{x}\|_q \leq \|\boldsymbol{x}\|_p \tag{2.12}$$

$$\|\boldsymbol{x}\|_p \leq n^{\frac{1}{p}-\frac{1}{q}}\|\boldsymbol{x}\|_q \tag{2.13}$$

when p and q are chosen from $\{1, 2, \infty\}$ and $p < q$. Here $\|\boldsymbol{x}\|_2$ is the Euclidean norm on R^n. These inequalities are sharp. Equality holds in inequality (2.12) when $\boldsymbol{x} = (1, 0, \ldots, 0)^*$, and equality holds in inequality (2.13) when $\boldsymbol{x} = (1, 1, \ldots, 1)^*$.

4. Show that $\|\boldsymbol{x}\|^2 \leq \|\boldsymbol{x}\|_\infty \|\boldsymbol{x}\|_1 \leq \sqrt{n}\|\boldsymbol{x}\|^2$ for any vector $\boldsymbol{x} \in \mathsf{R}^n$.

5. Prove that $1 \leq \|\boldsymbol{I}\|_\dagger$ and $\|\boldsymbol{M}\|_\dagger^{-1} \leq \|\boldsymbol{M}^{-1}\|_\dagger$ for any matrix norm on square matrices satisfying the defining properties (a) through (e) of Sect. 2.2.

6. Set $\|\boldsymbol{M}\|_{\max} = \max_{i,j} |m_{ij}|$ for $\boldsymbol{M} = (m_{ij})$. Show that this defines a vector norm but not a matrix norm on $n \times n$ matrices $\boldsymbol{M}$.

7. Let $\boldsymbol{M}$ be an $m \times n$ matrix. Prove that its spectral norm satisfies

$$\|\boldsymbol{M}\| \;=\; \sup_{\|\boldsymbol{v}\|=1} \|\boldsymbol{M}\boldsymbol{v}\| \;=\; \sup_{\|\boldsymbol{u}\|=1,\, \|\boldsymbol{v}\|=1} \boldsymbol{u}^* \boldsymbol{M} \boldsymbol{v}$$

for $\boldsymbol{u} \in \mathsf{R}^m$ and $\boldsymbol{v} \in \mathsf{R}^n$.

8. Demonstrate that the spectral norm on $m \times n$ matrices satisfies $\|\boldsymbol{U}\boldsymbol{M}\| = \|\boldsymbol{M}\| = \|\boldsymbol{M}\boldsymbol{V}\|$ for all orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ of the right dimensions. Show that the Frobenius norm satisfies the same orthogonal invariance principle.

9. Show that an $m \times n$ matrix $\boldsymbol{M} = (m_{ij})$ has the matrix norms

$$\begin{aligned} \|\boldsymbol{M}\|_1 &= \max_{1 \leq j \leq n} \sum_{i=1}^{m} |m_{ij}| \\ \|\boldsymbol{M}\|_\infty &= \max_{1 \leq i \leq m} \sum_{j=1}^{n} |m_{ij}| \end{aligned}$$

induced by the vector norms $\|\boldsymbol{x}\|_1$ and $\|\boldsymbol{x}\|_\infty$.

10. Let $\boldsymbol{M}$ be an $m \times n$ matrix of full column rank n. Prove that there exists a positive constant c such that $\|\boldsymbol{M}\boldsymbol{y}\| \geq c\|\boldsymbol{y}\|$ for all $\boldsymbol{y} \in \mathsf{R}^n$.

11. Demonstrate properties (2.4) and (2.5) of the limit superior and limit inferior. Also check that the sequence x_n has a limit if and only if equality holds in inequality (2.5).

12. Let $l = \limsup_{n \to \infty} x_n$. Show that:

    (a) $l = -\infty$ if and only if $\lim_{n \to \infty} x_n = -\infty$.

    (b) $l = +\infty$ if and only if for every positive integer m and real r there exists an $n \geq m$ with $x_n > r$.

    (c) l is finite if and only if (a) for every $\epsilon > 0$ there is an m such that $n \geq m$ implies $x_n < l + \epsilon$ and (b) for every $\epsilon > 0$ and every m there is an $n \geq m$ such that $x_n > l - \epsilon$.

    Similar properties hold for the limit inferior.

13. For any sequence of real numbers x_n, prove that

$$\inf_n x_n \quad \leq \quad \liminf_{n\to\infty} x_n \text{ and } \limsup_{n\to\infty} x_n \quad \leq \quad \sup_n x_n.$$

If y_n is a second sequence of real numbers, then prove that

$$\begin{aligned} \liminf_{n\to\infty} (x_n + y_n) &\geq \liminf_{n\to\infty} x_n + \liminf_{n\to\infty} y_n \\ \limsup_{n\to\infty} (x_n + y_n) &\leq \limsup_{n\to\infty} x_n + \limsup_{n\to\infty} y_n. \end{aligned}$$

Finally, if $x_n \leq y_n$ for all n, then prove that

$$\liminf_{n\to\infty} x_n \quad \leq \quad \liminf_{n\to\infty} y_n \text{ and } \limsup_{n\to\infty} x_n \quad \leq \quad \limsup_{n\to\infty} y_n.$$

14. Let x_n be a sequence of nonnegative real numbers with

$$x_{n+1} \quad \leq \quad x_n + \frac{1}{n^2}$$

for all $n \geq 1$. Show that $\lim_{n\to\infty} x_n$ exists [69].

15. Let $\boldsymbol{x}_m$ be a convergent sequence in R^n with limit $\boldsymbol{x}$. Prove that the sequence $\boldsymbol{s}_m = (\boldsymbol{x}_1 + \cdots + \boldsymbol{x}_m)/m$ of arithmetic means converges to $\boldsymbol{x}$.

16. Show that

$$\lim_{x\to\infty} p(x)e^{-x} \quad = \quad 0$$

for every polynomial $p(x)$.

17. Prove that the set of invertible matrices is open and that the sets of symmetric and orthogonal matrices are closed in R^{n^2}.

18. A square matrix is nilpotent if $\boldsymbol{A}^k = \boldsymbol{0}$ for some positive integer k. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are nilpotent, then show that $\boldsymbol{A}+\boldsymbol{B}$ need not be nilpotent. If we add the hypothesis that $\boldsymbol{A}$ and $\boldsymbol{B}$ commute, then show that $\boldsymbol{A} + \boldsymbol{B}$ is nilpotent. Use Example 2.3.3 to construct the inverses of the matrices $\boldsymbol{I} + \boldsymbol{A}$ and $\boldsymbol{I} - \boldsymbol{A}$ for $\boldsymbol{A}$ nilpotent [69].

19. Show that $e^{-\boldsymbol{M}}$ is the matrix inverse of $e^{\boldsymbol{M}}$. A skew symmetric matrix $\boldsymbol{M}$ satisfies $\boldsymbol{M}^* = -\boldsymbol{M}$. Show that $e^{\boldsymbol{M}}$ is orthogonal when $\boldsymbol{M}$ is skew symmetric.

20. Demonstrate that the matrix exponential function $\boldsymbol{M} \mapsto e^{\boldsymbol{M}}$ is continuous. (Hint: Apply the Weierstrass M-test.)

21. Demonstrate that the function

$$f(\boldsymbol{x}) \quad = \quad \begin{cases} \frac{x_1 x_2}{x_1^2 - x_2^2} & |x_1| \neq |x_2| \\ 0 & \text{otherwise} \end{cases}$$

on R^2 is discontinuous at $\boldsymbol{0}$.

22. Let $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ be real-valued continuous functions defined on the same domain. Prove that $\max\{f(\boldsymbol{x}), g(\boldsymbol{x})\}$ and $\min\{f(\boldsymbol{x}), g(\boldsymbol{x})\}$ are continuous functions. Prove that the function $f(\boldsymbol{x}) = \max_i x_i$ is continuous on R^n.

23. Define the function

$$f(x) \quad = \quad \begin{cases} x & x \text{ is rational} \\ 1-x & x \text{ is irrational} \end{cases}$$

on $[0,1]$. At what points is $f(x)$ continuous? What is the image $f([0,1])$?

24. Give an example of a continuous function that does not map an open set to an open set. Give another example of a continuous function that does not map a closed set to a closed set.

25. Show that the set of $n \times n$ orthogonal matrices is compact. (Hint: Show that every orthogonal matrix $\boldsymbol{O}$ has norm $\|\boldsymbol{O}\| = 1$.)

26. Let $f(\boldsymbol{x})$ be a continuous function from a compact set $S \subset \mathsf{R}^m$ into R^n. If $f(\boldsymbol{x})$ is one-to-one, then demonstrate that the inverse function $f^{-1}(\boldsymbol{y})$ is continuous from $f(S)$ to S.

27. Let $C = A \times B$ be the Cartesian product of two subsets $A \subset \mathsf{R}^m$ and $B \subset \mathsf{R}^n$. Prove that:

    (a) C is closed in R^{m+n} if both A and B are closed.

    (b) C is open in R^{m+n} if both A and B are open.

    (c) C is compact in R^{m+n} if both A and B are compact.

    (d) C is connected in R^{m+n} if both A and B are connected.

28. Prove the converse of each of the assertions in Problem 27.

29. Without appeal to Proposition 2.5.5, show that every polynomial on R is uniformly continuous on a compact interval $[a, b]$.

30. Let $f(x)$ be uniformly continuous on R and satisfy $f(0) = 0$. Demonstrate that there exists a nonnegative constant c such that

$$|f(x)| \quad \leq \quad 1 + c|x|$$

for all x [69].

31. Suppose that $f(x)$ is continuous on $[0, \infty)$ and $\lim_{x\to\infty} f(x)$ exists and is finite. Prove that $f(x)$ is uniformly continuous on $[0, \infty)$ [69].

32. Characterize those maps $f(\boldsymbol{x})$ from R^n into itself that have the property $\|f(\boldsymbol{y}) - f(\boldsymbol{x})\| = c\|\boldsymbol{y} - \boldsymbol{x}\|$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$. Here the constant c need not equal 1.

33. Prove that the multilinear map $(\boldsymbol{M}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) \mapsto \boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k]$ is jointly continuous in its $k+1$ variables. (Hint: Write

$$\begin{aligned} \boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k] - \boldsymbol{N}[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k] &= (\boldsymbol{M} - \boldsymbol{N})[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k] \\ &\quad + \boldsymbol{N}[\boldsymbol{u}_1 - \boldsymbol{v}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k] \\ &\quad + \boldsymbol{N}[\boldsymbol{v}_1, \boldsymbol{u}_2 - \boldsymbol{v}_2, \ldots, \boldsymbol{u}_k] \\ &\quad \vdots \\ &\quad + \boldsymbol{N}[\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{u}_k - \boldsymbol{v}_k] \end{aligned}$$

and take norms.)

34. Let $\boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k]$ be a symmetric k-linear map. Demonstrate that

$$\boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k] = \frac{1}{2^k k!} \sum \epsilon_1 \cdots \epsilon_k \boldsymbol{M}[(\epsilon_1 \boldsymbol{u}_1 + \cdots + \epsilon_k \boldsymbol{u}_k)^k],$$

where the sum ranges over all combinations of $\epsilon_1 = \pm 1, \ldots, \epsilon_k = \pm 1$. Hence, a symmetric k-linear map is determined by its values on the diagonal of its domain.

35. Continuing Problem 34, define the alternative norm

$$\|\boldsymbol{M}\|_{\text{sym}} = \sup_{\boldsymbol{u} \neq \boldsymbol{0}} \frac{\|\boldsymbol{M}[\boldsymbol{u}^k]\|}{\|\boldsymbol{u}\|^k} = \sup_{\|\boldsymbol{u}\|=1} \|\boldsymbol{M}[\boldsymbol{u}^k]\|.$$

Prove the inequalities

$$\|\boldsymbol{M}\|_{\text{sym}} \leq \|\boldsymbol{M}\| \leq \frac{k^k}{k!} \|\boldsymbol{M}\|_{\text{sym}}.$$

36. Show that the indicator function of an open set is lower semicontinuous and that the indicator function of a closed set is upper semicontinuous. Also show that floor function $f(x) = \lfloor x \rfloor$ is upper semicontinuous and that the ceiling function $f(x) = \lceil x \rceil$ is lower semicontinuous.

37. Suppose the n numbers $x_1, \ldots, x_n$ lie on $[0, 1]$. Prove that the function

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} |x - x_i|$$

attains the value $\frac{1}{2}$ for some $x \in [0, 1]$. (Hint: Consider $f(0)$ and $f(1)$.)

38. Show that a hyperplane $\{\boldsymbol{x} : \boldsymbol{z}^* \boldsymbol{x} = c\}$ in R^n is connected but that its complement is disconnected.

39. If the real-valued function $f(x)$ on $[a, b]$ is continuous and one-to-one, then prove that $f(x)$ is either strictly increasing or strictly decreasing.

40. Demonstrate that a polynomial of odd degree possesses at least one real root.

41. Prove that the closure of a connected set is connected.

42. Suppose T is a connected set in R^n. Define

$$\begin{aligned} U_\epsilon &= \{\boldsymbol{y} \in \mathsf{R}^n : \operatorname{dist}(\boldsymbol{y}, T) < \epsilon\} \\ V_\epsilon &= \{\boldsymbol{y} \in \mathsf{R}^n : \operatorname{dist}(\boldsymbol{y}, T) \leq \epsilon\} \end{aligned}$$

for $\epsilon > 0$ and $\operatorname{dist}(\boldsymbol{y}, T) = \inf_{\boldsymbol{x} \in T} \|\boldsymbol{y} - \boldsymbol{x}\|$. Demonstrate that U_ϵ and V_ϵ are connected. (Hints: V_ϵ is the closure of U_ϵ. For U_ϵ argue by contradiction using the definition of a connected set.)

43. On what domains do the sequences of functions

(a) $f_n(x) = (1 + x/n)^n$
(b) $f_n(x) = nx/(1 + n^2x^2)$
(c) $f_n(x) = n^x$
(d) $f_n(x) = x^{-1} \sin(nx)$
(e) $f_n(x) = xe^{-nx}$
(f) $f_n(x) = x^{2n}/(1 + x^{2n})$

converge [68]? On what domains do they converge uniformly?

44. Suppose that $f(x)$ is a function from the real line to itself satisfying $f(x + y) = f(x) + f(y)$ for all x and y. If $f(x)$ is continuous at a single point, then show that $f(x) = cx$ for some constant c. (Hints: Prove that $f(x)$ is continuous everywhere and that $f(q) = f(1)q$ for all rational numbers q.)

45. Suppose that $g(x)$ is a function from the real line to itself satisfying $g(x + y) = g(x)g(y)$ for all x and y. If $g(x)$ is continuous at a single point, then prove that either $g(x)$ is identically 0 or that there exists a positive constant d with $g(x) = d^x$. (Hint: Show that either $g(x)$ is identically 0 or that $g(x)$ is positive for all x. In the latter case, take logarithms and reduce to the previous problem.)

46. Suppose the real-valued function $f(\boldsymbol{x}, \boldsymbol{y})$ is jointly continuous in its two vector arguments and C is a compact set. Show that the functions

$$g(\boldsymbol{x}) = \inf_{\boldsymbol{y} \in C} f(\boldsymbol{x}, \boldsymbol{y}) \quad \text{and} \quad h(\boldsymbol{x}) = \sup_{\boldsymbol{y} \in C} f(\boldsymbol{x}, \boldsymbol{y})$$

are continuous.

3
The Gauge Integral

3.1 Introduction

Much of calculus deals with the interplay between differentiation and integration. The antiquated term "antidifferentiation" emphasizes the fact that differentiation and integration are inverses of one another. We will take it for granted that readers are acquainted with the mechanics of integration. The current chapter develops just enough integration theory to make our development of differentiation in Chap. 4 and the calculus of variations in Chap. 17 respectable. It is only fair to warn readers that in other chapters a few applications to probability and statistics will assume familiarity with properties of the expectation operator not covered here.

The first successful effort to put integration on a rigorous basis was undertaken by Riemann. In the early twentieth century, Lebesgue defined a more sophisticated integral that addresses many of the limitations of the Riemann integral. However, even Lebesgue's integral has its defects. In the past few decades, mathematicians such as Henstock and Kurzweil have expanded the definition of integration on the real line to include a wider variety of functions. The new integral emerging from these investigations is called the gauge integral or generalized Riemann integral [7, 68, 108, 193, 250, 255, 278]. The gauge integral subsumes the Riemann integral, the Lebesgue integral, and the improper integrals met in traditional advanced calculus courses. In contrast to the Lebesgue integral, the integrands of the gauge integral are not necessarily absolutely integrable.

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_3,

It would take us too far afield to develop the gauge integral in full generality. Here we will rest content with proving some of its elementary properties. One of the advantages of the gauge integral is that many theorems hold with fewer qualifications. The fundamental theorem of calculus is a case in point. The commonly stated version of the fundamental theorem concerns a differentiable function $f(x)$ on an interval $[a, b]$. As all students of calculus know,

$$\int_a^b f'(x)\,dx \quad = \quad f(b) - f(a).$$

Although this version is true for the gauge integral, it does not hold for the Lebesgue integral because the mere fact that $f'(x)$ exists throughout $[a, b]$ does not guarantee that it is Lebesgue integrable.

This quick description of the gauge integral is not intended to imply that the gauge integral is uniformly superior to the Lebesgue integral and its extensions. Certainly, probability theory would be severely handicapped without the full flexibility of modern measure theory. Furthermore, the advanced theory of the gauge integral is every bit as difficult as the advanced theory of the Lebesgue integral. For pedagogical purposes, however, one can argue that a student's first exposure to the theory of integration should feature the gauge integral. As we shall see, many of the basic properties of the gauge integral flow directly from its definition. As an added dividend, gauge functions provide an alternative approach to some of the material of Chap. 2.

3.2 Gauge Functions and δ-Fine Partitions

The gauge integral is defined through gauge functions. A gauge function is nothing more than a positive function $\delta(t)$ defined on a finite interval $[a, b]$. In approximating the integral of a function $f(t)$ over $[a, b]$ by a finite Riemann sum, it is important to sample the function most heavily in those regions where it changes most rapidly. Now by a Riemann sum we mean a sum

$$S(f, \pi) \quad = \quad \sum_{i=0}^{n-1} f(t_i)(s_{i+1} - s_i),$$

where the mesh points $a = s_0 < s_1 < \cdots < s_n = b$ form a partition π of $[a, b]$, and the tags t_i are chosen so that $t_i \in [s_i, s_{i+1}]$. If $\delta(t_i)$ measures the rapidity of change of $f(t)$ near t_i, then it makes sense to take $\delta(t)$ small in regions of rapid change and to force s_i and s_{i+1} to belong to the interval $(t_i - \delta(t_i), t_i + \delta(t_i))$. A tagged partition with this property is called a δ-fine partition. Our first proposition relieves our worry that δ-fine partitions exist.

Proposition 3.2.1 (Cousin's Lemma) *For every gauge $\delta(t)$ on a finite interval $[a, b]$ there is a δ-fine partition.*

Proof: Assume that $[a, b]$ lacks a δ-fine partition. Since we can construct a δ-fine partition of $[a, b]$ by appending a δ-fine partition of the half-interval $[(a + b)/2, b]$ to a δ-fine partition of the half-interval $[a, (a + b)/2]$, it follows that either $[a, (a+b)/2]$ or $[(a+b)/2, b]$ lacks a δ-fine partition. As in Example 2.3.1, we choose one of the half-intervals based on this failure and continue bisecting. This creates a nested sequence of intervals $[a_i, b_i]$ converging to a point x. If i is large enough, then $[a_i, b_i] \subset (x-\delta(x), x+\delta(x))$, and the interval $[a_i, b_i]$ with tag x is a δ-fine partition of itself. This contradicts the choice of $[a_i, b_i]$ and the assumption that the original interval $[a, b]$ lacks a δ-fine partition. ■

Before launching into our treatment of the gauge integral, we pause to gain some facility with gauge functions [108]. Here are three examples that illustrate their value.

Example 3.2.1 *A Gauge Proof of Weierstrass' Theorem*

Consider a real-valued continuous function $f(t)$ with domain $[a, b]$. Suppose that $f(t)$ does not attain its supremum on $[a, b]$. Then for each t there exists a point $x \in [a, b]$ with $f(t) < f(x)$. By continuity there exists $\delta(t) > 0$ such that $f(y) < f(x)$ for all $y \in [a, b]$ with $|y - t| < \delta(t)$. Using $\delta(t)$ as a gauge, select a δ-fine partition $a = s_0 < s_1 < \cdots < s_n = b$ with tags $t_i \in [s_i, s_{i+1}]$ and designated points x_i satisfying $f(t_i) < f(x_i)$. Let $x_{\max}$ be the point x_i having the largest value $f(x_i)$. Because $x_{\max}$ lies in some interval $[s_i, s_{i+1}]$, we have $f(x_{\max}) < f(x_i)$. This contradiction discredits our assumption that $f(x)$ does not attain its supremum. A similar argument applies to the infimum. ■

Example 3.2.2 *A Gauge Proof of the Heine-Borel Theorem*

One can use Cousin's lemma to prove the Heine-Borel Theorem on the real line [278]. This theorem states that if C is a compact set contained in the union $\cup_\alpha O_\alpha$ of a collection of open sets O_α, then C is actually contained in the union of a finite number of the O_α. Suppose $C \subset [a, b]$. Define a gauge $\delta(t)$ so that the interval $(t - \delta(t), t + \delta(t))$ does not intersect C when $t \notin C$ and $(t - \delta(t), t + \delta(t))$ is contained in some O_α when $t \in C$. Based on $\delta(t)$, select a δ-fine partition $a = s_0 < s_1 < \cdots < s_n = b$ with tags $t_i \in [s_i, s_{i+1}]$. By definition C is contained in the union $\cup_{t_i \in C} U_i$, where U_i is the set O_α covering t_i. The Heine-Borel theorem extends to compact sets in R^n.

Example 3.2.3 *A Gauge Proof of the Intermediate Value Theorem*

Under the assumption of the previous example, let c be an number strictly between $f(a)$ and $f(b)$. If we assume that there is no $t \in [a, b]$ with $f(t) = c$, then there exists a positive number $\delta(t)$ such that either $f(x) < c$ for all

$x \in [a,b]$ with $|x-t| < \delta(t)$ or $f(x) > c$ for all $x \in [a,b]$ with $|x-t| < \delta(t)$. We now select a δ-fine partition $a = s_0 < s_1 < \cdots < s_n = b$ and observe that throughout each interval $[s_i, s_{i+1}]$ either $f(t) < c$ or $f(t) > c$. If to start $f(s_0) = f(a) < c$, then $f(s_1) < c$, which implies $f(s_2) < c$ and so forth until we get to $f(s_n) = f(b) < c$. This contradicts the assumption that c lies strictly between $f(a)$ and $f(b)$. With minor differences, the same proof works when $f(a) > c$. ■

In preparation for our next example and for the fundamental theorem of calculus later in this chapter, we must define derivatives. A real-valued function $f(t)$ defined on an interval $[a,b]$ possesses a derivative $f'(c)$ at $c \in [a,b]$ provided the limit

$$\lim_{t \to c} \frac{f(t) - f(c)}{t - c} = f'(c) \tag{3.1}$$

exists. At the endpoints a and b, the limit is necessarily one sided. Taking a sequential view of convergence, definition (3.1) means that for every sequence t_m converging to c we must have

$$\lim_{m \to \infty} \frac{f(t_m) - f(c)}{t_m - c} = f'(c).$$

In calculus, we learn the following rules for computing derivatives:

Proposition 3.2.2 *If $f(t)$ and $g(t)$ are differentiable functions on (a,b), then*

$$\begin{aligned} \Big[\alpha f(t) + \beta g(t)\Big]' &= \alpha f'(t) + \beta g'(t) \\ \Big[f(t)g(t)\Big]' &= f'(t)g(t) + f(t)g'(t) \\ \Big[\frac{1}{f(t)}\Big]' &= -\frac{f'(t)}{f(t)^2}. \end{aligned}$$

In the third formula we must assume $f(t) \neq 0$. Finally, if $g(t)$ maps into the domain of $f(t)$, then the functional composition $f \circ g(t)$ has derivative

$$[f \circ g(t)]' = f' \circ g(t) g'(t).$$

Proof: We will prove the above sum, product, quotient, and chain rules in a broader context in Chap. 4. Our proofs will not rely on integration. ■

Example 3.2.4 *Strictly Increasing Functions*

Let $f(t)$ be a differentiable function on $[c,d]$ with strictly positive derivative. We now show that $f(t)$ is strictly increasing. For each $t \in [c,d]$ there exists $\delta(t) > 0$ such that

$$\frac{f(x) - f(t)}{x - t} > 0 \tag{3.2}$$

for all $x \in [a,b]$ with $|x-t| < \delta(t)$. According to Proposition 3.2.1, for any two points $a < b$ from $[c,d]$, there exists a δ-fine partition

$$a = s_0 < s_1 < \cdots < s_n = b$$

of $[a,b]$ with tags $t_i \in [s_i, s_{i+1}]$. In view of inequality (3.2), at least one of the two inequalities $f(s_i) \leq f(t_i) \leq f(s_{i+1})$ must be strict. Thus, the telescoping sum

$$f(b) - f(a) \quad = \quad \sum_{i=0}^{n-1} [f(s_{i+1}) - f(s_i)]$$

must be positive. ■

3.3 Definition and Basic Properties of the Integral

With later applications in mind, it will be convenient to define the gauge integral for vector-valued functions $f(x) : [a,b] \mapsto \mathsf{R}^n$. In this context, $f(x)$ is said to have integral $\boldsymbol{I}$ if for every $\epsilon > 0$ there exists a gauge $\delta(x)$ on $[a,b]$ such that

$$\|S(f,\pi) - \boldsymbol{I}\| \quad < \quad \epsilon \tag{3.3}$$

for all δ-fine partitions π. Our first order of business is to check that the integral is unique whenever it exists. Thus, suppose that the vector $\boldsymbol{J}$ is a second possible value of the integral. Given $\epsilon > 0$ choose gauges $\delta_{\boldsymbol{I}}(x)$ and $\delta_{\boldsymbol{J}}(x)$ leading to inequality (3.3). The minimum $\delta(x) = \min\{\delta_{\boldsymbol{I}}(x), \delta_{\boldsymbol{J}}(x)\}$ is also a gauge, and any partition π that is δ-fine is also $\delta_{\boldsymbol{I}}$ and $\delta_{\boldsymbol{J}}$-fine. Hence,

$$\|\boldsymbol{I} - \boldsymbol{J}\| \quad \leq \quad \|\boldsymbol{I} - S(f,\pi)\| + \|S(f,\pi) - \boldsymbol{J}\| \quad < \quad 2\epsilon.$$

Since ϵ is arbitrary, $\boldsymbol{J} = \boldsymbol{I}$.

One can also define $f(x)$ to be integrable if its Riemann sums are Cauchy in an appropriate sense.

Proposition 3.3.1 (Cauchy criterion) *A function $f(x) : [a,b] \mapsto \mathsf{R}^n$ is integrable if and only if for every $\epsilon > 0$ there exists a gauge $\delta(x) > 0$ such that*

$$\|S(f,\pi_1) - S(f,\pi_2)\| \quad < \quad \epsilon \tag{3.4}$$

for any two δ-fine partitions π_1 and π_2.

Proof: It is obvious that the Cauchy criterion is necessary for integrability. To show that it is sufficient, consider the sequence $\epsilon_m = m^{-1}$ and compatible sequence of gauges $\delta_m(x)$ determined by condition (3.4). We can force

the constraints $\delta_m(x) \le \delta_{m-1}(x)$ to hold by inductively replacing $\delta_m(x)$ by $\min\{\delta_{m-1}(x), \delta_m(x)\}$ whenever needed. Now select a δ_m-fine partition π_m for each m. Because the gauge sequence $\delta_m(x)$ is decreasing, every partition π that is δ_m-fine is also δ_{m-1}-fine. Hence, the sequence of Riemann sums $S(f, \pi_m)$ is Cauchy and has a limit $\boldsymbol{I}$ satisfying $\|S(f, \pi_m) - \boldsymbol{I}\| \le m^{-1}$. Finally, given the potential integral $\boldsymbol{I}$, we take an arbitrary $\epsilon > 0$ and choose m so that $m^{-1} < \epsilon$. If π is δ_m-fine, then the inequality

$$\|S(f, \pi) - \boldsymbol{I}\| \;\le\; \|S(f, \pi) - S(f, \pi_m)\| + \|S(f, \pi_m) - \boldsymbol{I}\| \;<\; 2\epsilon.$$

completes the proof. ■

For two integrable functions $f(x)$ and $g(x)$, the gauge integral inherits the linearity property

$$\int_a^b [\alpha f(x) + \beta g(x)]\, dx \;=\; \alpha \int_a^b f(x)\, dx + \beta \int_a^b g(x)\, dx$$

from its approximating Riemann sums. To prove this fact, take $\epsilon > 0$ and choose gauges $\delta_f(x)$ and $\delta_g(x)$ so that

$$\left\| S(f, \pi_f) - \int_a^b f(x)\, dx \right\| < \epsilon, \quad \left\| S(f, \pi_g) - \int_a^b g(x)\, dx \right\| < \epsilon$$

whenever π_f is δ_f-fine and π_g is δ_g-fine. If the tagged partition π is δ-fine for the gauge $\delta(x) = \min\{\delta_f(x), \delta_g(x)\}$, then

$$\begin{aligned} & \left\| S(\alpha f + \beta g, \pi) - \alpha \int_a^b f(x)\, dx - \beta \int_a^b g(x)\, dx \right\| \\ \le\; & |\alpha| \left\| S(f, \pi) - \int_a^b f(x)\, dx \right\| + |\beta| \left\| S(g, \pi) - \int_a^b g(x)\, dx \right\| \\ \le\; & (|\alpha| + |\beta|)\epsilon. \end{aligned}$$

The gauge integral also inherits obvious order properties. For example, $\int_a^b f(x)\, dx \ge 0$ whenever the integrand $f(x) \ge 0$ for all $x \in [a, b]$. In this case, the inequality $|S(f, \pi) - \int_a^b f(x)\, dx| < \epsilon$ implies

$$0 \;\le\; S(f, \pi) \;\le\; \int_a^b f(x)\, dx + \epsilon.$$

Since ϵ can be made arbitrarily small for $f(x)$ integrable, it follows that $\int_a^b f(x)\, dx \;\ge\; 0$. This nonnegativity property translates into the order property

$$\int_a^b f(x)\, dx \;\le\; \int_a^b g(x)\, dx$$

for two integrable functions $f(x) \leq g(x)$. In particular, when both $f(x)$ and $|f(x)|$ are both integrable, we have

$$\left| \int_a^b f(x)\, dx \right| \quad \leq \quad \int_a^b |f(x)|\, dx\,.$$

For vector-valued functions, the analogous rule

$$\left\| \int_a^b f(x)\, dx \right\| \quad \leq \quad \int_a^b \|f(x)\|\, dx \tag{3.5}$$

is also inherited from the approximating Riemann sums. The reader can easily supply the proof using the triangle inequality of the Euclidean norm. It does not take much imagination to extend the definition of the gauge integral to matrix-valued functions, and inequality (3.5) applies in this setting as well.

One of the nicest features of the gauge integral is that one can perturb an integrable function at a countable number of points without changing the value of its integral. This property fails for the Riemann integral but is exhibited by the Lebesgue integral. To validate the property, it suffices to prove that a function that equals $\mathbf{0}$ except at a countable number of points has integral $\mathbf{0}$. Suppose $f(x)$ is such a function with exceptional points $x_1, x_2, \ldots$ and corresponding exceptional values $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots$. We now define a gauge $\delta(x)$ with value 1 on the nonexceptional points and values

$$\delta(x_j) \quad = \quad \frac{\epsilon}{2^{j+2}[\|\boldsymbol{f}_j\| + 1]}$$

at the exceptional points. If π is a δ-fine partition, then x_j can serve as a tag for at most two intervals $[s_i, s_{i+1}]$ of π and each such interval has length less than $2\delta(x_j)$. It follows that

$$\|S(f,\pi)\| \quad \leq \quad 2\sum_j \|f(x_j)\| \frac{2\epsilon}{2^{j+2}[\|\boldsymbol{f}_j\| + 1]} \quad \leq \quad \epsilon \sum_{j=1}^{\infty} \frac{1}{2^j} \quad = \quad \epsilon$$

and therefore that $\int_a^b f(x)\, dx = \mathbf{0}$.

In practice, the interval additivity rule

$$\int_a^c f(x)\, dx \quad = \quad \int_a^b f(x)\, dx + \int_b^c f(x)\, dx \tag{3.6}$$

is obviously desirable. There are three separate issues in proving it. First, given the existence of the integral over $[a, c]$, do the integrals over $[a, b]$ and $[b, c]$ exist? Second, if the integrals over $[a, b]$ and $[b, c]$ exist, does the integral over $[a, c]$ exist? Third, if the integrals over $[a, b]$ and $[b, c]$ exist, are they additive? The first question is best approached through Proposition 3.3.1. For $\epsilon > 0$ there exists a gauge $\delta(x)$ such that

$$\|S(f,\pi_1) - S(f,\pi_2)\| \quad < \quad \epsilon$$

for any two δ-fine partitions π_1 and π_2 of $[a, c]$. Given $\delta(x)$, take any two δ-fine partitions γ_1 and γ_2 of $[a, b]$ and a single δ-fine partition ω of $[b, c]$. The concatenated partitions $\gamma_1 \cup \omega$ and $\gamma_2 \cup \omega$ are δ-fine throughout $[a, c]$ and satisfy

$$\|S(f, \gamma_1) - S(f, \gamma_2)\| \;=\; \|S(f, \gamma_1 \cup \omega) - S(f, \gamma_2 \cup \omega)\| \;<\; \epsilon.$$

According to the Cauchy criterion, the integral over $[a, b]$ therefore exists. A similar argument implies that the integral over $[b, c]$ also exists. Finally, the combination of these results shows that the integral exists over any interval $[u, v]$ contained within $[a, b]$.

For the converse, choose gauges $\delta_1(x)$ on $[a, b]$ and $\delta_2(x)$ on $[b, c]$ so that

$$\left\|S(f, \gamma) - \int_a^b f(x)\, dx\right\| \;<\; \epsilon, \quad \left\|S(f, \omega) - \int_b^c f(x)\, dx\right\| \;<\; \epsilon$$

for any δ_1-fine partition γ of $[a, b]$ and any δ_2-fine partition ω of $[b, c]$. The concatenated partition $\pi = \gamma \cup \omega$ satisfies

$$\begin{aligned} & \left\|S(f, \pi) - \int_a^b f(x)\, dx - \int_b^c f(x)\, dx\right\| \\ \leq\; & \left\|S(f, \gamma) - \int_a^b f(x)\right\| + \left\|S(f, \omega) - \int_b^c f(x)\, dx\right\| \qquad (3.7) \\ <\; & 2\epsilon \end{aligned}$$

because the Riemann sums satisfy $S(f, \pi) = S(f, \gamma) + S(f, \omega)$. This suggests defining a gauge $\delta(x)$ equal to $\delta_1(x)$ on $[a, b]$ and equal to $\delta_2(x)$ on $[b, c]$. The problem with this tactic is that some partitions of $[a, c]$ do not split at b. However, we can ensure a split by redefining $\delta(x)$ by

$$\tilde{\delta}(x) \;=\; \begin{cases} \min\{\delta_1(b), \delta_2(b)\} & x = b \\ \min\{\delta(x), \frac{1}{2}|x - b|\} & x \neq b\,. \end{cases}$$

This forces b to be the tag of its assigned interval, and we can if needed split this interval at b and retain b as tag of both subintervals. With $\delta(x)$ amended in this fashion, any δ-fine partition π can be viewed as a concatenated partition $\gamma \cup \omega$ splitting at b. As such π obeys inequality (3.7). This argument simultaneously proves that the integral over $[a, c]$ exists and satisfies the additivity property (3.6)

If the function $f(x)$ is vector-valued with n components, then the integrability of $f(x)$ should imply the integrability of each its components $f_i(x)$. Furthermore, we should be able to write

$$\int_a^b f(x)\, dx \;=\; \begin{pmatrix} \int_a^b f_1(x)\, dx \\ \vdots \\ \int_a^b f_n(x)\, dx \end{pmatrix}.$$

Conversely, if its components are integrable, then $f(x)$ should be integrable as well. The inequalities

$$\|S(f,\pi) - \boldsymbol{I}\| \;\leq\; \sum_{i=1}^{n} |S(f_i,\pi) - I_i| \;\leq\; \sqrt{n}\|S(f,\pi) - \boldsymbol{I}\|.$$

based on Example 2.5.6 and Problem 3 of Chap. 2 are instrumental in proving this logical equivalence. Given that we can integrate component by component, for the remainder of this chapter we will deal exclusively with real-valued functions.

We have not actually shown that any function is integrable. The most obvious possibility is a constant. Fortunately, it is trivial to demonstrate that

$$\int_a^b c\,dx \;=\; c(b-a).$$

Step functions are one rung up the hierarchy of functions. If

$$f(x) \;=\; \sum_{i=0}^{n-1} c_i 1_{(s_i,s_{i+1}]}(x)$$

for $a = s_0 < s_1 < \cdots < s_n = b$, then our nascent theory allows us to evaluate

$$\int_a^b f(x)\,dx \;=\; \sum_{i=0}^{n-1} \int_{s_i}^{s_{i+1}} c_i\,dx \;=\; \sum_{i=0}^{n-1} c_i(s_{i+1} - s_i).$$

This fact and the next technical proposition turn out to be the key to showing that continuous functions are integrable.

Proposition 3.3.2 *Let $f(x)$ be a function with domain $[a,b]$. Suppose for every $\epsilon > 0$ there exist two integrable functions $g(x)$ and $h(x)$ satisfying $g(x) \leq f(x) \leq h(x)$ for all x and*

$$\int_a^b h(x)\,dx \;\leq\; \int_a^b g(x)\,dx + \epsilon.$$

Then $f(x)$ is integrable.

Proof: For $\epsilon > 0$, choose gauges $\delta_g(x)$ and $\delta_h(x)$ on $[a,b]$ so that

$$\left|S(g,\pi_g) - \int_a^b g(x)\,dx\right| \;<\; \epsilon, \quad \left|S(h,\pi_h) - \int_a^b h(x)\,dx\right| \;<\; \epsilon$$

for any δ_g-fine partition π_g and any δ_h-fine partition π_h. If π is a δ-fine partition for $\delta(x) = \min\{\delta_g(x), \delta_h(x)\}$, then the inequalities

$$\begin{aligned}\int_a^b g(x)\,dx - \epsilon &< S(g,\pi)\\ &\le S(f,\pi)\\ &\le S(h,\pi)\\ &< \int_a^b h(x) + \epsilon\\ &\le \int_a^b g(x)\,dx + 2\epsilon\end{aligned}$$

trap $S(f,\pi)$ in an interval of length 3ϵ. Because the Riemann sum $S(f,\gamma)$ for any other δ-fine partition γ is trapped in the same interval, the integral of $f(x)$ exists by the Cauchy criterion. ■

Proposition 3.3.3 *Every continuous function $f(x)$ on $[a,b]$ is integrable.*

Proof: In view of the uniform continuity of $f(x)$ on $[a,b]$, for every $\epsilon > 0$ there exists a $\delta > 0$ with $|f(x) - f(y)| < \epsilon$ when $|x - y| < \delta$. For the constant gauge $\delta(x) = \delta$ and a corresponding δ-fine partition π with mesh points $s_0, \ldots, s_n$, let m_i be the minimum and M_i be the maximum of $f(x)$ on $[s_i, s_{i+1}]$. The step functions

$$g(x) = \sum_{i=1}^n m_i 1_{(s_i,s_{i+1}]}(x), \quad h(x) = \sum_{i=1}^n M_i 1_{(s_i,s_{i+1}]}(x)$$

then satisfy $g(x) \le f(x) \le h(x)$ except at the single point a. Furthermore,

$$\begin{aligned}\int_a^b h(x)\,dx - \int_a^b g(x)\,dx &\le \epsilon \sum_{i=1}^n (s_{i+1} - s_i)\\ &= \epsilon(b-a).\end{aligned}$$

Application of Proposition 3.3.2 now completes the proof. ■

3.4 The Fundamental Theorem of Calculus

The fundamental theorem of calculus divides naturally into two parts. For the gauge integral, the first and more difficult part is easily proved by invoking what is called the straddle inequality. Let $f(x)$ be differentiable at the point $t \in [a,b]$. Then there exists $\delta(t) > 0$ such that

$$\left|\frac{f(x) - f(t)}{x - t} - f'(t)\right| < \epsilon$$

for all $x \in [a,b]$ with $|x-t| < \delta(t)$. If $u < t < v$ are two points straddling t and located in $[a,b] \cap (t-\delta(t), t+\delta(t))$, then

$$\begin{aligned} |f(v) - f(u) - f'(t)(v-u)| &\leq |f(v) - f(t) - f'(t)(v-t)| \\ &\quad + |f(t) - f(u) - f'(t)(t-u)| \\ &\leq \epsilon(v-t) + \epsilon(t-u) \qquad (3.8) \\ &= \epsilon(v-u). \end{aligned}$$

Inequality (3.8) also clearly holds when either $u = t$ or $v = t$.

Proposition 3.4.1 (Fundamental Theorem I) *If $f(x)$ is differentiable throughout $[a,b]$, then*

$$\int_a^b f'(x)\,dx = f(b) - f(a).$$

Proof: Using the gauge $\delta(t)$ figuring in the straddle inequality (3.8), select a δ-fine partition π with mesh points $a = s_0 < s_1 < \cdots < s_n = b$ and tags $t_i \in [s_i, s_{i+1}]$. Application of the inequality and telescoping yield

$$\begin{aligned} |f(b) - f(a) - S(f', \pi)| &= \left| \sum_{i=0}^{n-1} [f(s_{i+1}) - f(s_i) - f'(t_i)(s_{i+1} - s_i)] \right| \\ &\leq \sum_{i=0}^{n-1} |f(s_{i+1}) - f(s_i) - f'(t_i)(s_{i+1} - s_i)| \\ &\leq \sum_{i=0}^{n-1} \epsilon(s_{i+1} - s_i) \\ &= \epsilon(b-a). \end{aligned}$$

This demonstrates that $f'(x)$ has integral $f(b) - f(a)$. ■

The first half of the fundamental theorem remains valid for a continuous function $f(x)$ that is differentiable except on a countable set N [250]. Since changing an integrand at a countable number of points does not alter its integral, it suffices to prove that

$$f(b) - f(a) = \int_a^b g(t)\,dt, \quad \text{where} \quad g(t) = \begin{cases} 0 & t \in N \\ f'(t) & t \notin N. \end{cases}$$

Suppose $\epsilon > 0$ is given. For $t \notin N$ define the gauge value $\delta(t)$ to satisfy the straddle inequality. Enumerate the points t_j of N, and define $\delta(t_j) > 0$ so that $|f(t_j) - f(t_j + s)| < 2^{-j-2}\epsilon$ whenever $|s| < \delta(t_j)$. Now select a δ-fine partition π with mesh points $a = s_0 < s_1 < \cdots < s_n = b$ and tags $r_i \in [s_i, s_{i+1}]$. Break the sum

$$f(b) - f(a) - S(g, \pi) = \sum_{i=0}^{n-1} \Big[f(s_{i+1}) - f(s_i) - g(r_i)(s_{i+1} - s_i) \Big]$$

into two parts. Let S' denote the sum of the terms with tags $r_i \notin N$, and let S'' denote the sum of the terms with tags $r_i \in N$. As noted earlier, $|S'| \leq \epsilon(b-a)$. Because a tag is attached to at most two subintervals, the second sum satisfies

$$\begin{aligned} |S''| &\leq \sum_{r_i \in N} |f(s_{i+1}) - f(s_i)| \\ &\leq \sum_{r_i \in N} \Big[|f(s_{i+1}) - f(r_i)| + |f(r_i) - f(s_i)| \Big] \\ &\leq 2 \sum_{j=1}^{\infty} 22^{-j-2} \epsilon \quad = \quad \epsilon. \end{aligned}$$

It follows that $|S' + S''| \leq \epsilon(b-a+1)$ and therefore that the stated integral exists and equals $f(b) - f(a)$.

In demonstrating the second half of the fundamental theorem, we will implicitly use the standard convention

$$\int_d^c f(x)\, dx \quad = \quad -\int_c^d f(x)\, dx$$

for $c < d$. This convention will also be in force in proving the substitution formula.

Proposition 3.4.2 (Fundamental Theorem II) *If a function $f(x)$ is integrable on $[a, b]$, then its indefinite integral*

$$F(t) \quad = \quad \int_a^t f(x)\, dx$$

has derivative $F'(t) = f(t)$ at any point t where $f(x)$ is continuous. The derivative is taken as one sided if $t = a$ or $t = b$.

Proof: In deriving the interval additivity rule (3.6), we showed that the integral $F(t)$ exists. At a point t where $f(x)$ is continuous, for any $\epsilon > 0$ there is a $\delta > 0$ such that $-\epsilon < f(x) - f(t) < \epsilon$ when $|x - t| < \delta$ and $x \in [a, b]$. Hence, the difference

$$\frac{F(t+s) - F(t)}{s} - f(t) \quad = \quad \frac{1}{s} \int_t^{t+s} [f(x) - f(t)]\, dx$$

is less than ϵ and greater than $-\epsilon$ for $|s| < \delta$. In the limit as s tends to 0, we recover $F'(t) = f(t)$. ■

The fundamental theorem of calculus has several important corollaries. These are covered in the next three propositions on the substitution rule, integration by parts, and finite Taylor expansions.

Proposition 3.4.3 (Substitution Rule) *Suppose $f(x)$ is differentiable on $[a,b]$, $g(x)$ is differentiable on $[c,d]$, and the image of $[c,d]$ under $g(x)$ is contained within $[a,b]$. Then*

$$\int_{g(c)}^{g(d)} f'(y)\,dy \quad = \quad \int_c^d f'[g(x)]g'(x)\,dx.$$

Proof: Part I of the fundamental theorem and the chain rule identity

$$\{f[g(x)]\}' \quad = \quad f'[g(x)]g'(x)$$

imply that both integrals have value $f[g(d)] - f[g(c)]$. ■

Proposition 3.4.4 (Integration by Parts) *Suppose $f(x)$ and $g(x)$ are differentiable on $[a,b]$. Then $f'(x)g(x)$ is integrable on $[a,b]$ if and only if $f(x)g'(x)$ is integrable on $[a,b]$. Furthermore, the two integrals are related by the identity*

$$\int_a^b f'(x)g(x)\,dx + \int_a^b f(x)g'(x)\,dx \quad = \quad f(b)g(b) - f(a)g(a),$$

Proof: The product rule for derivatives is

$$[f(x)g(x)]' \quad = \quad f'(x)g(x) + f(x)g'(x).$$

If two of three members of this identity are integrable, then the third is as well. Since part I of the fundamental theorem entails

$$\int_a^b [f(x)g(x)]'dx \quad = \quad f(b)g(b) - f(a)g(a),$$

the proposition follows. ■

The derivative of a function may itself be differentiable. Indeed, it makes sense to speak of the kth-order derivative of a function $f(x)$ if $f(x)$ is sufficiently smooth. Traditionally, the second-order derivative is denoted $f''(x)$ and an arbitrary kth-order derivative by $f^{(k)}(x)$. We can use these extra derivatives to good effect in approximating $f(x)$ locally. The next proposition makes this clear and offers an explicit estimate of the error in a finite Taylor expansion of $f(x)$.

Proposition 3.4.5 (Taylor Expansion) *Suppose $f(x)$ has a derivative of order $k+1$ on an open interval around the point y. Then for all x in the interval, we have*

$$f(x) \quad = \quad f(y) + \sum_{j=1}^{k} \frac{1}{j!} f^{(j)}(y)(x-y)^j + R_k(x), \tag{3.9}$$

where the remainder

$$R_k(x) \quad = \quad \frac{(x-y)^{k+1}}{k!} \int_0^1 f^{(k+1)}[y+t(x-y)](1-t)^k dt.$$

If $|f^{(k+1)}(z)| \leq b$ *for all* z *between* x *and* y*, then*

$$|R_k(x)| \quad \leq \quad \frac{b|x-y|^{k+1}}{(k+1)!}. \tag{3.10}$$

Proof: When $k = 0$, the Taylor expansion (3.9) reads

$$f(x) \quad = \quad f(y) + (x-y) \int_0^1 f'[y+t(x-y)]dt$$

and follows from the fundamental theorem of calculus and the chain rule. Induction and the integration-by-parts formula

$$\begin{aligned} & \int_0^1 f^{(k)}[y+t(x-y)](1-t)^{k-1} dt \\ = \quad & -\frac{1}{k} f^{(k)}[y+t(x-y)](1-t)^k \Big|_0^1 \\ & + \frac{x-y}{k} \int_0^1 f^{(k+1)}[y+t(x-y)](1-t)^k dt \\ = \quad & \frac{1}{k} f^{(k)}(y) + \frac{x-y}{k} \int_0^1 f^{(k+1)}[y+t(x-y)](1-t)^k dt \end{aligned}$$

now validate the general expansion (3.9). The error estimate follows directly from the bound $|f^{(k+1)}(z)| \leq b$ and the integral

$$\int_0^1 (1-t)^k dt \quad = \quad \frac{1}{k+1}.$$

3.5 More Advanced Topics in Integration

Within the confines of a single chapter, it is impossible to develop rigorously all of the properties of the gauge integral. In this section we will discuss briefly four topics: (a) integrals over unbounded intervals, (b) improper integrals and Hake's theorem, (c) the interchange of limits and integrals, and (d) multidimensional integrals and Fubini's theorem.

Defining the integral of a function over an unbounded interval requires several minor adjustments. First, the real line is extended to include the points $\pm\infty$. Second, a gauge function $\delta(x)$ is now viewed as mapping x to an open interval containing x. The associated interval may be infinite; indeed, it must be infinite if x equals $\pm\infty$. In a δ-fine partition π, the

interval I_j containing the tag x_j is contained in $\delta(x_j)$. The length of an infinite interval I_j is defined to be 0 in an approximating Riemann sum $S(f, \pi)$ to avoid infinite contributions to the sum. Likewise, the integrand $f(x)$ is assigned the value 0 at $x = \pm\infty$.

This extended definition carries with it all the properties we expect. Its most remarkable consequence is that it obliterates the distinction between proper and improper integrals. Hake's theorem provides the link. If we allow a and b to be infinite as well as finite, then Hake's theorem says a function $f(x)$ is integrable over (a, b) if and only if either of the two limits

$$\lim_{c\to a}\int_c^b f(x)\,dx \quad \text{or} \quad \lim_{c\to b}\int_a^c f(x)\,dx$$

exists. If either limit exists, then $\int_a^b f(x)\,dx$ equals that limit. For instance, the integral

$$\int_1^\infty \frac{1}{x^2}\,dx \;=\; \lim_{c\to\infty}\int_1^c \frac{1}{x^2}\,dx \;=\; \lim_{c\to\infty} -\frac{1}{x}\Big|_1^c \;=\; 1$$

exists and has the indicated limit by this reasoning.

Example 3.5.1 *Existence of* $\int_0^\infty \operatorname{sinc}(x)\,dx$

Consider the integral of $\operatorname{sinc}(x) = \sin(x)/x$ over the interval $(0, \infty)$. Because $\operatorname{sinc}(x)$ is continuous throughout $[0, 1]$ with limit 1 as x approaches 0, the integral over $[0, 1]$ is defined. Hake's theorem and integration by parts show that the integral

$$\begin{aligned}\int_1^\infty \frac{\sin x}{x}dx &= \lim_{c\to\infty}\int_1^c \frac{\sin x}{x}dx \\ &= \lim_{c\to\infty}\left(-\frac{\cos x}{x}\Big|_1^c - \int_1^c \frac{\cos x}{x^2}dx\right) \\ &= \cos 1 - \int_1^\infty \frac{\cos x}{x^2}dx\end{aligned}$$

exists provided the integral of $x^{-2}\cos x$ exists over $(1, \infty)$. We will demonstrate this fact in a moment. If we accept it, then it is clear that the integral of $\operatorname{sinc}(x)$ over $(0, \infty)$ exists as well. As we shall find in Example 3.5.4, this integral equals $\pi/2$. In contrast to these positive results, $\operatorname{sinc}(x)$ is not absolutely integrable over $(0, \infty)$. Finally, we note in passing that the substitution rule gives

$$\int_0^\infty \frac{\sin cx}{x}\,dx \;=\; \int_0^\infty \frac{\sin y}{c^{-1}y}c^{-1}\,dy \;=\; \int_0^\infty \frac{\sin y}{y}\,dy \;=\; \frac{\pi}{2}.$$

for any $c > 0$. ■

We now ask under what circumstances the formula

$$\lim_{n\to\infty}\int_a^b f_n(x)\,dx \quad = \quad \int_a^b \lim_{n\to\infty} f_n(x)\,dx \tag{3.11}$$

is valid. The two relevant theorems permitting the interchange of limits and integrals are the monotone convergence theorem and the dominated convergence theorem. In the monotone convergence theorem, we are given an increasing sequence $f_n(x)$ of integrable functions that converge to a finite limit for each x. Formula (3.11) is true in this setting provided

$$\sup_n \int_a^b f_n(x)\,dx \quad < \quad \infty.$$

In the dominated convergence theorem, we assume the sequence $f_n(x)$ is trapped between two integrable functions $g(x)$ and $h(x)$ in the sense that

$$g(x) \quad \leq \quad f_n(x) \quad \leq \quad h(x)$$

for all n and x. If $\lim_{n\to\infty} f_n(x)$ exists in this setting, then the interchange (3.11) is allowed. The choices

$$f_n(x) \quad = \quad 1_{[1,n]}(x)x^{-2}\cos x\,, \quad g(x) \quad = \quad -x^{-2}\,, \quad h(x) \quad = \quad x^{-2}$$

in the dominated convergence theorem validate the existence of

$$\int_1^\infty x^{-2}\cos x\,dx \quad = \quad \lim_{n\to\infty}\int_1^n x^{-2}\cos x\,dx.$$

We now consider two more substantive applications of the monotone and dominated convergence theorems.

Example 3.5.2 *Johann Bernoulli's Integral*

As example of delicate maneuvers in integration, consider the integral

$$\begin{aligned}\int_0^1 \frac{1}{x^x}dx \quad &= \quad \int_0^1 e^{-x\ln x}\,dx \\ &= \quad \int_0^1 \sum_{n=0}^\infty \frac{(-x\ln x)^n}{n!}\,dx \\ &= \quad \sum_{n=0}^\infty \frac{1}{n!}\int_0^1 (-x\ln x)^n dx\,.\end{aligned}$$

The reader will notice the application of the monotone convergence theorem in passing from the second to the third line above. Further progress can be made by applying the integration by parts result

$$\begin{aligned}\int_0^1 x^m \ln^n x\,dx \quad &= \quad -\frac{n}{m+1}\int_0^1 x^{m+1}\frac{\ln^{n-1} x}{x}dx \\ &= \quad -\frac{n}{m+1}\int_0^1 x^m \ln^{n-1} x\,dx\end{aligned}$$

recursively to evaluate

$$\int_0^1 (-x \ln x)^n dx = \frac{n!}{(n+1)^n} \int_0^1 x^n \, dx = \frac{n!}{(n+1)^{n+1}}.$$

The pleasant surprise

$$\int_0^1 \frac{1}{x^x} dx = \sum_{n=0}^{\infty} \frac{1}{(n+1)^{n+1}}$$

emerges. ■

Example 3.5.3 *Competing Definitions of the Gamma Function*

The dominated convergence theorem allows us to derive Gauss's representation

$$\Gamma(z) = \lim_{n \to \infty} \frac{n! n^z}{z(z+1) \cdots (z+n)}$$

of the gamma function from Euler's representation

$$\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} dx \, .$$

As students of statistics are apt to know from their exposure to the beta distribution, repeated integration by parts and the fundamental theorem of calculus show that

$$\int_0^1 x^{z-1} (1-x)^n \, dx = \frac{n!}{z(z+1) \cdots (z+n)}.$$

The substitution rule yields

$$n^z \int_0^1 x^{z-1} (1-x)^n \, dx = \int_0^n y^{z-1} \left(1 - \frac{y}{n}\right)^n dy \, .$$

Thus, it suffices to prove that

$$\int_0^{\infty} x^{z-1} e^{-x} dx = \lim_{n \to \infty} \int_0^n y^{z-1} \left(1 - \frac{y}{n}\right)^n dy \, .$$

Given the limit

$$\lim_{n \to \infty} \left(1 - \frac{y}{n}\right)^n = e^{-y},$$

we need an integrable function $h(y)$ that dominates the nonnegative sequence

$$f_n(y) = 1_{[0,n]}(y) y^{z-1} \left(1 - \frac{y}{n}\right)^n$$

from above in order to apply the dominated convergence theorem. In light of the inequality

$$\left(1 - \frac{y}{n}\right)^n \quad \leq \quad e^{-y},$$

the function $h(y) = y^{z-1}e^{-y}$ will serve. ■

Finally, the gauge integral extends to multiple dimensions, where a version of Fubini's theorem holds for evaluating multidimensional integrals via iterated integrals [278]. Consider a function $f(\boldsymbol{x}, \boldsymbol{y})$ defined over the Cartesian product $H \times K$ of two multidimensional intervals H and K. The intervals in question can be bounded or unbounded. If $f(\boldsymbol{x}, \boldsymbol{y})$ is integrable over $H \times K$, then Fubini's theorem asserts that the integrals $\int_H f(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{x}$ and $\int_K f(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{y}$ exist and can be integrated over the remaining variable to give the full integral. In symbols,

$$\int_{H\times K} f(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{x}\, d\boldsymbol{y} \;=\; \int_K \left[\int_J f(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{x}\right] d\boldsymbol{y} \;=\; \int_J \left[\int_K f(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{y}\right] d\boldsymbol{x}\,.$$

Conversely, if either iterated integral exists, one would like to conclude that the full integral exists as well. This is true whenever $f(\boldsymbol{x}, \boldsymbol{y})$ is nonnegative. Unfortunately, it is false in general, and two additional hypotheses introduced by Tonelli are needed to rescue the situation. One hypothesis is that $f(\boldsymbol{x}, \boldsymbol{y})$ is measurable. Measurability is a technical condition that holds except for very pathological functions. The other hypothesis is that $|f(\boldsymbol{x}, \boldsymbol{y})| \leq g(\boldsymbol{x}, \boldsymbol{y})$ for some nonnegative function $g(\boldsymbol{x}, \boldsymbol{y})$ for which the iterated integral exists. This domination condition is shared with the dominated convergence theorem and forces $f(\boldsymbol{x}, \boldsymbol{y})$ to be absolutely integrable.

Example 3.5.4 *Evaluation of* $\int_0^\infty \operatorname{sinc}(x)\, dx$

According to Fubini's theorem

$$\int_0^n \int_0^{n\pi} e^{-xy} \sin x\, dx\, dy \quad = \quad \int_0^{n\pi} \int_0^n e^{-xy} \sin x\, dy\, dx\,. \qquad (3.12)$$

The second of these iterated integrals

$$\int_0^{n\pi} \int_0^n e^{-xy} \sin x\, dy\, dx \quad = \quad \int_0^{n\pi} \frac{\sin x}{x} dx - \int_0^{n\pi} e^{-nx} \frac{\sin x}{x} dx$$

tends to $\int_0^\infty \operatorname{sinc}(x)\, dx$ as n tends to ∞ by a combination of Hake's theorem and the dominated convergence theorem. The inner integral of the left iterated integral in (3.12) equals

$$\begin{aligned}\int_0^{n\pi} e^{-xy} \sin x\, dx \quad &= \quad -e^{-xy} \cos x\Big|_0^{n\pi} - ye^{-xy} \sin x\Big|_0^{n\pi} \\ &\qquad - y^2 \int_0^{n\pi} e^{-xy} \sin x\, dx \\ &= \quad 1 - e^{-n\pi y} \cos n\pi - y^2 \int_0^{n\pi} e^{-xy} \sin x\, dx\end{aligned}$$

after two integrations by parts. It follows that

$$\int_0^{n\pi} e^{-xy} \sin x \, dx = \frac{1 - e^{-n\pi y} \cos n\pi}{1 + y^2}.$$

Finally, application of the dominated convergence theorem gives

$$\begin{aligned} \lim_{n\to\infty} \int_0^n \frac{1 - e^{-n\pi y} \cos n\pi}{1 + y^2} \, dy &= \int_0^\infty \frac{1}{1 + y^2} dy \\ &= \frac{\pi}{2}. \end{aligned}$$

Equating the limits of the right and left hand sides of the identity (3.12) therefore [278] yields the value of $\pi/2$ for $\int_0^\infty \operatorname{sinc}(x) \, dx$. ■

3.6 Problems

1. Give an alternative proof of Cousin's lemma by letting y be the supremum of the set of $x \in [a, b]$ such that $[a, x]$ possesses a δ-fine partition.

2. Use Cousin's lemma to prove that a continuous function $f(x)$ defined on an interval $[a, b]$ is uniformly continuous there [108]. (Hint: Given $\epsilon > 0$ define a gauge $\delta(x)$ by the requirement that $|f(y) - f(x)| < \frac{1}{2}\epsilon$ for all $y \in [a, b]$ with $|y - x| < 2\delta(x)$.)

3. A possibly discontinuous function $f(x)$ has one-sided limits at each point $x \in [a, b]$. Show by Cousin's lemma that $f(x)$ is bounded on $[a, b]$.

4. Suppose $f(x)$ has a nonnegative derivative $f'(x)$ throughout $[a, b]$. Prove that $f(x)$ is nondecreasing on $[a, b]$. Also prove that $f(x)$ is constant on $[a, b]$ if and only if $f'(x) = 0$ for all x. (Hint: These yield easily to the fundamental theorem of calculus. Alternatively for the first assertion, consider the function

$$f_\epsilon(x) = f(x) + \epsilon x$$

for $\epsilon > 0$.)

5. Using only the definition of the gauge integral, demonstrate that

$$\int_a^b f(t) \, dt = \int_{-b}^{-a} f(-t) \, dt$$

when either integral exists.

6. Based on the standard definition of the natural logarithm

$$\ln y \quad = \quad \int_1^y \frac{1}{x}\, dx,$$

prove that $\ln yz = \ln y + \ln z$ for all positive arguments y and z. Use this property to verify that $\ln y^{-1} = -\ln y$ and that $\ln y^r = r \ln y$ for every rational number r.

7. Apply Proposition 3.3.2 and demonstrate that every monotonic function defined on an interval $[a, b]$ is integrable on that interval.

8. Let $f(x)$ be a continuous real-valued function on $[a, b]$. Show that there exists $c \in [a, b]$ with

$$\int_a^b f(x)\, dx \quad = \quad f(c)(b - a).$$

9. In the Taylor expansion of Proposition 3.4.5, suppose $f^{(k+1)}(x)$ is continuous. Show that we can replace the remainder by

$$R_k(x) \quad = \quad \frac{(x - y)^{k+1}}{(k + 1)!} f^{(k+1)}(z)$$

for some z between x and y.

10. Suppose that $f(x)$ is infinitely differentiable and that c and r are positive numbers. If $|f^{(k)}(x)| \leq ck!r^k$ for all x near y and all nonnegative integers k, then use Proposition 3.4.5 to show that

$$f(x) \quad = \quad \sum_{k=0}^{\infty} \frac{f^{(k)}(y)}{k!}(x - y)^k$$

near y. Explicitly determine the infinite Taylor series expansion of the function $f(x) = (1 + x)^{-1}$ around $x = 0$ and justify its convergence.

11. Suppose the nonnegative continuous function $f(x)$ satisfies

$$\int_a^b f(x)\, dx = 0.$$

Prove that $f(x)$ is identically 0 on $[a, b]$.

12. Consider the function

$$f(x) = \begin{cases} x^2 \sin\left(x^{-2}\right) & x \neq 0 \\ 0 & x = 0\,. \end{cases}$$

Show that $\int_0^1 f'(x)\, dx = \sin(1)$ and $\lim_{t \downarrow 0} \int_t^1 |f'(x)|\, dx = \infty$. Hence, $f'(x)$ is integrable but not absolutely integrable on [0,1].

13. Prove that
$$\int_0^\infty x^\alpha e^{-x^\beta}\,dx \;=\; \frac{1}{\beta}\Gamma\left(\frac{\alpha+1}{\beta}\right)$$
for α and β positive [82].

14. Justify the formula
$$\int_0^1 \frac{\ln(1-x)}{x} \;=\; -\sum_{n=1}^\infty \frac{1}{n^2}.$$

15. Show that
$$\int_0^\infty \frac{x^{z-1}}{e^x-1}dx \;=\; \zeta(z)\Gamma(z),$$
where $\zeta(z) = \sum_{n=1}^\infty n^{-z}$.

16. Prove that the functions
$$\begin{aligned} f(x) &= \int_1^\infty \frac{\sin t}{x^2+t^2}\,dt \\ g(x) &= \int_0^\infty e^{-xt}\cos t\,dt, \quad x>0, \end{aligned}$$
are continuous.

17. Let $f_n(x)$ be a sequence of integrable functions on $[a,b]$ that converges uniformly to $f(x)$. Demonstrate that $f(x)$ is integrable and satisfies
$$\lim_{n\to\infty}\int_a^b f_n(x)\,dx \;=\; \int_a^b f(x)\,dx\,.$$
(Hints: For $\epsilon>0$ small take n large enough so that
$$f_n(x)-\frac{\epsilon}{2(b-a)} \;\le\; f(x) \;\le\; f_n(x)+\frac{\epsilon}{2(b-a)}$$
for all x.)

18. Let p and q be positive integers. Justify the series expansion
$$\int_0^1 \frac{x^{p-1}}{1+x^q}dx \;=\; \sum_{n=0}^\infty \frac{(-1)^n}{p+nq}$$
by the monotone convergence theorem. Be careful since the series does not converge absolutely [278].

19. Suppose $f(x)$ is a continuous function on R. Demonstrate that the sequence

$$f_n(x) \quad = \quad \frac{1}{n}\sum_{k=0}^{n-1} f\left(x+\frac{k}{n}\right)$$

converges uniformly to a continuous function on every finite interval $[a,b]$ [69].

20. Prove that

$$\int_0^1 \frac{x^b - x^a}{\ln x}dx \quad = \quad \ln\frac{b+1}{a+1}$$

for $0 < a < b$ [278] by showing that both sides equal the double integral

$$\int_{[0,1]\times[a,b]} x^y \, dx \, dy \, .$$

21. Integrate the function

$$f(x,y) \quad = \quad \frac{y^2 - x^2}{(x^2+y^2)^2}$$

over the unit square $[0,1]\times[0,1]$. Show that the two iterated integrals disagree, and explain why Fubini's theorem fails.

22. Suppose the two partial derivatives $\frac{\partial^2}{\partial x_1 \partial x_2} f(\boldsymbol{x})$ and $\frac{\partial^2}{\partial x_2 \partial x_1} f(\boldsymbol{x})$ exist and are continuous in a neighborhood of a point $\boldsymbol{y} \in \mathsf{R}^2$. Show that they are equal at the point. (Hints: If they are not equal, take a small box around the point where their difference has constant sign. Now apply Fubini's theorem.)

23. Demonstrate that

$$\int_0^\infty e^{-x^2} dx \quad = \quad \frac{\sqrt{\pi}}{2}$$

by evaluating the integral of $f(\boldsymbol{y}) = y_2 e^{-(1+y_1^2)y_2^2}$ over the rectangle $(0,\infty)\times(0,\infty)$.

4
Differentiation

4.1 Introduction

Differentiation and integration are the two pillars on which all of calculus rests. For real-valued functions of a real variable, all of the major issues surrounding differentiation were settled long ago. For multivariate differentiation, there are still some subtleties and snares. We adopt a definition of differentiability that avoids most of the pitfalls and makes differentiation of vectors and matrices relatively painless. In later chapters, this definition also improves the clarity of exposition.

The main theme of differentiation is the short-range approximation of curved functions by linear functions. A differential gives a recipe for carrying out such a linear approximation. Most linear approximations can be improved by adding more terms in a Taylor series expansion. Adding quadratic terms brings in second differentials. We will meet these in the next chapter after we have mastered first differentials. Our current treatment stresses theory and counterexamples rather than the nuts and bolts of differentiation.

4.2 Univariate Derivatives

In this section we explore univariate differentiation in more detail. The standard repertoire of differentiable functions includes the derivatives nx^{n-1}

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_4,

TABLE 4.1. Derivatives of some elementary functions

$f(x)$	$f'(x)$	$f(x)$	$f'(x)$	$f(x)$	$f'(x)$
x^n	nx^{n-1}	e^x	e^x	$\ln x$	$1/x$
$\sin x$	$\cos x$	$\cos x$	$-\sin x$	$\tan x$	$1+\tan^2 x$
$\sinh x$	$\cosh x$	$\cosh x$	$\sinh x$	$\tanh x$	$1-\tanh^2 x$
$\arcsin x$	$1/\sqrt{1-x^2}$	$\arccos x$	$-1/\sqrt{1-x^2}$	$\arctan x$	$1/(1+x^2)$
$\operatorname{arcsinh} x$	$1/\sqrt{x^2+1}$	$\operatorname{arccosh} x$	$1/\sqrt{x^2-1}$	$\operatorname{arctanh} x$	$1/(1-x^2)$

of the monomials x^n and, via the sum, product, and quotient rules, the derivatives of all polynomials and rational functions. These functions are supplemented by special functions such as $\ln x$, e^x, $\sin x$, and $\cos x$. Virtually all of the special functions can be defined by power series or as the solutions of differential equations. For instance, the system of differential equations

$$\begin{aligned}(\cos x)' &= -\sin x \\ (\sin x)' &= \cos x\end{aligned}$$

with the initial conditions $\cos 0 = 1$ and $\sin 0 = 0$ determines these trigonometric functions. We will take most of these facts for granted except to add in the case of $\cos x$ and $\sin x$ that the solution of the defining system of differential equations involves a particular matrix exponential. Table 4.1 lists the derivatives of the most important elementary functions.

It is worth emphasizing that differentiation is a purely local operation and that differentiability at a point implies continuity at the same point. The converse is clearly false. The functions

$$f_n(x) = \begin{cases} x^n & x \text{ rational} \\ 0 & x \text{ irrational} \end{cases}$$

illustrate the local character of continuity and differentiability. For $n > 0$ the functions $f_n(x)$ are continuous at the point 0 but discontinuous everywhere else. In contrast, $f_1'(0)$ fails to exist while $f_n'(0) = 0$ for all $n \geq 2$. In this instance, we must resort directly to the definition (3.1) to evaluate derivatives.

We have already mentioned Fermat's result that $f'(x)$ must vanish at any interior extreme point. For example, suppose that c is a local maximum of $f(x)$ on (a, b). If $f'(c) > 0$, then choose $\epsilon > 0$ such that $f'(c) - \epsilon > 0$. This choice then entails

$$\begin{aligned}f(x) &> f(c) + [f'(c) - \epsilon](x - c) \\ &> f(c)\end{aligned}$$

for all $x > c$ with $x - c$ sufficiently small, contradicting the assumption that c is a local maximum. If $f'(c) < 0$, we reach a similar contradiction using nearby points on the left of c.

Fermat's principle has some surprising implications. Among these is the mean value theorem.

Proposition 4.2.1 *Suppose $f(x)$ is continuous on $[a,b]$ and differentiable on (a,b). Then there exists a point $c \in (a,b)$ such that*

$$f(b) - f(a) \quad = \quad f'(c)(b-a).$$

Proof: Consider the function

$$g(x) \quad = \quad f(b) - f(x) + \frac{f(b)-f(a)}{b-a}(x-b).$$

Clearly, $g(x)$ is also continuous on $[a,b]$ and differentiable on (a,b). Furthermore, $g(a) = g(b) = 0$. It follows that $g(x)$ attains either a maximum or a minimum at some $c \in (a,b)$. At this point, $g'(c) = 0$, which is equivalent to the mean value property. ■

The mean value theorem has the following consequences:

(a) If $f'(x) \geq 0$ for all $x \in (a,b)$, then $f(x)$ is increasing.

(b) If $f'(x) = 0$ for all $x \in (a,b)$, then $f(x)$ is constant.

(c) If $f'(x) \leq 0$ for all $x \in (a,b)$, then $f(x)$ is decreasing.

For an alternative proof, one can build on Example 3.2.4 of Chap. 3. See Problem 4 of that chapter.

Example 4.2.1 *A Trigonometric Identity*

The function $f(x) = \cos^2 x + \sin^2 x$ has derivative

$$f'(x) \quad = \quad -2\cos x \sin x + 2 \sin x \cos x \quad = \quad 0.$$

Therefore, $f(x) = f(0) = 1$ for all x. ■

Here are two related matrix applications of univariate differentiation.

Example 4.2.2 *Differential Equations and the Matrix Exponential*

The derivative of a vector or matrix-valued function $f(x)$ with domain (a,b) is defined entry by entry. We have already met the matrix-valued differential equation $N'(t) = \boldsymbol{M}N(t)$ with initial condition $N(0) = \boldsymbol{I}$. To demonstrate that $N(t) = e^{t\boldsymbol{M}}$ is a solution, consider the difference quotient

$$\begin{aligned}
\frac{e^{(t+s)\boldsymbol{M}} - e^{t\boldsymbol{M}}}{s} &= \frac{1}{s}\sum_{j=1}^{\infty} \frac{(t+s)^j - t^j}{j!}\boldsymbol{M}^j \\
&= \boldsymbol{M}\sum_{j=1}^{\infty} \frac{t^{j-1}}{(j-1)!}\boldsymbol{M}^{j-1} + \sum_{j=1}^{\infty} \frac{(t+s)^j - t^j - jt^{j-1}s}{sj!}\boldsymbol{M}^j \\
&= \boldsymbol{M}e^{t\boldsymbol{M}} + \sum_{j=1}^{\infty} \frac{(t+s)^j - t^j - jt^{j-1}s}{sj!}\boldsymbol{M}^j.
\end{aligned}$$

We now apply the error estimate (3.10) of Chap. 3 for the first-order Taylor expansion of the function $f(t) = t^j$. If c bounds $|t+s|$ for s near 0, it follows that

$$\left|(t+s)^j - t^j - jt^{j-1}s\right| \leq j(j-1)c^{j-2}s^2/2$$

and that

$$\begin{aligned}\left\|\sum_{j=1}^{\infty}\frac{(t+s)^j - t^j - jt^{j-1}s}{sj!}\boldsymbol{M}^j\right\| &\leq \frac{|s|}{2}\sum_{j=2}^{\infty}\frac{j(j-1)c^{j-2}}{j!}\|\boldsymbol{M}\|^j \\ &= \frac{|s|}{2}\|\boldsymbol{M}\|^2 e^{c\|\boldsymbol{M}\|}.\end{aligned}$$

This is enough to show that

$$\lim_{s\to 0}\left\|\frac{e^{(t+s)\boldsymbol{M}} - e^{t\boldsymbol{M}}}{s} - \boldsymbol{M}e^{t\boldsymbol{M}}\right\| = 0.$$

One can demonstrate that $e^{t\boldsymbol{M}}$ is the unique solution of the differential equation $N'(t) = \boldsymbol{M}N(t)$ subject to $N(0) = \boldsymbol{I}$ by considering the matrix $P(t) = e^{-t\boldsymbol{M}}N(t)$ using any solution $N(t)$. Because the product rule of differentiation pertains to matrix multiplication as well as to ordinary multiplication,

$$P'(t) = -\boldsymbol{M}e^{-t\boldsymbol{M}}N(t) + e^{-t\boldsymbol{M}}\boldsymbol{M}N(t) = \boldsymbol{0}.$$

By virtue of part (b) of Proposition 4.2.1, $P(t)$ is the constant matrix $P(0) = \boldsymbol{I}$. If we take $N(t) = e^{t\boldsymbol{M}}$, then this argument demonstrates that $e^{-t\boldsymbol{M}}$ is the matrix inverse of $e^{t\boldsymbol{M}}$. If we take $N(t)$ to be an arbitrary solution of the differential equation, then multiplying both sides of $e^{-t\boldsymbol{M}}N(t) = \boldsymbol{I}$ on the left by $e^{t\boldsymbol{M}}$ implies that $N(t) = e^{t\boldsymbol{M}}$ as claimed. ■

Example 4.2.3 *Matrix Logarithm*

Let $\boldsymbol{M}$ be a square matrix with $\|\boldsymbol{M}\| < 1$. It is tempting to define the logarithm of $\boldsymbol{I} - \boldsymbol{M}$ by the series expansion

$$\ln(\boldsymbol{I} - \boldsymbol{M}) = -\sum_{k=1}^{\infty}\frac{\boldsymbol{M}^k}{k}$$

valid for scalars. This definition does not settle the question of whether

$$e^{\ln(\boldsymbol{I}-\boldsymbol{M})} = \boldsymbol{I} - \boldsymbol{M}. \tag{4.1}$$

The traditional approach to such issues relies on Jordan canonical forms [137]. Here we would like to sketch an analytic proof. Consider the matrix-valued functions

$$\begin{aligned}f(t) &= e^{\ln(\boldsymbol{I}-t\boldsymbol{M})} \\ f_n(t) &= e^{-\sum_{k=1}^{n}t^k\boldsymbol{M}^k/k} \\ &= e^{-t\boldsymbol{M}}e^{-t^2\boldsymbol{M}^2/2}\cdots e^{-t^n\boldsymbol{M}^n/n}\end{aligned}$$

of the scalar t. It is clear that $f_n(t)$ converges uniformly to $f(t)$ on every interval $[0, 1+\delta)$ for $\delta > 0$ small enough. Furthermore, the product rule, the chain rule, and the law of exponents show that

$$\begin{aligned} f_n'(t) &= -(\boldsymbol{M} + t\boldsymbol{M}^2 + \cdots + t^{n-1}\boldsymbol{M}^n) f_n(t) \\ &= -\boldsymbol{M}(\boldsymbol{I} - t^n\boldsymbol{M}^n)(\boldsymbol{I} - t\boldsymbol{M})^{-1} f_n(t). \end{aligned}$$

Because $\boldsymbol{I} - t^n\boldsymbol{M}^n$ tends to $\boldsymbol{I}$, it follows that

$$\lim_{n\to\infty} f_n'(t) = -\boldsymbol{M}(\boldsymbol{I} - t\boldsymbol{M})^{-1} f(t)$$

uniformly on $[0, 1+\delta)$. This in turn implies

$$\begin{aligned} f(t) - f(0) &= \lim_{n\to\infty} [f_n(t) - f_n(0)] \\ &= \lim_{n\to\infty} \int_0^t f_n'(s)\, ds \\ &= -\boldsymbol{M} \int_0^t (\boldsymbol{I} - s\boldsymbol{M})^{-1} f(s)\, ds \end{aligned}$$

by virtue of Problem 17 of Chap. 3. Differentiating this equation with respect to t produces the differential equation

$$f'(t) = -\boldsymbol{M}(\boldsymbol{I} - t\boldsymbol{M})^{-1} f(t) \tag{4.2}$$

with initial condition $f(0) = \boldsymbol{I}$. Clearly $f(t) = \boldsymbol{I} - t\boldsymbol{M}$ is one solution of the differential equation (4.2). In view of Problem 16, this solution is unique. Comparing the two formulas for $f(t)$ at the point $t = 1$ now gives the desired conclusion (4.1). ■

4.3 Partial Derivatives

There are several possible ways to extend differentiation to real-valued functions on R^n. The most familiar is the partial derivative

$$\partial_i f(\boldsymbol{x}) = \frac{\partial}{\partial x_i} f(\boldsymbol{x}) = \lim_{t\to 0} \frac{f(\boldsymbol{x} + t\boldsymbol{e}_i) - f(\boldsymbol{x})}{t},$$

where $\boldsymbol{e}_i$ is one of the standard unit vectors spanning R^n. There is nothing sacred about the coordinate directions. The directional derivative along the direction $\boldsymbol{v}$ is

$$d_{\boldsymbol{v}} f(\boldsymbol{x}) = \lim_{t\to 0} \frac{f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x})}{t}.$$

If we confine $t \geq 0$ in this limit, then we have a forward derivative along $\boldsymbol{v}$. Readers should be on the alert that we will use the same symbol $d_{\boldsymbol{v}} f(\boldsymbol{x})$ for both directional derivatives. The context will make it clear which concept is pertinent.

To illustrate these definitions, consider the function

$$f(\boldsymbol{x}) \quad = \quad \sqrt{|x_1 x_2|}$$

on R^2. It is clear that both partial derivatives are 0 at the origin. Along a direction $\boldsymbol{v} = (v_1, v_2)^*$ with neither $v_1 = 0$ nor $v_2 = 0$, consider the difference quotient

$$\frac{f(\mathbf{0} + t\boldsymbol{v}) - f(\mathbf{0})}{t} \quad = \quad \frac{|t|\sqrt{|v_1 v_2|}}{t}.$$

This has limit $\sqrt{|v_1 v_2|}$ as long as we restrict $t > 0$. Thus, the forward directional derivative exists, but the full directional derivative does not.

For another example, let

$$f(\boldsymbol{x}) \quad = \quad \begin{cases} x_1 + x_2 & \text{if } x_1 = 0 \text{ or } x_2 = 0 \\ 1 & \text{otherwise}. \end{cases}$$

This function is clearly discontinuous at the origin of R^2, but the partial derivatives $\partial_1 f(\mathbf{0}) = 1$ and $\partial_2 f(\mathbf{0}) = 1$ are well defined there. These and similar anomalies suggest the need for a carefully structured theory of differentiability. Such a theory is presented in the next section.

Second and higher-order partial derivatives are defined in the obvious way. For typographical convenience, we will occasionally employ such abbreviations as

$$\partial^2_{ij} f(\boldsymbol{x}) \quad = \quad \frac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x}).$$

Readers will doubtless recall from calculus the equality of mixed second partial derivatives. This property can fail. For example, suppose we define $f(\boldsymbol{x}) = g(x_1)$, where $g(x_1)$ is nowhere differentiable. Then $\partial_2 f(\boldsymbol{x})$ and $\partial^2_{12} f(\boldsymbol{x})$ are identically 0 while $\partial_1 f(\boldsymbol{x})$ does not even exist. The key to restoring harmony is to impose continuity in a neighborhood of the current point.

Proposition 4.3.1 *Suppose the real-valued function $f(\boldsymbol{y})$ on R^2 has partial derivatives $\partial_1 f(\boldsymbol{y})$, $\partial_2 f(\boldsymbol{y})$, and $\partial^2_{12} f(\boldsymbol{y})$ on some open set. If $\partial^2_{12} f(\boldsymbol{y})$ is continuous at a point $\boldsymbol{x}$ in the set, then $\partial^2_{21} f(\boldsymbol{x})$ exists and*

$$\frac{\partial^2}{\partial x_2 \partial x_1} f(\boldsymbol{x}) \quad = \quad \partial^2_{21} f(\boldsymbol{x}) \quad = \quad \partial^2_{12} f(\boldsymbol{x}) \quad = \quad \frac{\partial^2}{\partial x_1 \partial x_2} f(\boldsymbol{x}). \tag{4.3}$$

This result extends in the obvious way to the equality of second mixed partials for functions defined on open subsets of R^n for $n > 2$.

Proof: Consider the first difference by u_1

$$g(x_2) \;=\; \Delta_1 f(x_1, x_2) \;=\; f(x_1 + u_1, x_2) - f(x_1, x_2)$$

and the second difference by u_2

$$\begin{aligned} \Delta_{21} f(x_1, x_2) &= g(x_2 + u_2) - g(x_2) \\ &= f(x_1 + u_1, x_2 + u_2) - f(x_1, x_2 + u_2) \\ &\quad - f(x_1 + u_1, x_2) + f(x_1, x_2). \end{aligned}$$

Applying the mean value theorem twice gives

$$\begin{aligned} \Delta_{21} f(x_1, x_2) &= u_2 g'(x_2 + \theta_2 u_2) \\ &= u_2[\partial_2 f(x_1 + u_1, x_2 + \theta_2 u_2) - \partial_2 f(x_1, x_2 + \theta_2 u_2)] \\ &= u_1 u_2 \partial_{12}^2 f(x_1 + \theta_1 u_1, x_2 + \theta_2 u_2) \end{aligned}$$

for θ_1 and θ_2 in $(0, 1)$. In view of the continuity of $\partial_{12}^2 f(\boldsymbol{y})$ at $\boldsymbol{x}$, it follows that

$$\lim_{\|\boldsymbol{u}\| \to 0} \frac{\Delta_{21} f(x_1, x_2)}{u_1 u_2} \;=\; \partial_{12}^2 f(x_1, x_2)$$

regardless of how the limit is approached. This proves the existence of the iterated limit

$$\begin{aligned} \lim_{u_2 \to 0} \lim_{u_1 \to 0} \frac{\Delta_{21} f(x_1, x_2)}{u_1 u_2} &= \lim_{u_2 \to 0} \frac{\partial_1 f(x_1, x_2 + u_2) - \partial_1 f(x_1, x_2)}{u_2} \\ &= \partial_{21}^2 f(x_1, x_2) \end{aligned}$$

and simultaneously the equality of mixed partials. Problem 22 of Chap. 3 offers an alternative proof under stronger hypotheses. ■

4.4 Differentials

The question of when a real-valued function is differentiable is perplexing because of the variety of possible definitions. In choosing an appropriate definition, we are governed by several considerations. First, it should be consistent with the classical definition of differentiability on the real line. Second, continuity at a point should be a consequence of differentiability at the point. Third, all directional derivatives should exist. Fourth, the differential should vanish wherever the function attains a local maximum or minimum on the interior of its domain. Fifth, the standard rules for combining differentiable functions should apply. Sixth, the logical proofs of

the rules should be as transparent as possible. Seventh, the extension to vector-valued functions should be painless. Eighth and finally, our geometric intuition should be enhanced.

We now present a definition conceived by Constantin Carathéodory [40] and expanded by recent authors [1, 18, 29, 160] that fulfills these conditions. A real-valued function $f(\boldsymbol{y})$ on an open set $S \subset \mathsf{R}^m$ is said to be differentiable at $\boldsymbol{x} \in S$ if a function $s(\boldsymbol{y}, \boldsymbol{x})$ exists for $\boldsymbol{y}$ near $\boldsymbol{x}$ satisfying

$$\begin{aligned} f(\boldsymbol{y}) - f(\boldsymbol{x}) &= s(\boldsymbol{y}, \boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \\ \lim_{\boldsymbol{y} \to \boldsymbol{x}} s(\boldsymbol{y}, \boldsymbol{x}) &= s(\boldsymbol{x}, \boldsymbol{x}). \end{aligned} \tag{4.4}$$

The row vector $s(\boldsymbol{y}, \boldsymbol{x})$ is called a slope function. We will see in a moment that its limit $s(\boldsymbol{x}, \boldsymbol{x})$ defines the differential $df(\boldsymbol{x})$ of $f(\boldsymbol{y})$ at $\boldsymbol{x}$.

The standard definition of differentiability due to Fréchet reads

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = df(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) + o(\|\boldsymbol{y} - \boldsymbol{x}\|)$$

for $\boldsymbol{y}$ near the point $\boldsymbol{x}$. The row vector $df(\boldsymbol{x})$ appearing here is again termed the differential of $f(\boldsymbol{y})$ at $\boldsymbol{x}$. Fréchet's definition is less convenient than Carathéodory's because the former invokes approximate equality rather than true equality. Observe that the error $[s(\boldsymbol{y}, \boldsymbol{x}) - df(\boldsymbol{x})](\boldsymbol{y} - \boldsymbol{x})$ under Carathéodory's definition satisfies

$$|[s(\boldsymbol{y}, \boldsymbol{x}) - df(\boldsymbol{x})](\boldsymbol{y} - \boldsymbol{x})| \leq \|s(\boldsymbol{y}, \boldsymbol{x}) - df(\boldsymbol{x})\| \cdot \|\boldsymbol{y} - \boldsymbol{x}\|,$$

which is $o(\|\boldsymbol{y} - \boldsymbol{x}\|)$ as $\boldsymbol{y}$ tends to $\boldsymbol{x}$ in view of the continuity of $s(\boldsymbol{y}, \boldsymbol{x})$. Thus, Carathéodory's definition implies Fréchet's definition. The converse is trivial when the argument of $f(x)$ is a scalar, for then the difference quotient

$$s(y, x) = \frac{f(y) - f(x)}{y - x} \tag{4.5}$$

serves as a slope function. Proposition 4.4.1 addresses the general case.

Carathéodory's definition (4.4) has some immediate consequences. For example, it obviously compels $f(\boldsymbol{y})$ to be continuous at $\boldsymbol{x}$. Furthermore, if we send t to 0 in the equation

$$\frac{f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x})}{t} = s(\boldsymbol{x} + t\boldsymbol{v}, \boldsymbol{x})\boldsymbol{v}, \tag{4.6}$$

then it is clear that the directional derivative $d_{\boldsymbol{v}} f(\boldsymbol{x})$ exists and equals $s(\boldsymbol{x}, \boldsymbol{x})\boldsymbol{v}$. The special case $\boldsymbol{v} = \boldsymbol{e}_i$ shows that the ith component of $s(\boldsymbol{x}, \boldsymbol{x})$ reduces to the partial derivative $\partial_i f(\boldsymbol{x})$. Since $s(\boldsymbol{x}, \boldsymbol{x})$ and $df(\boldsymbol{x})$ agree component by component, they are equal, and, in general, we have the formula $d_{\boldsymbol{v}} f(\boldsymbol{x}) = df(\boldsymbol{x})\boldsymbol{v}$ for the directional derivative.

Fermat's principle that the differential of a function vanishes at an interior maximum or minimum point is also trivial to check in this context. Suppose $\boldsymbol{x}$ affords a local minimum of $f(\boldsymbol{y})$. Then

$$f(\boldsymbol{x}+t\boldsymbol{v}) - f(\boldsymbol{x}) \quad = \quad ts(\boldsymbol{x}+t\boldsymbol{v},\boldsymbol{x})\boldsymbol{v} \quad \geq \quad 0$$

for all $\boldsymbol{v}$ and small $t > 0$. Taking limits in the identity (4.6) now yields the conclusion $df(\boldsymbol{x})\boldsymbol{v} \geq 0$. The only way this can hold for all $\boldsymbol{v}$ is for $df(\boldsymbol{x}) = \boldsymbol{0}$. If $\boldsymbol{x}$ occurs on the boundary of the domain of $f(\boldsymbol{y})$, then we can still glean useful information. For example, if $f(y)$ is differentiable on the closed interval $[c, d]$ and c provides a local minimum, then the condition $f'(c) \geq 0$ must hold.

The extension of the definition of differentiability to vector-valued functions is equally simple. Suppose $f(\boldsymbol{y})$ maps an open subset $S \subset \mathsf{R}^m$ into R^n. Then $f(\boldsymbol{y})$ is said to be differentiable at $\boldsymbol{x} \in S$ if there exists an $n \times m$ matrix-valued function $s(\boldsymbol{y}, \boldsymbol{x})$ continuous at $\boldsymbol{x}$ and satisfying equation (4.4) for $\boldsymbol{y}$ near $\boldsymbol{x}$. The limit $\lim_{\boldsymbol{y}\to\boldsymbol{x}} s(\boldsymbol{y}, \boldsymbol{x}) = df(\boldsymbol{x})$ is again called the differential of $f(\boldsymbol{y})$ at $\boldsymbol{x}$. The rows of the differential are the differentials of the component functions of $f(\boldsymbol{x})$. Thus, $f(\boldsymbol{y})$ is differentiable at $\boldsymbol{x}$ if and only if each of its components is differentiable at $\boldsymbol{x}$. This characterization is also valid under Fréchet's definition of the differential and leads to a simple proof of the second half of the next proposition.

Proposition 4.4.1 *Carathéodory's definition and Fréchet's definition of the differential are logically equivalent.*

Proof: We have already proved that Carathéodory's definition implies Fréchet's definition. The converse is valid because it is valid for scalar-valued functions. For a matrix-oriented proof of the converse, suppose that $f(\boldsymbol{y})$ is Fréchet differentiable at $\boldsymbol{x}$. If we define the slope function

$$s(\boldsymbol{y}, \boldsymbol{x}) \quad = \quad \frac{1}{\|\boldsymbol{y}-\boldsymbol{x}\|^2}[f(\boldsymbol{y}) - f(\boldsymbol{x}) - df(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})](\boldsymbol{y}-\boldsymbol{x})^* + df(\boldsymbol{x})$$

for $\boldsymbol{y} \neq \boldsymbol{x}$, then the identity $f(\boldsymbol{y}) - f(\boldsymbol{x}) = s(\boldsymbol{y}, \boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})$ certainly holds. To show that $s(\boldsymbol{y}, \boldsymbol{x})$ tends to $df(\boldsymbol{x})$ as $\boldsymbol{y}$ tends to $\boldsymbol{x}$, we now observe that $s(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{u}\boldsymbol{v}^* + df(\boldsymbol{x})$ for vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. In view of the Cauchy-Schwarz inequality, the spectral norm of the matrix outer product $\boldsymbol{u}\boldsymbol{v}^*$ satisfies

$$\|\boldsymbol{u}\boldsymbol{v}^*\| \quad = \quad \sup_{\boldsymbol{w}\neq\boldsymbol{0}} \frac{\|\boldsymbol{u}\boldsymbol{v}^*\boldsymbol{w}\|}{\|\boldsymbol{w}\|} \quad = \quad \sup_{\boldsymbol{w}\neq\boldsymbol{0}} \frac{|\boldsymbol{v}^*\boldsymbol{w}|\|\boldsymbol{u}\|}{\|\boldsymbol{w}\|} \quad \leq \quad \|\boldsymbol{u}\|\|\boldsymbol{v}\|.$$

In the current setting this translates into the inequality

$$\|\boldsymbol{u}\boldsymbol{v}^*\| \quad \leq \quad \frac{\|f(\boldsymbol{y}) - f(\boldsymbol{x}) - df(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})\|}{\|\boldsymbol{y}-\boldsymbol{x}\|} \cdot \frac{\|\boldsymbol{y}-\boldsymbol{x}\|}{\|\boldsymbol{y}-\boldsymbol{x}\|}.$$

Hence, Fréchet's condition implies $\lim_{\boldsymbol{y}\to\boldsymbol{x}} \|s(\boldsymbol{y}, \boldsymbol{x}) - df(\boldsymbol{x})\| = 0$. ■

In many cases it is easy to identify a slope function. We have already mentioned slope functions defined by difference quotients. It is worth stressing that while differentials are uniquely determined, slope functions are not. The real-valued function $f(\boldsymbol{y}) = y_1 y_2$ is typical. Indeed, the identities

$$\begin{aligned} y_1 y_2 - x_1 x_2 &= x_2(y_1 - x_1) + y_1(y_2 - x_2) \\ y_1 y_2 - x_1 x_2 &= y_2(y_1 - x_1) + x_1(y_2 - x_2) \end{aligned}$$

define two equally valid slope functions at $\boldsymbol{x}$.

Example 4.4.1 *Differentials of Linear and Quadratic Functions*

A linear transformation $f(\boldsymbol{y}) = \boldsymbol{M}\boldsymbol{y}$ is differentiable with slope function $s(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{M}$. The real-valued coordinate functions y_i of $\boldsymbol{y} \in \mathsf{R}^n$ fall into this category. For a symmetric matrix $\boldsymbol{M}$, the quadratic form $g(\boldsymbol{x}) = \boldsymbol{x}^* \boldsymbol{M} \boldsymbol{x}$ has the difference

$$\boldsymbol{y}^* \boldsymbol{M} \boldsymbol{y} - \boldsymbol{x}^* \boldsymbol{M} \boldsymbol{x} = (\boldsymbol{y} + \boldsymbol{x})^* \boldsymbol{M} (\boldsymbol{y} - \boldsymbol{x}).$$

This gives the differential $dg(\boldsymbol{x}) = 2\boldsymbol{x}^* \boldsymbol{M}$ and gradient $\nabla g(\boldsymbol{x}) = 2\boldsymbol{M}\boldsymbol{x}$. ■

Example 4.4.2 *Differential of a Multilinear Map*

A multilinear map $\boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k]$ as defined in Example 2.5.10 is differentiable. The expansion

$$\begin{aligned} \boldsymbol{M}[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k] - \boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k] &= \boldsymbol{M}[\boldsymbol{v}_1 - \boldsymbol{u}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k] \\ &\quad + \boldsymbol{M}[\boldsymbol{u}_1, \boldsymbol{v}_2 - \boldsymbol{u}_2, \ldots, \boldsymbol{v}_k] \\ &\quad \vdots \\ &\quad + \boldsymbol{M}[\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{v}_k - \boldsymbol{u}_k] \end{aligned}$$

displays the slope function as a sum. The corresponding differential

$$\begin{aligned} d\boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k][\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k] &= \boldsymbol{M}[\boldsymbol{w}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k] \\ &\quad + \boldsymbol{M}[\boldsymbol{u}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{u}_k] \\ &\quad \vdots \\ &\quad + \boldsymbol{M}[\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{w}_k]. \end{aligned}$$

emerges in the limit. ■

The rules for calculating differentials of algebraic combinations of differentiable functions flow easily from Carathéodory's definition.

Proposition 4.4.2 *If the two functions $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ map the open set $S \subset \mathsf{R}^m$ differentiably into R^n, then the following combinations are differentiable and have the displayed differentials:*

(a) $d[\alpha f(\boldsymbol{x}) + \beta g(\boldsymbol{x})] = \alpha df(\boldsymbol{x}) + \beta dg(\boldsymbol{x})$ *for all constants* α *and* β.

(b) $d[f(\boldsymbol{x})^* g(\boldsymbol{x})] = f(\boldsymbol{x})^* dg(\boldsymbol{x}) + g(\boldsymbol{x})^* df(\boldsymbol{x})$.

(c) $d[f(\boldsymbol{x})^{-1}] = -f(\boldsymbol{x})^{-2} df(\boldsymbol{x})$ *when* $n = 1$ *and* $f(\boldsymbol{x}) \neq 0$.

Proof: Let the slope functions for $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ at $\boldsymbol{x}$ be $s_f(\boldsymbol{y}, \boldsymbol{x})$ and $s_g(\boldsymbol{y}, \boldsymbol{x})$. Rule (a) follows by taking the limit of the slope function identified in the equality

$$\alpha f(\boldsymbol{y}) + \beta g(\boldsymbol{y}) - \alpha f(\boldsymbol{x}) - \beta g(\boldsymbol{x}) \quad = \quad [\alpha s_f(\boldsymbol{y}, \boldsymbol{x}) + \beta s_g(\boldsymbol{y}, \boldsymbol{x})](\boldsymbol{y} - \boldsymbol{x}).$$

Rule (b) stems from the equality

$$\begin{aligned} & f(\boldsymbol{y})^* g(\boldsymbol{y}) - f(\boldsymbol{x})^* g(\boldsymbol{x}) \\ = \; & [f(\boldsymbol{y}) - f(\boldsymbol{x})]^* g(\boldsymbol{y}) + f(\boldsymbol{x})^* [g(\boldsymbol{y}) - g(\boldsymbol{x})] \\ = \; & g(\boldsymbol{y})^* s_f(\boldsymbol{y}, \boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) + f(\boldsymbol{x})^* s_g(\boldsymbol{y}, \boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}), \end{aligned}$$

and rule (c) stems from the equality

$$\begin{aligned} f(\boldsymbol{y})^{-1} - f(\boldsymbol{x})^{-1} &= -f(\boldsymbol{y})^{-1} f(\boldsymbol{x})^{-1} [f(\boldsymbol{y}) - f(\boldsymbol{x})] \\ &= -f(\boldsymbol{y})^{-1} f(\boldsymbol{x})^{-1} s_f(\boldsymbol{y}, \boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}). \end{aligned}$$

■

The chain rule also has an beautifully straightforward proof.

Proposition 4.4.3 *Suppose* $f(\boldsymbol{x})$ *maps the open set* $S \subset \mathsf{R}^k$ *differentiably into* R^m *and* $g(\boldsymbol{z})$ *maps the open set* $T \subset \mathsf{R}^m$ *differentiably into* R^n. *If the image* $f(S)$ *is contained in* T, *then the composition* $g \circ f(\boldsymbol{x})$ *is differentiable with differential* $dg \circ f(\boldsymbol{x})\, df(\boldsymbol{x})$.

Proof: Let $s_f(\boldsymbol{y}, \boldsymbol{x})$ be the slope function of $f(\boldsymbol{y})$ for $\boldsymbol{y}$ near $\boldsymbol{x}$ and $s_g(\boldsymbol{z}, \boldsymbol{w})$ be the slope function of $g(\boldsymbol{z})$ for $\boldsymbol{z}$ near $\boldsymbol{w} = f(\boldsymbol{x})$. The chain rule follows after taking the limit of the slope function identified in the equality

$$\begin{aligned} g \circ f(\boldsymbol{y}) - g \circ f(\boldsymbol{x}) &= s_g[f(\boldsymbol{y}), f(\boldsymbol{x})][f(\boldsymbol{y}) - f(\boldsymbol{x})] \\ &= s_g[f(\boldsymbol{y}), f(\boldsymbol{x})] s_f(\boldsymbol{y}, \boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}). \end{aligned}$$

The chain rule is much harder to prove under Fréchet's definition. ■

Of course, these rules do not exhaust the techniques for finding differentials. Here is an example where we must fall back on the definition. See Problem 17 for a generalization.

Example 4.4.3 *Differential of* $f_+(\boldsymbol{x})^2$

Suppose $f(\boldsymbol{x})$ is a real-valued differentiable function. Define the function $f_+(\boldsymbol{x}) = \max\{f(\boldsymbol{x}), 0\}$. In general, $f_+(\boldsymbol{x})$ is not a differentiable function,

but $g(\boldsymbol{x}) = f_+(\boldsymbol{x})^2$ is. This is obvious on the open set $\{\boldsymbol{x} : f(\boldsymbol{x}) < 0\}$, where $dg(\boldsymbol{x}) = \boldsymbol{0}$, and on the open set $\{\boldsymbol{x} : f(\boldsymbol{x}) > 0\}$, where

$$dg(\boldsymbol{x}) \quad = \quad 2f(\boldsymbol{x})df(\boldsymbol{x}).$$

The troublesome points are those with $f(\boldsymbol{x}) = 0$. Near such a point we have $f(\boldsymbol{y}) - 0 = s(\boldsymbol{y}, \boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})$, and

$$g(\boldsymbol{y}) - 0 \quad = \quad f_+(\boldsymbol{y})s(\boldsymbol{y}, \boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}).$$

It follows that $dg(\boldsymbol{x}) = f_+(\boldsymbol{x})df(\boldsymbol{x}) = \boldsymbol{0}^*$ when $f(\boldsymbol{x}) = 0$. In general, all three cases can be summarized by the same rule $dg(\boldsymbol{x}) = 2f_+(\boldsymbol{x})df(\boldsymbol{x})$. ■

Example 4.4.4 *Forward Directional Derivative of* $\max_{1 \le i \le p} g_i(\boldsymbol{x})$

As the example $|x| = \max\{-x, x\}$ illustrates, the maximum of two differentiable functions may not be differentiable. For many purposes in optimization, forward directional derivatives are adequate. Consider the maximum $f(\boldsymbol{x}) = \max_{1 \le i \le p} g_i(\boldsymbol{x})$ of a finite number of real-valued functions differentiable at the point $\boldsymbol{y}$. To show that $f(\boldsymbol{x})$ possesses all possible forward directional derivatives at $\boldsymbol{y}$, let $\boldsymbol{v} \neq \boldsymbol{0}$ be an arbitrary direction and t_n any sequence of positive numbers converging to 0. It suffices to prove that the difference quotients $t_n^{-1}[f(\boldsymbol{y} + t_n\boldsymbol{v}) - f(\boldsymbol{y})]$ tend to a limit $d_{\boldsymbol{v}}f(\boldsymbol{y})$ independent of the specific sequence t_n. Because the $g_i(\boldsymbol{x})$ are differentiable at $\boldsymbol{y}$, they are also continuous at $\boldsymbol{y}$. Those $g_i(\boldsymbol{x})$ with $g_i(\boldsymbol{y}) < f(\boldsymbol{y})$ play no role in determining $f(\boldsymbol{x})$ near $\boldsymbol{y}$ and can be discarded in calculating a directional derivative. Hence, we assume without loss of generality that all $g_i(\boldsymbol{y}) = f(\boldsymbol{y})$. With this proviso, we claim that $d_{\boldsymbol{v}}f(\boldsymbol{y}) = \max_{1 \le i \le p} dg_i(\boldsymbol{y})\boldsymbol{v}$. The inequality

$$\liminf_{n\to\infty} \frac{f(\boldsymbol{y} + t_n\boldsymbol{v}) - f(\boldsymbol{y})}{t_n} \quad \ge \quad \liminf_{n\to\infty} \frac{g_i(\boldsymbol{y} + t_n\boldsymbol{v}) - g_i(\boldsymbol{y})}{t_n} \quad = \quad dg_i(\boldsymbol{y})\boldsymbol{v}$$

for any i is obvious. Suppose that

$$\limsup_{n\to\infty} \frac{f(\boldsymbol{y} + t_n\boldsymbol{v}) - f(\boldsymbol{y})}{t_n} \quad > \quad \max_{1 \le i \le p} dg_i(\boldsymbol{y})\boldsymbol{v}. \tag{4.7}$$

In view of the definition of lim sup, there exists an $\epsilon > 0$ and a subsequence t_{n_m} along which

$$\frac{f(\boldsymbol{y} + t_{n_m}\boldsymbol{v}) - f(\boldsymbol{y})}{t_{n_m}} \quad \ge \quad \max_{1 \le i \le p} dg_i(\boldsymbol{y})\boldsymbol{v} + \epsilon.$$

Passing to a subsubsequence if necessary, we can choose a j such that

$$\frac{f(\boldsymbol{y} + t_{n_m}\boldsymbol{v}) - f(\boldsymbol{y})}{t_{n_m}} \quad = \quad \frac{g_j(\boldsymbol{y} + t_{n_m}\boldsymbol{v}) - g_j(\boldsymbol{y})}{t_{n_m}}$$

for all m. Taking limits now produces the contradiction

$$dg_j(\boldsymbol{y})\boldsymbol{v} \geq \max_{1\leq i\leq p} dg_i(\boldsymbol{y})\boldsymbol{v} + \epsilon.$$

Hence, inequality (4.7) is false, and the difference quotients tend to the claimed limit. Appendix A.6 treats this example in more depth. ■

Example 4.4.5 *Analytic Functions*

When we come to complex-valued functions $f(\boldsymbol{z})$ of a complex variable $\boldsymbol{z}$, we are confronted with a dilemma in defining differentiability. On the one hand, we can substitute complex arithmetic operations for real arithmetic operations in Carathéodory's definition. If we adopt this perspective, then the equations

$$\begin{aligned} f(\boldsymbol{z}) - f(\boldsymbol{w}) &= s(\boldsymbol{z},\boldsymbol{w})(\boldsymbol{z}-\boldsymbol{w}) \\ \lim_{\boldsymbol{z}\to\boldsymbol{w}} s(\boldsymbol{z},\boldsymbol{w}) &= s(\boldsymbol{w},\boldsymbol{w}) \end{aligned} \tag{4.8}$$

are summarized by saying that $f(\boldsymbol{z})$ is analytic (or holomorphic) at $\boldsymbol{w}$. Most of the results we have proved for differentiable functions carry over without change to analytic functions. The chain rule is a case in point. On the other hand, we can view the complex plane as R^2 and decompose $\boldsymbol{z} = x + iy$ and $f(\boldsymbol{z}) = g(\boldsymbol{z}) + ih(\boldsymbol{z})$ into their real and imaginary parts with $i = \sqrt{-1}$. In this context, $f(\boldsymbol{z})$ is differentiable at $\boldsymbol{w}$ if there exists a 2×2 slope matrix $m(\boldsymbol{z},\boldsymbol{w})$ satisfying

$$\begin{aligned} \begin{pmatrix} g(\boldsymbol{z}) - g(\boldsymbol{w}) \\ h(\boldsymbol{z}) - h(\boldsymbol{w}) \end{pmatrix} &= m(\boldsymbol{z},\boldsymbol{w}) \begin{pmatrix} x-u \\ y-v \end{pmatrix} \\ \lim_{\boldsymbol{z}\to\boldsymbol{w}} m(\boldsymbol{z},\boldsymbol{w}) &= m(\boldsymbol{w},\boldsymbol{w}) \end{aligned} \tag{4.9}$$

for $\boldsymbol{w} = u + iv$. A function $f(\boldsymbol{z})$ analytic at $\boldsymbol{w}$ is differentiable at $\boldsymbol{w}$. Indeed, if we decompose $s(\boldsymbol{z},\boldsymbol{w}) = r(\boldsymbol{z},\boldsymbol{w}) + it(\boldsymbol{z},\boldsymbol{w})$ in equation (4.8), then identifying real and imaginary parts demonstrates that the matrix

$$m(\boldsymbol{z},\boldsymbol{w}) = \begin{pmatrix} r(\boldsymbol{z},\boldsymbol{w}) & -t(\boldsymbol{z},\boldsymbol{w}) \\ t(\boldsymbol{z},\boldsymbol{w}) & r(\boldsymbol{z},\boldsymbol{w}) \end{pmatrix}$$

satisfies equation (4.9). Hence, analyticity is stronger than differentiability. Taking limits on $\boldsymbol{z}$ in this definition of $m(\boldsymbol{z},\boldsymbol{w})$ furthermore shows that $f(\boldsymbol{z})$ satisfies the Cauchy-Riemann equations

$$\begin{aligned} \frac{\partial}{\partial x} g(\boldsymbol{z}) &= \frac{\partial}{\partial y} h(\boldsymbol{z}) \\ -\frac{\partial}{\partial y} g(\boldsymbol{z}) &= \frac{\partial}{\partial x} h(\boldsymbol{z}) \end{aligned}$$

at $\boldsymbol{z} = \boldsymbol{w}$. A beautiful theory with surprising consequences can be constructed for analytic functions [129, 244]. For example, every analytic function on an open domain is infinitely differentiable on that domain. The price we pay for such powerful results is that the class of analytic functions is much smaller than the class of differentiable functions of R^2 into itself. For example the failure of the Cauchy-Riemann equations implies that the complex conjugate function $x + iy \mapsto x - iy$ is not analytic even though it is differentiable. ■

4.5 Multivariate Mean Value Theorem

The mean value theorem is one of the most useful tools of the differential calculus. Here is a simple generalization to multiple dimensions.

Proposition 4.5.1 *Let the function $f(\boldsymbol{y})$ map an open subset S of R^m to R^n. If $f(\boldsymbol{y})$ is differentiable on a neighborhood of $\boldsymbol{x} \in S$, then*

$$f(\boldsymbol{y}) \;=\; f(\boldsymbol{x}) + \int_0^1 df[\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})]\, dt\, (\boldsymbol{y} - \boldsymbol{x}) \tag{4.10}$$

for $\boldsymbol{y}$ near $\boldsymbol{x}$. If S is convex, then identity (4.10) holds for all $\boldsymbol{y} \in S$.

Proof: Integrating component by component, we need only consider the case $n = 1$. According to the chain rule stated in Proposition 4.4.3, the real-valued function $g(t) = f[\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})]$ of the scalar t has differential $dg(t) = df[\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})](\boldsymbol{y}-\boldsymbol{x})$. Because differentials and derivatives coincide on the real line, equality (4.10) follows from the fundamental theorem of calculus applied to $g(t)$. ■

The notion of continuous differentiability is ambiguous. On the one hand, we could say that $f(\boldsymbol{y})$ is continuously differentiable around $\boldsymbol{x}$ if it possesses a slope function $s(\boldsymbol{y}, \boldsymbol{z})$ that is jointly continuous in its two arguments. This implies the continuity of $df(\boldsymbol{y}) = s(\boldsymbol{y}, \boldsymbol{y})$ around $\boldsymbol{x}$. On the other hand, continuous differentiability suggests that we postulate the continuity of $df(\boldsymbol{y})$ to start with. In this case, equation (4.10) yields the slope function

$$s(\boldsymbol{y}, \boldsymbol{z}) \;=\; \int_0^1 df[\boldsymbol{z} + t(\boldsymbol{y} - \boldsymbol{z})]\, dt. \tag{4.11}$$

If $df(\boldsymbol{y})$ is continuous near $\boldsymbol{x}$, then this choice of $s(\boldsymbol{y}, \boldsymbol{z})$ is jointly continuous in its arguments. Hence, the two definitions of continuous differentiability coincide. It is noteworthy that the particular slope function (4.11) is symmetric in the sense that $s(\boldsymbol{y}, \boldsymbol{z}) = s(\boldsymbol{z}, \boldsymbol{y})$.

Example 4.5.1 *Characterization of Constant Functions*

If the differential $df(\boldsymbol{x})$ of a real-valued function $f(\boldsymbol{x})$ vanishes on an open connected set $S \subset \mathsf{R}^n$, then $f(\boldsymbol{x})$ is constant there. To establish this fact, let $\boldsymbol{z}$ be an arbitrary point of S and define $T = \{\boldsymbol{x} \in S : f(\boldsymbol{x}) = f(\boldsymbol{z})\}$. Given the continuity of $f(\boldsymbol{x})$, it is obvious that T is closed relative to S. To prove that $T = S$, it suffices to show that T is also open relative to S. Indeed, if T and $S \setminus T$ are both nonempty open sets, then they disconnect S. Now any point $\boldsymbol{x} \in T$ is contained in a ball $B(\boldsymbol{x}, r) \subset S$. For $\boldsymbol{y} \in B(\boldsymbol{x}, r)$, formula (4.10) and the vanishing of the differential show that $f(\boldsymbol{y}) = f(\boldsymbol{x})$. Thus, T is open. ■

Example 4.5.2 *Failure of Proposition 4.2.1*

Consider the function $f(x) = (\cos x, \sin x)^*$ from R to R^2. The obvious generalization of the mean value theorem stated in Proposition 4.2.1 fails because there is no $x \in (0, 2\pi)$ satisfying

$$\mathbf{0} \;=\; f(2\pi) - f(0) \;=\; \begin{pmatrix} -\sin x \\ \cos x \end{pmatrix} (2\pi - 0).$$

In this regard recall Example 4.2.1. ■

In spite of this counterexample, the bound

$$\begin{aligned} \|f(\boldsymbol{y}) - f(\boldsymbol{x})\| &\leq \left\| \int_0^1 df[\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})]\, dt \right\| \cdot \|\boldsymbol{y} - \boldsymbol{x}\| \\ &\leq \sup_{t \in [0,1]} \|df[\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})]\| \cdot \|\boldsymbol{y} - \boldsymbol{x}\| \end{aligned} \tag{4.12}$$

is often an adequate substitute for theoretical purposes for functions with a convex domain. See inequality (3.5) of Chap. 3 for a proof.

4.6 Inverse and Implicit Function Theorems

Two of the harder theorems involving differentials are the inverse and implicit function theorems. The definition of differentials through slope functions tends to make the proofs easier to understand. The current proof of the inverse function theorem also features an interesting optimization argument [1].

Proposition 4.6.1 *Let $f(\boldsymbol{x})$ map an open set $U \subset \mathsf{R}^n$ into R^n. If $f(\boldsymbol{x})$ is continuously differentiable on U and the square matrix $df(\boldsymbol{x})$ is invertible at the point $\boldsymbol{z}$, then there exist neighborhoods V of $\boldsymbol{z}$ and W of $f(\boldsymbol{z})$ such that the inverse function $g(\boldsymbol{y})$ satisfying $g \circ f(\boldsymbol{x}) = \boldsymbol{x}$ exists and maps W onto V. Furthermore, $g(\boldsymbol{x})$ is continuously differentiable with differential $dg(\boldsymbol{x}) = df[g(\boldsymbol{x})]^{-1}$.*

Proof: Let $f(\boldsymbol{x})$ have continuous slope function $s(\boldsymbol{y},\boldsymbol{x})$. If $s(\boldsymbol{y},\boldsymbol{x})$ is invertible as a square matrix, then the relations

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \quad = \quad s(\boldsymbol{y},\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})$$

and

$$\boldsymbol{y}-\boldsymbol{x} \quad = \quad s(\boldsymbol{y},\boldsymbol{x})^{-1}[f(\boldsymbol{y}) - f(\boldsymbol{x})]$$

are equivalent. Now suppose we know $f(\boldsymbol{x})$ has functional inverse $g(\boldsymbol{y})$. Exchanging $g(\boldsymbol{y})$ for $\boldsymbol{y}$ and $g(\boldsymbol{x})$ for $\boldsymbol{x}$ in the second relation above produces

$$g(\boldsymbol{y}) - g(\boldsymbol{x}) \quad = \quad s[g(\boldsymbol{y}), g(\boldsymbol{x})]^{-1}(\boldsymbol{y}-\boldsymbol{x}), \tag{4.13}$$

and taking limits gives the claimed differential, provided $g(\boldsymbol{y})$ is continuous. To prove the continuity of $g(\boldsymbol{y})$ and therefore the joint continuity of $s[g(\boldsymbol{y}), g(\boldsymbol{x})]^{-1}$, it suffices to show that $\|s[g(\boldsymbol{y}), g(\boldsymbol{x})]^{-1}\|$ is locally bounded. Continuity in this circumstance is then a consequence of the bound

$$\|g(\boldsymbol{y}) - g(\boldsymbol{x})\| \quad \leq \quad \|s[g(\boldsymbol{y}), g(\boldsymbol{x})]^{-1}\| \cdot \|\boldsymbol{y}-\boldsymbol{x}\|$$

flowing from equation (4.13). In view of these remarks, the difficult part of the proof consists in proving that $g(\boldsymbol{y})$ exists.

Given the continuous differentiability of $f(\boldsymbol{x})$, there is some neighborhood V of $\boldsymbol{z}$ such that $s(\boldsymbol{y},\boldsymbol{x})$ is invertible for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in V. Furthermore, we can take V small enough so that the norm $\|s(\boldsymbol{y},\boldsymbol{x})^{-1}\|$ is bounded there. On V, the equality $f(\boldsymbol{y}) - f(\boldsymbol{x}) = s(\boldsymbol{y},\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})$ shows that $f(\boldsymbol{x})$ is one-to-one. Hence, all that remains is to show that we can shrink V so that $f(\boldsymbol{x})$ maps V onto an open subset W containing $f(\boldsymbol{z})$.

For some $r > 0$, the ball $B(\boldsymbol{z}, r)$ of radius r centered at $\boldsymbol{z}$ is contained in V. The sphere $S(\boldsymbol{z}, r) = \{\boldsymbol{x} : \|\boldsymbol{x}-\boldsymbol{z}\| = r\}$ and the ball $B(\boldsymbol{z}, r)$ are disjoint and must have disjoint images under $f(\boldsymbol{x})$ because $f(\boldsymbol{x})$ is one-to-one on V. In particular, $f(\boldsymbol{z})$ is not contained in $f[S(\boldsymbol{z}, r)]$. The latter set is compact because $S(\boldsymbol{z}, r)$ is compact and $f(\boldsymbol{x})$ is continuous. Let $d > 0$ be the distance from $f(\boldsymbol{z})$ to $f[S(\boldsymbol{z}, r)]$.

We now define the set W mentioned in the statement of the proposition to be the ball $B[f(\boldsymbol{z}), d/2]$ and show that W is contained in the image of $B(\boldsymbol{z}, r)$ under $f(\boldsymbol{x})$. Take any $\boldsymbol{y} \in W = B[f(\boldsymbol{z}), d/2]$. The particular function $h(\boldsymbol{x}) = \|\boldsymbol{y} - f(\boldsymbol{x})\|^2$ is differentiable and attains its minimum on the closed ball $C(\boldsymbol{z}, r) = \{\boldsymbol{x} : \|\boldsymbol{x}-\boldsymbol{z}\| \leq r\}$. This minimum is strictly less than $(d/2)^2$ because $\boldsymbol{z}$ certainly performs this well. Furthermore, the minimum cannot be reached at a point $\boldsymbol{u} \in S(\boldsymbol{z}, r)$, for then

$$\|f(\boldsymbol{u}) - f(\boldsymbol{z})\| \quad \leq \quad \|f(\boldsymbol{u}) - \boldsymbol{y}\| + \|\boldsymbol{y} - f(\boldsymbol{z})\| \quad < \quad 2d/2,$$

contradicting the choice of d. Thus, $h(\boldsymbol{x})$ reaches its minimum at some point $\boldsymbol{u}$ in the open set $B(\boldsymbol{z}, r)$. Fermat's principle requires that the differential

$$dh(\boldsymbol{x}) \quad = \quad -2[\boldsymbol{y} - f(\boldsymbol{x})]^* df(\boldsymbol{x}) \tag{4.14}$$

vanish at $\boldsymbol{u}$. Given the invertibility of $df(\boldsymbol{u})$, we therefore have $f(\boldsymbol{u}) = \boldsymbol{y}$.

Finally replace V by the open set $B(\boldsymbol{z}, r) \cap f^{-1}\{B[f(\boldsymbol{z}), d/2]\}$ contained within it. Our arguments have shown that $f(\boldsymbol{x})$ is one-to-one from V onto $W = B[f(\boldsymbol{z}), d/2]$. This allows us to define the inverse function $g(\boldsymbol{x})$ from W onto V and completes the proof. ■

Example 4.6.1 *Polar Coordinates*

Consider the transformation

$$\begin{aligned} r &= \|\boldsymbol{x}\| \\ \theta &= \arctan(x_2/x_1) \end{aligned}$$

to polar coordinates in R^2. A brief calculation shows that this transformation has differential

$$\begin{pmatrix} \frac{\partial}{\partial x_1} r & \frac{\partial}{\partial x_2} r \\ \frac{\partial}{\partial x_1} \theta & \frac{\partial}{\partial x_2} \theta \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\frac{1}{r}\sin\theta & \frac{1}{r}\cos\theta \end{pmatrix}.$$

Because the determinant of the differential is $1/r$, the transformation is locally invertible wherever $r \neq 0$. Excluding the half axis $\{x_1 \leq 0, x_2 = 0\}$, the polar transformation maps the plane one-to-one and onto the open set $(0, \infty) \times (-\pi, \pi)$. Here the inverse transformation reduces to

$$\begin{aligned} x_1 &= r\cos\theta \\ x_2 &= r\sin\theta \end{aligned}$$

with differential

$$\begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\frac{1}{r}\sin\theta & \frac{1}{r}\cos\theta \end{pmatrix}^{-1}.$$

■

We now turn to the implicit function theorem.

Proposition 4.6.2 *Let $f(\boldsymbol{x}, \boldsymbol{y})$ map an open set $S \subset \mathsf{R}^{m+n}$ into R^m. Suppose that $f(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{0}$ and that $f(\boldsymbol{x}, \boldsymbol{y})$ is continuously differentiable on a neighborhood of $(\boldsymbol{a}, \boldsymbol{b}) \in S$. If we split the differential*

$$df(\boldsymbol{x}, \boldsymbol{y}) = [\partial_{\boldsymbol{1}} f(\boldsymbol{x}, \boldsymbol{y}), \partial_{\boldsymbol{2}} f(\boldsymbol{x}, \boldsymbol{y})]$$

into an $m \times m$ block and an $m \times n$ block, and if $\partial_{\boldsymbol{1}} f(\boldsymbol{a}, \boldsymbol{b})$ is invertible, then there exists a neighborhood U of $\boldsymbol{b} \in \mathsf{R}^n$ and a continuously differentiable function $g(\boldsymbol{y})$ from U into R^m such that $f[g(\boldsymbol{y}), \boldsymbol{y}] = \boldsymbol{0}$. Furthermore, $g(\boldsymbol{y})$ is unique and has differential

$$dg(\boldsymbol{y}) = -\partial_{\boldsymbol{1}} f[g(\boldsymbol{y}), \boldsymbol{y}]^{-1} \partial_{\boldsymbol{2}} f[g(\boldsymbol{y}), \boldsymbol{y}].$$

Proof: The function

$$h(\boldsymbol{x},\boldsymbol{y}) \;=\; \begin{pmatrix} f(\boldsymbol{x},\boldsymbol{y}) \\ \boldsymbol{y} \end{pmatrix}$$

from S to R^{m+n} is continuously differentiable with differential

$$dh(\boldsymbol{x},\boldsymbol{y}) \;=\; \begin{pmatrix} \partial_{\boldsymbol{1}} f(\boldsymbol{x},\boldsymbol{y}) & \partial_{\boldsymbol{2}} f(\boldsymbol{x},\boldsymbol{y}) \\ \boldsymbol{0} & \boldsymbol{I}_n \end{pmatrix}.$$

To apply the inverse function theorem to $h(\boldsymbol{x},\boldsymbol{y})$, we must check that $dh(\boldsymbol{a},\boldsymbol{b})$ is invertible. This is straightforward because

$$\begin{pmatrix} \partial_{\boldsymbol{1}} f(\boldsymbol{a},\boldsymbol{b}) & \partial_{\boldsymbol{2}} f(\boldsymbol{a},\boldsymbol{b}) \\ \boldsymbol{0} & \boldsymbol{I}_n \end{pmatrix} \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} \;=\; \boldsymbol{0}$$

can only occur if $\boldsymbol{v} = \boldsymbol{0}$ and $\partial_{\boldsymbol{1}} f(\boldsymbol{a},\boldsymbol{b})\boldsymbol{u} = \boldsymbol{0}$. In view of the invertibility of $\partial_{\boldsymbol{1}} f(\boldsymbol{a},\boldsymbol{b})$, the second of these equalities entails $\boldsymbol{u} = \boldsymbol{0}$.

Given the invertibility of $dh(\boldsymbol{a},\boldsymbol{b})$, the inverse function theorem implies that $h(\boldsymbol{x},\boldsymbol{y})$ possesses a continuously differentiable inverse that maps an open set W containing $(\boldsymbol{0},\boldsymbol{b})$ onto an open set V containing $(\boldsymbol{a},\boldsymbol{b})$. The inverse function takes a point $(\boldsymbol{z},\boldsymbol{y})$ into the point $[k(\boldsymbol{z},\boldsymbol{y}),\boldsymbol{y}]$. Consider the function $g(\boldsymbol{y}) = k(\boldsymbol{0},\boldsymbol{y})$ defined for $(\boldsymbol{0},\boldsymbol{y}) \in W$. Being open, W contains a ball of radius r around $(\boldsymbol{0},\boldsymbol{b})$. There is no harm in restricting the domain of $g(\boldsymbol{y})$ to the ball $U = \{\boldsymbol{y} \in \mathsf{R}^n : \|\boldsymbol{y}-\boldsymbol{b}\| < r\}$. On this domain, $g(\boldsymbol{y})$ is continuously differentiable and $f[g(\boldsymbol{y}),\boldsymbol{y}] = \boldsymbol{0}$. Because $h(\boldsymbol{x},\boldsymbol{y})$ is one-to-one, $g(\boldsymbol{y})$ is uniquely determined.

Finally, we can identify the differential of $g(\boldsymbol{y})$ by constructing a slope function. If we let $f(\boldsymbol{x},\boldsymbol{y})$ have slope function $[s_{\boldsymbol{1}}(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y}), s_{\boldsymbol{2}}(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x},\boldsymbol{y})]$ at $(\boldsymbol{x},\boldsymbol{y})$, then

$$\begin{aligned} \boldsymbol{0} &= f[g(\boldsymbol{v}),\boldsymbol{v}] - f[g(\boldsymbol{y}),\boldsymbol{y}] \\ &= s_{\boldsymbol{1}}[g(\boldsymbol{v}),\boldsymbol{v},g(\boldsymbol{y}),\boldsymbol{y}][g(\boldsymbol{v}) - g(\boldsymbol{y})] + s_{\boldsymbol{2}}[g(\boldsymbol{v}),\boldsymbol{v},g(\boldsymbol{y}),\boldsymbol{y}](\boldsymbol{v}-\boldsymbol{y}). \end{aligned}$$

The invertibility of $s_{\boldsymbol{1}}[g(\boldsymbol{v}),\boldsymbol{v},g(\boldsymbol{y}),\boldsymbol{y}]$ for $\boldsymbol{v}$ near $\boldsymbol{y}$ therefore gives

$$g(\boldsymbol{v}) - g(\boldsymbol{y}) \;=\; -s_{\boldsymbol{1}}[g(\boldsymbol{v}),\boldsymbol{v},g(\boldsymbol{y}),\boldsymbol{y}]^{-1} s_{\boldsymbol{2}}[g(\boldsymbol{v}),\boldsymbol{v},g(\boldsymbol{y}),\boldsymbol{y}](\boldsymbol{v}-\boldsymbol{y}).$$

In the limit as $\boldsymbol{v}$ approaches $\boldsymbol{y}$, we recover the differential $dg(\boldsymbol{y})$. ■

Example 4.6.2 *Circles and the Implicit Function Theorem*

The function $f(x,y) = x^2 + y^2 - r^2$ has differential $df(x,y) = (2x, 2y)$. Unless $a = 0$, the conditions of the implicit function theorem apply at (a,b). The choice

$$g(y) \;=\; \operatorname{sgn}(a)\sqrt{r^2 - y^2} \quad \text{with} \quad g'(y) \;=\; -\frac{y}{x}.$$

clearly satisfies $g(b) = a$ and $g(y)^2 + y^2 - r^2 = 0$. ■

Example 4.6.3 *Tangent Vectors and Tangent Curves*

The intuitive discussion of tangent vectors in Sect. 1.4 can be made rigorous by introducing the notion of a tangent curve at the point $\boldsymbol{x}$. This is simply a differentiable curve $v(s)$ having a neighborhood of the scalar 0 as its domain and satisfying $v(0) = \boldsymbol{x}$ and $g_i[v(s)] = 0$ for all equality constraints and all s sufficiently close to 0. If we apply the chain rule to the composite function $g_i[v(s)] = 0$, then the identity $dg_i(\boldsymbol{x})v'(0) = 0$ emerges. The vector $\boldsymbol{w} = v'(0)$ is said to be a tangent vector at $\boldsymbol{x}$. Conversely, if $\boldsymbol{w}$ satisfies $dg_i(\boldsymbol{x})\boldsymbol{w} = 0$ for all i, then we can construct a tangent curve at $\boldsymbol{x}$ with tangent vector $\boldsymbol{w}$. This application of the implicit function theorem requires a little notation. Let $G(\boldsymbol{x})$ be the vector-valued function with ith component $g_i(\boldsymbol{x})$. The differential $dG(\boldsymbol{x})$ is the Jacobi matrix whose ith row is the differential $dg_i(\boldsymbol{x})$. In agreement with our earlier notation, $\nabla G(\boldsymbol{x})$ is the transpose of $dG(\boldsymbol{x})$.

Now consider the relationship

$$h(\boldsymbol{u}, s) \;=\; G[\boldsymbol{x} + \nabla G(\boldsymbol{x})\boldsymbol{u} + s\boldsymbol{w}] \;=\; \boldsymbol{0}.$$

Applying the chain rule to the function $h(\boldsymbol{u}, s)$ gives

$$\begin{aligned} \partial_{\boldsymbol{u}} h(\boldsymbol{0}, 0) &= dG[\boldsymbol{x} + \nabla G(\boldsymbol{x})\boldsymbol{u} + s\boldsymbol{w}]\nabla G(\boldsymbol{x})\Big|_{(\boldsymbol{u},s)=(\boldsymbol{0},0)} \\ &= dG(\boldsymbol{x})\nabla G(\boldsymbol{x}). \end{aligned}$$

Since $G(\boldsymbol{x}) = \boldsymbol{0}$ and $dG(\boldsymbol{x})\nabla G(\boldsymbol{x})$ is invertible when $dG(\boldsymbol{x})$ has full row rank, the implicit function theorem implies that we can solve for $\boldsymbol{u}$ as a function of s in a neighborhood of 0. If we denote the resulting continuously differentiable function by $u(s)$, then our tangent curve is

$$v(s) \;=\; \boldsymbol{x} + \nabla G(\boldsymbol{x})u(s) + s\boldsymbol{w}.$$

By definition $u(0) = \boldsymbol{0}$, $v(0) = \boldsymbol{x}$, and $G[v(s)] = \boldsymbol{0}$ for all s close to 0. Thus, we need only check that $v'(0) = \boldsymbol{w}$. Because

$$v'(0) \;=\; \nabla G(\boldsymbol{x})u'(0) + \boldsymbol{w},$$

it suffices to check that $\nabla G(\boldsymbol{x})u'(0) = \boldsymbol{0}$. However, in view of the equality

$$0 \;=\; u'(0)^*\boldsymbol{0} \;=\; u'(0)^*\frac{d}{ds}h[u(0), 0] \;=\; u'(0)^* dG(\boldsymbol{x})[\nabla G(\boldsymbol{x})u'(0) + \boldsymbol{w}]$$

and the assumption $dG(\boldsymbol{x})\boldsymbol{w} = \boldsymbol{0}$, this fact is obvious. ■

4.7 Differentials of Matrix-Valued Functions

To define the differential of a matrix-valued function [184], it helps to unroll Carathéodory's definition of differentiability for a vector-valued function $f(\boldsymbol{x})$. Let us rewrite the slope expansion $f(\boldsymbol{y}) - f(\boldsymbol{x}) = s(\boldsymbol{y}, \boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})$ as

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = \sum_{j=1}^{m} s_j(\boldsymbol{y}, \boldsymbol{x})(y_j - x_j) \tag{4.15}$$

using the columns $s_j(\boldsymbol{y}, \boldsymbol{x})$ of the slope matrix $s(\boldsymbol{y}, \boldsymbol{x})$. This notational retreat retains the linear dependence of the difference $f(\boldsymbol{y}) - f(\boldsymbol{x})$ on the increment $\boldsymbol{y} - \boldsymbol{x}$ and suggests how to deal with matrix-valued functions. The key step is simply to re-interpret equation (4.15) by replacing the vector-valued function $f(\boldsymbol{x})$ by a matrix-valued function $f(\boldsymbol{x})$ and the vector-valued slope $s_j(\boldsymbol{y}, \boldsymbol{x})$ by a matrix-valued slope $s_j(\boldsymbol{y}, \boldsymbol{x})$. We retain the requirement that $\lim_{\boldsymbol{y}\to\boldsymbol{x}} s_j(\boldsymbol{y}, \boldsymbol{x}) = s_j(\boldsymbol{x}, \boldsymbol{x})$ for each j. The partial differential matrices $s_j(\boldsymbol{x}, \boldsymbol{x}) = \partial_j f(\boldsymbol{x})$ collectively constitute the differential of $f(\boldsymbol{x})$. The gratifying thing about this revised definition of differentiability is that it applies to scalars, vectors, and matrices in a unified way. Furthermore, the components of the differential match the scalar, vector, or matrix nature of the original function. We now illustrate the virtue of this perspective by several examples involving matrix differentials.

Example 4.7.1 *The Sum and Transpose Rules*

The rules

$$d[f(\boldsymbol{x}) + g(\boldsymbol{x})] = df(\boldsymbol{x}) + dg(\boldsymbol{x}), \qquad df(\boldsymbol{x})^* = [df(\boldsymbol{x})]^*$$

flow directly from the above definition of a differential. ■

Example 4.7.2 *The Chain Rule*

Consider the composition $g \circ f(\boldsymbol{x})$ of a matrix-valued function with a vector-valued function. Let $g(\boldsymbol{y})$ have the slope expansion

$$g(\boldsymbol{y}) - g(\boldsymbol{x}) = \sum_j t_j(\boldsymbol{y}, \boldsymbol{x})(y_j - x_j),$$

and let the jth component $f_j(\boldsymbol{v})$ of $f(\boldsymbol{v})$ have the slope expansion

$$f_j(\boldsymbol{v}) - f_j(\boldsymbol{u}) = \sum_k s_{jk}(\boldsymbol{v}, \boldsymbol{u})(v_k - u_k).$$

Then the identity

$$\begin{aligned} g \circ f(\boldsymbol{v}) - g \circ f(\boldsymbol{u}) &= \sum_j t_j[f(\boldsymbol{v}), f(\boldsymbol{u})][f_j(\boldsymbol{v}) - f_j(\boldsymbol{u})] \\ &= \sum_j t_j[f(\boldsymbol{v}), f(\boldsymbol{u})] \sum_k s_{jk}(\boldsymbol{v}, \boldsymbol{u})(v_k - u_k) \end{aligned}$$

shows that $g \circ f(\boldsymbol{v})$ has a differential with $\sum_j \partial_j g[f(\boldsymbol{u})]\partial_k f_j(\boldsymbol{u})$ as its kth component. ■

Example 4.7.3 *The Product Rule*

The matrix product $f(\boldsymbol{y})g(\boldsymbol{y})$ satisfies the identities

$$\begin{aligned} f(\boldsymbol{y})g(\boldsymbol{y}) - f(\boldsymbol{x})g(\boldsymbol{x}) &= [f(\boldsymbol{y}) - f(\boldsymbol{x})]g(\boldsymbol{y}) + f(\boldsymbol{x})[g(\boldsymbol{y}) - g(\boldsymbol{x})] \\ &= f(\boldsymbol{y})[g(\boldsymbol{y}) - g(\boldsymbol{x})] + [f(\boldsymbol{y}) - f(\boldsymbol{x})]g(\boldsymbol{x}). \end{aligned}$$

Substituting the slope formulas

$$\begin{aligned} f(\boldsymbol{y}) - f(\boldsymbol{x}) &= \sum_k s_k(\boldsymbol{y}, \boldsymbol{x})(y_k - x_k) \\ g(\boldsymbol{y}) - g(\boldsymbol{x}) &= \sum_k t_k(\boldsymbol{y}, \boldsymbol{x})(y_k - x_k), \end{aligned}$$

in either identity demonstrates that $f(\boldsymbol{y})g(\boldsymbol{y})$ possesses a differential with kth component $\partial_k f(\boldsymbol{x})g(\boldsymbol{x}) + f(\boldsymbol{x})\partial_k g(\boldsymbol{x})$ at $\boldsymbol{x}$.

When the product $f(\boldsymbol{x})g(\boldsymbol{x})$ is a square matrix, we can apply the trace operator to the difference $f(\boldsymbol{y})g(\boldsymbol{y}) - f(\boldsymbol{x})g(\boldsymbol{x})$ and conclude that

$$\partial_k \operatorname{tr}[f(\boldsymbol{x})g(\boldsymbol{x})] = \operatorname{tr}[\partial_k f(\boldsymbol{x})g(\boldsymbol{x}) + f(\boldsymbol{x})\partial_k g(\boldsymbol{x})].$$

If $f(\boldsymbol{x})$ is a square matrix, then we can differentiate polynomials in the argument $f(\boldsymbol{x})$. In view of the sum rule, it suffices to show how to differentiate powers of $f(\boldsymbol{x})$. For example,

$$\begin{aligned} \partial_k[f(\boldsymbol{x})]^3 &= \partial_k f(\boldsymbol{x})[f(\boldsymbol{x})]^2 + f(\boldsymbol{x})\partial_k[f(\boldsymbol{x})]^2 \\ &= \partial_k f(\boldsymbol{x})[f(\boldsymbol{x})]^2 + f(\boldsymbol{x})\partial_k f(\boldsymbol{x})f(\boldsymbol{x}) + [f(\boldsymbol{x})]^2\partial_k f(\boldsymbol{x}). \end{aligned}$$

In general,

$$\partial_k[f(\boldsymbol{x})]^n = \partial_k f(\boldsymbol{x})f(\boldsymbol{x})^{n-1} + \cdots + f(\boldsymbol{x})^{n-1}\partial_k f(\boldsymbol{x}).$$

Because matrix multiplication is not commutative, further simplification is impossible. Differentiation of polynomials constitutes a special case of differentiation of power series as discussed in a moment. ■

Example 4.7.4 *The Inverse Rule*

Suppose $f(\boldsymbol{y})$ is an invertible matrix with slope expansion

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = \sum_k s_k(\boldsymbol{y}, \boldsymbol{x})(y_k - x_k).$$

Either of the identities

$$\begin{aligned} f(\boldsymbol{y})^{-1} - f(\boldsymbol{x})^{-1} &= -f(\boldsymbol{x})^{-1}[f(\boldsymbol{y}) - f(\boldsymbol{x})]f(\boldsymbol{y})^{-1} \\ &= f(\boldsymbol{x})^{-1}\sum_k s_k(\boldsymbol{y}, \boldsymbol{x})(y_k - x_k)f(\boldsymbol{y})^{-1} \end{aligned}$$

$$\begin{aligned} f(\boldsymbol{y})^{-1} - f(\boldsymbol{x})^{-1} &= -f(\boldsymbol{y})^{-1}[f(\boldsymbol{y}) - f(\boldsymbol{x})]f(\boldsymbol{x})^{-1} \\ &= f(\boldsymbol{y})^{-1} \sum_k s_k(\boldsymbol{y}, \boldsymbol{x})(y_k - x_k) f(\boldsymbol{x})^{-1} \end{aligned}$$

entail a differential of $f(\boldsymbol{y})^{-1}$ with kth component $-f(\boldsymbol{x})^{-1}\partial_k f(\boldsymbol{x}) f(\boldsymbol{x})^{-1}$ at the point $\boldsymbol{x}$. ■

Example 4.7.5 *Matrix Power Series*

Let $p(x) = \sum_{n=0}^{\infty} c_n x^n$ denote a power series with radius of convergence r. As we have seen for the choices $p(x) = e^x$, $p(x) = (1-x)^{-1}$, and $p(x) = \ln(1-x)$, one defines the matrix power series $p(\boldsymbol{M})$ by substituting a square matrix $\boldsymbol{M}$ with norm $\|\boldsymbol{M}\| < r$ for the scalar argument x [128]. On any disc $\{x : |x| \leq s < r\}$ strictly inside its circle of convergence, a power series $p(x)$ is analytic and converges uniformly and absolutely [134, 223]. Furthermore, $p(x)$ can be differentiated term by term; its derivative $p'(x)$ retains r as its radius of convergence.

An obvious question of interest is whether $p(\boldsymbol{M})$ is differentiable. The easiest route to an affirmative answer exploits forward directional derivatives. Appendix A.6 introduces the notion of a semidifferentiable function. In the current context, this is a function $f(\boldsymbol{M})$ possessing all possible forward directional derivatives in the strong sense of Hadamard. Thus, we demand the existence of the uniform limit

$$\lim_{t \downarrow 0, \boldsymbol{U} \to \boldsymbol{V}} \frac{f(\boldsymbol{M} + t\boldsymbol{U}) - f(\boldsymbol{M})}{t} = d_{\boldsymbol{V}} f(\boldsymbol{M})$$

for all $\boldsymbol{V}$. For a semidifferentiable function, Proposition A.6.3 proves that the ordinary differential $df(\boldsymbol{M})$ exists whenever the map $\boldsymbol{V} \mapsto d_{\boldsymbol{V}} f(\boldsymbol{M})$ is linear. Our experience with polynomials suggests that

$$d_{\boldsymbol{V}} p(\boldsymbol{M}) = \sum_{n=1}^{\infty} c_n (\boldsymbol{V}\boldsymbol{M}^{n-1} + \cdots + \boldsymbol{M}^{n-1}\boldsymbol{V}).$$

This series certainly qualifies as linear in $\boldsymbol{V}$. It converges because

$$\|\boldsymbol{V}\boldsymbol{M}^{n-1} + \cdots + \boldsymbol{M}^{n-1}\boldsymbol{V}\| \leq n\|\boldsymbol{V}\| \cdot \|\boldsymbol{M}\|^{n-1}$$

and the series $\sum_{n=1}^{\infty} n c_n x^{n-1}$ for $p'(x)$ converges absolutely.

Consider the difference quotient

$$\frac{p(\boldsymbol{M} + t\boldsymbol{U}) - p(\boldsymbol{M})}{t} = \sum_{n=1}^{\infty} c_n (\boldsymbol{U}\boldsymbol{M}^{n-1} + \cdots + \boldsymbol{M}^{n-1}\boldsymbol{U}) + \frac{1}{t} R(\boldsymbol{U}).$$

The series displayed on the right of this equality converges to $d_{\boldsymbol{V}} p(\boldsymbol{M})$. Furthermore, the norm of the remainder is bounded above by

$$\begin{aligned}\frac{1}{t}\|R(\boldsymbol{U})\| &\leq \frac{1}{t}\sum_{n=2}^{\infty} c_n \sum_{j=2}^{n}\binom{n}{j}\|\boldsymbol{M}\|^{n-j}\|t\boldsymbol{U}\|^j \\ &= t\|\boldsymbol{U}\|^2\sum_{n=2}^{\infty} c_n \sum_{j=2}^{n}\frac{n(n-1)}{j(j-1)}\binom{n-2}{j-2}\|\boldsymbol{M}\|^{n-j}\|t\boldsymbol{U}\|^{j-2} \\ &\leq t\|\boldsymbol{U}\|^2\sum_{n=2}^{\infty} n(n-1)c_n(\|\boldsymbol{M}\|+t\|\boldsymbol{U}\|)^{n-2}.\end{aligned}$$

Comparison with the absolutely convergent series $\sum_{n=2}^{\infty} n(n-1)c_n x^{n-2}$ for $p''(x)$ shows that the remainder tends uniformly in norm to $\mathbf{0}$. ■

Example 4.7.6 *Differential of a Determinant*

Sometimes it is simpler to calculate the old-fashioned way with partial derivatives. For example, consider $\det M(\boldsymbol{x})$. In this case, we exploit the determinant expansion

$$\det \boldsymbol{M} = \sum_j m_{ij}M_{ij} \tag{4.16}$$

of a square matrix $\boldsymbol{M}$ in terms of the entries m_{ij} and corresponding cofactors M_{ij} of its ith row. If we ignore the dependence of $M(\boldsymbol{x})$ on $\boldsymbol{x}$ and view $\boldsymbol{M}$ exclusively as a function of its entries m_{ij}, then the expansion (4.16) gives

$$\frac{\partial}{\partial m_{ij}}\det \boldsymbol{M} = M_{ij}.$$

The chain rule therefore implies

$$\begin{aligned}\frac{\partial}{\partial x_i}\det M(\boldsymbol{x}) &= \sum_j\sum_k \frac{\partial}{\partial m_{jk}}\det M(\boldsymbol{x})\frac{\partial}{\partial x_i}m_{jk}(\boldsymbol{x}) \\ &= \sum_j\sum_k M_{jk}(\boldsymbol{x})\frac{\partial}{\partial x_i}m_{jk}(\boldsymbol{x}).\end{aligned}$$

According to Cramer's rule, this can be simplified by noting that the matrix with entry $(\det \boldsymbol{M})^{-1}M_{jk}$ in row k and column j is $\boldsymbol{M}^{-1}$. It follows that

$$\frac{\partial}{\partial x_i}\det M(\boldsymbol{x}) = \det M(\boldsymbol{x})\,\mathrm{tr}\left[M(\boldsymbol{x})^{-1}\frac{\partial}{\partial x_i}M(\boldsymbol{x})\right]$$

when $M(\boldsymbol{x})$ is invertible. If $\det M(\boldsymbol{x})$ is positive, for instance if $M(\boldsymbol{x})$ is positive definite, then we have the even cleaner formula

$$\frac{\partial}{\partial x_i}\ln\det M(\boldsymbol{x}) = \mathrm{tr}\left[M(\boldsymbol{x})^{-1}\frac{\partial}{\partial x_i}M(\boldsymbol{x})\right].$$

This is sometimes expressed in the ambiguous but suggestive form

$$d\left[\ln\det M(\boldsymbol{x})\right] \;=\; \operatorname{tr}\left[M(\boldsymbol{x})^{-1}dM(\boldsymbol{x})\right]. \tag{4.17}$$

As an application of these results, consider minimization of the function

$$f(\boldsymbol{M}) \;=\; a\ln\det\boldsymbol{M} + b\operatorname{tr}(S\boldsymbol{M}^{-1}), \tag{4.18}$$

where a and b are positive constants and $\boldsymbol{S}$ and $\boldsymbol{M}$ are positive definite matrices. On its open domain $f(\boldsymbol{M})$ is obviously differentiable. Because $\boldsymbol{M}$ is symmetric, we parameterize it by its lower triangle, including of course its diagonal. At a local minimum, the differential of $f(\boldsymbol{M})$ vanishes. In view of the cyclic permutation property of the trace function, this differential has components

$$\begin{aligned}\partial_k f(\boldsymbol{M}) &= a\operatorname{tr}(\boldsymbol{M}^{-1}\partial_k\boldsymbol{M}) - b\operatorname{tr}(\boldsymbol{S}\boldsymbol{M}^{-1}\partial_k\boldsymbol{M}\boldsymbol{M}^{-1})\\ &= \operatorname{tr}[\partial_k\boldsymbol{M}(a\boldsymbol{M}^{-1} - b\boldsymbol{M}^{-1}\boldsymbol{S}\boldsymbol{M}^{-1})]\\ &= \operatorname{tr}(\partial_k\boldsymbol{M}\boldsymbol{N})\end{aligned}$$

in obvious notation. If k corresponds to the ith diagonal entry of $\boldsymbol{M}$, then $\partial_k\boldsymbol{M}$ has all entries 0 except for a 1 in the ith diagonal entry. If k corresponds to an off-diagonal entry in row i and column j, then $\partial_k\boldsymbol{M}$ has all entries 0 except for a 1 in the off-diagonal entries in the symmetric positions (i,j) and (j,i). An easy calculation based on $\partial_k f(\boldsymbol{M}) = \operatorname{tr}(\partial_k\boldsymbol{M}\boldsymbol{N})$ now shows that

$$\partial_k f(\boldsymbol{M}) \;=\; \begin{cases} n_{ii} & k=(i,i)\\ 2n_{ij} & k=(i,j),\quad i>j.\end{cases}$$

Because the components of the differential vanish, the matrix $\boldsymbol{N}$ must vanish as well. In other words $\boldsymbol{N} = a\boldsymbol{M}^{-1} - b\boldsymbol{M}^{-1}\boldsymbol{S}\boldsymbol{M}^{-1} = \boldsymbol{0}$. The one and only solution to this equation is $\boldsymbol{M} = \frac{b}{a}\boldsymbol{S}$. In Example 6.3.12 we will prove that this stationary point provides the minimum of $f(\boldsymbol{M})$. ■

4.8 Problems

1. Verify the entries in Table 4.1 not derived in the text.

2. For each positive integer n and real number x, find the derivative, if possible, of the function

$$f_n(x) \;=\; \begin{cases} x^n\sin\left(\frac{1}{x}\right) & x\neq 0\\ 0 & x=0\,.\end{cases}$$

   Pay particular attention to the point 0.

3. Show that the function

$$f(x) \quad = \quad x \ln(1 + x^{-1})$$

is strictly increasing on $(0, \infty)$ and satisfies $\lim_{x \to 0} f(x) = 0$ and $\lim_{x \to \infty} f(x) = 1$ [69].

4. Let $h(x) = f(x)g(x)$ be the product of two functions that are each k times differentiable. Derive Leibnitz's formula

$$h^{(k)}(x) \quad = \quad \sum_{j=0}^{k} \binom{k}{j} f^{(j)}(x) g^{(k-j)}(x).$$

5. Assume that the real-valued functions $f(y)$ and $g(y)$ are differentiable at the real point x. If (a) $f(x) = g(x) = 0$, (b) $g(y) \neq 0$ for y near x, and (c) $g'(x) \neq 0$, then demonstrate L'Hôpital's rule

$$\lim_{y \to x} \frac{f(y)}{g(y)} \quad = \quad \frac{f'(x)}{g'(x)}.$$

6. Let $f(x)$ and $g(x)$ be continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b). Prove that there exists a point $x \in (a, b)$ such that

$$[f(b) - f(a)]g'(x) \quad = \quad [g(b) - g(a)]f'(x).$$

(Hint: Consider $h(x) = [f(b) - f(a)]g(x) - [g(b) - g(a)]f(x)$.)

7. Show that $\frac{\sin x}{x} < 1$ for $x \neq 0$. Use this fact to prove $1 - \frac{x^2}{2} < \cos x$ for $x \neq 0$.

8. Suppose the positive function $f(x)$ satisfies the inequality

$$f'(x) \quad \leq \quad cf(x)$$

for some constant $c \geq 0$ and all $x \geq 0$. Prove that $f(x) \leq e^{cx} f(0)$ for $x \geq 0$ [69].

9. Abbreviate the scalar exponential and logarithmic functions by $e(t)$ and $l(t)$. Based on the defining differential equations $e'(t) = e(t)$ and $l'(t) = t^{-1}$ and the initial conditions $e(0) = 1$ and $l(1) = 0$, prove that $e \circ l(t) = t$ for $t > 0$ and $l \circ e(t) = t$ for all t. Thus, $e(t)$ and $l(t)$ are functional inverses. (Hint: The derivatives of $l \circ e(t)$ and $t^{-1} e \circ l(t)$ are constant.)

10. Prove the identities $\cos x = \cos(-x)$ and $\sin x = -\sin(-x)$ and the identities

$$\begin{aligned}\cos(x+y) &= \cos x \cos y - \sin x \sin y \\ \sin(x+y) &= \sin x \cos y + \cos x \sin y.\end{aligned}$$

(Hint: The differential equation

$$f'(x) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} f(x)$$

with $f(0)$ fixed has a unique solution. In each case demonstrate that both sides of the two proposed identities satisfy the differential equation. The initial condition is $f(0) = (1,0)^*$ in the first case and $f(0) = (\cos y, \sin y)^*$ in the second case.)

11. Use the defining differential equations for $\cos x$ and $\sin x$ to show that there is a smallest positive root $\frac{\pi}{2}$ of the equation $\cos x = 0$. Applying the trigonometric identities of the previous problem, deduce that $\cos \pi = -1$ and that $\cos(x+2\pi) = \cos x$ and $\sin(x+2\pi) = \sin x$. (Hint: Argue by contradiction that $\cos x > 0$ cannot hold for all positive x.)

12. Let $f(x,y)$ be an integrable function of x for each y. If the partial derivative $\frac{\partial}{\partial y} f(x,y)$ exists, it makes sense to ask when the interchange

$$\frac{d}{dy} \int_a^b f(x,y)\, dx = \int_a^b \frac{\partial}{\partial y} f(x,y)\, dx$$

is permissible. Demonstrate that a sufficient condition is the existence of integrable functions $g(x)$ and $h(x)$ satisfying

$$g(x) \le \frac{\partial}{\partial y} f(x,y) \le h(x)$$

for all x and y. Show how to construct $g(x)$ and $h(x)$ when $\frac{\partial}{\partial y} f(x,y)$ is jointly continuous in x and y. (Hint: Apply the mean value theorem to the difference quotient. Then invoke the dominated convergence theorem.)

13. Let $f(x)$ be a differentiable curve mapping the interval $[a,b]$ into R^n. Show that $\|f(x)\|$ is constant if and only if $f(x)$ and $f'(x)$ are orthogonal for all x.

14. Suppose the constants $c_0, \ldots, c_k$ satisfy the condition

$$c_0 + \frac{c_1}{2} + \cdots + \frac{c_k}{k+1} = 0.$$

Demonstrate that the polynomial $p(x) = c_0 + c_1 x + \cdots + c_k x^k$ has a root on the interval $[0,1]$.

15. Prove that the function

$$f(x) = \begin{cases} e^{-x^{-2}} & x \neq 0 \\ 0 & x = 0 \end{cases}$$

is infinitely differentiable and has derivative $f^{(n)}(0) = 0$ for every n.

16. Consider the ordinary differential equation $M'(t) = N(t)M(t)$ with initial condition $M(0) = \boldsymbol{A}$ for $n \times n$ matrices. If $\boldsymbol{A}$ is invertible, then demonstrate that any two solutions coincide in a neighborhood of 0. (Hint: If $P(t)$ and $Q(t)$ are two solutions, then differentiate the product $P(t)^{-1}Q(t)$ using the product and inverse rules.)

17. Let $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ be real-valued functions defined on a neighborhood of $\boldsymbol{y}$ in R^n. If $f(\boldsymbol{x})$ is differentiable at $\boldsymbol{y}$ and has $f(\boldsymbol{y}) = 0$, and $g(\boldsymbol{x})$ is continuous at $\boldsymbol{y}$, then prove that $f(\boldsymbol{x})g(\boldsymbol{x})$ is differentiable at $\boldsymbol{y}$.

18. Consider a continuous function $f(x, y)$ defined on a compact interval $I = [a, b] \times [c, d]$ of R^2. Assume that the partial derivative $\partial_2 f(x, y)$ is continuous on I and that $p(y)$ and $q(y)$ are differentiable functions mapping $[c, d]$ into $[a, b]$. In this setting, prove that the integral

$$F(y) = \int_{p(y)}^{q(y)} f(x, y)\, dx$$

has derivative

$$F'(y) = \int_{p(y)}^{q(y)} \partial_2 f(x, y)\, dx + f[q(y), y]q'(y) - f[p(y), y]p'(y).$$

(Hint: Use Problem 12.)

19. A real-valued function $f(\boldsymbol{x})$ on R^n is said to be homogeneous of integer order $k \geq 1$ if $f(t\boldsymbol{x}) = t^k f(\boldsymbol{x})$ for every scalar t and $\boldsymbol{x} \in \mathsf{R}^n$. For instance, $f(x) = x^3$ is homogeneous of order 3 on R. Demonstrate that a homogeneous differentiable function $f(\boldsymbol{x})$ satisfies $df(\boldsymbol{x})\boldsymbol{x} = kf(\boldsymbol{x})$.

20. As a converse to the chain rule stated in Proposition 4.4.3, suppose (a) $f(\boldsymbol{y})$ is continuous at $\boldsymbol{x}$, (b) $g(\boldsymbol{z})$ is differentiable at $f(\boldsymbol{x})$, (c) $g \circ f(\boldsymbol{y})$ is differentiable at $\boldsymbol{x}$, and (d) the matrix $dg[f(\boldsymbol{x})]$ is invertible. Show that $f(\boldsymbol{y})$ is differentiable at $\boldsymbol{x}$. (Hint: Equate the two slope expansions of $g \circ f(\boldsymbol{y}) - g \circ f(\boldsymbol{x})$.)

21. A real-valued function $f(\boldsymbol{y})$ is said to be Gâteaux differentiable at $\boldsymbol{x}$ if there exists a vector $\boldsymbol{g}$ such that

$$\lim_{t \to 0} \frac{f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x})}{t} = \boldsymbol{g}^* \boldsymbol{v}$$

for all unit vectors $\boldsymbol{v}$. In other words, all directional derivatives exist and depend linearly on the direction $\boldsymbol{v}$. In general, Gâteaux differentiability does not imply differentiability. However, if $f(\boldsymbol{y})$ satisfies a Lipschitz condition $|f(\boldsymbol{y}) - f(\boldsymbol{z})| \leq c\|\boldsymbol{y} - \boldsymbol{z}\|$ in a neighborhood of $\boldsymbol{x}$, then prove that Gâteaux differentiability at $\boldsymbol{x}$ implies differentiability at $\boldsymbol{x}$ with $\nabla f(\boldsymbol{x}) = \boldsymbol{g}$. In Chap. 6, the proof of Proposition 6.4.1 shows that a convex function is locally Lipschitz around each of its interior points. Thus, Gâteaux differentiability and differentiability are equivalent for a convex function at an interior point of its domain. Consult Appendix A.6 for a fuller treatment of this topic. (Hint: Every $\boldsymbol{u}$ on the unit sphere of R^n is within $\epsilon > 0$ of some member of a finite set $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m\}$ of points on the sphere. Now write

$$\begin{aligned} f(\boldsymbol{x}+\boldsymbol{w}) - f(\boldsymbol{x}) - \boldsymbol{g}^*\boldsymbol{w} &= f(\boldsymbol{x} + \|\boldsymbol{w}\|\boldsymbol{u}_k) - f(\boldsymbol{x}) - \|\boldsymbol{w}\|\boldsymbol{g}^*\boldsymbol{u}_k \\ &\quad + f(\boldsymbol{x}+\boldsymbol{w}) - f(\boldsymbol{x} + \|\boldsymbol{w}\|\boldsymbol{u}_k) \\ &\quad + \|\boldsymbol{w}\|\boldsymbol{g}^*\Big(\boldsymbol{u}_k - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\Big) \end{aligned}$$

for an appropriate choice of $\boldsymbol{u}_k$.)

22. Continuing Problem 21, show that the function

$$f(\boldsymbol{x}) = \begin{cases} 1 & \text{if } x_1 = x_2^2 \text{ and } x_2 \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

is Gâteaux differentiable at the origin $\mathbf{0}$ of R^2 but not differentiable or even continuous there.

23. Demonstrate that the function

$$f(\boldsymbol{x}) = \begin{cases} \frac{x_1 x_2}{x_1^2 + x_2} & x_1^2 + x_2 \neq 0 \\ 0 & x_1^2 + x_2 = 0 \end{cases}$$

is discontinuous at $\mathbf{0}$, yet the directional derivative $d_{\boldsymbol{v}} f(\mathbf{0})$ exists for all $\boldsymbol{v}$. In this instance, show that $\boldsymbol{v} \mapsto d_{\boldsymbol{v}} f(\mathbf{0})$ is nonlinear in $\boldsymbol{v}$.

24. On the set $\{\boldsymbol{x} \neq \mathbf{0}\}$, demonstrate that $\|\boldsymbol{x}\|$ is differentiable with differential $d\|\boldsymbol{x}\| = \|\boldsymbol{x}\|^{-1}\boldsymbol{x}^*$.

25. Prove that the directional derivative $d_{\boldsymbol{v}}\|\boldsymbol{x}\|$ fails to exists when $\boldsymbol{x} = \mathbf{0}$ and $\boldsymbol{v} \neq \mathbf{0}$. Note that the forward directional derivative

$$\lim_{t \downarrow 0} \frac{\|t\boldsymbol{v}\| - 0}{t} = \|\boldsymbol{v}\|$$

does exist for all $\boldsymbol{v}$.

26. Suppose the vector-valued function $\boldsymbol{x}(t)$ is differentiable in the scalar t. Show that $\|\boldsymbol{x}(t)\|' \leq \|\boldsymbol{x}'(t)\|$ whenever $\boldsymbol{x}(t) \neq \mathbf{0}$.

27. Continuing Problem 2, show that $f_4(x)$ is continuously differentiable on R, that its image is an open subset of R, and yet $f_4(x)$ is not one-to-one on any interval around 0. What bearing does this have on the inverse function theorem?

28. Demonstrate that the function
$$f(x) \quad = \quad \begin{cases} x + 2x^2 \sin\left(\frac{1}{x}\right) & x \neq 0 \\ 0 & x = 0 \end{cases}$$
has a bounded derivative $f'(x)$ for all x and $f'(0) \neq 0$, yet $f(x)$ is not one-to-one on any interval around 0. Why does the inverse function theorem fail in this case?

29. Consider the equation $f(x) = tg(x)$ determined by the continuously differentiable functions $f(x)$ and $g(x)$ from R into R. If $f(0) = 0$ and $f'(0) \neq 0$, then show that in a suitably small interval $|t| < \delta$ there is a unique continuously differentiable function $x(t)$ solving the equation and satisfying $x(0) = 0$. Prove that $x'(0) = g(0)/f'(0)$.

30. Suppose that the differential $df(\boldsymbol{x})$ of the continuously differentiable function $f(\boldsymbol{x})$ from R^m to R^n has full rank at $\boldsymbol{y}$. Show that $f(\boldsymbol{x})$ is one-to-one in a neighborhood of $\boldsymbol{y}$. Note that $m \leq n$ must hold.

31. The mean value inequality (4.12) can be improved. Suppose that along the line segment $\{\boldsymbol{u} = \boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}) : t \in [0, 1]\}$ the gradient $\nabla f(\boldsymbol{u})$ satisfies the Lipschitz inequality
$$\|\nabla f(\boldsymbol{u}) - \nabla f(\boldsymbol{v})\| \quad \leq \quad \lambda \|\boldsymbol{u} - \boldsymbol{v}\| \tag{4.19}$$
for some constant $\lambda \geq 0$. This is the case if the second differential $d^2 f(\boldsymbol{u})$ exists and is continuous in $\boldsymbol{u}$. Prove that
$$\|f(\boldsymbol{y}) - f(\boldsymbol{x}) - df(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})\| \quad \leq \quad \frac{\lambda}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2.$$
(Hint: Invoke the fundamental theorem of calculus, and bound the norm of the integrand by inequality (4.19).)

32. Demonstrate that the bound in Problem 31 can be generalized to
$$\|f(\boldsymbol{u}) - f(\boldsymbol{v}) - df(\boldsymbol{x})(\boldsymbol{u} - \boldsymbol{v})\| \quad \leq \quad \frac{\lambda}{2}(\|\boldsymbol{u} - \boldsymbol{x}\| + \|\boldsymbol{v} - \boldsymbol{x}\|)\|\boldsymbol{u} - \boldsymbol{v}\|$$
for any triple of points $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{x}$ contained in the convex domain of the function $f(\boldsymbol{x})$. Note the similarity to the straddle inequality (3.8).

33. Given the result of Problem 32, suppose that the domain and range of $f(\boldsymbol{x})$ are both contained in R^n and that the matrix $df(\boldsymbol{x})$ is invertible. Show that there exist positive constants α, β, and ϵ such that
$$\alpha\|\boldsymbol{u} - \boldsymbol{v}\| \quad \leq \quad \|f(\boldsymbol{u}) - f(\boldsymbol{v})\| \quad \leq \quad \beta\|\boldsymbol{u} - \boldsymbol{v}\|$$

for all $\boldsymbol{u}$ and $\boldsymbol{v}$ with $\max\{\|\boldsymbol{u}-\boldsymbol{x}\|, \|\boldsymbol{v}-\boldsymbol{x}\|\} \leq \epsilon$. (Hints: Write

$$f(\boldsymbol{u}) - f(\boldsymbol{v}) \;=\; f(\boldsymbol{u}) - f(\boldsymbol{v}) - df(\boldsymbol{x})(\boldsymbol{u}-\boldsymbol{v}) + df(\boldsymbol{x})(\boldsymbol{u}-\boldsymbol{v}),$$

and apply the bound $\|\boldsymbol{u}-\boldsymbol{v}\| \leq \|df(\boldsymbol{x})^{-1}\| \cdot \|df(\boldsymbol{x})(\boldsymbol{u}-\boldsymbol{v})\|$.)

34. Consider the function $f(\boldsymbol{M}) = \boldsymbol{M} + \boldsymbol{M}^2$ defined on $n \times n$ matrices. Show that the range of $f(\boldsymbol{M})$ contains a neighborhood of the trivial matrix $\boldsymbol{0}$ [69]. (Hint: Compute the differential of $f(\boldsymbol{M})$ and apply the inverse function theorem.)

35. Calculate the differentials of the matrix-valued functions $e^{\boldsymbol{M}}$ and $\ln(\boldsymbol{I}-\boldsymbol{M})$. What are the values of these differentials at $\boldsymbol{M} = \boldsymbol{0}$?

36. For a differentiable matrix $M(\boldsymbol{x})$, verify the partial derivatives

$$\begin{aligned}
\partial_k(MM^*) &= (\partial_k M)M^* + M(\partial_k M)^* \\
\partial_k(M^*M) &= (\partial_k M)^*M + M^*(\partial_k M) \\
\partial_k M^p &= \sum_{j=1}^{p} M^{j-1}\partial_k M M^{p-j}, \quad p > 0 \\
\partial_k M^{-p} &= -\sum_{j=1}^{p} M^{-j}\partial_k M M^{-p+j-1}, \quad p > 0 \\
\partial_k \operatorname{tr}(MM^*) &= 2\operatorname{tr}(M^*\partial_k M) \\
\partial_k \operatorname{tr}(M^*M) &= 2\operatorname{tr}(M^*\partial_k M) \\
\partial_k \operatorname{tr}(M^p) &= p\operatorname{tr}(M^{p-1}\partial_k M) \\
\partial_k \det(MM^*) &= 2\det(MM^*)\operatorname{tr}[M^*(MM^*)^{-1}\partial_k M] \\
\partial_k \det(M^*M) &= 2\det(M^*M)\operatorname{tr}[(M^*M)^{-1}M^*\partial_k M] \\
\partial_k \det(M^p) &= p\det(M^p)\operatorname{tr}(M^{-1}\partial_k M),
\end{aligned}$$

where p is an integer and $\partial_k = \frac{\partial}{\partial x_k}$.

37. A random $m \times m$ matrix $\boldsymbol{W}$ with density

$$f(\boldsymbol{w}) \;=\; \frac{|\det \boldsymbol{w}|^{(n-m-1)/2} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{w})}}{2^{nm/2}\pi^{m(m-1)/4}|\det \boldsymbol{\Sigma}|^{n/2}\prod_{i=1}^{m}\Gamma[(n+1-i)/2]}$$

is said to be Wishart distributed [3]. Here $\boldsymbol{\Sigma}$ is positive definite and n is an integer with $n > m$. It turns out that the inverse matrix $\boldsymbol{V} = \boldsymbol{W}^{-1}$ has density

$$g(\boldsymbol{v}) \;=\; \frac{|\det \boldsymbol{v}|^{-(n+m+1)/2} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{v}^{-1})}}{2^{nm/2}\pi^{m(m-1)/4}|\det \boldsymbol{\Sigma}|^{n/2}\prod_{i=1}^{m}\Gamma[(n+1-i)/2]}.$$

This naturally is called the inverse Wishart density. Demonstrate that the modes of the Wishart and inverse Wishart densities occur at $\boldsymbol{w} = (n - m - 1)\boldsymbol{\Sigma}$ and $\boldsymbol{v} = (n + m + 1)^{-1}\boldsymbol{\Sigma}^{-1}$, respectively. (Hints: Show that these points are stationary points of the log densities by considering the function (4.18).)

5
Karush-Kuhn-Tucker Theory

5.1 Introduction

In the current chapter, we study the problem of minimizing a real-valued function $f(\boldsymbol{x})$ subject to the constraints

$$\begin{aligned} g_i(\boldsymbol{x}) &= 0, \qquad 1 \le i \le p \\ h_j(\boldsymbol{x}) &\le 0, \qquad 1 \le j \le q. \end{aligned}$$

All of these functions share some open set $U \subset \mathsf{R}^n$ as their domain. Maximizing $f(\boldsymbol{x})$ is equivalent to minimizing $-f(\boldsymbol{x})$, so there is no loss of generality in considering minimization. The function $f(\boldsymbol{x})$ is called the objective function, the functions $g_i(\boldsymbol{x})$ are called equality constraints, and the functions $h_j(\boldsymbol{x})$ are called inequality constraints. Any point $\boldsymbol{x} \in U$ satisfying all of the constraints is said to be feasible. A constraint $h_j(\boldsymbol{x})$ is active at the feasible point $\boldsymbol{x}$ provided $h_j(\boldsymbol{x}) = 0$; it is inactive if $h_j(\boldsymbol{x}) < 0$. In general, we will assume that the feasible region is nonempty. The case $p = 0$ of no equality constraints and the case $q = 0$ of no inequality constraints are both allowed.

In exploring solutions to the above constrained minimization problem, we will meet a generalization of the Lagrange multiplier rule fashioned independently by Karush [149] and Kuhn and Tucker [161]. Under fairly weak regularity conditions, the rule holds at all extrema. In contrast to this necessary condition, sufficient conditions for an extremum involve second derivatives. To state and prove the most useful sufficient condition, we

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_5,

must confront second differentials and what it means for a function to be twice differentiable. The matter is straightforward conceptually but computationally messy. Fortunately, we can build on the material presented in Chap. 4.

5.2 The Multiplier Rule

Before embarking on the long and interesting proof of the multiplier rule, we turn to linear programming as a specific example of constrained optimization. A huge literature has grown up around this single application.

Example 5.2.1 *Linear Programming*

If the objective function $f(\boldsymbol{x})$ and the constraints $g_i(\boldsymbol{x})$ and $h_j(\boldsymbol{x})$ are all affine functions $\boldsymbol{z}^*\boldsymbol{x} + c$, then the constrained minimization problem is termed linear programming. In the literature on linear programming, the standard linear program is posed as one of minimizing $f(\boldsymbol{x}) = \boldsymbol{z}^*\boldsymbol{x}$ subject to the linear equality constraints $\boldsymbol{V}\boldsymbol{x} = \boldsymbol{d}$ and the nonnegativity constraints $x_i \geq 0$ for all $1 \leq i \leq n$. The inequality constraints are collectively abbreviated as $\boldsymbol{x} \geq \boldsymbol{0}$. To show that the standard linear program encompasses our apparently more general version of linear programming, we note first that we can omit the affine constant in the objective function $f(\boldsymbol{x})$. The p linear equality constraints

$$0 \;=\; g_i(\boldsymbol{x}) \;=\; \boldsymbol{v}_i^*\boldsymbol{x} - d_i$$

are already in the form $\boldsymbol{V}\boldsymbol{x} = \boldsymbol{d}$ if we define $\boldsymbol{V}$ to be the $p \times n$ matrix with ith row $\boldsymbol{v}_i^*$ and $\boldsymbol{d}$ to be the $p \times 1$ vector with ith entry d_i. The inequality constraint $h_j(\boldsymbol{x}) \leq 0$ can be elevated to an equality constraint $h_j(\boldsymbol{x}) + y_j = 0$ by introducing an additional variable y_j called a slack variable with the stipulation that $y_j \geq 0$. If any of the variables x_i is not already constrained by $x_i \geq 0$, then we can introduce what are termed free variables $u_i \geq 0$ and $w_i \geq 0$ so that $x_i = u_i - w_i$ and replace x_i everywhere by this difference. ■

In proving the multiplier rule, it turns out that one must restrict the behavior of the constraints at a local extremum to avoid redundant constraints. There are several constraint qualifications.

Definition 5.2.1 *Mangasarian-Fromovitz Constraint Qualification*

This condition holds at a feasible point $\boldsymbol{x}$ provided the differentials $dg_i(\boldsymbol{x})$ are linearly independent and there exists a vector v with $dg_i(\boldsymbol{x})\boldsymbol{v} = 0$ for all i and $dh_j(\boldsymbol{x})\boldsymbol{v} < 0$ for all inequality constraints $h_j(\boldsymbol{x})$ active at $\boldsymbol{x}$ [186]. The vector $\boldsymbol{v}$ is a tangent vector in the sense that infinitesimal motion from $\boldsymbol{x}$ along $\boldsymbol{v}$ stays within the feasible region. ■

Because the Mangasarian-Fromovitz condition is difficult to check, we will consider the simpler sufficient condition of Kuhn and Tucker [161] in the next section. In the meantime, we state and prove the Lagrange multiplier rule extended to inequality constraints by Karush and Kuhn and Tucker. Our proof reproduces McShane's lovely argument, which substitutes penalties for constraints [116, 194].

Proposition 5.2.1 *Suppose the objective function $f(\boldsymbol{y})$ of the constrained optimization problem has a local minimum at the feasible point $\boldsymbol{x}$. If $f(\boldsymbol{y})$ and the various constraint functions are continuously differentiable near $\boldsymbol{x}$, then there exists a unit vector of Lagrange multipliers $\lambda_0, \ldots, \lambda_p$ and $\mu_1, \ldots, \mu_q$ such that*

$$\lambda_0 \nabla f(\boldsymbol{x}) + \sum_{i=1}^{p} \lambda_i \nabla g_i(\boldsymbol{x}) + \sum_{j=1}^{q} \mu_j \nabla h_j(\boldsymbol{x}) \quad = \quad \mathbf{0}. \tag{5.1}$$

Moreover, each of the multipliers λ_0 and μ_j is nonnegative, and $\mu_j = 0$ whenever $h_j(\boldsymbol{x}) < 0$. If the constraint functions satisfy the Mangasarian-Fromovitz constraint qualification at $\boldsymbol{x}$, then we can take $\lambda_0 = 1$.

Proof: Without loss of generality, we assume $\boldsymbol{x} = \mathbf{0}$ and $f(\mathbf{0}) = 0$. By renumbering the inequality constraints if necessary, we also suppose that the first r of them are active at $\mathbf{0}$ and the last $q - r$ of them are inactive at $\mathbf{0}$. Now choose $\delta > 0$ so that (a) the closed ball

$$C(\mathbf{0}, \delta) \quad = \quad \{\boldsymbol{y} \in \mathsf{R}^n : \|\boldsymbol{y}\| \leq \delta\}$$

is contained in the open domain U, (b) $\mathbf{0}$ is the minimum point of $f(\boldsymbol{y})$ in $C(\mathbf{0}, \delta)$ subject to the constraints, (c) the objective and constraint functions are continuously differentiable in $C(\mathbf{0}, \delta)$, and (d) the constraints $h_j(\boldsymbol{x})$ inactive at $\mathbf{0}$ are inactive throughout $C(\mathbf{0}, \delta)$.

On the road to our ultimate goal, consider the functions

$$h_{j+}(\boldsymbol{y}) \quad = \quad \max\{h_j(\boldsymbol{y}), 0\}.$$

Using these functions, we now prove that for each $0 < \epsilon \leq \delta$, there exists an $\alpha > 0$ such that

$$f(\boldsymbol{y}) + \|\boldsymbol{y}\|^2 + \alpha \sum_{i=1}^{p} g_i(\boldsymbol{y})^2 + \alpha \sum_{j=1}^{r} h_{j+}(\boldsymbol{y})^2 \quad > \quad 0 \tag{5.2}$$

for all $\boldsymbol{y}$ with $\|\boldsymbol{y}\| = \epsilon$. This is not an entirely trivial claim to prove because $f(\boldsymbol{y})$ can be negative on $C(\mathbf{0}, \delta)$ outside the feasible region.

Suppose the claim is false. Then there is a sequence of points $\boldsymbol{y}_m$ with $\|\boldsymbol{y}_m\| = \epsilon$ and a sequence of numbers α_m tending to ∞ such that

$$f(\boldsymbol{y}_m) + \|\boldsymbol{y}_m\|^2 \quad \leq \quad -\alpha_m \sum_{i=1}^{p} g_i(\boldsymbol{y}_m)^2 - \alpha_m \sum_{j=1}^{r} h_{j+}(\boldsymbol{y}_m)^2. \tag{5.3}$$

Because the boundary of $C(\mathbf{0},\epsilon)$ is compact, the sequence $\boldsymbol{y}_m$ has a convergent subsequence, which without loss of generality we take to be the sequence itself. The limit $\boldsymbol{z}$ of the sequence clearly has norm $\|\boldsymbol{z}\| = \epsilon$. Dividing both sides of inequality (5.3) by $-\alpha_m$ and sending m to ∞ produce

$$\sum_{i=1}^{p} g_i(\boldsymbol{z})^2 + \sum_{j=1}^{r} h_{j+}(\boldsymbol{z})^2 \quad = \quad 0. \tag{5.4}$$

It follows that $\boldsymbol{z}$ is feasible with $f(\boldsymbol{z}) \geq f(\mathbf{0}) = 0$. However, inequality (5.3) requires that each $f(\boldsymbol{y}_m) \leq -\epsilon^2$. Since this last relation is preserved in the limit, we reach a contradiction and consequently establish the validity of inequality (5.2).

Our next goal is to prove that there exists a point $\boldsymbol{u}$ and a unit vector $(\lambda_0, \lambda_1, \ldots, \lambda_p, \mu_1, \ldots, \mu_r)^*$ such that (a) $\|\boldsymbol{u}\| < \epsilon$, (b) each of the multipliers λ_0 and μ_j is nonnegative, and (c)

$$\lambda_0[\nabla f(\boldsymbol{u}) + 2\boldsymbol{u}] + \sum_{i=1}^{p} \lambda_i \nabla g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \mu_j \nabla h_j(\boldsymbol{u}) \quad = \quad \mathbf{0}. \tag{5.5}$$

Observe here that the distinction between active and inactive constraints comes into play again. To prove the Lagrange multiplier rule (5.5), define

$$F(\boldsymbol{y}) \quad = \quad f(\boldsymbol{y}) + \|\boldsymbol{y}\|^2 + \alpha \sum_{i=1}^{p} g_i(\boldsymbol{y})^2 + \alpha \sum_{j=1}^{r} h_{j+}(\boldsymbol{y})^2$$

using the α satisfying condition (5.2). For typographical reasons, we omit the dependence of α on ϵ.

Given that $F(\boldsymbol{y})$ is continuous, there is a point $\boldsymbol{u}$ giving the unconstrained minimum of $F(\boldsymbol{y})$ on the compact set $C(\mathbf{0},\epsilon)$. Because this point satisfies $F(\boldsymbol{u}) \leq F(\mathbf{0}) = 0$, it is impossible that $\|\boldsymbol{u}\| = \epsilon$ in view of inequality (5.2). Thus, $\boldsymbol{u}$ falls in the interior of $C(\mathbf{0},\epsilon)$ where $\nabla F(\boldsymbol{u}) = \mathbf{0}$ must occur. The gradient condition $\nabla F(\boldsymbol{u}) = \mathbf{0}$ can be expressed as

$$\nabla f(\boldsymbol{u}) + 2\boldsymbol{u} + \alpha \sum_{i=1}^{p} 2g_i(\boldsymbol{u}) \nabla g_i(\boldsymbol{u}) + \alpha \sum_{j=1}^{r} 2h_{j+}(\boldsymbol{u}) \nabla h_j(\boldsymbol{u}) \quad = \quad \mathbf{0}, \tag{5.6}$$

invoking the differentiability of the functions $h_{j+}(\boldsymbol{y})^2$ derived in Example 4.4.3 of Chap. 4. If we divide equality (5.6) by the norm of the vector

$$\boldsymbol{v} \quad = \quad [1, 2\alpha g_1(\boldsymbol{u}), \ldots, 2\alpha g_p(\boldsymbol{u}), 2\alpha h_{1+}(\boldsymbol{u}), \ldots, 2\alpha h_{r+}(\boldsymbol{u})]^*$$

and redefine the Lagrange multipliers accordingly, then the multiplier rule (5.5) holds with each of the multipliers λ_0 and μ_j nonnegative.

Now choose a sequence $\epsilon_m > 0$ tending to 0 and corresponding points $\boldsymbol{u}_m$ where the Lagrange multiplier rule (5.5) holds. The sequence of unit

vectors $(\lambda_{m0}, \ldots, \lambda_{mp}, \mu_{m1}, \ldots, \mu_{mr})^*$ has a convergent subsequence with limit $(\lambda_0, \ldots, \lambda_p, \mu_1, \ldots, \mu_r)^*$ that is also a unit vector. Replacing the sequence $\boldsymbol{u}_m$ by the corresponding subsequence $\boldsymbol{u}_{m_k}$ allows us to take limits along $\boldsymbol{u}_{m_k}$ in equality (5.5) and achieve equality (5.1). Observe that $\boldsymbol{u}_{m_k}$ converges to $\mathbf{0}$ because $\|\boldsymbol{u}_{m_k}\| \leq \epsilon_{m_k}$.

Finally, suppose the constraint qualification holds at the local minimum $\mathbf{0}$ and that $\lambda_0 = 0$. If all of the nonnegative multipliers μ_j are 0, then at least one of the λ_i with $1 \leq i \leq p$ is not 0. But this contradicts the linear independence of the $dg_i(\mathbf{0})$. Now consider the vector $\boldsymbol{v}$ guaranteed by the constraint qualification. Taking its inner product with both sides of equation (5.1) gives

$$\sum_{j=1}^{r} \mu_j dh_j(\mathbf{0})\boldsymbol{v} \quad = \quad 0,$$

contradicting the assumption that $dh_j(\mathbf{0})\boldsymbol{v} < 0$ for all $1 \leq j \leq r$ and the fact that at least one $\mu_j > 0$. Thus, $\lambda_0 > 0$, and we can divide equation (5.1) by λ_0. ■

Example 5.2.2 *Application to an Inequality*

Let us demonstrate the inequality

$$\frac{x_1^2 + x_2^2}{4} \quad \leq \quad e^{x_1+x_2-2}$$

subject to the constraints $x_1 \geq 0$ and $x_2 \geq 0$ [69]. It suffices to show that the minimum of $f(\boldsymbol{x}) = -(x_1^2 + x_2^2)e^{-x_1-x_2}$ is $-4e^{-2}$. According to Proposition 5.2.1 with $h_1(\boldsymbol{x}) = -x_1$ and $h_2(\boldsymbol{x}) = -x_2$, a minimum point entails the conditions

$$\begin{aligned} -\frac{\partial}{\partial x_1} f(\boldsymbol{x}) \quad &= \quad (2x_1 - x_1^2 - x_2^2)e^{-x_1-x_2} \quad = \quad -\mu_1 \\ -\frac{\partial}{\partial x_2} f(\boldsymbol{x}) \quad &= \quad (2x_2 - x_1^2 - x_2^2)e^{-x_1-x_2} \quad = \quad -\mu_2, \end{aligned}$$

where the multipliers μ_1 and μ_2 are nonnegative and satisfy $\mu_1 x_1 = 0$ and $\mu_2 x_2 = 0$. In this problem, the Mangasarian-Fromovitz constraint qualification is trivial to check using the vector $\boldsymbol{v} = \mathbf{1}$. If neither x_1 nor x_2 vanishes, then

$$2x_1 - x_1^2 - x_2^2 \quad = \quad 2x_2 - x_1^2 - x_2^2 \quad = \quad 0.$$

This forces $x_1 = x_2$ and $2x_1 - 2x_1^2 = 0$. It follows that $x_1 = x_2 = 1$, where $f(\mathbf{1}) = -2e^{-2}$. We can immediately eliminate the origin $\mathbf{0}$ from contention because $f(\mathbf{0}) = 0$. If $x_1 = 0$ and $x_2 > 0$, then $\mu_2 = 0$ and $2x_2 - x_2^2 = 0$. This implies that $x_2 = 2$ and $(0, 2)$ is a candidate minimum point. By symmetry, $(2, 0)$ is also a candidate minimum point. At these two boundary points, $f(2, 0) = f(0, 2) = -4e^{-2}$, and this verifies the claimed minimum value. ■

Example 5.2.3 *Application to Linear Programming*

The gradient of the Lagrangian

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad = \quad \boldsymbol{z}^*\boldsymbol{x} + \sum_{i=1}^{p} \lambda_i \Big(\sum_{j=1}^{n} v_{ij} x_j - d_i \Big) - \sum_{j=1}^{q} \mu_j x_j$$

vanishes at the minimum of the linear function $f(\boldsymbol{x}) = \boldsymbol{z}^*\boldsymbol{x}$ subject to the constraints $\boldsymbol{V}\boldsymbol{x} = \boldsymbol{d}$ and $\boldsymbol{x} \geq \boldsymbol{0}$. Differentiating $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ with respect to x_j and setting the result equal to 0 gives $z_j + \sum_{i=1}^{p} \lambda_i v_{ij} - \mu_j = 0$. In vector notation this is just

$$\boldsymbol{z} + \boldsymbol{V}^*\boldsymbol{\lambda} - \boldsymbol{\mu} \quad = \quad \boldsymbol{0}$$

subject to the restrictions $\boldsymbol{\mu} \geq \boldsymbol{0}$ and $\boldsymbol{\mu}^*\boldsymbol{x} = 0$. We will revisit linear programming later and discuss an interior point algorithm and dual linear programs. ■

Example 5.2.4 *A Counterexample to the Multiplier Rule*

The Lagrange multiplier condition is necessary but not sufficient for a point to furnish a minimum. For example, consider the function $f(\boldsymbol{x}) = x_1^3 - x_2$ subject to the constraint $h(\boldsymbol{x}) = x_2 \leq 0$. The Lagrange multiplier condition

$$\nabla f(\boldsymbol{0}) \quad = \quad \begin{pmatrix} 0 \\ -1 \end{pmatrix} \quad = \quad -\nabla h(\boldsymbol{0})$$

holds, but the origin $\boldsymbol{0}$ fails to minimize $f(\boldsymbol{x})$. Indeed, the one-dimensional slice $x_1 \mapsto f(x_1, 0)$ has a saddle point at $x_1 = 0$. This function has no minimum subject to the inequality constraint. ■

Example 5.2.5 *Shadow Values*

In economic applications the objective function $f(\boldsymbol{x})$ is often viewed as a profit or a cost, and the constraints equation $g_i(\boldsymbol{x}) = 0$ is rephrased as $g_i(\boldsymbol{x}) = c_i$, where c_i is the amount of some available resource. In the absence of inequality constraints, the negative Lagrange multiplier $-\lambda_i$ is called a shadow value or price and measures the rate of change of the optimal value of $f(\boldsymbol{x})$ relative to c_i. Let $x(\boldsymbol{c})$ denote the optimal point as a function of the constraint vector $c = (c_1, \ldots, c_p)^*$. If we assume that $x(\boldsymbol{c})$ is differentiable, then we can multiply the equation

$$df[x(\boldsymbol{c})] + \sum_{i=1}^{p} \lambda_i dg_i[x(\boldsymbol{c})] \quad = \quad \boldsymbol{0}^*$$

on the right by $dx(\boldsymbol{c})$ and recover via the chain rule an equation relating the differentials of $f[x(\boldsymbol{c})]$ and the $g_i[x(\boldsymbol{c})]$ with respect to $\boldsymbol{c}$. Because it is obvious that

$$\frac{\partial}{\partial c_j} g_i[x(\boldsymbol{c})] \quad = \quad 1_{\{j=i\}},$$

it follows that

$$\frac{\partial}{\partial c_j} f[x(\boldsymbol{c})] + \lambda_j = 0.$$

Of course, this result is valid generally and transcends its narrow economic origin. ■

Example 5.2.6 *Quadratic Programming with Equality Constraints*

To minimize the quadratic function $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x} + c$ subject to the linear equality constraints $\boldsymbol{V}\boldsymbol{x} = \boldsymbol{d}$, we introduce the Lagrangian

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) &= \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x} + \sum_{i=1}^{p} \lambda_i[\boldsymbol{v}_i^*\boldsymbol{x} - d_i] \\ &= \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x} + \boldsymbol{\lambda}^*(\boldsymbol{V}\boldsymbol{x} - \boldsymbol{d}).\end{aligned}$$

A stationary point of $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ is determined by the equations

$$\begin{aligned}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{V}^*\boldsymbol{\lambda} &= \boldsymbol{0} \\ \boldsymbol{V}\boldsymbol{x} &= \boldsymbol{d},\end{aligned}$$

whose formal solution amounts to

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{V}^* \\ \boldsymbol{V} & \boldsymbol{0} \end{pmatrix}^{-1} \begin{pmatrix} -\boldsymbol{b} \\ \boldsymbol{d} \end{pmatrix}.$$

The next proposition shows that the indicated matrix inverse exists when $\boldsymbol{A}$ is positive definite. ■

Proposition 5.2.2 *Let $\boldsymbol{A}$ be an $n \times n$ positive definite matrix and $\boldsymbol{V}$ be a $p \times n$ matrix. Then the matrix*

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{V}^* \\ \boldsymbol{V} & \boldsymbol{0} \end{pmatrix}$$

is invertible if and only if $\boldsymbol{V}$ has linearly independent rows $\boldsymbol{v}_1^*, \ldots, \boldsymbol{v}_p^*$. When this condition holds, $\boldsymbol{M}$ has inverse

$$\boldsymbol{M}^{-1} = \begin{pmatrix} \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{V}^*(\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{V}^*)^{-1}\boldsymbol{V}\boldsymbol{A}^{-1} & \boldsymbol{A}^{-1}\boldsymbol{V}^*(\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{V}^*)^{-1} \\ (\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{V}^*)^{-1}\boldsymbol{V}\boldsymbol{A}^{-1} & -(\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{V}^*)^{-1} \end{pmatrix}.$$

Proof: We first show that the symmetric matrix $\boldsymbol{M}$ is invertible with the specified inverse if and only if $(\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{V}^*)^{-1}$ exists. If $\boldsymbol{M}^{-1}$ exists, it is necessarily symmetric. Indeed, taking the transpose of $\boldsymbol{M}\boldsymbol{M}^{-1} = \boldsymbol{I}$ gives $(\boldsymbol{M}^{-1})^*\boldsymbol{M} = I$. Suppose $\boldsymbol{M}^{-1}$ has block form $\begin{pmatrix} \boldsymbol{B} & \boldsymbol{C}^* \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}$. Then the identity

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{V}^* \\ \boldsymbol{V} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{B} & \boldsymbol{C}^* \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_p \end{pmatrix}$$

implies that $\boldsymbol{V}\boldsymbol{C}^* = \boldsymbol{I}_p$ and $\boldsymbol{A}\boldsymbol{C}^* + \boldsymbol{V}^*\boldsymbol{D} = \boldsymbol{0}$. Multiplying the last equality by $\boldsymbol{V}\boldsymbol{A}^{-1}$ gives $\boldsymbol{I}_p = -\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{V}^*\boldsymbol{D}$. Thus, $(\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{V}^*)^{-1}$ exists. Conversely, if $(\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{V}^*)^{-1}$ exists, then one can check by direct multiplication that $\boldsymbol{M}$ has the claimed inverse.

If $(\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{V}^*)^{-1}$ exists, then $\boldsymbol{V}$ must have full row rank p. Conversely, if $\boldsymbol{V}$ has full row rank p, take any nontrivial $\boldsymbol{u} \in \mathsf{R}^p$. The fact

$$\boldsymbol{u}^*\boldsymbol{V} \;=\; u_1\boldsymbol{v}_1^* + \cdots u_p\boldsymbol{v}_p^* \;\neq\; \boldsymbol{0}^*$$

and the positive definiteness of $\boldsymbol{A}$ imply $\boldsymbol{u}^*\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{V}^*\boldsymbol{u} > 0$. Thus, $\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{V}^*$ is positive definite and invertible. ■

Example 5.2.7 *Smallest Matrix Subject to Secant Conditions*

In some situations covered by Example 5.2.6, the answer can be radically simplified. Consider the problem of minimizing the Frobenius norm of a matrix $\boldsymbol{M}$ subject to the linear constraints $\boldsymbol{M}\boldsymbol{u}_i = \boldsymbol{v}_i$ for $i = 1, \ldots, q$. It is helpful to rewrite the constraints in matrix form as $\boldsymbol{M}\boldsymbol{U} = \boldsymbol{V}$ for $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q)$ and $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q)$. Provided $\boldsymbol{U}$ has full column rank q, the minimum of the squared norm $\|\boldsymbol{M}\|_F^2$ subject to the constraints is attained by the choice $\boldsymbol{M} = \boldsymbol{V}(\boldsymbol{U}^*\boldsymbol{U})^{-1}\boldsymbol{U}^*$. We can prove this assertion by taking the partial derivative of the Lagrangian

$$\begin{aligned}
\mathcal{L} &= \frac{1}{2}\|\boldsymbol{M}\|_F^2 + \sum_i \sum_k \lambda_{ik}\Big(\sum_j m_{ij}u_{jk} - v_{ik}\Big) \\
&= \frac{1}{2}\sum_i \sum_j m_{ij}^2 + \sum_i \sum_k \lambda_{ik}\Big(\sum_j m_{ij}u_{jk} - v_{ik}\Big)
\end{aligned}$$

with respect to m_{ij} and equating it to 0. This gives the Lagrange multiplier equation

$$0 \;=\; m_{ij} + \sum_k \lambda_{ik}u_{jk},$$

which we collectively express in matrix notation as $\boldsymbol{0} = \boldsymbol{M} + \boldsymbol{\Lambda}\boldsymbol{U}^*$. This equation and the constraint equation $\boldsymbol{M}\boldsymbol{U} = \boldsymbol{V}$ uniquely determine the minimum of the objective function. Indeed, straightforward substitution shows that $\boldsymbol{M} = \boldsymbol{V}(\boldsymbol{U}^*\boldsymbol{U})^{-1}\boldsymbol{U}^*$ and $\boldsymbol{\Lambda} = -\boldsymbol{V}(\boldsymbol{U}^*\boldsymbol{U})^{-1}$ constitute the solution. This result will come in handy later when we discuss accelerating various algorithms. ■

5.3 Constraint Qualification

Fortunately, the Mangasarian-Fromovitz constraint qualification is a consequence of a stronger condition suggested by Kuhn and Tucker. This alternative condition requires the differentials $dg_i(\boldsymbol{x})$ of the equality constraints and the differentials $dh_j(\boldsymbol{x})$ of the active inequality constraints to

be collectively linearly independent at $\boldsymbol{x}$. When the Kuhn-Tucker condition is true, we can find an $n \times n$ invertible matrix $\boldsymbol{A}$ whose first p rows are the vectors $dg_i(\boldsymbol{x})$ and whose next r rows are the vectors $dh_1(\boldsymbol{x}), \ldots, dh_r(\boldsymbol{x})$ corresponding to the active inequality constraints. For example, we can choose the last $n - p - r$ rows of $\boldsymbol{A}$ to be a basis for the orthogonal complement of the subspace spanned by the first $p + r$ rows of $\boldsymbol{A}$. Given that $\boldsymbol{A}$ is invertible, there certainly exists a vector $\boldsymbol{v} \in \mathsf{R}^n$ with

$$\boldsymbol{A}\boldsymbol{v} \quad = \quad -\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{1} \\ \boldsymbol{0} \end{pmatrix},$$

where the column vectors on the right side of this equality have p, r, and $n-p-r$ rows, respectively. Clearly, the vector $\boldsymbol{v}$ satisfies the Mangasarian-Fromovitz constraint qualification.

Even if the Kuhn-Tucker condition fails, there is still a chance that the constraint qualification holds. Let $\boldsymbol{G}$ be the $p \times n$ matrix with rows $dg_i(\boldsymbol{x})$. Assume that $\boldsymbol{G}$ has full rank. Taking the orthogonal complement of the subspace spanned by the p rows of $\boldsymbol{G}$, we can construct an $n \times (n-p)$ matrix $\boldsymbol{K}$ of full rank whose columns are orthogonal to the rows of $\boldsymbol{G}$. This fact can be expressed as $\boldsymbol{G}\boldsymbol{K} = \boldsymbol{0}$ and implies that the image of the linear transformation $\boldsymbol{K}$ from R^{n-p} to R^n is the kernel or null space of $\boldsymbol{G}$. In other words, any $\boldsymbol{v}$ satisfying $\boldsymbol{G}\boldsymbol{v} = \boldsymbol{0}$ can be expressed as $\boldsymbol{v} = \boldsymbol{K}\boldsymbol{u}$ and vice versa. If we let $\boldsymbol{z}_j^* = dh_j(\boldsymbol{x})\boldsymbol{K}$ for each of the active inequality constraints, then the Mangasarian-Fromovitz constraint qualification is equivalent to the existence of a vector $\boldsymbol{u} \in \mathsf{R}^{n-p}$ satisfying $\boldsymbol{z}_j^*\boldsymbol{u} < 0$ for all $1 \leq j \leq r$. If the number of equality constraints $p = 0$, then we take $\boldsymbol{z}_j^* = dh_j(\boldsymbol{x})$.

The next proposition paves the way for proving a necessary and sufficient condition for this equality-free form of the Mangasarian-Fromovitz constraint qualification. The result described by the proposition is of independent interest; its proof illustrates the fact that adding small penalties as opposed to large penalties is sometimes helpful [84].

Proposition 5.3.1 (Ekeland) *Suppose the real-valued function $f(\boldsymbol{x})$ is defined and differentiable on R^n. If $f(\boldsymbol{x})$ is bounded below, then there are points where $\|\nabla f(\boldsymbol{x})\|$ is arbitrarily close to 0.*

Proof: Take any small $\epsilon > 0$ and define the continuous function

$$f_\epsilon(\boldsymbol{x}) \quad = \quad f(\boldsymbol{x}) + \epsilon\|\boldsymbol{x}\|.$$

In view of the boundedness condition, for any $\boldsymbol{y}$ the set

$$\{\boldsymbol{x} : f_\epsilon(\boldsymbol{x}) \leq f_\epsilon(\boldsymbol{y})\}$$

is compact. Hence, Proposition 2.5.4 implies that $f_\epsilon(\boldsymbol{x})$ has a global minimum $\boldsymbol{z}$ depending on ϵ. We now prove that $\boldsymbol{v} = \nabla f(\boldsymbol{z})$ satisfies $\|\boldsymbol{v}\| \leq \epsilon$. If the opposite is true, then the limit relation

$$\begin{aligned}\lim_{t\downarrow 0}\frac{f(\boldsymbol{z}-t\boldsymbol{v})-f(\boldsymbol{z})}{t} &= -df(\boldsymbol{z})\boldsymbol{v}\\ &= -\|\boldsymbol{v}\|^2\\ &< -\epsilon\|\boldsymbol{v}\|,\end{aligned}$$

the choice of $\boldsymbol{z}$, and the triangle inequality together entail

$$\begin{aligned}-t\epsilon\|\boldsymbol{v}\| &> f(\boldsymbol{z}-t\boldsymbol{v})-f(\boldsymbol{z})\\ &= f_\epsilon(\boldsymbol{z}-t\boldsymbol{v})-f_\epsilon(\boldsymbol{z})+\epsilon\|\boldsymbol{z}\|-\epsilon\|\boldsymbol{z}-t\boldsymbol{v}\|\\ &\geq \epsilon\|\boldsymbol{z}\|-\epsilon\|\boldsymbol{z}-t\boldsymbol{v}\|\\ &\geq -\epsilon t\|\boldsymbol{v}\|\end{aligned}$$

for sufficiently small $t>0$. This contradiction implies that $\|\boldsymbol{v}\|\leq\epsilon$. ■

We are now in position to characterize the Mangasarian-Fromovitz constraint qualification in the absence of equality constraints. As the next proposition indicates, the constraint qualification holds if and only if the convex set generated by the active inequality constraints does not contain the origin. The geometric nature of this result will be clearer after we consider convex sets in the next chapter.

Proposition 5.3.2 (Gordon) *Given r vectors $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_r$ in R^n, define the function $f(\boldsymbol{x})=\ln\left[\sum_{j=1}^r\exp(\boldsymbol{z}_j^*\boldsymbol{x})\right]$. Then the following three conditions are logically equivalent:*

(a) *The function $f(\boldsymbol{x})$ is bounded below on R^n.*

(b) *There are nonnegative constants $\mu_1,\ldots,\mu_r$ such that*

$$\sum_{i=1}^r\mu_i\boldsymbol{z}_i = \boldsymbol{0},\qquad \sum_{i=1}^r\mu_i = 1.$$

(c) *There is no vector $\boldsymbol{u}$ such that $\boldsymbol{z}_j^*\boldsymbol{u}<0$ for all j.*

Proof: It is trivial to check that (b) implies (c) and (c) implies (a). To demonstrate that (a) implies (b), first observe that

$$\nabla f(\boldsymbol{x}) = \frac{1}{\sum_{i=1}^r e^{\boldsymbol{z}_i^*\boldsymbol{x}}}\sum_{j=1}^r e^{\boldsymbol{z}_j^*\boldsymbol{x}}\boldsymbol{z}_j.$$

According to Proposition 5.3.1, it is possible to choose for each k a point $\boldsymbol{u}_k$ at which

$$\|\nabla f(\boldsymbol{u}_k)\| = \left\|\sum_{j=1}^r\mu_{kj}\boldsymbol{z}_j\right\| \leq \frac{1}{k},$$

where

$$\mu_{kj} = \frac{e^{\boldsymbol{z}_j^* \boldsymbol{u}_k}}{\sum_{i=1}^r e^{\boldsymbol{z}_i^* \boldsymbol{u}_k}}.$$

Because the μ_{kj} form a vector $\boldsymbol{\mu}_k$ with nonnegative components summing to 1, it is possible to find a subsequence of the sequence $\boldsymbol{\mu}_k$ that converges to a vector $\boldsymbol{\mu}$ having the same properties. This vector satisfies the requirements of condition (b). ■

When the equality constraints in nonlinear programming are affine, the forgoing discussion is easy to summarize. One first eliminates the equality constraints via the reparameterization $\boldsymbol{x} = \boldsymbol{K}\boldsymbol{y}$. Then one defines the differentials $\boldsymbol{z}_j^* = dh_j(\boldsymbol{x})\boldsymbol{K}$ of the inequality constraints $h_j(\boldsymbol{K}\boldsymbol{y})$ in the new parameterization. The Mangasarian-Fromovitz constraint qualification postulates the existence of a vector $\boldsymbol{u}$ with $\boldsymbol{u}^*\boldsymbol{z}_j < 0$ for all active constraints $h_j(\boldsymbol{x})$. The existence of $\boldsymbol{u}$ rules out the possibility that $\boldsymbol{0}$ is a convex combination of the vectors $\boldsymbol{z}_j$. Hence, the Lagrange multiplier λ_0 of $f(\boldsymbol{x})$ in the $\boldsymbol{y}$ parameterization cannot equal 0.

5.4 Taylor-Made Higher-Order Differentials

Roughly speaking, the higher-order differentials of a function $f(\boldsymbol{x})$ on R^n arise from its multiple partial derivatives. Thus, $\partial_{ij}^2 f(\boldsymbol{x})$ contributes to a second differential and $\partial_{ijk}^3 f(\boldsymbol{x})$ to a third differential. Equality of mixed partial derivatives involves simple identities such as $\partial_{ij}^2 f(\boldsymbol{x}) = \partial_{ji}^2 f(\boldsymbol{x})$ and $\partial_{ijk}^3 f(\boldsymbol{x}) = \partial_{kji}^3 f(\boldsymbol{x})$. The right notation efficiently exposes these symmetries. Let $\boldsymbol{j}$ be a vector whose m components j_i are drawn from the set $\{1, \ldots, n\}$. In this notation we write a mixed partial derivative of the function $f(\boldsymbol{x})$ as

$$\partial_{\boldsymbol{j}}^m f(\boldsymbol{x}) = \partial_{j_1} \cdots \partial_{j_m} f(\boldsymbol{x}).$$

If the partial derivatives of $f(\boldsymbol{x})$ of order m are continuous at $\boldsymbol{x} \in \mathsf{R}^n$, then Proposition 4.3.1 shows that equality of mixed partials holds and the order of the components of $\boldsymbol{j}$ is irrelevant. The multinomial coefficient

$$\binom{m}{\boldsymbol{k}} = \binom{m}{k_1 \ldots k_n}$$

counts the number of mixed partial derivatives of order m in which the partial differential operator ∂_i appears k_i times.

Suppose $f(\boldsymbol{y})$ is a real-valued function possessing all partial derivatives of order p or less in a neighborhood of the point $\boldsymbol{x}$. The first-order Taylor expansion

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \sum_{i=1}^n \int_0^1 \partial_i f[t\boldsymbol{y} + (1-t)\boldsymbol{x}]\, dt (y_i - x_i)$$

follows directly from the fundamental theorem of calculus and the chain rule. To pass to the second-order Taylor expansion, one integrates by parts and replaces the integral

$$\int_0^1 \partial_i f[t\boldsymbol{y} + (1-t)\boldsymbol{x}]\,dt\,(y_i - x_i)$$

by

$$\partial_i f(\boldsymbol{x})(y_i - x_i) + \sum_{j=1}^n \int_0^1 \partial_j \partial_i f[t\boldsymbol{y} + (1-t)\boldsymbol{x}](1-t)\,dt(y_i - x_i)(y_j - x_j).$$

Repeated integration by parts, differentiating $f(\boldsymbol{x})$ and integrating successive powers of $(1-t)$, ultimately leads to the Taylor expansion

$$\begin{aligned} f(\boldsymbol{y}) &= \sum_{m=0}^{p-1} \frac{1}{m!} d^m f(\boldsymbol{x})[(\boldsymbol{y}-\boldsymbol{x})^m] + R(\boldsymbol{y},\boldsymbol{x}) \qquad (5.7) \\ R(\boldsymbol{y},\boldsymbol{x}) &= \frac{1}{(p-1)!} \int_0^1 d^p f[t\boldsymbol{y} + (1-t)\boldsymbol{x}][(\boldsymbol{y}-\boldsymbol{x})^p](1-t)^{p-1}\,dt \end{aligned}$$

of order p. Here we employ the notation of multilinear maps sketched in Example 2.5.10. The abstract entity $d^m f(\boldsymbol{x})[(\boldsymbol{y}-\boldsymbol{x})^m]$ is a m-linear form evaluated along its diagonal whose coefficients relative to the standard basis of R^n reduce to the mixed partial derivatives $\partial_{\boldsymbol{j}}^m f(\boldsymbol{x})$ of order m. Equality of mixed partials makes $d^m f(\boldsymbol{x})$ a symmetric m-linear form. All of this sounds complicated, but it simply amounts to local approximation of $f(\boldsymbol{y})$ by a polynomial in the components of the difference vector $\boldsymbol{y} - \boldsymbol{x}$. Our derivation of the Taylor expansion (5.7) remains valid for $f(\boldsymbol{y})$ vector or matrix valued if we operate entry by entry.

The remainder $R(\boldsymbol{y},\boldsymbol{x})$ in the expansion (5.7) can be recast in two useful ways. Assuming the partial derivatives of $f(\boldsymbol{x})$ of order p and less are continuous at $\boldsymbol{x}$, it is obvious that

$$s^p(\boldsymbol{y},\boldsymbol{x})[\boldsymbol{u}_1,\ldots,\boldsymbol{u}_p] = p\int_0^1 d^p f[t\boldsymbol{y} + (1-t)\boldsymbol{x}][\boldsymbol{u}_1,\ldots,\boldsymbol{u}_p](1-t)^{p-1}\,dt$$

is an p-linear map with $\lim_{\boldsymbol{y}\to\boldsymbol{x}} s^p(\boldsymbol{y},\boldsymbol{x}) = s^p(\boldsymbol{x},\boldsymbol{x}) = d^p f(\boldsymbol{x})$. This suggests that the expansion

$$f(\boldsymbol{y}) = \sum_{m=0}^{p-1} \frac{1}{m!} d^m f(\boldsymbol{x})[(\boldsymbol{y}-\boldsymbol{x})^m] + \frac{1}{p!} s^p(\boldsymbol{y},\boldsymbol{x})[(\boldsymbol{y}-\boldsymbol{x})^p] \qquad (5.8)$$

be taken as the definition of Carathéodory differentiability of order p, with the understanding that the slope $s^p(\boldsymbol{y},\boldsymbol{x})$ tends to $s^p(\boldsymbol{x},\boldsymbol{x}) = d^p f(\boldsymbol{x})$ as $\boldsymbol{y}$ tends to $\boldsymbol{x}$. There is no harm in assuming that $s^p(\boldsymbol{y},\boldsymbol{x})$ and $d^p f(\boldsymbol{x})$ are

symmetric multilinear maps since they operate only on diagonal arguments. Indeed, any multilinear map $\boldsymbol{M}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p]$ agrees with its symmetrization

$$\boldsymbol{M}_{\text{sym}}[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p] \quad = \quad \frac{1}{p!} \sum_{\sigma} \boldsymbol{M}[\boldsymbol{u}_{\sigma_1}, \ldots, \boldsymbol{u}_{\sigma_p}]$$

along its diagonal $\boldsymbol{u}_1 = \boldsymbol{u}_2 = \cdots = \boldsymbol{u}_p$. Here the sum ranges over all permutations σ of $\{1, \ldots, p\}$.

Alternatively, if we let $r^p(\boldsymbol{y}, \boldsymbol{x}, t) = d^p f[t\boldsymbol{y} + (1-t)\boldsymbol{x}] - d^p(\boldsymbol{x})$, then the identity

$$\begin{aligned} R(\boldsymbol{y}, \boldsymbol{x}) \quad &= \quad \frac{1}{(p-1)!} \int_0^1 r^p(\boldsymbol{y}, \boldsymbol{x}, t)[(\boldsymbol{y} - \boldsymbol{x})^p](1-t)^{p-1}\, dt \\ &\quad + \frac{1}{p!} d^p f(\boldsymbol{x})[(\boldsymbol{y} - \boldsymbol{x})^p] \end{aligned}$$

and the inequality

$$\|r^p(\boldsymbol{y}, \boldsymbol{x}, t)[(\boldsymbol{y} - \boldsymbol{x})^p]\| \quad \leq \quad \|r^p(\boldsymbol{y}, \boldsymbol{x}, t)\| \cdot \|\boldsymbol{y} - \boldsymbol{x}\|^p$$

suggest that the expansion

$$f(\boldsymbol{y}) \quad = \quad \sum_{m=0}^{p} \frac{1}{m!} d^m f(\boldsymbol{x})[(\boldsymbol{y} - \boldsymbol{x})^m] + o(\|\boldsymbol{y} - \boldsymbol{x}\|^p)$$

be taken as the definition of Fréchet differentiability of order p. These two axiomatic definitions of higher-order differentials are logically equivalent. Given Carathéodory's definition, we set $d^p f(\boldsymbol{x}) = s^p(\boldsymbol{x}, \boldsymbol{x})$ and note that

$$\|[s^p(\boldsymbol{y}, \boldsymbol{x}) - s^p(\boldsymbol{x}, \boldsymbol{x})][(\boldsymbol{y} - \boldsymbol{x})^p]\| \quad \leq \quad \|s^p(\boldsymbol{y}, \boldsymbol{x}) - s^p(\boldsymbol{x}, \boldsymbol{x})\| \|\boldsymbol{y} - \boldsymbol{x}\|^p.$$

Fréchet's definition follows directly. Proof of the converse is more subtle, and we omit it. Both definitions require the existence of all partial derivatives of $f(\boldsymbol{y})$ of order $p-1$ and less in a neighborhood of $\boldsymbol{x}$.

Perhaps just as important as the equivalence of the two definition of differentiability is the following simple result.

Proposition 5.4.1 *The symmetric multilinear maps $d^m f(\boldsymbol{x})$ and $s^p(\boldsymbol{x}, \boldsymbol{x})$ appearing in the Taylor expansion (5.8) are unique. The slope function $s^p(\boldsymbol{y}, \boldsymbol{x})$ is not unique as noted in Chap. 4.*

Proof: Abbreviate $d^m f(\boldsymbol{x}) = \boldsymbol{M}^m$. The constant $\boldsymbol{M}^0$ equals $\lim_{\boldsymbol{y} \to \boldsymbol{x}} f(\boldsymbol{y})$. Consider a second expansion of $f(\boldsymbol{y})$ with $\boldsymbol{N}^m$ replacing $\boldsymbol{M}^m$ and $t^p(\boldsymbol{y}, \boldsymbol{x})$ replacing $s^p(\boldsymbol{y}, \boldsymbol{x})$. Suppose by induction that $\boldsymbol{M}^k = \boldsymbol{N}^k$ for $k < m < p$, and let $\boldsymbol{y}_j$ be the sequence $\boldsymbol{x} + \frac{1}{j}\boldsymbol{u}$ for some fixed vector $\boldsymbol{u}$. Taking limits on j in the equality

$$\begin{aligned} \boldsymbol{0} \quad &= \quad \frac{1}{m!}(\boldsymbol{M}^m - \boldsymbol{N}^m)[\boldsymbol{u}^m] + \sum_{k=m+1}^{p-1} \frac{j^{m-k}}{k!}(\boldsymbol{M}^k - \boldsymbol{N}^k)[\boldsymbol{u}^k] \\ &\quad + j^{m-p}[s^p(\boldsymbol{y}_j, \boldsymbol{x}) - t^p(\boldsymbol{y}_j, \boldsymbol{x})][\boldsymbol{u}^p] \end{aligned}$$

gives $\boldsymbol{M}^m[\boldsymbol{u}^m] = \boldsymbol{N}^m[\boldsymbol{u}^m]$. If we now interpret $\boldsymbol{M}^m[\boldsymbol{u}^m]$ and $\boldsymbol{N}^m[\boldsymbol{u}^m]$ as polynomials in the entries of $\boldsymbol{u}$, then their coefficients must coincide. In view of the symmetry of $\boldsymbol{M}^m$ and $\boldsymbol{N}^m$, the coefficients $\boldsymbol{M}^m[\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_m}]$ and $\boldsymbol{N}^m[\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_m}]$ therefore also coincide for all choices $\boldsymbol{e}_{i_1}, \ldots, \boldsymbol{e}_{i_m}$. This argument advances the induction one step. The proof of the equality $s^p(\boldsymbol{x}, \boldsymbol{x}) = t^p(\boldsymbol{x}, \boldsymbol{x})$ follows essentially the same pattern. ■

Example 5.4.1 *Second Differential of a Bilinear Map*

Consider a bilinear map $f(\boldsymbol{x}) = M(\boldsymbol{x}_1, \boldsymbol{x}_2)$ of the concatenated vector $\boldsymbol{x}$ with left block $\boldsymbol{x}_1$ and right block $\boldsymbol{x}_2$. If we take $\boldsymbol{u}_i = \boldsymbol{y}_i - \boldsymbol{x}_i$, then the expansion

$$\begin{aligned} M(\boldsymbol{y}_1, \boldsymbol{y}_2) &= M(\boldsymbol{x}_1 + \boldsymbol{u}_1, \boldsymbol{x}_2 + \boldsymbol{u}_2) \\ &= M(\boldsymbol{x}_1, \boldsymbol{x}_2) + M(\boldsymbol{x}_1, \boldsymbol{u}_2) + M(\boldsymbol{u}_1, \boldsymbol{x}_2) + M(\boldsymbol{u}_1, \boldsymbol{u}_2) \end{aligned}$$

identifies the bilinear map $2M(\boldsymbol{u}_1, \boldsymbol{u}_2)$ as the second differential of $f(\boldsymbol{x})$. Similar expansions hold for higher-order multilinear maps. ■

The most important applications in optimization theory of higher-order differentials involve scalar-valued functions $f(\boldsymbol{x})$ and their first and second differentials $df(\boldsymbol{x})$ and $d^2 f(\boldsymbol{x})$. We retain our convention that the gradient $\nabla f(\boldsymbol{x})$ equals the transpose of the differential $df(\boldsymbol{x})$. The coefficients of the linear form $df(\boldsymbol{x})[\boldsymbol{u}]$ are the first partials $\partial_i f(\boldsymbol{x})$. The second-order Taylor expansion of $f(\boldsymbol{y})$ around $\boldsymbol{x}$ reads

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + df(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^* s^2(\boldsymbol{y}, \boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}). \quad (5.9)$$

Here the second-order slope $s^2(\boldsymbol{y}, \boldsymbol{x})$ has limit $d^2 f(\boldsymbol{x})$; this Hessian matrix incorporates the second partials $\partial_{ij}^2 f(\boldsymbol{x})$. Consequently, one is justified in viewing $d^2 f(\boldsymbol{x})$ as the differential of $\nabla f(\boldsymbol{x})$.

Example 5.4.2 *Second Differential of a Quadratic Function*

Consider the quadratic function $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^* \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^* \boldsymbol{x} + c$ defined by an $n \times n$ symmetric matrix $\boldsymbol{A}$ and an $n \times 1$ vector $\boldsymbol{b}$. The gradient of $f(\boldsymbol{y})$ is $\nabla f(\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}$. The second differential emerges from the slope equation

$$\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) = \boldsymbol{A}(\boldsymbol{y} - \boldsymbol{x})$$

or the simple calculation $d^2 f(\boldsymbol{x}) = d\nabla f(\boldsymbol{x}) = (d\boldsymbol{A})\boldsymbol{y} + \boldsymbol{A}d\boldsymbol{y} + d\boldsymbol{b} = \boldsymbol{A}$. Either perspective is consistent with the exact expansion

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + (\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})^*(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^* \boldsymbol{A}(\boldsymbol{y} - \boldsymbol{x}).$$

Uniqueness of the differentials $df(\boldsymbol{x})$ and $d^2 f(\boldsymbol{x})$ is guaranteed by Proposition 5.4.1. ■

Example 5.4.3 *A Pathological Example*

On R^2 define the indicator function $1_Q(\boldsymbol{x})$ to be 1 when both coordinates x_1 and x_2 are rational numbers and to be 0 otherwise. The function

$$f(\boldsymbol{x}) \quad = \quad 1_Q(\boldsymbol{x})(x_1^3 + x_2^3)$$

is differentiable at the origin $\mathbf{0}$ but nowhere else on R^2. The differential $df(\mathbf{0}) = \mathbf{0}$ together with the choice

$$s^2(\boldsymbol{y}, \mathbf{0}) \quad = \quad 1_Q(\boldsymbol{y}) \begin{pmatrix} 2y_1 & 0 \\ 0 & 2y_2 \end{pmatrix}.$$

satisfy expansion (5.9). Furthermore, $s^2(\boldsymbol{y}, \mathbf{0})$ tends to the $\mathbf{0}$ matrix as $\boldsymbol{y}$ tends to the origin. However, claiming that $f(\boldsymbol{x})$ is twice differentiable at the origin seems extreme. To eliminate such pathologies, we demand that $f(\boldsymbol{x})$ be differentiable in a neighborhood of a point before we allow it to be twice differentiable at the point. ■

Example 5.4.4 *Second Differential of the Inverse Polar Transformation*

Continuing Example 4.6.1, it is straightforward to calculate that the inverse polar transformation

$$g\left[\begin{pmatrix} r \\ \theta \end{pmatrix}\right] \quad = \quad \begin{pmatrix} r\cos\theta \\ r\sin\theta \end{pmatrix}$$

has second differential

$$d^2 g\left[\begin{pmatrix} r \\ \theta \end{pmatrix}\right] \quad = \quad \begin{pmatrix} d^2\, r\cos\theta \\ d^2\, r\sin\theta \end{pmatrix} \quad = \quad \begin{pmatrix} 0 & -\sin\theta \\ -\sin\theta & -r\cos\theta \\ 0 & \cos\theta \\ \cos\theta & -r\sin\theta \end{pmatrix}.$$

Here we stack the differentials corresponding to the two components for easy visualization. ■

Example 5.4.5 *First and Second Differentials of the Inverse of a Matrix*

Taylor expansions take different shapes depending on how one displays the results. For instance, the identity

$$\boldsymbol{Y}^{-1} - \boldsymbol{X}^{-1} \quad = \quad -\boldsymbol{X}^{-1}(\boldsymbol{Y} - \boldsymbol{X})\boldsymbol{Y}^{-1}$$

is a disguised slope expansion of the inverse of a matrix as a function of its entries. The corresponding first differential is the linear map

$$\boldsymbol{U} \quad \mapsto \quad -\boldsymbol{X}^{-1}\boldsymbol{U}\boldsymbol{X}^{-1}.$$

The more elaborate second-order expansion

$$\begin{aligned}\boldsymbol{Y}^{-1} - \boldsymbol{X}^{-1} &= -\boldsymbol{X}^{-1}(\boldsymbol{Y}-\boldsymbol{X})\boldsymbol{X}^{-1} + \frac{1}{2}\boldsymbol{Y}^{-1}(\boldsymbol{Y}-\boldsymbol{X})\boldsymbol{X}^{-1}(\boldsymbol{Y}-\boldsymbol{X})\boldsymbol{X}^{-1} \\ &\quad + \frac{1}{2}\boldsymbol{X}^{-1}(\boldsymbol{Y}-\boldsymbol{X})\boldsymbol{X}^{-1}(\boldsymbol{Y}-\boldsymbol{X})\boldsymbol{Y}^{-1}\end{aligned}$$

shows that the second differential is the bilinear map

$$(\boldsymbol{U},\boldsymbol{V}) \quad \mapsto \quad \boldsymbol{X}^{-1}\boldsymbol{U}\boldsymbol{X}^{-1}\boldsymbol{V}\boldsymbol{X}^{-1} + \boldsymbol{X}^{-1}\boldsymbol{V}\boldsymbol{X}^{-1}\boldsymbol{U}\boldsymbol{X}^{-1}$$

evaluated along $\boldsymbol{U} = \boldsymbol{V}$. Despite the odd appearance of the second differential, it clearly fulfills its quadratic approximation responsibility. ■

The rules for calculating second differentials are naturally more complicated than those for calculating first differentials.

Proposition 5.4.2 *If the two functions $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ map the open set $S \subset \mathsf{R}^p$ twice differentiably into R^q, then the following functional combinations are twice differentiable and have the indicated second differentials:*

(a) *For all constants α and β,*

$$d^2[\alpha f(\boldsymbol{x}) + \beta g(\boldsymbol{x})] \quad = \quad \alpha d^2 f(\boldsymbol{x}) + \beta d^2 g(\boldsymbol{x}).$$

(b) *The inner product $f(\boldsymbol{x})^* g(\boldsymbol{x})$ satisfies*

$$\begin{aligned}d^2[f(\boldsymbol{x})^* g(\boldsymbol{x})] &= \sum_{i=1}^{q} \left[f_i(\boldsymbol{x}) d^2 g_i(\boldsymbol{x}) + g_i(\boldsymbol{x}) d^2 f_i(\boldsymbol{x})\right] \\ &\quad + \sum_{i=1}^{q} \left[\nabla f_i(\boldsymbol{x}) dg_i(\boldsymbol{x}) + \nabla g_i(\boldsymbol{x}) df_i(\boldsymbol{x})\right].\end{aligned}$$

(c) *For $q = 1$ and $f(\boldsymbol{x}) \neq 0$,*

$$d^2[f(\boldsymbol{x})^{-1}] \quad = \quad 2f(\boldsymbol{x})^{-3}\nabla f(\boldsymbol{x}) df(\boldsymbol{x}) - f(\boldsymbol{x})^{-2} d^2 f(\boldsymbol{x}).$$

Proof: Rule (a) follows directly from the linearity implicit in formula (5.9) covering the scalar case. For rule (b) it also suffices to consider the scalar case in view of rule (a). Applying the sum and product rules of differentiation to the gradient of $f(\boldsymbol{x})g(\boldsymbol{x})$ gives

$$\begin{aligned}d[\nabla g(\boldsymbol{x}) f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}) g(\boldsymbol{x})] &= d^2 g(\boldsymbol{x}) f(\boldsymbol{x}) + \nabla g(\boldsymbol{x}) df(\boldsymbol{x}) \\ &\quad + d^2 f(\boldsymbol{x}) g(\boldsymbol{x}) + \nabla f(\boldsymbol{x}) dg(\boldsymbol{x}).\end{aligned}$$

To verify rule (c), we apply the product, quotient, and chain rules to the gradient of $f(\boldsymbol{y})^{-1}$ identified in the proof of Proposition 4.4.2. This gives

$$-d\left[\nabla f(\boldsymbol{x}) \frac{1}{f(\boldsymbol{x})^2}\right] \quad = \quad -d^2 f(\boldsymbol{x}) \frac{1}{f(\boldsymbol{x})^2} + \nabla f(\boldsymbol{x}) \frac{2}{f(\boldsymbol{x})^3} df(\boldsymbol{x}).$$

Note the care exercised here in ordering the different factors prior to differentiation. ■

The chain rule for second differentials also is more complex.

Proposition 5.4.3 *Suppose $f(\boldsymbol{x})$ maps the open set $S \subset \mathsf{R}^p$ twice differentiably into R^q and $g(\boldsymbol{y})$ maps the open set $T \subset \mathsf{R}^q$ twice differentiably into R^r. If the image $f(S)$ is contained in T, then the composite function $h(\boldsymbol{x}) = g \circ f(\boldsymbol{x})$ is twice differentiable with second partial derivatives*

$$\partial_{kl}^2 h_m(\boldsymbol{x}) = \sum_{i=1}^q \partial_i g_m \circ f(\boldsymbol{x}) \partial_{kl}^2 f_i(\boldsymbol{x}) + \sum_{i=1}^q \sum_{j=1}^q \partial_k f_i(\boldsymbol{x}) \partial_{ij}^2 g_m \circ f(\boldsymbol{x}) \partial_l f_j(\boldsymbol{x}).$$

Proof: It suffices to prove the result when $r = 1$ and $g(\boldsymbol{x})$ is scalar valued. The function $h(\boldsymbol{x})$ has first differential $dh(\boldsymbol{x}) = (dg) \circ f(\boldsymbol{x}) df(\boldsymbol{x})$ and gradient $df(\boldsymbol{x})^* (\nabla g) \circ f(\boldsymbol{x})$. The matrix transpose, chain, and product rules of differentiation derived in Examples 4.7.1, 4.7.2, and 4.7.3 show that $\nabla h(\boldsymbol{x})$ has differential components

$$\begin{aligned} \partial_k [df(\boldsymbol{x})^* \nabla g \circ f(\boldsymbol{x})] &= [\partial_k df(\boldsymbol{x})]^* \nabla g \circ f(\boldsymbol{x}) \\ &\quad + df(\boldsymbol{x})^* \sum_{j=1}^q (\partial_j \nabla g) \circ f(\boldsymbol{x}) \partial_k f_j(\boldsymbol{x}). \end{aligned}$$

Alternatively, one can calculate the conventional way with explicit partial derivatives and easily verify the claimed formula for $d^2 h(\boldsymbol{x})$. ■

5.5 Applications of Second Differentials

Our first proposition, mentioned informally in Chap. 1, emphasizes the importance of second differentials in optimization theory.

Proposition 5.5.1 *Consider a real-valued function $f(\boldsymbol{y})$ with domain an open set $U \subset \mathsf{R}^p$. If $f(\boldsymbol{y})$ has a local minimum at $\boldsymbol{x}$ and is twice differentiable there, then the second differential $d^2 f(\boldsymbol{x})$ is positive semidefinite. Conversely, if $\boldsymbol{x}$ is a stationary point of $f(\boldsymbol{y})$ and $d^2 f(\boldsymbol{x})$ is positive definite, then $\boldsymbol{x}$ is a local minimum of $f(\boldsymbol{y})$. Similar statements hold for local maxima if we replace the modifiers positive semidefinite and positive definite by the modifiers negative semidefinite and negative definite. Finally, if $\boldsymbol{x}$ is a stationary point of $f(\boldsymbol{y})$ and $d^2 f(\boldsymbol{x})$ possesses both positive and negative eigenvalues, then $\boldsymbol{x}$ is neither a local minimum nor a local maximum of $f(\boldsymbol{y})$.*

Proof: Suppose $\boldsymbol{x}$ provides a local minimum of $f(\boldsymbol{y})$. For any unit vector $\boldsymbol{v}$ and $t > 0$ sufficiently small, the point $\boldsymbol{y} = \boldsymbol{x} + t\boldsymbol{v}$ belongs to U and satisfies $f(\boldsymbol{y}) \geq f(\boldsymbol{x})$. If we divide the expansion

$$\begin{aligned} 0 &\leq f(\boldsymbol{y}) - f(\boldsymbol{x}) \\ &= \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^* s^2(\boldsymbol{y}, \boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \end{aligned}$$

by $t^2 = \|\boldsymbol{y} - \boldsymbol{x}\|^2$ and send t to 0, then it follows that

$$0 \leq \frac{1}{2}\boldsymbol{v}^* d^2 f(\boldsymbol{x})\boldsymbol{v}.$$

Because $\boldsymbol{v}$ is an arbitrary unit vector, the quadratic form $d^2 f(\boldsymbol{x})$ must be positive semidefinite.

On the other hand, suppose $\boldsymbol{x}$ is a stationary point of $f(\boldsymbol{y})$, $d^2 f(\boldsymbol{x})$ is positive definite, and $\boldsymbol{x}$ fails to be a local minimum. Then there exists a sequence of points $\boldsymbol{y}_m$ tending to $\boldsymbol{x}$ and satisfying

$$0 > f(\boldsymbol{y}_m) - f(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{y}_m - \boldsymbol{x})^* s^2(\boldsymbol{y}_m, \boldsymbol{x})(\boldsymbol{y}_m - \boldsymbol{x}). \tag{5.10}$$

Passing to a subsequence if necessary, we may assume that the unit vectors $\boldsymbol{v}_m = (\boldsymbol{y}_m - \boldsymbol{x})/\|\boldsymbol{y}_m - \boldsymbol{x}\|$ converge to a unit vector $\boldsymbol{v}$. Dividing inequality (5.10) by $\|\boldsymbol{y}_m - \boldsymbol{x}\|^2$ and sending m to ∞ consequently yields $0 \geq \boldsymbol{v}^* d^2 f(\boldsymbol{x})\boldsymbol{v}$, contrary to the hypothesis that $d^2 f(\boldsymbol{x})$ is positive definite. This contradiction shows that $\boldsymbol{x}$ represents a local minimum.

To prove the final claim of the proposition, let μ be a nonzero eigenvalue with corresponding eigenvector $\boldsymbol{v}$. Then the difference

$$\begin{aligned} f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x}) &= \frac{t^2}{2}\left\{\boldsymbol{v}^* d^2 f(\boldsymbol{x})\boldsymbol{v} + \boldsymbol{v}^*[s^2(\boldsymbol{x} + t\boldsymbol{v}, \boldsymbol{x}) - d^2 f(\boldsymbol{x})]\boldsymbol{v}\right\} \\ &= \frac{t^2}{2}\left\{\mu\|\boldsymbol{v}\|^2 + \boldsymbol{v}^*[s^2(\boldsymbol{x} + t\boldsymbol{v}, \boldsymbol{x}) - d^2 f(\boldsymbol{x})]\boldsymbol{v}\right\} \end{aligned}$$

has the same sign as μ for t small. ■

Example 5.5.1 *Distinguishing Extrema from Saddle Points*

Consider the function

$$f(\boldsymbol{x}) = \frac{1}{4}x_1^4 + \frac{1}{2}x_2^2 - x_1 x_2 + x_1 - x_2$$

on R^2. It is obvious that

$$\nabla f(\boldsymbol{x}) = \begin{pmatrix} x_1^3 - x_2 + 1 \\ x_2 - x_1 - 1 \end{pmatrix}, \quad d^2 f(\boldsymbol{x}) = \begin{pmatrix} 3x_1^2 & -1 \\ -1 & 1 \end{pmatrix}.$$

Adding the two rows of the stationarity equation $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$ gives the equation $x_1^3 - x_1 = 0$ with solutions $0, \pm 1$. Solving for x_2 in each case yields

the stationary points $(0,1)$, $(-1,0)$, and $(1,2)$. The last two points are local minima because $d^2 f(\boldsymbol{x})$ is positive definite. The first point is a saddle point because

$$d^2 f(0,1) \;=\; \begin{pmatrix} 0 & -1 \\ -1 & 1 \end{pmatrix}$$

has characteristic polynomial $\lambda^2 - \lambda - 1$ and eigenvalues $\frac{1}{2}(1 \pm \sqrt{5})$. One of these eigenvalues is positive, and one is negative. ■

We now state and prove a sufficient condition for a point $\boldsymbol{x}$ to be a constrained local minimum of the objective function $f(\boldsymbol{y})$. Even in the absence of constraints, inequality (5.11) below represents an improvement over the qualitative claims of Proposition 5.5.1.

Proposition 5.5.2 *Suppose the objective function $f(\boldsymbol{y})$ of the constrained optimization problem satisfies the multiplier rule (5.1) at the point $\boldsymbol{x}$ with $\lambda_0 = 1$. Let $f(\boldsymbol{y})$ and the various constraint functions be twice differentiable at $\boldsymbol{x}$, and let $\mathcal{L}(\boldsymbol{y})$ be the Lagrangian*

$$\mathcal{L}(\boldsymbol{y}) \;=\; f(\boldsymbol{y}) + \sum_{i=1}^{p} \lambda_i g_i(\boldsymbol{y}) + \sum_{j=1}^{q} \mu_j h_j(\boldsymbol{y}).$$

If $\boldsymbol{v}^* d^2 \mathcal{L}(\boldsymbol{x}) \boldsymbol{v} > 0$ for every vector $\boldsymbol{v} \neq \boldsymbol{0}$ satisfying $dg_i(\boldsymbol{x})\boldsymbol{v} = 0$ and $dh_j(\boldsymbol{x})\boldsymbol{v} \leq 0$ for all active constraints, then $\boldsymbol{x}$ provides a local minimum of $f(\boldsymbol{y})$. Furthermore, there exists a constant $c > 0$ such that

$$f(\boldsymbol{y}) \;\geq\; f(\boldsymbol{x}) + c\|\boldsymbol{y} - \boldsymbol{x}\|^2 \tag{5.11}$$

for all feasible $\boldsymbol{y}$ in a neighborhood of $\boldsymbol{x}$.

Proof: Because of the sign restrictions on the μ_j, we have $f(\boldsymbol{y}) \geq \mathcal{L}(\boldsymbol{y})$ for all feasible $\boldsymbol{y}$ in addition to $f(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x})$. If $\mathcal{L}(\boldsymbol{y}) \geq \mathcal{L}(\boldsymbol{x})$, then $f(\boldsymbol{y}) \geq f(\boldsymbol{x})$, and if $\mathcal{L}(\boldsymbol{y}) \geq \mathcal{L}(\boldsymbol{x}) + c\|\boldsymbol{y} - \boldsymbol{x}\|^2$, then $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + c\|\boldsymbol{y} - \boldsymbol{x}\|^2$. Therefore, it suffices to prove these inequalities for $\mathcal{L}(\boldsymbol{y})$ rather than $f(\boldsymbol{y})$. The second inequality $\mathcal{L}(\boldsymbol{y}) \geq \mathcal{L}(\boldsymbol{x}) + c\|\boldsymbol{y} - \boldsymbol{x}\|^2$ is stronger than the first inequality $\mathcal{L}(\boldsymbol{y}) \geq \mathcal{L}(\boldsymbol{x})$, so it also suffices to focus on the second inequality.

With this end in mind, let $s^2_{\mathcal{L}}(\boldsymbol{y}, \boldsymbol{x})$ be a second slope function for $\mathcal{L}(\boldsymbol{y})$ at $\boldsymbol{x}$. If the second inequality is false, then there exists a sequence of feasible points $\boldsymbol{y}_m$ converging to $\boldsymbol{x}$ and a sequence of positive constants c_m converging to 0 such that

$$\begin{aligned} \mathcal{L}(\boldsymbol{y}_m) - \mathcal{L}(\boldsymbol{x}) \;&=\; \frac{1}{2}(\boldsymbol{y}_m - \boldsymbol{x})^* s^2_{\mathcal{L}}(\boldsymbol{y}_m, \boldsymbol{x})(\boldsymbol{y}_m - \boldsymbol{x}) \\ &<\; c_m \|\boldsymbol{y}_m - \boldsymbol{x}\|^2. \end{aligned} \tag{5.12}$$

Here $d\mathcal{L}(\boldsymbol{x})$ vanishes by virtue of the multiplier condition. As usual, we suppose that the sequence of unit vectors

$$\boldsymbol{v}_m \;=\; \frac{1}{\|\boldsymbol{y}_m - \boldsymbol{x}\|}(\boldsymbol{y}_m - \boldsymbol{x})$$

converges to a unit vector $\boldsymbol{v}$ by extracting a subsequence if necessary. Dividing inequality (5.12) by $\|\boldsymbol{y}_m - \boldsymbol{x}\|^2$ and taking limits then yields $\boldsymbol{v}^* d^2\mathcal{L}(\boldsymbol{x})\boldsymbol{v} \leq 0$. This contradicts our supposition about $d^2\mathcal{L}(\boldsymbol{x})$ provided we can demonstrate that the tangent conditions $dg_i(\boldsymbol{x})\boldsymbol{v} = 0$ and $dh_j(x)\boldsymbol{v} \leq 0$ hold for all active constraints. These follow by dividing the equations

$$0 \;=\; g_i(\boldsymbol{y}_m) - g_i(\boldsymbol{x}) \;=\; s_{g_i}(\boldsymbol{y}_m, \boldsymbol{x})(\boldsymbol{y}_m - \boldsymbol{x})$$

and

$$0 \;\geq\; h_j(\boldsymbol{y}_m) - h_j(\boldsymbol{x}) \;=\; s_{h_j}(\boldsymbol{y}_m, \boldsymbol{x})(\boldsymbol{y}_m - \boldsymbol{x})$$

by $\|\boldsymbol{y}_m - \boldsymbol{x}\|$ and taking limits. Recall here that $h_j(\boldsymbol{x}) = 0$ at an active constraint. ■

Example 5.5.2 *Minimum Eigenvalue of a Symmetric Matrix*

Example 1.4.3 demonstrated how each eigenvector-eigenvalue pair $(\boldsymbol{x}, \alpha)$ of a symmetric matrix $\boldsymbol{M}$ provides a stationary point of the Lagrangian

$$\mathcal{L}(\boldsymbol{x}) \;=\; \frac{1}{2}\boldsymbol{x}^*\boldsymbol{M}\boldsymbol{x} - \frac{\alpha}{2}(\|\boldsymbol{x}\|^2 - 1).$$

Suppose that the eigenvalues are arranged so that $\alpha_1 \leq \cdots \leq \alpha_n$ and $\boldsymbol{x}_i$ is the unit eigenvector corresponding to α_i. We expect that $\boldsymbol{x}_1$ furnishes the minimum of $\frac{1}{2}\boldsymbol{y}^*\boldsymbol{M}\boldsymbol{y}$ subject to $g_1(\boldsymbol{y}) = \frac{1}{2} - \frac{1}{2}\|\boldsymbol{y}\|^2 = 0$. To check that this is indeed the case, we note that $d^2\mathcal{L}(\boldsymbol{y}) = \boldsymbol{M} - \alpha_1\boldsymbol{I}_n$. The condition $dg_1(\boldsymbol{x}_1)\boldsymbol{v} = 0$ is equivalent to $\boldsymbol{x}_1^*\boldsymbol{v} = 0$. Because the eigenvectors constitute an orthonormal basis, the equality $\boldsymbol{x}_1^*\boldsymbol{v} = 0$ can hold only if

$$\boldsymbol{v} \;=\; \sum_{i=2}^{n} c_i\boldsymbol{x}_i.$$

For such a choice of $\boldsymbol{v}$, the quadratic form

$$\boldsymbol{v}^* d^2\mathcal{L}(\boldsymbol{x}_1)\boldsymbol{v} \;=\; \sum_{i=2}^{n} c_i^2(\alpha_i - \alpha_1) \;>\; 0$$

so long as $\alpha_1 < \alpha_2$ and $\boldsymbol{v} \neq \boldsymbol{0}$. Thus, the condition cited in the statement of Proposition 5.5.2 holds, and the point $\boldsymbol{x}_1$ minimizes $\boldsymbol{y}^*\boldsymbol{M}\boldsymbol{y}$ subject to the constraint $\|\boldsymbol{y}\| = 1$.

The case $\alpha_1 = \alpha_2$ is not covered by Proposition 5.5.2. However, we can fall back on an alternative and simpler proof. Consider any unit vector

$$\boldsymbol{w} \;=\; \sum_{i=1}^{n} b_i\boldsymbol{x}_i.$$

The inequality

$$\boldsymbol{w}^*\boldsymbol{M}\boldsymbol{w} \;=\; \sum_{i=1}^n \alpha_i b_i^2 \|\boldsymbol{x}_i\|^2 \;\geq\; \alpha_1 \sum_{i=1}^n b_i^2 \;=\; \alpha_1.$$

immediately establishes the claimed property once we note that $\boldsymbol{w} = \boldsymbol{x}_1$ achieves the lower bound α_1. ■

Example 5.5.3 *Minimum of a Linear Reciprocal Function*

Consider minimizing the nonlinear function $f(\boldsymbol{x}) = \sum_{i=1}^n c_i x_i^{-1}$ subject to the linear inequality constraint $\sum_{i=1}^n a_i x_i \leq b$. Here all indicated variables and parameters are positive. Differentiating the Lagrangian

$$\mathcal{L}(\boldsymbol{x}) \;=\; \sum_{i=1}^n c_i x_i^{-1} + \mu\left(\sum_{i=1}^n a_i x_i - b\right)$$

gives the multiplier equations

$$-\frac{c_i}{x_i^2} + \mu a_i \;=\; 0.$$

It follows that $\mu > 0$, that the constraint is active, and that

$$\begin{aligned} x_i &= \sqrt{\frac{c_i}{\mu a_i}}, \qquad 1 \leq i \leq n \\ \mu &= \left(\frac{1}{b}\sum_{i=1}^n \sqrt{a_i c_i}\right)^2. \end{aligned} \tag{5.13}$$

The second differential $d^2\mathcal{L}(\boldsymbol{x})$ is diagonal with ith diagonal entry $2c_i/x_i^3$. This matrix is certainly positive definite, and Proposition 5.5.2 confirms that the stationary point (5.13) provides the minimum of $f(\boldsymbol{x})$ subject to the constraint. ■

When there are only equality constraints, one can say more about the sufficient criterion described in Proposition 5.5.2. Following the discussion in Sect. 5.3, let $\boldsymbol{G}$ be the $p \times n$ matrix $\boldsymbol{G}$ with rows $dg_i(\boldsymbol{x})$ and $\boldsymbol{K}$ an $n \times (n-p)$ matrix of full rank satisfying $\boldsymbol{G}\boldsymbol{K} = \boldsymbol{0}$. On the kernel of $\boldsymbol{G}$ the matrix $\boldsymbol{A} = d^2\mathcal{L}(\boldsymbol{x})$ is positive definite. Since every $\boldsymbol{v}$ in the kernel equals some image point $\boldsymbol{K}\boldsymbol{u}$, we can establish the validity of the sufficient condition of Proposition 5.5.2 by checking whether the matrix $\boldsymbol{K}^*\boldsymbol{A}\boldsymbol{K}$ of the quadratic form $\boldsymbol{u}^*\boldsymbol{K}^*\boldsymbol{A}\boldsymbol{K}\boldsymbol{u}$ is positive definite. There are many practical methods of making this determination. For instance, the sweep operator from computational statistics performs such a check easily in the process of inverting $\boldsymbol{K}^*\boldsymbol{A}\boldsymbol{K}$ [166].

If we want to work directly with the matrix $\boldsymbol{G}$, there is another interesting criterion involving the relation between the positive semidefinite matrix

$\boldsymbol{B} = \boldsymbol{G}^*\boldsymbol{G}$ and the second differential $\boldsymbol{A} = d^2\mathcal{L}(\boldsymbol{x})$. One can rephrase the sufficient condition of Proposition 5.5.2 by saying that $\boldsymbol{v}^*\boldsymbol{A}\boldsymbol{v} > 0$ whenever $\boldsymbol{v}^*\boldsymbol{B}\boldsymbol{v} = 0$ and $\boldsymbol{v} \neq \boldsymbol{0}$. We claim that this condition is equivalent to the existence of some constant $\gamma > 0$ such that the matrix $\boldsymbol{A} + \gamma\boldsymbol{B}$ is positive definite [58]. Clearly, if such a γ exists, then the condition holds. Conversely, suppose the condition holds and that no such γ exists. Then there is a sequence of unit vectors $\boldsymbol{v}_m$ and a sequence of scalars α_m tending to ∞ such that

$$\boldsymbol{v}_m^*\boldsymbol{A}\boldsymbol{v}_m + \alpha_m\boldsymbol{v}_m^*\boldsymbol{B}\boldsymbol{v}_m \quad \leq \quad 0. \tag{5.14}$$

By passing to a subsequence if needed, we may assume that the sequence $\boldsymbol{v}_m$ converges to a unit vector $\boldsymbol{v}$. On the one hand, because $\boldsymbol{B}$ is positive semidefinite, inequality (5.14) compels the conclusions $\boldsymbol{v}_m^*\boldsymbol{A}\boldsymbol{v}_m \leq 0$, which must carry over to the limit. On the other hand, dividing inequality (5.14) by α_m and taking limits imply $\boldsymbol{v}^*\boldsymbol{B}\boldsymbol{v} \leq 0$ and therefore $\boldsymbol{v}^*\boldsymbol{B}\boldsymbol{v} = 0$. Because the limit vector $\boldsymbol{v}$ violates the condition $\boldsymbol{v}^*\boldsymbol{A}\boldsymbol{v} > 0$, the required $\gamma > 0$ exists.

5.6 Problems

1. Find a minimum of $f(\boldsymbol{x}) = x_1^2 + x_2^2$ subject to the inequality constraints $h_1(\boldsymbol{x}) = -2x_1 - x_2 + 10 \leq 0$ and $h_2(\boldsymbol{x}) = -x_1 \leq 0$ on R^2. Prove that it is the global minimum.

2. Minimize the function $f(\boldsymbol{x}) = e^{-(x_1+x_2)}$ subject to the constraints $h_1(\boldsymbol{x}) = e^{x_1} + e^{x_2} - 20 \leq 0$ and $h_2(\boldsymbol{x}) = -x_1 \leq 0$ on R^2.

3. Find the minimum and maximum of the function $f(\boldsymbol{x}) = x_1 + x_2$ over the subset of R^2 defined by the constraints

$$\begin{aligned} h_1(\boldsymbol{x}) &= -x_1 \\ h_2(\boldsymbol{x}) &= -x_2 \\ h_3(\boldsymbol{x}) &= 1 - x_1x_2. \end{aligned}$$

4. Consider the problem of minimizing $f(\boldsymbol{x}) = (x_1 + 1)^2 + x_2^2$ subject to the inequality constraint $h(\boldsymbol{x}) = -x_1^3 + x_2^2 \leq 0$ on R^2. Solve the problem by sketching the feasible region and using a little geometry. Show that the multiplier rule of Proposition 5.2.1 with $\lambda_0 = 1$ fails and explain why.

5. In the multiplier rule of Proposition 5.2.1, suppose that the Kuhn-Tucker constraint qualification holds and that $\lambda_0 = 1$. Prove that the remaining multipliers are unique.

6. Consider the inequality constraint functions

$$\begin{aligned} h_1(\boldsymbol{x}) &= -x_1 \\ h_2(\boldsymbol{x}) &= -x_2 \\ h_3(\boldsymbol{x}) &= x_1^2 + 4x_2^2 - 4 \\ h_4(\boldsymbol{x}) &= (x_1 - 2)^2 + x_2^2 - 5 \end{aligned}$$

on R^2. Show that the Kuhn-Tucker constraint qualification fails but the Mangasarian-Fromovitz constraint qualification succeeds at the point $x = (0, 1)^*$. For the inequality constraint functions

$$\begin{aligned} h_1(\boldsymbol{x}) &= x_1^2 - x_2 \\ h_2(\boldsymbol{x}) &= -3x_1^2 + x_2, \end{aligned}$$

show that both constraint qualifications fail at the point $\boldsymbol{x} = \mathbf{0}$ [96].

7. Consider the two functions $f(\boldsymbol{x}) = x_1$ and

$$g(\boldsymbol{x}) = \begin{cases} x_2 & \text{if } x_1 \geq 0 \\ x_2 - x_1^2 & \text{if } x_1 < 0,\ x_2 \leq 0 \\ x_2 + x_1^2 & \text{if } x_1 < 0,\ x_2 > 0 \end{cases}$$

defined on R^2. Demonstrate that:

   (a) $f(\boldsymbol{x})$ has differential $df(\boldsymbol{x}) = (1, 0)$,
   (b) $g(\boldsymbol{x})$ is continuous except on the half line $\{(x_1, 0)^* : x_1 < 0\}$,
   (c) $g(\boldsymbol{x})$ has differential $dg(\mathbf{0}) = (0, 1)$ at the origin.

   With these functions in hand, it is possible to show that the Lagrange multiplier rule can fail due to lack of continuity of an equality constraint. The optimization problem we have in mind is minimizing $f(\boldsymbol{x})$ subject to $g(\boldsymbol{x}) = 0$. Prove that the origin is the unique solution of this problem. In addition prove that no nontrivial pair (μ, λ) satisfies the multiplier condition

$$\mu \nabla f(\mathbf{0}) + \lambda \nabla g(\mathbf{0}) = \mathbf{0}.$$

8. For a real 2×2 matrix

$$\boldsymbol{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

   define the Frobenius norm $\|\boldsymbol{M}\|_F = \sqrt{a^2 + b^2 + c^2 + d^2}$. Let S denote the set of matrices $\boldsymbol{M}$ with $\det(\boldsymbol{M}) = 0$. Find the minimum distance from S to the matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix},$$

and exhibit a matrix attaining this distance [69]. (Hints: Introduce a Lagrange multiplier λ in minimizing $\frac{1}{2}\|\boldsymbol{M}\|_F^2$ subject to $\boldsymbol{M} \in S$. From the multiplier conditions deduce that $\lambda = \pm 1$ if $b \neq 0$. Show that the assumption $\lambda = \pm 1$ leads to a contradiction. Thus, $b = 0$ and consequently $c = 0$. Express a and d as functions of λ and find the λ's for which $\det(\boldsymbol{M}) = 0$.)

9. The equation

$$\sum_{i=1}^{n} a_i x_i^2 \quad = \quad c$$

defines an ellipse in R^n whenever all $a_i > 0$. The problem of Apollonius is to find the closest point on the ellipse from an external point $\boldsymbol{y}$ [30]. Demonstrate that the solution has coordinates

$$x_i \quad = \quad \frac{y_i}{1 + \lambda a_i},$$

where λ is chosen to satisfy

$$\sum_{i=1}^{n} a_i \left(\frac{y_i}{1 + \lambda a_i} \right)^2 \quad = \quad c.$$

Show how you can adapt this solution to solve the problem with the more general ellipse $(\boldsymbol{x} - \boldsymbol{z})^* \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{z}) = c$ for $\boldsymbol{A}$ a positive definite matrix.

10. Let $\boldsymbol{A}$ be a positive definite matrix. For a given vector $\boldsymbol{y}$, find the maximum of $f(\boldsymbol{x}) = \boldsymbol{y}^* \boldsymbol{x}$ subject to $h(\boldsymbol{x}) = \boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x} - 1 \leq 0$. Use your result to prove the inequality $|\boldsymbol{y}^* \boldsymbol{x}|^2 \leq (\boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x})(\boldsymbol{y}^* \boldsymbol{A}^{-1} \boldsymbol{y})$.

11. Let $\boldsymbol{A}$ be a full rank $m \times n$ matrix and $\boldsymbol{b}$ be an $m \times 1$ vector with $m < n$. The set $S = \{\boldsymbol{x} \in \mathsf{R}^n : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ defines a plane in R^n. If $m = 1$, S is a hyperplane. Given $\boldsymbol{y} \in \mathsf{R}^n$, prove that the closest point to $\boldsymbol{y}$ in S is

$$P(\boldsymbol{y}) \quad = \quad \boldsymbol{y} - \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}\boldsymbol{A}\boldsymbol{y} + \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}\boldsymbol{b}.$$

12. If $\boldsymbol{A}$ is a matrix and $\boldsymbol{y}$ is a compatible vector, then $\boldsymbol{A}\boldsymbol{y} \geq \boldsymbol{0}$ means that all entries of the vector $\boldsymbol{A}\boldsymbol{y}$ are nonnegative. Farkas' lemma says that $\boldsymbol{x}^* \boldsymbol{y} \geq 0$ for all vectors $\boldsymbol{y}$ with $\boldsymbol{A}\boldsymbol{y} \geq \boldsymbol{0}$ if and only if $\boldsymbol{x}$ is a nonnegative linear combination of the rows of $\boldsymbol{A}$. Prove Farkas' lemma assuming that the rows of $\boldsymbol{A}$ are linearly independent.

13. It is possible to give an elementary proof of the necessity of the Lagrange multiplier rule under a set of alternative conditions [28].

In the setting of Proposition 5.2.1, suppose that the equality constraints $g_i(\boldsymbol{y})$ are affine and that $f(\boldsymbol{y})$ and the inequality constraints $h_j(\boldsymbol{x})$ are merely differentiable at the constrained minimum $\boldsymbol{x}$. The constraint qualification is simplified by taking all $h_j(\boldsymbol{y})$ to be active at $\boldsymbol{x}$ and requiring the matrix

$$\boldsymbol{M} \;=\; \begin{pmatrix} dg(\boldsymbol{x}) \\ dh(\boldsymbol{x}) \end{pmatrix}$$

to have full row rank, where $dg(\boldsymbol{x})$ and $dh(\boldsymbol{x})$ stack the differentials of the $g_i(\boldsymbol{x})$ and $h_j(\boldsymbol{x})$ row by row. The number of columns of $\boldsymbol{M}$ should exceed the number of rows of $\boldsymbol{M}$. To validate the Lagrange multiplier rule, first argue that the equation

$$\begin{pmatrix} df(\boldsymbol{x}) \\ \boldsymbol{M} \end{pmatrix} \boldsymbol{u} \;=\; \begin{pmatrix} c \\ \boldsymbol{0} \\ \boldsymbol{v} \end{pmatrix} \tag{5.15}$$

can have no solution $\boldsymbol{u}$ when c and the entries of $\boldsymbol{v}$ are all negative numbers. Indeed, if a solution exists, then show that $\boldsymbol{x}+t\boldsymbol{u}$ is feasible for $t > 0$ small enough and satisfies $f(\boldsymbol{x}+t\boldsymbol{u}) < f(\boldsymbol{x})$. Next argue that the full rank assumption implies that equation (5.1) holds for some set of Lagrange multipliers. The real question is whether all μ_j are nonnegative. Suppose otherwise and construct a vector $\boldsymbol{v}$ with all entries negative such that the inner product $\boldsymbol{\mu}^*\boldsymbol{v}$ is positive. The full rank assumption then implies that the equation

$$\boldsymbol{M}\boldsymbol{u} \;=\; \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{v} \end{pmatrix}$$

has a solution $\boldsymbol{u}$. Finally, demonstrate that taking $c = -\boldsymbol{\mu}^*\boldsymbol{v}$ forces $\boldsymbol{u}$ to solve equation (5.15). This contradiction proves that all μ_j are nonnegative. (Hints: Full row rank implies full column rank. Hence, if the matrix appearing on the left side of equation (5.15) has full row rank, then the equation is solvable for any choice of c and $\boldsymbol{v}$. This impossibility implies that $df(\boldsymbol{x})$ can be expressed as a linear combination of the rows of $\boldsymbol{M}$.)

14. A random variable takes the value x_i with probability p_i for i ranging from 1 to n. Maximize the entropy $-\sum_{i=1}^n p_i \ln p_i$ subject to a fixed mean $m = \sum_{i=1}^n x_i p_i$. Show that $p_i = \alpha e^{\lambda x_i}$ for constants α and λ. Argue that λ is determined by the equation

$$\sum_{i=1}^n x_i e^{\lambda x_i} \;=\; m \sum_{i=1}^n e^{\lambda x_i}.$$

15. Continuing the previous problem, suppose that each $x_j = j$. At the maximum, show that

$$p_i = \frac{p^{i-1}(1-p)}{1-p^n}$$

for some $p > 0$ and all $1 \leq i \leq n$. Argue that p exists and is unique for $n > 1$.

16. Establish the bound

$$\sum_{m=1}^{n} \left(p_m + \frac{1}{p_m} \right)^2 \geq n^3 + 2n + \frac{1}{n}$$

for a discrete probability density $p_1, \ldots, p_n$. Determine a necessary and sufficient condition for equality to hold [243].

17. Consider the problem of minimizing the continuously differentiable function $f(\boldsymbol{x})$ subject to the constraint $\boldsymbol{x} \geq \boldsymbol{0}$. At a local minimum $\boldsymbol{y}$ demonstrate that the partial derivative $\partial_i f(\boldsymbol{y}) = 0$ when $y_i > 0$ and $\partial_i f(\boldsymbol{y}) \geq 0$ when $y_i = 0$.

18. As a variation on Problem 17, consider minimizing the continuously differentiable function $f(\boldsymbol{x})$ subject to the constraints $\sum_{i=1}^{n} x_i = 1$ and $\boldsymbol{x} \geq \boldsymbol{0}$. At a local minimum $\boldsymbol{y}$ demonstrate that there exists a number λ such that the partial derivative $\partial_i f(\boldsymbol{y}) = \lambda$ when $y_i > 0$ and $\partial_i f(\boldsymbol{y}) \geq \lambda$ when $y_i = 0$. This result is known as Gibbs' lemma.

19. Prove Nesbitt's inequality

$$\sum_{k=1}^{n} \frac{p_k}{\sum_{j \neq k} p_j} \geq \frac{n}{n-1}$$

for a discrete probability density $p_1, \ldots, p_n$. Determine a necessary and sufficient condition for equality to hold [243].

20. Find the minimum value of $f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$ subject to the constraints $\sum_{i=1}^{n} x_i = 1$ and $\boldsymbol{x} \geq \boldsymbol{0}$. Interpret the result geometrically.

21. For $p > 1$ define the norm $\|\boldsymbol{x}\|_p$ on R^n satisfying $\|\boldsymbol{x}\|_p^p = \sum_{i=1}^{n} |x_i|^p$. For a fixed vector $\boldsymbol{z}$, maximize $f(\boldsymbol{x}) = \boldsymbol{z}^* \boldsymbol{x}$ subject to $\|\boldsymbol{x}\|_p^p \leq 1$. Deduce Hölder's inequality $|\boldsymbol{z}^* \boldsymbol{x}| \leq \|\boldsymbol{x}\|_p \|\boldsymbol{z}\|_q$ for q defined by the equation $p^{-1} + q^{-1} = 1$.

22. Suppose $\boldsymbol{A}$ is an $n \times n$ positive definite matrix. Find the minimum of $f(\boldsymbol{x}) = \frac{1}{2} \boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x}$ subject to the constraint $\boldsymbol{z}^* \boldsymbol{x} - c \leq 0$. It may help to consider the cases $c \geq 0$ and $c < 0$ separately.

23. Suppose that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ are orthogonal eigenvectors of the $n \times n$ symmetric matrix $\boldsymbol{M}$. Subject to the constraints

$$\|\boldsymbol{x}\|^2 = 1, \quad \boldsymbol{v}_i^* \boldsymbol{x} = 0, \quad 1 \leq i \leq m < n,$$

show that a minimum of $\boldsymbol{x}^* \boldsymbol{M} \boldsymbol{x}$ must coincide with an eigenvector of $\boldsymbol{M}$. Under what circumstances is there a unique minimum of $\boldsymbol{x}^* \boldsymbol{M} \boldsymbol{x}$ subject to the constraints?

24. In the context of Proposition 5.2.1, suppose at the feasible point $\boldsymbol{x}$ one has $df(\boldsymbol{x})\boldsymbol{v} > 0$ for every nontrivial tangent vector $\boldsymbol{v}$. Recall that $\boldsymbol{v}$ satisfies $dg_i(\boldsymbol{x})\boldsymbol{v} = 0$ for all equality constraints and $dh_j(\boldsymbol{x})\boldsymbol{v} \leq 0$ for all active inequality constraints. Prove that there exists a positive constant c such that

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + c\|\boldsymbol{y} - \boldsymbol{x}\|$$

for all feasible points $\boldsymbol{y}$ close enough to $\boldsymbol{x}$. Thus, $\boldsymbol{x}$ represents a local minimum of $f(\boldsymbol{y})$. (Hints: If the contrary is true, then there exists a sequence of feasible points $\boldsymbol{x}_m$ such that

$$\|\boldsymbol{x}_m - \boldsymbol{x}\| < \frac{1}{m} \text{ and } f(\boldsymbol{x}_m) < f(\boldsymbol{x}) + \frac{1}{m}\|\boldsymbol{x}_m - \boldsymbol{x}\|.$$

Pass to a subsequence so that $\|\boldsymbol{x}_m - \boldsymbol{x}\|^{-1}(\boldsymbol{x}_m - \boldsymbol{x})$ converges.)

25. Continuing Problem 24, let $f(\boldsymbol{x}) = 3x_1 - x_2 + x_1 x_2$ and $h_1(\boldsymbol{x}) = -x_1$, $h_2(\boldsymbol{x}) = x_1 - x_2$, and $h_3(\boldsymbol{x}) = x_2 - 2x_1$. Show that $\boldsymbol{0}$ represents a local minimum by demonstrating the condition $df(\boldsymbol{0})\boldsymbol{v} > 0$ for every nontrivial tangent vector $\boldsymbol{v}$.

26. In Problem 24, the strict inequality $df(\boldsymbol{x})\boldsymbol{v} > 0$ cannot be relaxed to simple inequality. As an example take $f(\boldsymbol{x}) = x_2$ subject to the constraint $h_1(\boldsymbol{x}) = -x_1^2 - x_2 \leq 0$. Demonstrate that $df(\boldsymbol{0})\boldsymbol{v} \geq 0$ for every vector $\boldsymbol{v}$ satisfying $dh_1(\boldsymbol{0})\boldsymbol{v} \leq 0$, yet $\boldsymbol{0}$ is not a local minimum.

27. Assume that the objective function $f(\boldsymbol{x})$ and the equality constraints $g_i(\boldsymbol{x})$ in an equality constrained minimization problem are continuously differentiable. If the gradients $\nabla g_i(\boldsymbol{y})$ at a local minimum $\boldsymbol{y}$ are linearly independent, then the standard Lagrange multiplier rule holds at $\boldsymbol{y}$. If in addition $f(\boldsymbol{x})$ and the $g_i(\boldsymbol{x})$ possess second differentials at $\boldsymbol{y}$, then show that the second differential $d^2\mathcal{L}(\boldsymbol{y})$ of the Lagrangian satisfies $\boldsymbol{v}^* d^2\mathcal{L}(\boldsymbol{x})\boldsymbol{v} \geq 0$ for every tangent direction $\boldsymbol{v}$. (Hints: As demonstrated in Example 4.6.3, there is a one-to-one correspondence between tangent vectors and tangent curves. Expand $\mathcal{L}(\boldsymbol{x})$ to second order around $\boldsymbol{y}$, and use the fact that it coincides with $f(\boldsymbol{x})$ at feasible points.)

28. Demonstrate that the quadratic function $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x} + c$ is unbounded below if either (a) $\boldsymbol{A}$ is not positive semidefinite, or (b) $\boldsymbol{A}$ is positive semidefinite and $\boldsymbol{A}\boldsymbol{x} = -\boldsymbol{b}$ has no solution. Why does $f(\boldsymbol{x})$ attain its minimum when it is bounded below? (Hints: Diagonalize $\boldsymbol{A}$ in the form $\boldsymbol{O}\boldsymbol{D}\boldsymbol{O}^*$, where $\boldsymbol{O}$ is orthogonal and $\boldsymbol{D}$ is diagonal. Consider the transformed function $g(\boldsymbol{z}) = f(\boldsymbol{x})$ with $\boldsymbol{z} = \boldsymbol{O}^*\boldsymbol{x}$.)

29. Sylvester's criterion states that an $n \times n$ symmetric matrix $\boldsymbol{A}$ is positive definite if and only if its leading principal minors are positive. To prove this result by induction, verify it in the scalar case $n = 1$. Now consider the $(n+1) \times (n+1)$ symmetric block matrix

$$\boldsymbol{A} \;=\; \begin{pmatrix} \boldsymbol{B} & \boldsymbol{b} \\ \boldsymbol{b}^* & c \end{pmatrix},$$

where $\boldsymbol{B}$ is $n \times n$ and positive definite. Define the function

$$f(\boldsymbol{x}) \;=\; (\boldsymbol{x}^*, 1)\boldsymbol{A}\begin{pmatrix} \boldsymbol{x} \\ 1 \end{pmatrix} \;=\; \boldsymbol{x}^*\boldsymbol{B}\boldsymbol{x} + 2\boldsymbol{b}^*\boldsymbol{x} + c,$$

and show that $\boldsymbol{x} = -\boldsymbol{B}^{-1}\boldsymbol{b}$ furnishes its minimum. At that point the function has value $\boldsymbol{c} - \boldsymbol{b}^*\boldsymbol{B}^{-1}\boldsymbol{b}$. Thus, $f(\boldsymbol{x})$ is positive provided $\boldsymbol{c} - \boldsymbol{b}^*\boldsymbol{B}^{-1}\boldsymbol{b}$ is positive. Next verify that

$$\det \boldsymbol{A} \;=\; (c - \boldsymbol{b}^*\boldsymbol{B}^{-1}\boldsymbol{b}) \det \boldsymbol{B}$$

by checking that

$$\begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{d}^* & 1 \end{pmatrix}\begin{pmatrix} \boldsymbol{B} & \boldsymbol{b} \\ \boldsymbol{b}^* & c \end{pmatrix}\begin{pmatrix} \boldsymbol{I} & \boldsymbol{d} \\ \boldsymbol{0}^* & 1 \end{pmatrix} \;=\; \begin{pmatrix} \boldsymbol{B} & \boldsymbol{0} \\ \boldsymbol{0}^* & c - \boldsymbol{b}^*\boldsymbol{B}^{-1}\boldsymbol{b} \end{pmatrix}$$

for an appropriate vector $\boldsymbol{d}$. Put all of these hints together, and advance the induction from n to $n+1$.

30. Let $\boldsymbol{M}$ be an $m \times n$ matrix. Show that there exist unit vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ such that $\boldsymbol{M}\boldsymbol{v} = \|\boldsymbol{M}\|\boldsymbol{u}$ and $\boldsymbol{M}^*\boldsymbol{u} = \|\boldsymbol{M}\|\boldsymbol{v}$, where $\|\boldsymbol{M}\|$ is the matrix norm induced by the Euclidean norms on R^m and R^n. (Hint: See Problem 7 in Chap. 2.)

31. Consider the set of $n \times n$ matrices $\boldsymbol{M} = (m_{ij})$. Demonstrate that $\det \boldsymbol{M}$ has maximum value $\prod_{i=1}^n d_i$ subject to the constraints

$$\sqrt{\sum_{j=1}^n m_{ij}^2} \;=\; d_i$$

for $1 \leq i \leq n$. This is Hadamard's inequality. (Hints: Use the Lagrange multiplier condition and the identities $\det \boldsymbol{M} = \sum_{j=1}^n m_{ij}M_{ij}$ and $(\boldsymbol{M}^{-1})_{ij} = M_{ji}/\det \boldsymbol{M}$ to show that $\boldsymbol{M}$ can be written as the product $\boldsymbol{D}\boldsymbol{R}$ of a diagonal matrix $\boldsymbol{D}$ with diagonal entries d_i and an orthogonal matrix $\boldsymbol{R}$ with $\det \boldsymbol{R} = 1$.)

32. For m a positive integer, verify the explicit second slope

$$s^2(y,x) = 2[y^{m-2} + 2y^{m-3}x + 3y^{m-4}x^2 + \cdots + (m-1)x^{m-2}]$$

of the function $f(x) = x^m$ on R. Show that

$$\lim_{y \to x} s^2(y,x) = m(m-1)x^{m-2}.$$

33. Supply the missing algebraic steps in Example 5.4.5.

34. Let $f(y)$ be a real-valued function of the real variable y. Suppose that $f''(y)$ exists at a point x. Prove that

$$f''(x) = \lim_{u \to 0} \frac{f(x+u) - 2f(x) + f(x-u)}{u^2}.$$

Use Problem 2 of Chap. 4 to devise an example where this limit quotient exists but $f''(x)$ does not exist.

35. Suppose that $f(x)$ is twice differentiable on the interval $(0, \infty)$. If $m_j = \sup_x |f^{(j)}(x)|$, then show that $m_1^2 \leq 4m_0m_2$. (Hints: Expanding $f(x)$ in a second-order Taylor series, demonstrate that

$$|f'(x)| \leq \frac{2m_0}{h} + \frac{m_2 h}{2}$$

for all positive h. Choose h to minimize the right-hand side.)

36. Show that the function $f(\boldsymbol{x}) = e^{x_1} \ln(1 + x_2)$ on R^2 has the second-order Taylor expansion

$$f(\boldsymbol{x}) = x_2 + x_1x_2 - \frac{1}{2}x_2^2 + R_2(\boldsymbol{x})$$

around $\mathbf{0}$ with remainder $R_2(\boldsymbol{x})$.

37. Assume the functions $f(y)$ and $g(y)$ mapping R into R are differentiable of order p at the point x. If $f^{(m)}(x) = g^{(m)}(x) = 0$ for all $m < p$ but $g^{(p)}(x) \neq 0$, then demonstrate L'Hôpital's rule

$$\lim_{y \to x} \frac{f(y)}{g(y)} = \frac{f^{(p)}(x)}{g^{(p)}(x)}.$$

Find the limit of the ratio $\sin^2 x/(e^{x^2} - 1)$ as x tends to 0. (Hint: Consider the Taylor expansion (5.8) for $f(y)$ and the analogous expansion for $g(y)$.)

38. Suppose the function $f(y) : \mathsf{R} \mapsto \mathsf{R}$ is differentiable of order $p > 1$ in a neighborhood of a point x where $f^{(m)}(x) = 0$ for $1 \leq m < p$ and $f^{(p)}(x) \neq 0$. If p is odd, then show that x is a saddlepoint. If p is even, then show that x is a minimum point when $f^{(p)}(x) > 0$ and a maximum point when $f^{(p)}(x) < 0$. (Hint: Invoke the Taylor expansion (5.8).)

6
Convexity

6.1 Introduction

Convexity is one of the cornerstones of mathematical analysis and has interesting consequences for optimization theory, statistical estimation, inequalities, and applied probability. Despite this fact, students seldom see convexity presented in a coherent fashion. It always seems to take a backseat to more pressing topics. The current chapter is intended as a partial remedy to this pedagogical gap.

We start with convex sets and proceed to convex functions. These intertwined concepts define and illuminate all sorts of inequalities. It is helpful to have a variety of tests to recognize convex functions. We present such tests and discuss the important class of log-convex functions. A strictly convex function has at most one minimum point. This property tremendously simplifies optimization. For a few functions, we are fortunate enough to be able to find their optima explicitly. For other functions, we must iterate.

The definition of a convex function can be extended in various ways. The quasi-convex functions mentioned in the current chapter serve as substitutes for convex functions in many optimization arguments. Later chapters will extend the notion of a convex function to include functions with infinite values. Mathematicians by nature seek to isolate the key properties that drive important theories. However, too much abstraction can be a hinderance in learning. For now we stick to the concrete setting of ordinary convex functions.

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_6,

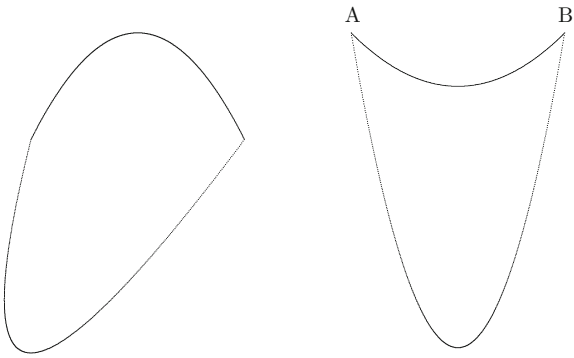

FIGURE 6.1. A convex set on the *left* and a non-convex set on the *right*

The concluding section of this chapter rigorously treats several inequalities from the perspective of probability theory. Our inclusion of Bernstein's proof of the Weierstrass approximation theorem provides a surprising application of Chebyshev's inequality and illustrates the role of probability theory in solving problems outside its usual sphere of influence. The less familiar inequalities of Jensen, Schlömilch, and Hölder find numerous applications in optimization theory and functional analysis.

6.2 Convex Sets

A set $S \subset \mathsf{R}^n$ is said to be convex if the line segment between any two points $\boldsymbol{x}$ and $\boldsymbol{y}$ of S lies entirely within S. Formally, this means that whenever $\boldsymbol{x}, \boldsymbol{y} \in S$ and $\alpha \in [0,1]$, the point $\boldsymbol{z} = \alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y} \in S$ as well. In general, any convex combination $\sum_{i=1}^m \alpha_i \boldsymbol{x}_i$ of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ from S must also reside in S. Here, the coefficients α_i are nonnegative and sum to 1. Figure 6.1 depicts two sets S in R^2, one convex and the other non-convex. The set on the right fails the line segment test for the segment connecting its two cusps A and B.

It is easy to concoct other examples of convex sets. For example, every interval on the real line is convex; every ball in R^n, either open or closed, is convex; and every multidimensional rectangle, either open, closed, or neither, is convex. Halfspaces and affine subspaces are convex. The former can be open or closed; the latter are always closed. Finitely generated cones as described in Example 2.4.1 are closed convex cones. The set of $n \times n$ positive semidefinite matrices is a closed convex cone in the space of symmetric matrices. The set of $n \times n$ positive definite matrices treated in Example 2.5.4 is an open convex set in the same space. It is not a convex cone because it excludes the $\mathbf{0}$ matrix.

These examples suggest several of the important properties listed in the next proposition.

Proposition 6.2.1 *Convex sets in* R^n *enjoy the following properties:*

(a) *The closure of a convex set is convex.*

(b) *The interior of a convex set is convex.*

(c) *A convex set is connected.*

(d) *The intersection of an arbitrary number of convex sets is convex.*

(e) *The image and inverse image of a convex set under an affine map* $f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ *are convex.*

(f) *The Cartesian product* $S \times T$ *of two convex sets* S *and* T *is convex.*

Proof: To prove assertion (a), consider two points $\boldsymbol{x}$ and $\boldsymbol{y}$ in the closure of the convex set S. There exist sequences $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ from S converging to $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. The convex combination $\boldsymbol{w}_k = \alpha\boldsymbol{u}_k + (1-\alpha)\boldsymbol{v}_k$ is in S as well and converges to $\boldsymbol{z} = \alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}$ in the closure of S. To verify assertion (b), suppose $\boldsymbol{x}$ and $\boldsymbol{y}$ lie in the interior of S. For r sufficiently small, S contains the two balls $\boldsymbol{x} + B(\boldsymbol{0}, r)$ and $\boldsymbol{y} + B(\boldsymbol{0}, r)$ of radius r. Consider a point $\boldsymbol{w} = \alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y} + \boldsymbol{z}$ with $\boldsymbol{z} \in B(\boldsymbol{0}, r)$. The decomposition

$$\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y} + \boldsymbol{z} \quad = \quad \alpha(\boldsymbol{x} + \boldsymbol{z}) + (1-\alpha)(\boldsymbol{y} + \boldsymbol{z})$$

makes it obvious that $\boldsymbol{w}$ also lies in S. Assertion (c) is a consequence of the fact that an arcwise connected set is connected. Assertions (d), (e), and (f) follow directly from the definitions. ■

Some obvious corollaries can be drawn from items (e) and (f) of Proposition 6.2.1. For example, suppose S and T are convex sets and λ is any real number. Then we can assert that the sets λS and $S + T$ are convex. Convex sets have many other crucial properties. Here is one that figures prominently in optimization theory.

Proposition 6.2.2 *For a convex set* S *of* R^n*, there is at most one point* $\boldsymbol{y} \in S$ *attaining the minimum distance* $\operatorname{dist}(\boldsymbol{x}, S)$ *from* $\boldsymbol{x}$ *to* S*. If* S *is closed, there is exactly one point.*

Proof: These claims are obvious if $\boldsymbol{x}$ is in S. Suppose that $\boldsymbol{x}$ is not in S and that $\boldsymbol{y}$ and $\boldsymbol{z}$ in S both attain the minimum. Then $\frac{1}{2}(\boldsymbol{y} + \boldsymbol{z}) \in S$ and

$$\begin{aligned} \operatorname{dist}(\boldsymbol{x}, S) &\leq \|\boldsymbol{x} - \frac{1}{2}(\boldsymbol{y} + \boldsymbol{z})\| \\ &\leq \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\| + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\| \\ &= \operatorname{dist}(\boldsymbol{x}, S). \end{aligned}$$

Hence, equality must hold in the displayed triangle inequality. This is possible if and only if $\boldsymbol{x} - \boldsymbol{y} = c(\boldsymbol{x} - \boldsymbol{z})$ for some positive number c. In view of the fact that $\operatorname{dist}(\boldsymbol{x}, S) = \|\boldsymbol{x} - \boldsymbol{y}\| = \|\boldsymbol{x} - \boldsymbol{z}\|$, the value of c is 1, and $\boldsymbol{y}$ and $\boldsymbol{z}$ coincide. The second claim follows from Example 2.5.5 of Chap. 2. ■

Another important property relates to separation by hyperplanes.

Proposition 6.2.3 *Consider a closed convex set S of R^n and a point $\boldsymbol{x}$ outside S. There exists a unit vector $\boldsymbol{v}$ and real number c such that*

$$\boldsymbol{v}^*\boldsymbol{x} \quad > \quad c \quad \geq \quad \boldsymbol{v}^*\boldsymbol{z} \tag{6.1}$$

for all $\boldsymbol{z} \in S$. As a consequence, S equals the intersection of all closed halfspaces containing it. If $\boldsymbol{x}$ is a boundary point of a convex set S, then there exists a unit vector $\boldsymbol{v}$ such that $\boldsymbol{v}^*\boldsymbol{x} \geq \boldsymbol{v}^*\boldsymbol{z}$ for all $\boldsymbol{z} \in S$.

Proof: Let $\boldsymbol{y}$ be the closest point to $\boldsymbol{x}$ in S. Suppose that we can prove the obtuse angle criterion

$$(\boldsymbol{x} - \boldsymbol{y})^*(\boldsymbol{z} - \boldsymbol{y}) \quad \leq \quad 0 \tag{6.2}$$

for all $\boldsymbol{z} \in S$. If we take $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{y}$, then $c = \boldsymbol{v}^*\boldsymbol{y} \geq \boldsymbol{v}^*\boldsymbol{z}$ for all $\boldsymbol{z} \in S$. Furthermore, $\boldsymbol{v}^*\boldsymbol{x} > \boldsymbol{v}^*\boldsymbol{y} = c$ because $\boldsymbol{v}^*\boldsymbol{v} = \|\boldsymbol{v}\|^2 > 0$. One can clearly replace $\boldsymbol{v}$ by $\|\boldsymbol{v}\|^{-1}\boldsymbol{v}$ without disrupting the separation inequalities (6.1).

To prove inequality (6.2), suppose it fails. Then $(\boldsymbol{x} - \boldsymbol{y})^*(\boldsymbol{z} - \boldsymbol{y}) > 0$ for some $\boldsymbol{z} \in S$. For each $0 < \alpha < 1$, the point $\alpha\boldsymbol{z} + (1 - \alpha)\boldsymbol{y}$ is in S, and

$$\begin{aligned} \|\boldsymbol{x} - \alpha\boldsymbol{z} - (1 - \alpha)\boldsymbol{y}\|^2 &= \|\boldsymbol{x} - \boldsymbol{y} - \alpha(\boldsymbol{z} - \boldsymbol{y})\|^2 \\ &= \|\boldsymbol{x} - \boldsymbol{y}\|^2 - \alpha\left[2(\boldsymbol{x} - \boldsymbol{y})^*(\boldsymbol{z} - \boldsymbol{y}) - \alpha\|\boldsymbol{z} - \boldsymbol{y}\|^2\right]. \end{aligned}$$

For α sufficiently small, the term above in square brackets is positive, so $\alpha\boldsymbol{z} + (1 - \alpha)\boldsymbol{y}$ improves on the choice of $\boldsymbol{y}$. This contradiction demonstrates inequality (6.2).

If $\boldsymbol{x}$ is a boundary point of S, then there exists a sequence of points $\boldsymbol{x}_i$ outside S that converge to x. Let $\boldsymbol{v}_i$ be the unit vector defining the hyperplane separating $\boldsymbol{x}_i$ from S. Without loss of generality, we can assume that some subsequence $\boldsymbol{v}_{i_j}$ converges to a unit vector $\boldsymbol{v}$. Taking limits in the strict inequality $\boldsymbol{v}_{i_j}^*\boldsymbol{x}_{i_j} > \boldsymbol{v}_{i_j}^*\boldsymbol{z}$ for $\boldsymbol{z} \in S$ then yields the desired result $\boldsymbol{v}^*\boldsymbol{x} \geq \boldsymbol{v}^*\boldsymbol{z}$. ■

Example 6.2.1 *Farkas' Lemma and Markov Chains*

As a continuation of Example 2.4.1, consider the finitely generated cone

$$C \quad = \quad \Big\{\sum_{i=1}^{m} \alpha_i \boldsymbol{v}_i : \alpha_i \geq 0,\ i = 1, \ldots, m\Big\}.$$

Because C is closed and convex, any point $\boldsymbol{x}$ outside C can be separated from C by a hyperplane. Thus, there exists a vector $\boldsymbol{w}$ and a constant

b with $\boldsymbol{w}^*\boldsymbol{x} > b \geq \boldsymbol{w}^*\boldsymbol{y}$ for all $\boldsymbol{y}$ in C. Given the origin $\mathbf{0}$ belongs to C, the constant $b \geq 0$. In fact, b must equal 0. If $\boldsymbol{w}^*\boldsymbol{y} > 0$ for some $\boldsymbol{y}$ in C, then $r\boldsymbol{w}^*\boldsymbol{y} > b$ for some $r > 0$. Since $r\boldsymbol{y}$ belongs to C, we reach a contradiction. Farkas' lemma summarizes these findings by posing two mutually exclusive alternatives, one of which must hold. Either the point $\boldsymbol{x}$ satisfies $\boldsymbol{x} = \sum_{i=1}^m \alpha_i \boldsymbol{v}_i$ for nonnegative constants α_i, or there exists a vector $\boldsymbol{w}$ with $\boldsymbol{w}^*\boldsymbol{x} > 0$ and $0 \geq \boldsymbol{w}^*\boldsymbol{v}_i$ for all i.

Farkas' lemma has a clever application in Markov chain theory [97]. Recall that a Markov chain on n states is governed by a transition probability matrix $\boldsymbol{P} = (p_{ij})$ with nonnegative entries and row sums equal to 1. A stationary distribution is a row vector $\boldsymbol{\alpha}$ with nonnegative entries α_i that sum equal to 1 and satisfy the equilibrium condition $\boldsymbol{\alpha} = \boldsymbol{\alpha}\boldsymbol{P}$. Farkas' lemma gives an easy proof that such vectors $\boldsymbol{\alpha}$ exist. The stationarity conditions can be restated as the vector equation

$$\begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix} = \sum_{i=1}^n \alpha_i \boldsymbol{v}_i$$

for vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ with entries $v_{ij} = p_{ij} - 1_{\{j=i\}}$ and $v_{i,n+1} = 1$ for all i and j between 1 and n. The Farkas alternative postulates the existence of a vector $\boldsymbol{w} = (w_1, \ldots, w_{n+1})^*$ with

$$\begin{aligned} \boldsymbol{w}^* \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix} &= w_{n+1} > 0 \\ \boldsymbol{w}^*\boldsymbol{v}_i &= \sum_{j=1}^n w_j p_{ij} - w_i + w_{n+1} \leq 0 \end{aligned}$$

for $1 \leq i \leq n$. Choosing $w_i = \min_{1 \leq j \leq n} w_j$, we find that

$$0 \geq \sum_{j=1}^n w_j p_{ij} - w_i + w_{n+1} \geq w_{n+1},$$

contradicting the assumption $w_{n+1} > 0$. Thus, the stationarity conditions must hold. ■

We now turn to a famous convexity result of Carathéodory and its application to compact sets. The convex hull of a set S, denoted $\operatorname{conv} S$, is the smallest convex set containing S. Equivalently, $\operatorname{conv} S$ consists of all convex combinations $\sum_i \alpha_i \boldsymbol{v}_i$ of points $\boldsymbol{v}_i$ from S. The convex hull displayed in Fig. 6.2 consists of the boundary plus all points internal to it.

Proposition 6.2.4 (Carathéodory) *Let S be a nonempty subset of R^n. Every vector from $\operatorname{conv} S$ can be represented as a convex combination of at most $n+1$ vectors from S. Furthermore, if S is compact, then $\operatorname{conv} S$ is also compact.*

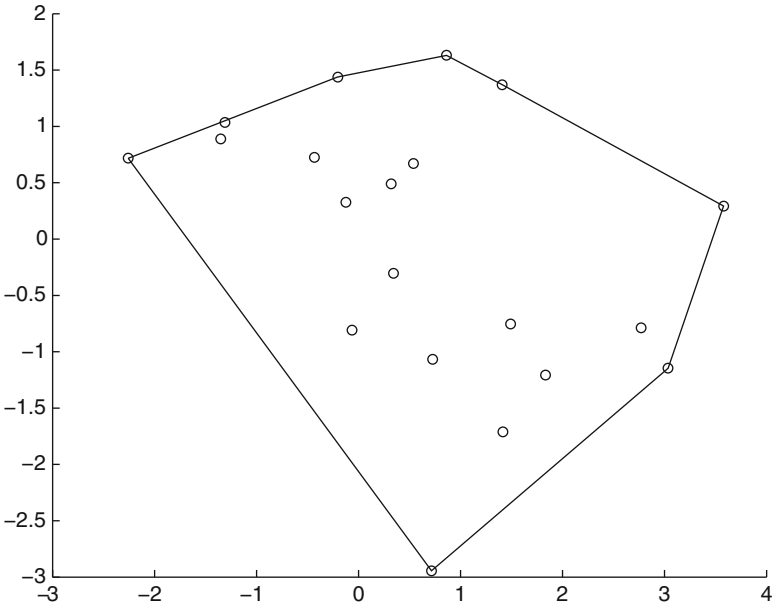

FIGURE 6.2. Convex hull of 20 random points in the plane

Proof: Consider the set $T = \{(\boldsymbol{y}, 1) : \boldsymbol{y} \in S\}$. A point $(\boldsymbol{x}, 1)$ in $\operatorname{conv} T$ can be represented as a convex combination $(\boldsymbol{x}, 1) = \sum_i \alpha_i(\boldsymbol{v}_i, 1)$. The point $(\boldsymbol{x}, 1)$ also belongs to the cone generated by the vectors $(\boldsymbol{v}_i, 1)$. As noted in Example 2.4.1, we can eliminate all but $n+1$ linearly independent vectors $(\boldsymbol{v}_{i_j}, 1)$ in this representation. It follows that $\boldsymbol{x} = \sum_j \beta_j \boldsymbol{v}_{i_j}$ with $1 = \sum_j \beta_j$ and all $\beta_j \geq 0$.

When S is compact, consider a sequence $\boldsymbol{x}_j = \sum_i \beta_{ji} \boldsymbol{v}_{ji}$ from $\operatorname{conv} S$. By the first part of the proposition, one can assume that the sum over i runs from 1 to $n+1$ at most. It suffices to prove that $\boldsymbol{x}_j$ possesses a subsequence that converges to a point in $\operatorname{conv} S$. By passing to successive subsequences as needed, one can assume that each β_{ji} converges to $\beta_i \geq 0$ and each $\boldsymbol{v}_{ji}$ converges to $\boldsymbol{v}_i \in S$. It follows that $\boldsymbol{x}_j$ converges to the convex combination $\sum_i \beta_i \boldsymbol{v}_i$ of points from S. ■

6.3 Convex Functions

Convex functions are defined on convex sets. A real-valued function $f(\boldsymbol{x})$ defined on a convex set S is convex provided

$$f[\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}] \quad \leq \quad \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) \tag{6.3}$$

for all $\boldsymbol{x}, \boldsymbol{y} \in S$ and $\alpha \in [0, 1]$. Figure 6.3 depicts how in one dimension definition (6.3) requires the chord connecting two points on the curve $f(x)$

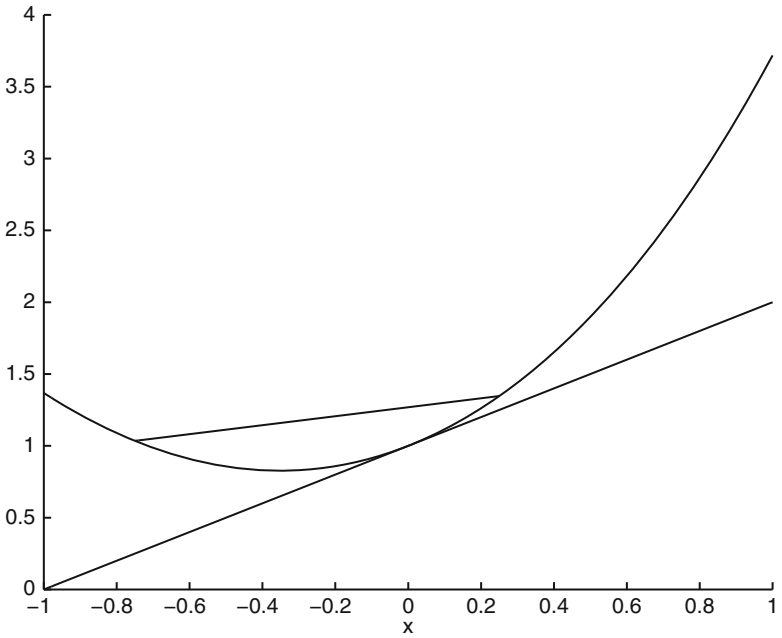

FIGURE 6.3. Plot of the convex function $e^x + x^2$

to lie above the curve. If strict inequality holds in inequality (6.3) for every $\boldsymbol{x} \neq \boldsymbol{y}$ and $\alpha \in (0,1)$, then $f(\boldsymbol{x})$ is said to be strictly convex. One can prove by induction that inequality (6.3) extends to

$$f\Big(\sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i\Big) \quad \leq \quad \sum_{i=1}^{m} \alpha_i f(\boldsymbol{x}_i)$$

for any convex combination of points from S. This is the finite form of Jensen's inequality. Proposition 6.6.1 discusses an integral form. A concave function satisfies the reverse of inequality (6.3).

Example 6.3.1 *Affine Functions Are Convex*

For an affine function $f(\boldsymbol{x}) = \boldsymbol{a}^* \boldsymbol{x} + b$, equality holds in inequality (6.3). ■

Example 6.3.2 *Norms Are Convex*

The Euclidean norm $f(\boldsymbol{x}) = \|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2}$ satisfies the triangle inequality and the homogeneity condition $\|c\boldsymbol{x}\| = |c| \, \|\boldsymbol{x}\|$. Thus,

$$\|\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}\| \quad \leq \quad \|\alpha\boldsymbol{x}\| + \|(1-\alpha)\boldsymbol{y}\| \quad = \quad \alpha\|\boldsymbol{x}\| + (1-\alpha)\|\boldsymbol{y}\|$$

for every $\alpha \in [0,1]$. The same argument works for any norm. The choice $\boldsymbol{y} = 2\boldsymbol{x}$ gives equality in inequality (6.3) and shows that no norm is strictly convex. ■

Example 6.3.3 *The Distance to a Convex Set Is Convex*

The distance $\operatorname{dist}(\boldsymbol{x}, S)$ from a point $\boldsymbol{x} \in \mathsf{R}^n$ to a convex set S is convex in $\boldsymbol{x}$. Indeed, for any convex combination $\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}$, take sequences $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ from S such that

$$\begin{aligned} \operatorname{dist}(\boldsymbol{x}, S) &= \lim_{k\to\infty} \|\boldsymbol{x} - \boldsymbol{u}_k\| \\ \operatorname{dist}(\boldsymbol{y}, S) &= \lim_{k\to\infty} \|\boldsymbol{y} - \boldsymbol{v}_k\|. \end{aligned}$$

The points $\alpha\boldsymbol{u}_k + (1-\alpha)\boldsymbol{v}_k$ lie in S, and taking limits in the inequality

$$\begin{aligned} \operatorname{dist}[\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}, S] &\leq \|\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y} - \alpha\boldsymbol{u}_k - (1-\alpha)\boldsymbol{v}_k\| \\ &\leq \alpha\|\boldsymbol{x} - \boldsymbol{u}_k\| + (1-\alpha)\|\boldsymbol{y} - \boldsymbol{v}_k\| \end{aligned}$$

yields $\operatorname{dist}[\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}, S] \leq \alpha \operatorname{dist}(\boldsymbol{x}, S) + (1-\alpha)\operatorname{dist}(\boldsymbol{y}, S)$. ■

Example 6.3.4 *Convex Functions Generate Convex Sets*

Consider a convex function $f(\boldsymbol{x})$ defined on R^n. Examination of definition (6.3) shows that the sublevel sets $\{\boldsymbol{x} : f(\boldsymbol{x}) \leq c\}$ and $\{\boldsymbol{x} : f(\boldsymbol{x}) < c\}$ are convex for any constant c. They may be empty. Conversely, a closed convex set S can be represented as $\{\boldsymbol{x} : f(\boldsymbol{x}) \leq 0\}$ using the continuous convex function $f(\boldsymbol{x}) = \operatorname{dist}(\boldsymbol{x}, S)$. This result does not preclude the possibility that a convex set is a sublevel set of a nonconvex function. For instance, the set $\{\boldsymbol{x} : 1 - x_1x_2 \leq 0,\ x_1 \geq 0,\ x_2 \geq 0\}$ is convex while the function $1 - x_1x_2$ is nonconvex on the domain $\{\boldsymbol{x} : x_1 \geq 0,\ x_2 \geq 0\}$. ■

Example 6.3.5 *A Convex Function Has a Convex Epigraph*

The epigraph of a real-valued function $f(\boldsymbol{x})$ is defined as the set of points $(\boldsymbol{y}, r)$ with $f(\boldsymbol{y}) \leq r$. Roughly speaking, the epigraph is the region lying above the graph of $f(\boldsymbol{x})$. Consider two points $(\boldsymbol{y}, r)$ and $(\boldsymbol{z}, s)$ in the epigraph of $f(\boldsymbol{x})$. If $f(\boldsymbol{x})$ is convex, then

$$\begin{aligned} f[\alpha\boldsymbol{y} + (1-\alpha)\boldsymbol{z}] &\leq \alpha f(\boldsymbol{y}) + (1-\alpha)f(\boldsymbol{z}) \\ &\leq \alpha r + (1-\alpha)s, \end{aligned}$$

and the convex combination $\alpha(\boldsymbol{y}, r) + (1-\alpha)(\boldsymbol{z}, s)$ occurs in the epigraph of $f(\boldsymbol{x})$. Conversely, if the epigraph of $f(\boldsymbol{x})$ is a convex set, then $f(\boldsymbol{x})$ must be a convex function. ■

Figure 6.3 illustrates how a tangent line to a convex curve lies below the curve. This property characterizes convex differentiable functions.

Proposition 6.3.1 *Let $f(\boldsymbol{x})$ be a differentiable function on the open convex set $S \subset \mathsf{R}^n$. Then $f(\boldsymbol{x})$ is convex if and only if*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + df(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \tag{6.4}$$

for all $\boldsymbol{x}, \boldsymbol{y} \in S$. Furthermore, $f(\boldsymbol{x})$ is strictly convex if and only if strict inequality prevails in inequality (6.4) when $\boldsymbol{y} \neq \boldsymbol{x}$.

Proof: If $f(\boldsymbol{x})$ is convex, then we can rearrange inequality (6.3) to give

$$\begin{aligned}\frac{f[\boldsymbol{x}+(1-\alpha)(\boldsymbol{y}-\boldsymbol{x})]-f(\boldsymbol{x})}{(1-\alpha)} &= \frac{f[\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y}]-f(\boldsymbol{x})}{1-\alpha}\\ &\leq f(\boldsymbol{y})-f(\boldsymbol{x}).\end{aligned}$$

Letting α tend to 1 proves inequality (6.4). To demonstrate the converse, let $\boldsymbol{z}=\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y}$. Then with obvious notational changes, inequality (6.4) implies

$$\begin{aligned}f(\boldsymbol{x}) &\geq f(\boldsymbol{z})+df(\boldsymbol{z})(\boldsymbol{x}-\boldsymbol{z})\\ f(\boldsymbol{y}) &\geq f(\boldsymbol{z})+df(\boldsymbol{z})(\boldsymbol{y}-\boldsymbol{z}).\end{aligned}$$

Multiplying the first of these inequalities by α and the second by $1-\alpha$ and adding the results produce

$$\alpha f(\boldsymbol{x})+(1-\alpha)f(\boldsymbol{y}) \geq f(\boldsymbol{z})+df(\boldsymbol{z})(\boldsymbol{z}-\boldsymbol{z}) = f(\boldsymbol{z}),$$

which is just inequality (6.3). The claims about strict convexity are left to the reader. ■

It is useful to have simpler tests for convexity than inequalities (6.3) and (6.4). One such test involves the second differential $d^2f(\boldsymbol{x})$ of a function $f(\boldsymbol{x})$.

Proposition 6.3.2 *Consider a twice differentiable function $f(\boldsymbol{x})$ on the open convex set $S\subset\mathsf{R}^n$. If its second differential $d^2f(\boldsymbol{x})$ is positive semidefinite for all $\boldsymbol{x}$, then $f(\boldsymbol{x})$ is convex. If $d^2f(\boldsymbol{x})$ is positive definite for all $\boldsymbol{x}$, then $f(\boldsymbol{x})$ is strictly convex.*

Proof: The expansion

$$\begin{aligned}f(\boldsymbol{y}) &= f(\boldsymbol{x})+df(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})\\ &\quad +(\boldsymbol{y}-\boldsymbol{x})^*\int_0^1 d^2f[\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})](1-t)\,dt\,(\boldsymbol{y}-\boldsymbol{x})\end{aligned}$$

for $\boldsymbol{y}\neq\boldsymbol{x}$ shows that

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x})+df(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}),$$

with strict inequality when $d^2f(\boldsymbol{x})$ is positive definite for all $\boldsymbol{x}$. ■

Example 6.3.6 *Generalized Arithmetic-Geometric Mean Inequality*

The second derivative test shows that the function e^x is strictly convex. Taking $y_i=e^{x_i}$, $\sum_{i=1}^n\alpha_i=1$, and all $\alpha_i\geq 0$ produces the generalized arithmetic-geometric mean inequality

$$\prod_{i=1}^n y_i^{\alpha_i} \leq \sum_{i=1}^n \alpha_i y_i. \tag{6.5}$$

Equality holds if all y_i coincide or all but one α_i equals 0. ■

Example 6.3.7 *Strictly Convex Quadratic Functions*

If the matrix $\boldsymbol{A}$ is positive definite, then Proposition 6.3.2 implies that the quadratic function $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x} + c$ is strictly convex. ■

Even Proposition 6.3.2 can be difficult to apply. The next proposition helps us to recognize convex functions by their closure properties.

Proposition 6.3.3 *Convex functions satisfy the following:*

(a) *If $f(\boldsymbol{x})$ is convex and $g(\boldsymbol{x})$ is convex and increasing, then the functional composition $g \circ f(\boldsymbol{x})$ is convex.*

(b) *If $f(\boldsymbol{x})$ is convex, then the functional composition $f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ of $f(\boldsymbol{x})$ with an affine function $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ is convex.*

(c) *If $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are convex and α and β are nonnegative constants, then $\alpha f(\boldsymbol{x}) + \beta g(\boldsymbol{x})$ is convex.*

(d) *If $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are convex, then $\max\{f(\boldsymbol{x}), g(\boldsymbol{x})\}$ is convex.*

(e) *If $f_m(\boldsymbol{x})$ is a sequence of convex functions, then $\lim_{m\to\infty} f_m(\boldsymbol{x})$ is convex whenever it exists.*

Proof: To prove assertion (a), we calculate

$$\begin{aligned} g \circ f[\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}] &\leq g[\alpha f(\boldsymbol{x}) + (1-\alpha)f(\boldsymbol{y})] \\ &\leq \alpha g \circ f(\boldsymbol{x}) + (1-\alpha) g \circ f(\boldsymbol{y}). \end{aligned}$$

The remaining assertions are left to the reader. ■

Part (a) of Proposition 6.3.3 implies that $e^{f(\boldsymbol{x})}$ is convex when $f(\boldsymbol{x})$ is convex and that $f(\boldsymbol{x})^\beta$ is convex when $f(\boldsymbol{x})$ is nonnegative and convex and $\beta > 1$. One case not covered by the Proposition is products. The counterexample $x^3 = x^2 x$ shows that the product of two convex functions is not necessarily convex. In some situations the limit of a sequence of convex functions is no longer finite. Many authors consider $+\infty$ to be a legitimate value for a convex function while $-\infty$ is illegitimate. For the sake of simplicity, we prefer to deal with functions having only finite values. In Chap. 14 we relax this restriction.

Example 6.3.8 *Largest Eigenvalue of a Symmetric Matrix*

Part (d) of Proposition 6.3.3 can be generalized. Suppose the function $f(\boldsymbol{x}, \boldsymbol{y})$ is convex in $\boldsymbol{y}$ for each fixed $\boldsymbol{x}$. Then, provided it is finite, the function $\sup_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})$ is convex in $\boldsymbol{y}$. For a specific application, recall that Example 1.4.3 proves the formula

$$\lambda_{\max}(\boldsymbol{M}) = \max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^*\boldsymbol{M}\boldsymbol{x}$$

for the largest eigenvalue of a symmetric matrix. Because the map taking $\boldsymbol{M}$ into $\boldsymbol{x}^*\boldsymbol{M}\boldsymbol{x}$ is linear, it follows that $\lambda_{\max}(\boldsymbol{M})$ is convex in $\boldsymbol{M}$. Applying the same reasoning to $-\boldsymbol{M}$, we deduce that the minimum eigenvalue $\lambda_{\min}(\boldsymbol{M})$ is concave in $\boldsymbol{M}$. ■

Example 6.3.9 *Differences of Convex Functions*

Although the class of convex functions is rather narrow, most well-behaved functions can be expressed as the difference of two convex functions. For example, consider a polynomial $p(x) = \sum_{m=0}^{n} p_m x^m$. The second derivative test shows that x^m is convex whenever m is even. If m is odd, then x^m is convex on $[0, \infty)$, and $-x^m$ is convex on $(-\infty, 0)$. Therefore,

$$x^m \quad = \quad \max\{x^m, 0\} - \max\{-x^m, 0\}$$

is the difference of two convex functions. Because the class of differences of convex functions is closed under the formation of linear combinations, it follows that $p(x)$ belongs to this larger class. ■

A positive function $f(\boldsymbol{x})$ is said to be log-convex if $\ln f(\boldsymbol{x})$ is convex. Log-convex functions have excellent closure properties as documented by the next proposition.

Proposition 6.3.4 *Log-convex functions satisfy the following:*

(a) *If $f(\boldsymbol{x})$ is log-convex, then $f(\boldsymbol{x})$ is convex.*

(b) *If $f(\boldsymbol{x})$ is convex and $g(\boldsymbol{x})$ is log-convex and increasing, then the functional composition $g \circ f(\boldsymbol{x})$ is log-convex.*

(c) *If $f(\boldsymbol{x})$ is log-convex, then the functional composition $f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ of $f(\boldsymbol{x})$ with an affine function $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ is log-convex.*

(d) *If $f(\boldsymbol{x})$ is log-convex, then $f(\boldsymbol{x})^\alpha$ and $\alpha f(\boldsymbol{x})$ are log-convex for any $\alpha > 0$.*

(e) *If $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are log-convex, then $f(\boldsymbol{x}) + g(\boldsymbol{x})$, $\max\{f(\boldsymbol{x}), g(\boldsymbol{x})\}$, and $f(\boldsymbol{x})g(\boldsymbol{x})$ are log-convex.*

(f) *If $f_m(\boldsymbol{x})$ is a sequence of log-convex functions, then $\lim_{m\to\infty} f_m(\boldsymbol{x})$ is log-convex whenever it exists and is positive.*

Proof: Assertion (a) follows from part (a) of Proposition 6.3.3 after composing the functions e^x and $\ln f(\boldsymbol{x})$. To prove that the sum of log-convex functions is log-convex, we let $h(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$ and apply Hölder's inequality as stated in Problem 21 of Chap. 5 and in Example 6.6.3 later in this chapter. Taking $\alpha = 1/p$ and $1 - \alpha = 1/q$ consequently implies that

$$\begin{aligned} h[\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}] &= f[\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}] + g[\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}] \\ &\leq f(\boldsymbol{x})^\alpha f(\boldsymbol{y})^{1-\alpha} + g(\boldsymbol{x})^\alpha g(\boldsymbol{y})^{1-\alpha} \\ &\leq [f(\boldsymbol{x}) + g(\boldsymbol{x})]^\alpha [f(\boldsymbol{y}) + g(\boldsymbol{y})]^{1-\alpha} \\ &= h(\boldsymbol{x})^\alpha h(\boldsymbol{y})^{1-\alpha}. \end{aligned}$$

The remaining assertions are left to the reader. ■

Example 6.3.10 *The Convex Function of Gordon's Theorem*

In Proposition 5.3.2, we encountered the function

$$f(\boldsymbol{x}) = \ln \left[\sum_{j=1}^{r} \exp(\boldsymbol{z}_j^* \boldsymbol{x}) \right].$$

Given the log-convexity of the functions $\exp(\boldsymbol{z}_j^* \boldsymbol{x})$, we now recognize $f(\boldsymbol{x})$ as convex. This is one of the reasons for its success in Gordon's theorem. ■

Example 6.3.11 *Gamma Function*

Gauss's representation of the gamma function

$$\Gamma(z) = \lim_{n \to \infty} \frac{n! n^z}{z(z+1)\cdots(z+n)} \tag{6.6}$$

shows that it is log-convex on $(0, \infty)$ [132]. Indeed, one can easily check that n^z and $(z+k)^{-1}$ are log-convex and then apply the closure of the set of log-convex functions under the formation of products and limits. Note that invoking convexity in this argument is insufficient because the set of convex functions is not closed under the formation of products. Alternatively, one can deduce log-convexity from Euler's definition

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$$

by viewing the integral as the limit of Riemann sums, each of which is log-convex. ■

Example 6.3.12 *Log-concavity of* $\det \boldsymbol{\Sigma}$ *for* $\boldsymbol{\Sigma}$ *Positive Definite*

Let $\boldsymbol{\Omega}$ be an $n \times n$ positive definite matrix. According to Appendix A.2, the function

$$f(\boldsymbol{y}) = \left(\frac{1}{2\pi}\right)^{n/2} |\det \boldsymbol{\Omega}|^{-1/2} e^{-\boldsymbol{y}^* \boldsymbol{\Omega}^{-1} \boldsymbol{y}/2}$$

is a probability density. Integrating over all $\boldsymbol{y} \in \mathsf{R}^n$ produces the identity

$$|\det \boldsymbol{\Omega}|^{1/2} = \frac{1}{(2\pi)^{n/2}} \int e^{-\boldsymbol{y}^* \boldsymbol{\Omega}^{-1} \boldsymbol{y}/2} d\boldsymbol{y}.$$

We can restate this identity in terms of the inverse matrix $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$ as

$$\ln \det \boldsymbol{\Sigma} \quad = \quad n \ln(2\pi) - 2 \ln \int e^{-\boldsymbol{y}^* \boldsymbol{\Sigma} \boldsymbol{y}/2} d\boldsymbol{y}.$$

By the reasoning of the last two examples, the integral on the right is log-convex. Because $\boldsymbol{\Sigma}$ is positive definite if and only if $\boldsymbol{\Omega}$ is positive definite, it follows that $\ln \det \boldsymbol{\Sigma}$ is concave in the positive definite matrix $\boldsymbol{\Sigma}$. ■

6.4 Continuity, Differentiability, and Integrability

In this section we discuss some continuity, differentiability, and integrability properties of convex functions. Let us start with a sobering counterexample involving the closed unit ball $C(\mathbf{0}, 1)$ of R^n and a positive function $g(\boldsymbol{y})$ defined on its boundary. One can extend $g(\boldsymbol{y})$ to a convex function $f(\boldsymbol{x})$ with domain $C(\mathbf{0}, 1)$ by setting

$$f(\boldsymbol{x}) \quad = \quad \begin{cases} 0 & \|\boldsymbol{x}\| < 1 \\ g(\boldsymbol{x}) & \|\boldsymbol{x}\| = 1. \end{cases}$$

Even though $f(\boldsymbol{x})$ is convex throughout $C(\mathbf{0}, 1)$, it is discontinuous everywhere on the boundary. Even worse, $f(\boldsymbol{x})$ is not even lower semicontinuous. As the next proposition demonstrates, matters improve considerably if we restrict our attention to the interior of the domain of a convex function.

Proposition 6.4.1 *A convex function $f(\boldsymbol{x})$ is continuous on the interior of its domain and locally Lipschitz around every interior point. In other words, there exists a constant c such that $|f(\boldsymbol{z}) - f(\boldsymbol{y})| \leq c\|\boldsymbol{z} - \boldsymbol{y}\|$ for all $\boldsymbol{y}$ and $\boldsymbol{z}$ near $\boldsymbol{x}$.*

Proof: Let $\boldsymbol{y}$ be an interior point and $C(\boldsymbol{y}, r)$ be a closed ball of radius r around $\boldsymbol{y}$ contained within the domain of $f(\boldsymbol{x})$. Without loss of generality, we may assume that $\boldsymbol{y} = \mathbf{0}$. We first demonstrate that $f(\boldsymbol{x})$ is bounded above near $\mathbf{0}$. Define the $n + 1$ points

$$\boldsymbol{v}_0 \quad = \quad -\frac{r}{2n}\mathbf{1}, \quad \boldsymbol{v}_i \quad = \quad r\boldsymbol{e}_i - \frac{r}{2n}\mathbf{1}, \quad 1 \leq i \leq n,$$

using the standard basis $\boldsymbol{e}_i$ of R^n. It is easy to check that all of these points lie in $C(\mathbf{0}, r)$. Hence, any convex combination $\sum_{i=0}^n \alpha_i \boldsymbol{v}_i$ also lies in $C(\mathbf{0}, r)$. Even more surprising, any point $\boldsymbol{x}$ in the open interval

$$J \quad = \quad \prod_{i=1}^n \Big(-\frac{r}{2n}, \frac{r}{2n}\Big)$$

can be represented as such a convex combination. This assertion follows from the component-by-component equation

$$x_i \quad = \quad r\Big(\alpha_i - \frac{1}{2n}\Big)$$

with $\alpha_i \in (0, 1/n)$ and the identity

$$\begin{aligned}
\boldsymbol{x} &= \sum_{i=1}^{n} x_i \boldsymbol{e}_i \\
&= r \sum_{i=1}^{n} \alpha_i \boldsymbol{e}_i - \frac{r}{2n}\mathbf{1} \\
&= r \sum_{i=1}^{n} \alpha_i \Big(\boldsymbol{e}_i - \frac{1}{2n}\mathbf{1}\Big) - \Big(1 - \sum_{i=1}^{n} \alpha_i\Big)\frac{r}{2n}\mathbf{1}.
\end{aligned}$$

The boundedness of $f(\boldsymbol{x})$ on J now follows from the inequalities

$$f\Big(\sum_{i=0}^{n} \alpha_i \boldsymbol{v}_i\Big) \;\le\; \sum_{i=0}^{n} \alpha_i f(\boldsymbol{v}_i) \;\le\; \max\{f(\boldsymbol{v}_0), \ldots, f(\boldsymbol{v}_n)\}.$$

Without affecting its convexity, we now rescale and translate $f(\boldsymbol{x})$ so that $f(\mathbf{0}) = 0$ and $f(\boldsymbol{x}) \le 1$ on J. We also rescale $\boldsymbol{x}$ so that J contains the open ball $B(\mathbf{0}, 2)$. For any $\boldsymbol{x}$ in $B(\mathbf{0}, 2)$, we have

$$0 \;=\; f(\mathbf{0}) \;\le\; \frac{1}{2} f(\boldsymbol{x}) + \frac{1}{2} f(-\boldsymbol{x}).$$

It follows that $f(\boldsymbol{x})$ is bounded below by -1 on $B(\mathbf{0}, 2)$. The final step of the proof proceeds by choosing two distinct points $\boldsymbol{x}$ and $\boldsymbol{z}$ from the unit ball $B(\mathbf{0}, 1)$. If we define $\boldsymbol{w} = \boldsymbol{z} + t^{-1}(\boldsymbol{z} - \boldsymbol{x})$ with $t = \|\boldsymbol{z} - \boldsymbol{x}\|$, then $\boldsymbol{w} \in B(\mathbf{0}, 2)$,

$$\boldsymbol{z} \;=\; \frac{t}{1+t}\boldsymbol{w} + \frac{1}{1+t}\boldsymbol{x},$$

and

$$\begin{aligned}
f(\boldsymbol{z}) - f(\boldsymbol{x}) &\le \frac{t}{1+t} f(\boldsymbol{w}) + \frac{1}{1+t} f(\boldsymbol{x}) - f(\boldsymbol{x}) \\
&= \frac{t}{1+t} f(\boldsymbol{w}) - \frac{t}{1+t} f(\boldsymbol{x}) \\
&\le \frac{2t}{1+t} \\
&\le 2\|\boldsymbol{z} - \boldsymbol{x}\|.
\end{aligned}$$

Switching the roles of $\boldsymbol{x}$ and $\boldsymbol{z}$ gives $|f(\boldsymbol{z}) - f(\boldsymbol{x})| \le 2\|\boldsymbol{z} - \boldsymbol{x}\|$. This Lipschitz inequality establishes the continuity of $f(\boldsymbol{x})$ throughout $B(\mathbf{0}, 1)$. ■

We next turn to derivatives. The simplest place to start is forward directional derivatives. In the case of a convex function $f(x)$ defined on an interval (a, b), we will prove the existence of one-sided derivatives by establishing the inequalities

$$\frac{f(y) - f(x)}{y - x} \;\le\; \frac{f(z) - f(x)}{z - x} \;\le\; \frac{f(z) - f(y)}{z - y} \tag{6.7}$$

for all points $x < y < z$ drawn from (a, b). If we write

$$y = \frac{z-y}{z-x}x + \frac{y-x}{z-x}z,$$

then both of these inequalities are rearrangements of the inequality

$$f(y) \leq \frac{z-y}{z-x}f(x) + \frac{y-x}{z-x}f(z).$$

Careful examination of the inequalities (6.7) with relabeling of points as necessary leads to the conclusion that the slope

$$\frac{f(y)-f(x)}{y-x}$$

is bounded below and increasing in y for x fixed. Similarly, this same slope is bounded above and increasing in x for y fixed. It follows that both one-sided derivatives exist at y and satisfy

$$f'_-(y) = \lim_{x \uparrow y} \frac{f(y)-f(x)}{y-x} \leq \lim_{z \downarrow y} \frac{f(z)-f(y)}{z-y} = f'_+(y).$$

In view of the monotonicity properties of the slope, any number d between these two limits satisfies the supporting hyperplane inequalities

$$\begin{aligned} f(x) &\geq f(y) + d(x-y) \\ f(z) &\geq f(y) + d(z-y). \end{aligned}$$

Such a number is termed a subgradient. The existence of subgradients is closely tied to the fact that $f(x)$ is locally Lipschitz.

Our reasoning for convex functions defined on the real lines proves the existence of forward directional derivatives on higher-dimensional domains. Indeed, for any point $\boldsymbol{y}$ in the domain of $f(\boldsymbol{x})$, all one must do is focus on the function $g(t) = f(\boldsymbol{y} + t\boldsymbol{v})$ of the nonnegative scalar t. To define the difference quotient of $g(t)$ at 0, the point $\boldsymbol{y} + t\boldsymbol{v}$ must belong to the domain of $f(\boldsymbol{x})$ for all t sufficiently small. This is certainly possible when $\boldsymbol{y}$ occurs on the interior of the domain, but it is also possible for boundary points $\boldsymbol{y}$ and directions $\boldsymbol{v}$ that point into the domain of $f(\boldsymbol{x})$. In either case, the convexity of $f(\boldsymbol{x})$ carries over to $g(t)$. Furthermore, the directional derivative

$$d_{\boldsymbol{v}} f(\boldsymbol{y}) = \lim_{t \downarrow 0} \frac{f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x})}{t} = \lim_{t \downarrow 0} \frac{g(t) - g(0)}{t}.$$

is well defined and satisfies $d_{\boldsymbol{v}} f(\boldsymbol{y}) < \infty$. The value $d_{\boldsymbol{v}} f(\boldsymbol{y}) = -\infty$ can occur for boundary points $\boldsymbol{y}$ but can be ruled out for interior points. At an interior point $g(t)$ is defined for t small and negative.

Finally, let us consider the integrability of a convex function $f(x)$ defined on a closed interval $[a, b]$. On the interior of the interval, $f(x)$ is continuous. Continuity can fail at the endpoints, but typically we can restore it by replacing $f(a)$ by $\lim_{x\to a} f(x)$ and $f(b)$ by $\lim_{x\to b} f(x)$. If we assume these limits are finite, then $f(x)$ is continuous throughout the interval and hence integrable. More surprising is the fact that the fundamental theorem of calculus applies. The right-hand and left-hand derivatives $f'_+(x)$ and $f'_-(x)$ exist throughout (a, b), and

$$f(b) - f(a) \quad = \quad \int_a^b f'_+(x)\,dx \quad = \quad \int_a^b f'_-(x)\,dx. \tag{6.8}$$

Because $f'_-(x) \leq f'_+(x) \leq f'_-(y) \leq f'_+(y)$ when $x < y$, the interiors of the intervals $[f'_-(x), f'_+(x)]$ are disjoint. Choosing a rational number from each nonempty interior shows that the set of points where $f'_-(x) \neq f'_+(x)$ is countable. The value of an integral is insensitive to the value of its integrand at a countable number of points, and the discussion following Proposition 3.4.1 demonstrates that the fundamental theorem of calculus holds as stated in equation (6.8).

6.5 Minimization of Convex Functions

Optimization theory is much simpler for convex functions than for ordinary functions. The continuity and differentiability properties of convex functions certainly support this contention. Even more relevant are the following theoretical results.

Proposition 6.5.1 *Suppose that $f(\boldsymbol{y})$ is a convex function on the convex set $S \subset \mathsf{R}^n$. If $\boldsymbol{x}$ is a local minimum of $f(\boldsymbol{y})$, then it is a global minimum of $f(\boldsymbol{y})$, and the set $\{\boldsymbol{y} \in S : f(\boldsymbol{y}) = f(\boldsymbol{x})\}$ is convex. If $f(\boldsymbol{y})$ is strictly convex and $\boldsymbol{x}$ is a global minimum, then the solution set $\{\boldsymbol{y} \in S : f(\boldsymbol{y}) = f(\boldsymbol{x})\}$ consists of $\boldsymbol{x}$ alone.*

Proof: If $f(\boldsymbol{y}) \leq f(\boldsymbol{x})$ and $f(\boldsymbol{z}) \leq f(\boldsymbol{x})$, then

$$f[\alpha\boldsymbol{y} + (1-\alpha)\boldsymbol{z}] \quad \leq \quad \alpha f(\boldsymbol{y}) + (1-\alpha)f(\boldsymbol{z}) \quad \leq \quad f(\boldsymbol{x}) \tag{6.9}$$

for any $\alpha \in [0, 1]$. This shows that the set $\{\boldsymbol{y} \in S : f(\boldsymbol{y}) \leq f(\boldsymbol{x})\}$ is convex. Now suppose that $f(\boldsymbol{y}) < f(\boldsymbol{x})$. Strict inequality then prevails between the extreme members of inequality (6.9) provided $\alpha > 0$. Taking $\boldsymbol{z} = \boldsymbol{x}$ and α close to 0 shows that $\boldsymbol{x}$ cannot serve as a local minimum. This contradiction demonstrates that $\boldsymbol{x}$ must be a global minimum. Finally, if $f(\boldsymbol{y})$ is strictly convex, then strict inequality holds in the first half of equality (6.9) for all $\alpha \in (0, 1)$. This leads to another contradiction when $\boldsymbol{y} = \boldsymbol{x} \neq \boldsymbol{z}$, and both are minimum points. ■

Example 6.5.1 *Piecewise Linear Functions*

The function $f(x) = |x|$ on the real line is piecewise linear. It attains its minimum of 0 at the point $x = 0$. The convex function $f(x) = \max\{1, |x|\}$ is also piecewise linear, but it attains its minimum throughout the interval $[-1, 1]$. In both cases the set $\{y : f(y) = \min_x f(x)\}$ is convex. In higher dimensions, the convex function $f(\boldsymbol{x}) = \max\{1, \|\boldsymbol{x}\|\}$ attains its minimum of 1 throughout the closed ball $\|\boldsymbol{x}\| \leq 1$. ■

Proposition 6.5.2 *Let $f(\boldsymbol{y})$ be a convex function defined on a convex set $S \subset \mathsf{R}^n$. A point $\boldsymbol{x} \in S$ furnishes a global minimum of $f(\boldsymbol{y})$ if and only if the forward directional derivative $d_{\boldsymbol{v}} f(\boldsymbol{x})$ exists and is nonnegative for all tangent vectors $\boldsymbol{v} = \boldsymbol{z} - \boldsymbol{x}$ defined by $\boldsymbol{z} \in S$. In particular, a stationary point of $f(\boldsymbol{y})$ represents a global minimum.*

Proof: Suppose the condition holds and $\boldsymbol{z} \in S$. Then taking $t = 1$ and $\boldsymbol{v} = \boldsymbol{z} - \boldsymbol{x}$ in the inequality

$$\frac{f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x})}{t} \quad \geq \quad d_{\boldsymbol{v}} f(\boldsymbol{x})$$

shows that $f(\boldsymbol{z}) \geq f(\boldsymbol{x})$. Conversely, if $\boldsymbol{x}$ represents the minimum, then the displayed difference quotient is nonnegative. Sending t to 0 now gives $d_{\boldsymbol{v}} f(\boldsymbol{x}) \geq 0$. ■

Example 6.5.2 *Minimum of y on $[0, \infty)$*

The convex function $f(y) = y$ has derivative $f'(y) = 1$. On the convex set $[0, \infty)$, we have $f'(0)(z - 0) = z \geq 0$ for any $z \in [0, \infty)$. Hence, 0 provides the minimum of y. Of course, this is consistent with the Lagrange multiplier rule $f'(0) - 1 = 0$. ■

Example 6.5.3 *The Obtuse Angle Criterion*

As a continuation of Propositions 6.2.2 and 6.2.3, define $f(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2$ for $\boldsymbol{x} \notin S$. If $\boldsymbol{z}$ is the projection of $\boldsymbol{x}$ onto S, then the necessary and sufficient condition for a minimum given by Proposition 6.5.2 reads

$$d_{\boldsymbol{v}} f(\boldsymbol{z}) \quad = \quad (\boldsymbol{z} - \boldsymbol{x})^* \boldsymbol{v} \quad \geq \quad 0$$

for every direction $\boldsymbol{v} = \boldsymbol{y} - \boldsymbol{z}$ defined by another point $\boldsymbol{y} \in S$. This inequality can be rephrased as the obtuse angle criterion $(\boldsymbol{x} - \boldsymbol{z})^*(\boldsymbol{y} - \boldsymbol{z}) \leq 0$ for all such $\boldsymbol{y}$. A simple diagram makes this conclusion visually obvious. ■

In Chap. 5 we found that the multiplier rule (5.1) is a necessary condition for a feasible point $\boldsymbol{x}$ to be a local minimum of the objective function $f(\boldsymbol{y})$ subject to the constraints

$$\begin{aligned} g_i(\boldsymbol{y}) &= 0, \qquad 1 \leq i \leq p \\ h_j(\boldsymbol{y}) &\leq 0, \qquad 1 \leq j \leq q. \end{aligned}$$

In the presence of convexity, the multiplier rule is also a sufficient condition. We now revisit the entire issue under a combination of weaker and stronger hypotheses. The stronger hypotheses amount to: (a) $f(\boldsymbol{y})$ is a convex function, (b) the $g_i(\boldsymbol{y})$ are affine functions, and (c) the feasible region S is a convex set. Instead of assuming that $f(\boldsymbol{y})$ and the $h_j(\boldsymbol{y})$ are continuously differentiable, we now require them to be simply differentiable at the point of interest $\boldsymbol{x}$. Note that we do not require the $h_j(\boldsymbol{y})$ to be convex. Of course if they are, then S is automatically convex.

Proposition 6.5.3 *Under the above conditions, suppose the feasible point $\boldsymbol{x}$ satisfies the multiplier rule (5.1) with $\lambda_0 = 1$. Then $\boldsymbol{x}$ furnishes a global minimum of $f(\boldsymbol{y})$. Conversely, if $\boldsymbol{x}$ is a minimum point satisfying the Mangasarian-Fromovitz constraint qualification, then the multiplier rule holds with $\lambda_0 = 1$.*

Proof: Suppose $\boldsymbol{x}$ satisfies the multiplier rule. Take the inner product of the multiplier formula

$$\nabla f(\boldsymbol{x}) + \sum_{i=1}^{p} \lambda_i \nabla g_i(\boldsymbol{x}) + \sum_{j=1}^{q} \mu_j \nabla h_j(\boldsymbol{x}) \quad = \quad \mathbf{0}$$

with a vector $\boldsymbol{v} = \boldsymbol{z} - \boldsymbol{x}$ defined by a second feasible point $\boldsymbol{z}$. Because the equality constraint $g_i(\boldsymbol{y})$ is affine, $dg_i(\boldsymbol{x})\boldsymbol{v} = 0$. It follows that

$$df(\boldsymbol{x})\boldsymbol{v} \quad = \quad -\sum_{j=1}^{q} \mu_j dh_j(\boldsymbol{x})\boldsymbol{v}.$$

This representation puts us into position to apply Proposition 6.5.2. If some $h_j(\boldsymbol{x}) < 0$, then the multiplier $\mu_j = 0$. If $h_j(\boldsymbol{x}) = 0$, then the difference quotient inequality

$$\frac{h_j(\boldsymbol{x} + t\boldsymbol{v}) - h_j(\boldsymbol{x})}{t} \quad \leq \quad 0$$

holds for all $t \in (0, 1)$. Note here that the convexity of S subtly enters the argument. In any case, sending t to 0 produces $dh_j(\boldsymbol{x})\boldsymbol{v} \leq 0$. Because $\mu_j \geq 0$, we conclude that $df(\boldsymbol{x})\boldsymbol{v} \geq 0$, and this suffices to establish the claim that $\boldsymbol{x}$ is a minimum point.

Proposition 14.7.1 proves the converse under relaxed differentiability assumptions. Here we limit ourselves to the case of no equality constraints ($p = 0$). In this setting we consider the convex function

$$m(\boldsymbol{y}) \quad = \quad \max\{f(\boldsymbol{y}) - f(\boldsymbol{x}), h_j(\boldsymbol{y}), 1 \leq j \leq q\}.$$

It is clear that the minimum of $m(\boldsymbol{y})$ occurs at $\boldsymbol{x}$ because $m(\boldsymbol{x}) = 0$ and for all remaining $\boldsymbol{y}$ either a constraint is violated or $f(\boldsymbol{y}) \geq f(\boldsymbol{x})$. Proposition 6.5.2 therefore implies $d_{\boldsymbol{v}} m(\boldsymbol{x}) \geq 0$ for all directions $\boldsymbol{v}$. Now let J

denote the index set $\{j : h_j(\boldsymbol{x}) = 0\}$. Since all of the functions defining $m(\boldsymbol{x})$ except the inactive inequality constraints achieve the maximum of 0, Example 4.4.4 yields the forward directional derivative

$$d_{\boldsymbol{v}} m(\boldsymbol{x}) \quad = \quad \max\{df(\boldsymbol{x})\boldsymbol{v}, dh_j(\boldsymbol{x})\boldsymbol{v}, 1 \leq i \leq p, j \in J\}.$$

Because $d_{\boldsymbol{v}} m(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{v}$, Proposition 5.3.2 shows that there exist a convex combination of the relevant gradients with

$$\lambda_0 \nabla f(\boldsymbol{x}) + \sum_{j=1}^{q} \mu_j \nabla h_j(\boldsymbol{x}) \quad = \quad \boldsymbol{0}.$$

To eliminate the possibility $\lambda_0 = 0$, now invoke the argument in the last paragraph of the proof of Proposition 5.2.1 . ■

Example 6.5.4 *Slater's Constraint Qualification*

The Mangasarian-Fromovitz constraint qualification is implied by a simpler condition called the Slater constraint qualification under affine equality constraints and convex inequality constraints. Slater's condition postulates the existence of a feasible point $\boldsymbol{z}$ such that $h_j(\boldsymbol{z}) < 0$ for all j. If $\boldsymbol{x}$ is a candidate minimum point and the row vectors $dg_i(\boldsymbol{x})$ are linearly independent, then the Mangasarian-Fromovitz constraint qualification involves finding a vector $\boldsymbol{v}$ with $dg_i(\boldsymbol{x})\boldsymbol{v} = 0$ for all i and $dh_j(\boldsymbol{x})\boldsymbol{v} < 0$ for all inequality constraints active at $\boldsymbol{x}$. If $h_j(\boldsymbol{y})$ is active at $\boldsymbol{x}$, then the inequalities

$$0 \quad > \quad h_j(\boldsymbol{z}) - h_j(\boldsymbol{x}) \quad \geq \quad dh_j(\boldsymbol{x})(\boldsymbol{z} - \boldsymbol{x})$$

demonstrate that the vector $\boldsymbol{v} = \boldsymbol{z} - \boldsymbol{x}$ satisfies the Mangasarian-Fromovitz constraint qualification. Nothing in this argument depends on the objective function $f(\boldsymbol{y})$ being convex. ■

Example 6.5.5 *Concave Constraints*

Consider once again the general nonlinear programming problem covered by Proposition 5.2.1 with the proviso that the equality constraints $g_i(\boldsymbol{y})$ are affine and the inequality constraints $h_j(\boldsymbol{y})$ are concave rather than convex. One can eliminate the equality constraint $g_i(\boldsymbol{y}) = 0$ and replace it by the two inequality constraints $g_i(\boldsymbol{y}) \leq 0$ and $-g_i(\boldsymbol{y}) \leq 0$. Under this substitution all inequality constraints remain concave. In the multiplier rule (5.1) at a local minimum $\boldsymbol{x}$, concavity allows us to rule out the possibility that the multiplier λ_0 of $\nabla f(\boldsymbol{x})$ is 0. To prove this fact, assume that the first r inequality constraints are active at $\boldsymbol{x}$ and the subsequent $q - r$ inequality constraints are inactive. Fortunately, Farkas' lemma poses a relevant dichotomy. Either

$$-\nabla f(\boldsymbol{x}) \quad = \quad \sum_{j=1}^{r} \mu_j \nabla h_j(\boldsymbol{x})$$

for nonnegative multipliers μ_j, or there is a vector $\boldsymbol{w}$ with $-\boldsymbol{w}^*\nabla f(\boldsymbol{x}) > 0$ and $\boldsymbol{w}^*\nabla h_j(\boldsymbol{x}) \leq 0$ for all $j \leq r$. Suppose such a vector $\boldsymbol{w}$ exists. If $\boldsymbol{x}$ is an interior point of the common domain of $f(\boldsymbol{x})$ and the constraints $h_j(\boldsymbol{x})$, then on the one hand the inequality

$$h_j(\boldsymbol{x}+t\boldsymbol{w}) \quad \leq \quad h_j(\boldsymbol{x}) + tdh_j(\boldsymbol{x})\boldsymbol{w} \quad \leq \quad h_j(\boldsymbol{x}) \quad \leq \quad 0$$

shows that the point $\boldsymbol{x}+t\boldsymbol{w}$ is feasible for small $t > 0$. On the other hand,

$$f(\boldsymbol{x}+t\boldsymbol{w}) \quad = \quad f(\boldsymbol{x}) + tdf(\boldsymbol{x})\boldsymbol{w} + o(t) \quad < \quad f(\boldsymbol{x})$$

for small $t > 0$. This contradicts the assumption that $\boldsymbol{x}$ is a local minimum, and the multiplier rule holds with $\lambda_0 = 1$. Linear programming is the most important application. Here the multiplier rule is a necessary and sufficient condition for a global minimum regardless of whether the active affine constraints are linearly independent. ■

Example 6.5.6 *Minimum of a Positive Definite Quadratic Function*

The quadratic function $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x} + c$ has gradient

$$\nabla f(\boldsymbol{x}) \quad = \quad \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$$

for $\boldsymbol{A}$ symmetric. Assuming that $\boldsymbol{A}$ is positive definite and affine equality constraints are present, Proposition 6.5.3 demonstrates that the candidate minimum point identified in Example 5.2.6 furnishes the global minimum of $f(\boldsymbol{x})$. When $\boldsymbol{A}$ is merely positive semidefinite, Example 6.5.5 shows that the multiplier rule with $\lambda_0 = 1$ is still a necessary and sufficient condition for a minimum. ■

Example 6.5.7 *Maximum Likelihood for the Multivariate Normal*

The sample mean and sample variance

$$\begin{aligned} \bar{\boldsymbol{y}} &= \frac{1}{k}\sum_{j=1}^{k} \boldsymbol{y}_j \\ \boldsymbol{S} &= \frac{1}{k}\sum_{j=1}^{k} (\boldsymbol{y}_j - \bar{\boldsymbol{y}})(\boldsymbol{y}_j - \bar{\boldsymbol{y}})^* \end{aligned}$$

are also the maximum likelihood estimates of the theoretical mean $\boldsymbol{\mu}$ and theoretical variance $\boldsymbol{\Omega}$ of a random sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ from a multivariate normal distribution. (See Appendix A.2 for a review of the multivariate normal.) To prove this fact, we first note that maximizing the loglikelihood function

$$-\frac{k}{2}\ln\det\boldsymbol{\Omega} - \frac{1}{2}\sum_{j=1}^{k}(\boldsymbol{y}_j - \boldsymbol{\mu})^*\boldsymbol{\Omega}^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu})$$

$$\begin{aligned} &= -\frac{k}{2}\ln\det\boldsymbol{\Omega} - \frac{k}{2}\boldsymbol{\mu}^*\boldsymbol{\Omega}^{-1}\boldsymbol{\mu} + \Big(\sum_{j=1}^{k}\boldsymbol{y}_j\Big)^*\boldsymbol{\Omega}^{-1}\boldsymbol{\mu} - \frac{1}{2}\sum_{j=1}^{k}\boldsymbol{y}_j^*\boldsymbol{\Omega}^{-1}\boldsymbol{y}_j \\ &= -\frac{k}{2}\ln\det\boldsymbol{\Omega} - \frac{1}{2}\operatorname{tr}\Big[\boldsymbol{\Omega}^{-1}\sum_{j=1}^{k}(\boldsymbol{y}_j-\boldsymbol{\mu})(\boldsymbol{y}_j-\boldsymbol{\mu})^*\Big] \end{aligned}$$

with respect to $\boldsymbol{\mu}$ constitutes a special case of the previous example with $\boldsymbol{A} = k\boldsymbol{\Omega}^{-1}$ and $\boldsymbol{b} = -\boldsymbol{\Omega}^{-1}\sum_{j=1}^{k}\boldsymbol{y}_j$. This leads to the same estimate $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$ regardless of the value of $\boldsymbol{\Omega}$. Once we fix $\hat{\boldsymbol{\mu}}$, we left with the problem of estimating $\boldsymbol{\Omega}$. Fortunately, this is a special case of the problem treated at the end of Example 4.7.6. The point $\hat{\boldsymbol{\Omega}} = \boldsymbol{S}$ corresponds to a maximum because the function $-\ln\det\boldsymbol{\Omega}$ is log-concave in $\boldsymbol{\Omega}^{-1}$ and the function $\operatorname{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{S})$ is linear in $\boldsymbol{\Omega}^{-1}$. Here we implicitly assume that $\boldsymbol{S}$ is invertible.

Alternatively, we can estimate $\boldsymbol{\Omega}$ by exploiting the Cholesky decompositions $\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{L}^*$ and $\boldsymbol{S} = \boldsymbol{M}\boldsymbol{M}^*$. (See Problems 37 and 38 for a development of the Cholesky decomposition of a positive definite matrix.) In view of the identities $\boldsymbol{\Omega}^{-1} = (\boldsymbol{L}^{-1})^*\boldsymbol{L}^{-1}$ and $\det\boldsymbol{\Omega} = (\det\boldsymbol{L})^2$, the loglikelihood becomes

$$\begin{aligned} & k\ln\det\boldsymbol{L}^{-1} - \frac{k}{2}\operatorname{tr}\Big[(\boldsymbol{L}^{-1})^*\boldsymbol{L}^{-1}\boldsymbol{M}\boldsymbol{M}^*\Big] \\ = \;& k\ln\det\big(\boldsymbol{L}^{-1}\boldsymbol{M}\big) - \frac{k}{2}\operatorname{tr}\Big[(\boldsymbol{L}^{-1}\boldsymbol{M})(\boldsymbol{L}^{-1}\boldsymbol{M})^*\Big] - k\ln\det\boldsymbol{M} \end{aligned}$$

using the cyclic permutation property of the matrix trace function. Because products and inverses of lower triangular matrices are lower triangular, the matrix $\boldsymbol{R} = \boldsymbol{L}^{-1}\boldsymbol{M}$ ranges over the set of lower triangular matrices with positive diagonal entries as $\boldsymbol{L}$ ranges over the same set. This permits us to reparameterize and estimate $\boldsymbol{R} = (r_{ij})$ instead of $\boldsymbol{L}$. Up to an irrelevant additive constant, the loglikelihood reduces to

$$k\ln\det\boldsymbol{R} - \frac{k}{2}\operatorname{tr}(\boldsymbol{R}\boldsymbol{R}^*) = k\sum_i \ln r_{ii} - \frac{k}{2}\sum_i\sum_{j=1}^{i} r_{ij}^2.$$

Clearly, this is maximized by taking $r_{ij} = 0$ for $j \neq i$. Differentiation of the concave function $k\ln r_{ii} - \frac{k}{2}r_{ii}^2$ shows that it is maximized by taking $r_{ii} = 1$. In other words, the maximum likelihood estimator $\hat{\boldsymbol{R}}$ is the identity matrix $\boldsymbol{I}$. This implies that $\hat{\boldsymbol{L}} = \boldsymbol{M}$ and consequently that $\hat{\boldsymbol{\Omega}} = \boldsymbol{S}$. ■

Example 6.5.8 *Geometric Programming*

The function

$$f(\boldsymbol{t}) = \sum_{i=1}^{j} c_i \prod_{k=1}^{n} t_k^{\beta_{ik}}$$

is called a posynomial if all components $t_1, \ldots, t_n$ of the argument $\boldsymbol{t}$ and all coefficients $c_1, \ldots, c_j$ are positive. The powers β_{ik} may be positive, negative, or zero. For instance, $t_1^{-1} + 2t_1^3 t_2^{-2}$ is a posynomial on R^2. Geometric programming deals with the minimization of a posynomial $f(\boldsymbol{t})$ subject to posynomial inequality constraints of the form $h_j(\boldsymbol{t}) \leq 1$ for $1 \leq j \leq q$.

Better understanding of geometric programming can be achieved by making the change of variables $t_k = e^{x_k}$. This eliminates the constraint $t_k > 0$ and shows that

$$g(\boldsymbol{x}) \;=\; \sum_{i=1}^{j} c_i \prod_{k=1}^{n} t_k^{\beta_{ik}} \;=\; \sum_{i=1}^{j} c_i e^{\boldsymbol{\beta}_i^* \boldsymbol{x}}$$

is log-convex in the transformed parameters. The reparameterized constraint functions are likewise log-convex and define a convex feasible region S. If the vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_j$ span R^n, then the expression

$$d^2 g(\boldsymbol{x}) \;=\; \sum_{i=1}^{j} c_i e^{\boldsymbol{\beta}_i^* \boldsymbol{x}} \boldsymbol{\beta}_i \boldsymbol{\beta}_i^*$$

for the second differential proves that $g(\boldsymbol{x})$ is strictly convex. It follows that if $g(\boldsymbol{x})$ possesses a minimum, then it is achieved at a single point. ■

Example 6.5.9 *Quasi-Convexity*

If $f(\boldsymbol{x})$ is convex and $g(z)$ is an increasing function of the real variable z, then the inequality

$$\begin{aligned} f[\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}] &\leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) \\ &\leq \max\{f(\boldsymbol{x}), f(\boldsymbol{y})\} \end{aligned}$$

implies that the function $h(\boldsymbol{x}) = g \circ f(\boldsymbol{x})$ satisfies

$$h[\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}] \;\leq\; \max\{h(\boldsymbol{x}), h(\boldsymbol{y})\} \tag{6.10}$$

for any $\alpha \in [0,1]$. Satisfaction of inequality (6.10) is sometimes taken as the definition of quasi-convexity for an arbitrary function $h(\boldsymbol{x})$. If $f(\boldsymbol{x})$ is strictly convex and $g(z)$ is strictly increasing, then strict inequality prevails in inequality (6.10) when $\alpha \in (0,1)$ and $\boldsymbol{y} \neq \boldsymbol{x}$. Once again this implication can be turned into a definition. The importance of strict quasi-convexity lies in the fact that a strictly quasi-convex function possesses at most one local minimum, and if a local minimum exists, then it is necessarily the global minimum.

Similar considerations apply to concave and quasi-concave functions. For example, the function $h(x) = e^{-(x-\mu)^2}$ is strictly quasi-concave because it is the composition of the strictly increasing function $g(y) = e^y$ with the strictly concave function $f(x) = -(x-\mu)^2$. It is clear that $h(x)$ has a global maximum at $x = \mu$. ■

6.6 Moment Inequalities

In this section we assume that readers have a good grasp of probability theory. For those with limited background, most of the material can be comprehended by restricting attention to discrete random variables.

Inequalities give important information about the magnitude of probabilities and expectations without requiring their exact calculation. The Cauchy-Schwarz inequality $|\mathrm{E}(XY)| \leq \mathrm{E}(X^2)^{1/2}\,\mathrm{E}(Y^2)^{1/2}$ is one of the most useful of the classical inequalities. (The reader should check that this is just a disguised form of the Cauchy-Schwarz inequality of Chap. 1 applied to random variables X and Y.) It is also one of the easiest to remember because it is equivalent to the fact that a correlation coefficient must lie on the interval $[-1, 1]$. Equality occurs in the Cauchy-Schwarz inequality if and only if X is proportional to Y or vice versa.

Markov's inequality is another widely applied bound. Let $g(x)$ be a nonnegative, increasing function, and let X be a random variable such that $g(X)$ has finite expectation. Then Markov's inequality

$$\Pr(X \geq c) \quad \leq \quad \frac{\mathrm{E}[g(X)]}{g(c)}$$

holds for any constant c for which $g(c) > 0$ and follows logically by taking expectations in the inequality $g(c)1_{\{X \geq c\}} \leq g(X)$. Chebyshev's inequality is the special case of Markov's inequality with $g(x) = x^2$ applied to the random variable $|X - \mathrm{E}(X)|$. Chebyshev's inequality reads

$$\Pr[|X - \mathrm{E}(X)| \geq c] \quad \leq \quad \frac{\mathrm{Var}(X)}{c^2}.$$

In large deviation theory, we take $g(x) = e^{tx}$ and $c > 0$ and choose $t > 0$ to minimize the right-hand side of the inequality $\Pr(X \geq c) \leq e^{-ct}\,\mathrm{E}(e^{tX})$ involving the moment generating function of X. As an example, suppose X follows a standard normal distribution. The moment generating function $e^{t^2/2}$ of X is derived by a minor variation of the argument given in Appendix A.1 for the characteristic function of X. The large deviation inequality

$$\Pr[X \geq c] \quad \leq \quad \inf_t e^{-ct} e^{t^2/2} \quad = \quad e^{-c^2/2}$$

is called a Chernoff bound. Problem 43 discusses another typical Chernoff bound.

Our next example involves a nontrivial application of Chebyshev's inequality. In preparation for the example, we recall that a binomially distributed random variable S_n has distribution

$$\Pr(S_n = k) \quad = \quad \binom{n}{k} x^k (1-x)^{n-k}.$$

Here S_n is interpreted as the number of successes in n independent trials with success probability x per trial [89]. The mean and variance of S_n are $\mathrm{E}(S_n) = nx$ and $\mathrm{Var}(S_n) = nx(1-x)$.

Example 6.6.1 *Weierstrass's Approximation Theorem*

Weierstrass showed that a continuous function $f(x)$ on $[0,1]$ can be uniformly approximated to any desired degree of accuracy by a polynomial. Bernstein's lovely proof of this fact relies on applying Chebyshev's inequality to the random variable S_n/n derived from the binomial random variable S_n just discussed. The corresponding candidate polynomial is defined by the expectation

$$\mathrm{E}\left[f\left(\frac{S_n}{n}\right)\right] = \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k}x^k(1-x)^{n-k}.$$

Note that $\mathrm{E}(S_n/n) = x$ and

$$\mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{x(1-x)}{n} \leq \frac{1}{4n}.$$

Now given an arbitrary $\epsilon > 0$, one can find by the uniform continuity of $f(x)$ a $\delta > 0$ such that $|f(u) - f(v)| < \epsilon$ whenever $|u - v| < \delta$. If $\|f\|_\infty = \sup|f(x)|$ on $[0,1]$, then Chebyshev's inequality implies

$$\begin{aligned}
&\left|\mathrm{E}\left[f\left(\frac{S_n}{n}\right)\right] - f(x)\right| \\
\leq\ &\mathrm{E}\left[\left|f\left(\frac{S_n}{n}\right) - f(x)\right|\right] \\
\leq\ &\epsilon \Pr\left(\left|\frac{S_n}{n} - x\right| < \delta\right) + 2\|f\|_\infty \Pr\left(\left|\frac{S_n}{n} - x\right| \geq \delta\right) \\
\leq\ &\epsilon + \frac{2\|f\|_\infty x(1-x)}{n\delta^2} \\
\leq\ &\epsilon + \frac{\|f\|_\infty}{2n\delta^2}.
\end{aligned}$$

Taking $n \geq \|f\|_\infty/(2\epsilon\delta^2)$ then gives $\left|\mathrm{E}\left[f\left(\frac{S_n}{n}\right)\right] - f(x)\right| \leq 2\epsilon$ regardless of the chosen $x \in [0,1]$. ■

Proposition 6.6.1 (Jensen) *Let the values of the random variable W be confined to the possibly infinite interval (a,b). If $h(w)$ is convex on (a,b), then $\mathrm{E}[h(W)] \geq h[\mathrm{E}(W)]$, provided both expectations exist. For a strictly convex function $h(w)$, equality holds in Jensen's inequality if and only if $W = \mathrm{E}(W)$ almost surely.*

Proof: For the sake of simplicity, assume that $h(w)$ is differentiable at the point $v = \mathrm{E}(W)$. Then Jensen's inequality follows from Proposition 6.3.1

after taking expectations in the inequality

$$h(W) \geq h(v) + dh(v)(W - v). \tag{6.11}$$

If $h(w)$ is strictly convex, and W is not constant, then inequality (6.11) is strict with positive probability. Hence, strict inequality prevails in Jensen's inequality. As we will see later, the differentiability assumption on $h(w)$ can be relaxed by substituting a subgradient for the gradient. ■

Jensen's inequality is the key to a host of other inequalities. Here is one important example.

Example 6.6.2 *Schlömilch's Inequality for Weighted Means*

If X is a nonnegative random variable, then we define the weighted mean function $\mathrm{M}(p) = \mathrm{E}(X^p)^{\frac{1}{p}}$. For the sake of argument, we assume that $\mathrm{M}(p)$ exists and is finite for all real p. Typical values of $\mathrm{M}(p)$ are $\mathrm{M}(1) = \mathrm{E}(X)$ and $\mathrm{M}(-1) = 1/\mathrm{E}(X^{-1})$. To make $\mathrm{M}(p)$ continuous at $p = 0$, it turns out that we should set $\mathrm{M}(0) = e^{\mathrm{E}(\ln X)}$. The reader is asked to check this fact in Problem 50. Here we are more concerned with proving Schlömilch's assertion that $\mathrm{M}(p)$ is an increasing function of p. If $0 < p < q$, then the function $x \mapsto x^{q/p}$ is convex, and Jensen's inequality says

$$\mathrm{E}(X^p)^{q/p} \leq \mathrm{E}(X^q).$$

Taking the qth root of both sides of this inequality yields $\mathrm{M}(p) \leq \mathrm{M}(q)$. On the other hand if $p < q < 0$, then the function $x \mapsto x^{q/p}$ is concave, and Jensen's inequality says

$$\mathrm{E}(X^p)^{q/p} \geq \mathrm{E}(X^q).$$

Taking the qth root reverses the inequality and again yields $\mathrm{M}(p) \leq \mathrm{M}(q)$. When either p or q is 0, we have to change tactics. One approach is to invoke the continuity of $\mathrm{M}(p)$ at $p = 0$. Another approach is to exploit the concavity of $\ln x$. Jensen's inequality now gives

$$\mathrm{E}(\ln X^p) \leq \ln \mathrm{E}(X^p),$$

which on exponentiation becomes

$$e^{p\,\mathrm{E}(\ln X)} \leq \mathrm{E}(X^p).$$

If $p > 0$, then taking the pth root produces

$$M(0) = e^{\mathrm{E}(\ln X)} \leq \mathrm{E}(X^p)^{\frac{1}{p}},$$

and if $p < 0$, then taking the pth root produces the opposite inequality

$$M(0) = e^{\mathrm{E}(\ln X)} \geq \mathrm{E}(X^p)^{\frac{1}{p}}.$$

When the random variable X is confined to the space $\{1, \ldots, n\}$ equipped with the uniform probabilities $p_i = 1/n$, Schlömilch's inequalities for the values $p = -1$, 0, and 1 reduce to the classical inequalities

$$\frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \cdots \frac{1}{x_n}\right)} \quad \leq \quad \left(x_1 \cdots x_n\right)^{\frac{1}{n}} \quad \leq \quad \frac{1}{n}\left(x_1 + \cdots + x_n\right)$$

relating the harmonic, geometric, and arithmetic means. ■

Example 6.6.3 *Hölder's Inequality*

Consider two random variables X and Y and two numbers $p > 1$ and $q > 1$ such that $p^{-1} + q^{-1} = 1$. Then Hölder's inequality

$$|\operatorname{E}(XY)| \quad \leq \quad \operatorname{E}(|X|^p)^{\frac{1}{p}} \operatorname{E}(|Y|^q)^{\frac{1}{q}} \tag{6.12}$$

generalizes the Cauchy-Schwarz inequality whenever the indicated expectations on its right exist. To prove (6.12), it clearly suffices to assume that X and Y are nonnegative. It also suffices to take $\operatorname{E}(X^p) = \operatorname{E}(Y^q) = 1$ once we divide the left-hand side of (6.12) by its right-hand side. To complete the proof, substitute the random variables X and Y for the scalars x and y in Young's inequality (1.3) and take expectations. ■

6.7 Problems

1. Suppose S and T are nonempty closed convex sets with empty intersection. Prove that there exists a unit vector $\boldsymbol{v}$ such that

$$\sup_{\boldsymbol{x} \in S} \boldsymbol{v}^* \boldsymbol{x} \quad \leq \quad \inf_{\boldsymbol{y} \in T} \boldsymbol{v}^* \boldsymbol{y}.$$

   If either S or T is bounded, then demonstrate further that strict inequality prevails in this inequality. (Hints: The set $S - T$ is convex and does not contain $\mathbf{0}$. If S is compact, then $S - T$ is closed.)

2. Let C be a convex set situated in R^{m+n}. Show that the projected set $\{\boldsymbol{x} \in \mathsf{R}^m : (\boldsymbol{x}, \boldsymbol{y}) \in C \text{ for some } \boldsymbol{y} \in \mathsf{R}^n\}$ is convex.

3. Demonstrate that the set

$$S \quad = \quad \{\boldsymbol{x} \in \mathsf{R}^2 : x_1 x_2 \geq 1, \, x_1 \geq 0, \, x_2 \geq 0\}$$

   is closed and convex. Further show that $P(S)$ is convex but not closed, where $P\boldsymbol{x} = x_1$ denotes projection onto the first coordinate of $\boldsymbol{x}$.

4. The function $P(\boldsymbol{x}, t) = t^{-1}\boldsymbol{x}$ from $\mathsf{R}^n \times (0, \infty)$ to R^n is called the perspective map. Show that the image $P(C)$ and inverse image $P^{-1}(D)$ of convex sets C and D are convex. (Hint: Prove that $P(\boldsymbol{x}, t)$ maps line segments onto line segment.)

5. The topological notions of open, closed, and bounded can be recast for a convex set $C \subset \mathsf{R}^n$ [172]. Validate the necessary and sufficient conditions in each of the following assertions:

   (a) C is open if and only if for every $\boldsymbol{x} \in C$ and $\boldsymbol{y} \in \mathsf{R}^n$, the point $\boldsymbol{x} + t\boldsymbol{y}$ belongs to C for all sufficiently small positive t.

   (b) C is closed if and only if whenever C contains the open segment $\{\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y} : \alpha \in (0,1)\}$, it also contains the endpoints $\boldsymbol{x}$ and $\boldsymbol{y}$ of the segment.

   (c) C is bounded if and only if it contains no ray $\{\boldsymbol{x}+t\boldsymbol{y} : t \in [0,\infty)\}$.

6. Demonstrate that the convex hull of an open set is open.

7. The Gauss-Lucas theorem says that the roots of the derivative $p'(z)$ of a polynomial $p(z)$ are contained in the convex hull of the roots of $p(z)$. Prove this claim by exploiting the expansion

$$0 \;=\; \frac{p'(y)}{p(y)} \;=\; \sum_i \frac{1}{y - z_i} \;=\; \sum_i \frac{\bar{y} - \bar{z}_i}{\|y - z_i\|^2},$$

   where y is a root of $p'(z)$ differing from each of the roots z_i of $p(z)$, and the overbar sign denotes complex conjugation.

8. Deduce Farkas' lemma from Proposition 5.3.2.

9. On which intervals are the following functions convex: e^x, e^{-x}, x^n for n an integer, $|x|^p$ for $p \geq 1$, $\sqrt{1+x^2}$, $x \ln x$, and $\cosh x$? On these intervals, which functions are log-convex?

10. Demonstrate that the function $f(x) = x^n - na \ln x$ is convex on $(0,\infty)$ for any positive real number a and nonnegative integer n. Where does its minimum occur?

11. Show that Riemann's zeta function

$$\zeta(s) \;=\; \sum_{n=1}^{\infty} \frac{1}{n^s}$$

    is log-convex for $s > 1$.

12. Show that the function

$$f(x) \;=\; \begin{cases} \frac{1-e^{-xt}}{x} & x \neq 0 \\ t & x = 0 \end{cases}$$

    is log-convex for $t > 0$ fixed. (Hints: Use either the second derivative test, or express $f(x)$ as the integral

$$f(x) \;=\; \int_0^t e^{-xs} ds,$$

    and use the closure properties of log-convex functions.)

13. Use Proposition 6.3.1 to prove the Cauchy-Schwarz inequality.

14. Prove the strict convexity assertions of Proposition 6.3.1.

15. Prove parts (b), (c), (d), and (e) of Proposition 6.3.3.

16. Prove the unproved assertions of Proposition 6.3.4.

17. Let $f(x)$ and $g(x)$ be positive functions defined on an interval of the real line. Prove that:

    (a) If $f(x)$ and $g(x)$ are both convex and increasing or both convex and decreasing, then the product $f(x)g(x)$ is convex.

    (b) If $f(x)$ and $g(x)$ are both concave, one is increasing, and the other is decreasing, then the product $f(x)g(x)$ is concave.

    (c) If $f(x)$ is convex and increasing and $g(x)$ is concave and decreasing, then the ratio $f(x)/g(x)$ is convex.

    Here "increasing" means nondecreasing and similarly for "decreasing." Your proofs should not assume that either $f(x)$ or $g(x)$ is differentiable.

18. Let $f(x)$ be a convex function on the real line that is bounded above. Demonstrate that $f(x)$ is constant.

19. Suppose $f(\boldsymbol{x}, \boldsymbol{y})$ is jointly convex in its two arguments and C is a convex set. Show that the function

$$g(\boldsymbol{x}) \quad = \quad \inf_{\boldsymbol{y} \in C} f(\boldsymbol{x}, \boldsymbol{y})$$

    is convex. Assume here that the infimum is finite. As a special case, demonstrate that the distance function $\operatorname{dist}(\boldsymbol{x}, C) = \inf_{\boldsymbol{y} \in C} \|\boldsymbol{y} - \boldsymbol{x}\|_*$ is convex for any norm $\|\cdot\|_*$. (Hint: For $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and $\epsilon > 0$ choose points $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ in C satisfying $f(\boldsymbol{x}_i, \boldsymbol{y}_i) \leq g(\boldsymbol{x}_i) + \epsilon$.)

20. Let the function $f(x, y)$ be convex in y for each fixed x. Replace x by a random variable X and take expectations. Show that $\mathrm{E}[f(X, y)]$ is convex in y, assuming the expectations in question exist. For n an even positive integer with $\mathrm{E}(X^n) < \infty$, use this result to prove that $y \mapsto \mathrm{E}[(X - y)^n]$ is convex in y.

21. Suppose $f(x)$ is absolutely integrable over every compact interval $[a, b]$ and $\phi(x)$ is nonnegative, bounded, vanishes outside a symmetric interval $[-c, c]$, and has total mass $\int \phi(x)\, dx = 1$. The convolution of f and $\phi(x)$ is defined by

$$f * \phi(x) \quad = \quad \int f(x - y)\phi(y)\, dy \quad = \quad \int f(y)\phi(x - y)\, dy.$$

Prove the following claims:

(a) $f * \phi(x)$ is continuous if $\phi(x)$ is continuous.

(b) $f * \phi(x)$ is k times continuously differentiable if $\phi(x)$ is k times continuously differentiable.

(c) $f * \phi(x)$ is nonnegative if $f(x)$ is nonnegative.

(d) $f * \phi(x)$ is increasing (decreasing) if $f(x)$ is increasing (decreasing).

(e) $f * \phi(x)$ is convex (concave) if $f(x)$ is convex (concave).

(f) $f * \phi(x)$ is log-convex if $f(x)$ is log-convex.

Define $\phi_n(s) = n\phi(nx)$ and prove as well that $f * \phi_n(x)$ converges to $f(x)$ at every point of continuity of $f(x)$. This convergence is uniform on every compact interval on which $f(x)$ is continuous. These facts imply that convexity properties established by differentiation carry over to convex functions in general. Can you supply any examples?

22. Suppose the polynomial $p(x)$ has only real roots. Show that $1/p(x)$ is log-convex on any interval where $p(x)$ is positive.

23. Demonstrate that the Kullback-Leibler (cross-entropy) distance

$$f(\boldsymbol{x}) \quad = \quad x_1 \ln \frac{x_1}{x_2} + x_2 - x_1$$

is convex on the set $\{\boldsymbol{x} = (x_1, x_2) : x_1 > 0, x_2 > 0\}$.

24. Show that the function $f(\boldsymbol{x}) = x_1^2 + x_2^4$ on R^2 is strictly convex even though $d^2 f(\boldsymbol{x})$ is singular along the line $x_2 = 0$.

25. Prove that:

(a) $f(\boldsymbol{x}) = x_1^2/x_2$ is convex for x_2 positive,

(b) $f(\boldsymbol{x}) = \left(\prod_{i=1}^n x_i \right)^{1/n}$ is concave when all $x_i > 0$,

(c) $f(\boldsymbol{x}) = \sum_{i=1}^m x_{[i]}$ is concave, where $x_{[1]} \leq \cdots \leq x_{[n]}$ are the order statistics of $\boldsymbol{x} = (x_1, \ldots, x_n)^*$,

(d) $f(\boldsymbol{x}) = \left(\sum_{i=1}^n \sqrt{x_i} \right)^{-1}$ is convex when all $x_i \geq 0$ and at least one is positive.

(Hints: Use the second derivative test for (a), (b), and (d). Write

$$f(\boldsymbol{x}) \quad = \quad \min_{i_1 < \cdots < i_m} [x_{i_1} + \cdots + x_{i_m}]$$

for (c).)

26. Let $f(x)$ be a continuous function on the real line satisfying

$$f\left[\frac{1}{2}(x+y)\right] \quad \leq \quad \frac{1}{2}f(x)+\frac{1}{2}f(y).$$

Demonstrate that $f(x)$ is convex.

27. If $f(x)$ is a nondecreasing function on the interval $[a,b]$, then show that $g(x)=\int_a^x f(y)dy$ is a convex function on $[a,b]$.

28. The Bohr-Mollerup theorem asserts that $\Gamma(z)$ is the only log-convex function on the interval $(0,\infty)$ that satisfies $\Gamma(1)=1$ and the factorial identity $\Gamma(z+1)=z\Gamma(z)$ for all z. We have seen that $\Gamma(z)$ has these properties. Prove conversely that any function $G(z)$ with these properties coincides with $\Gamma(z)$. (Hints: Check the inequalities

$$\begin{aligned} G(n+z) &\leq G(n)^{1-z}G(n)^z n^z &= (n-1)!n^z \\ G(n+1) &\leq G(n+z)^z G(n+1+z)^{1-z} &= G(n+z)(n+z)^{1-z} \end{aligned}$$

for all positive integers n and real numbers $z\in(0,1)$. These in turn yield the inequalities

$$\frac{n!n^z}{z(z+1)\cdots(z+n)} \quad \leq \quad G(z) \quad \leq \quad \frac{n!n^z}{z(z+1)\cdots(z+n)}\cdot\frac{z+n}{n}.$$

Taking limits on n shows that $G(z)$ equals Gauss's infinite product expansion of $\Gamma(z)$. Note that this proof simultaneously validates Gauss's expansion (6.6).)

29. Let $f(\boldsymbol{x})$ be a real-valued differentiable function on R^n. If $f(\boldsymbol{x})$ is strictly convex, prove that $df(\boldsymbol{x})=df(\boldsymbol{y})$ if and only if $\boldsymbol{x}=\boldsymbol{y}$.

30. Suppose $f(\boldsymbol{x})$ is convex on R^n and $f(\boldsymbol{y})=0$. Prove that $f(\boldsymbol{x})^2$ is differentiable at $\boldsymbol{y}$ with differential $\mathbf{0}^*$. (Hint: Invoke Proposition 6.4.1 and Fréchet's definition of differentiability.)

31. Let $f(\boldsymbol{x})$ be a convex differentiable function on R^n. Show that the function

$$g(\boldsymbol{x}) \quad = \quad f(\boldsymbol{x})+\epsilon\|\boldsymbol{x}\|^2$$

is strictly convex for $\epsilon>0$ and that the set $\{\boldsymbol{x}\in\mathsf{R}^n : g(\boldsymbol{x})\leq g(\boldsymbol{y})\}$ is compact for any $\boldsymbol{y}$.

32. A square matrix $\boldsymbol{M}$ is said to be primitive if its entries are nonnegative and all of the entries of some power $\boldsymbol{M}^p$ of $\boldsymbol{M}$ are positive. The Perron-Frobenius theorem [136, 150, 234] asserts that a primitive matrix $\boldsymbol{M}$ has a dominant eigenvalue $\lambda>0$ possessing unique right and

left eigenvectors $\boldsymbol{u}$ and $\boldsymbol{v}^*$ with positive entries. Furthermore, if we choose $\boldsymbol{u}$ and $\boldsymbol{v}^*$ so that $\boldsymbol{v}^*\boldsymbol{u} = 1$, then

$$\lim_{m\to\infty} \lambda^{-m}\boldsymbol{M}^m \;=\; \boldsymbol{u}\boldsymbol{v}^*.$$

From this limit deduce that λ is log-convex in $\boldsymbol{x}$ if the entries of $\boldsymbol{M}$ are either 0 or log-convex functions of $\boldsymbol{x}$.

33. Consider minimizing the quadratic function $f(\boldsymbol{y}) = \frac{1}{2}\boldsymbol{y}^*\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}^*\boldsymbol{y} + c$ subject to the vector constraint $\boldsymbol{y} \geq \boldsymbol{0}$ for a positive semidefinite matrix $\boldsymbol{A}$. Show that the three conditions $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \geq \boldsymbol{0}$, $\boldsymbol{x} \geq \boldsymbol{0}$, and $\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x} = 0$ are necessary and sufficient for $\boldsymbol{x}$ to represent a minimum.

34. Let C be a convex set. Proposition 6.5.2 declares that a point $\boldsymbol{x} \in C$ minimizes a convex function $f(\boldsymbol{y})$ on C provided $d_{\boldsymbol{v}} f(\boldsymbol{x}) \geq 0$ for every tangent vector $\boldsymbol{v} = \boldsymbol{y} - \boldsymbol{x}$ constructed from a point $\boldsymbol{y} \in C$. If C is a convex cone, and $f(\boldsymbol{y})$ is convex and differentiable, then show that this condition is equivalent to the conditions $df(\boldsymbol{x})\boldsymbol{x} = 0$ and $df(\boldsymbol{x})\boldsymbol{y} \geq 0$ for every $\boldsymbol{y} \in C$.

35. The posynomial $f(\boldsymbol{x}) = \prod_{i=1}^n x_i^{\alpha_i}$ achieves a unique maximum on the unit simplex $\{\boldsymbol{x} \in \mathsf{R}^n : \boldsymbol{x} \geq \boldsymbol{0}, \;\; \sum_{i=1}^n x_i = 1\}$ whenever the powers α_i are positive. Find this maximum, and show that it is global. (Hint: Minimize $-\ln f(\boldsymbol{x})$.)

36. Show that the functions $\sqrt{|x|}$, $\ln x$, and $\lfloor x \rfloor$ are quasi-convex. To the extent possible, state and prove the quasi-convex analogues of the convex closure properties covered in Proposition 6.3.3.

37. Let $\boldsymbol{A}$ be an $n \times n$ positive definite matrix. The Cholesky decomposition $\boldsymbol{B}$ of $\boldsymbol{A}$ is a lower-triangular matrix with positive diagonal entries such that $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}^*$. To prove that such a decomposition exists we can argue by induction. Why is the case of a 1×1 matrix trivial? Now suppose $\boldsymbol{A}$ is partitioned as

$$\boldsymbol{A} \;=\; \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}.$$

Applying the induction hypothesis, there exist matrices $\boldsymbol{C}_{11}$ and $\boldsymbol{D}_{22}$ such that

$$\begin{aligned} \boldsymbol{C}_{11}\boldsymbol{C}_{11}^* &= \boldsymbol{A}_{11} \\ \boldsymbol{D}_{22}\boldsymbol{D}_{22}^* &= \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}. \end{aligned}$$

Prove that

$$\boldsymbol{B} \;=\; \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{0} \\ \boldsymbol{A}_{21}(\boldsymbol{C}_{11}^*)^{-1} & \boldsymbol{D}_{22} \end{pmatrix}$$

gives the desired decomposition. Extend this argument to show that $\boldsymbol{B}$ is uniquely determined.

38. Continuing Problem 37, show that one can compute the Cholesky decomposition $\boldsymbol{B} = (b_{ij})$ of $\boldsymbol{A} = (a_{ij})$ by the recurrence relations

$$\begin{aligned} b_{jj} &= \sqrt{a_{jj} - \sum_{k=1}^{j-1} b_{jk}^2} \\ b_{ij} &= \frac{a_{ij} - \sum_{k=1}^{j-1} b_{ik} b_{jk}}{b_{jj}}, \quad i > j \end{aligned}$$

for columns $j = 1$, $j = 2$, and so forth until column $j = n$. How can you compute $\det \boldsymbol{A}$ in terms of the entries of $\boldsymbol{B}$?

39. Prove that the set of lower triangular matrices with positive diagonal entries is closed under matrix multiplication and matrix inversion.

40. Let $X_1, \ldots, X_n$ be n independent random variables from a common distributional family. Suppose the variance $\sigma^2(\mu)$ of a generic member of this family is a function of the mean μ. Now consider the sum $S = X_1 + \cdots + X_n$. If the mean $\omega = \mathrm{E}(S)$ is fixed, it is of some interest to determine whether taking $\mathrm{E}(X_i) = \mu_i = \omega/n$ minimizes or maximizes $\mathrm{Var}(S)$. Show that the minimum occurs when $\sigma^2(\mu)$ is convex in μ and the maximum occurs when $\sigma^2(\mu)$ is concave in μ. What do you deduce in the special cases where the family is binomial, Poisson, and exponential [201]?

41. If the random variable X has values in the interval $[a, b]$, then show that $\mathrm{Var}(X) \leq (b-a)^2/4$ and that this bound is sharp. (Hints: Reduce to the case $[a, b] = [0, 1]$. If $\mathrm{E}(X) = p$, then demonstrate that $\mathrm{Var}(X) \leq p(1-p)$.)

42. Suppose $g(x)$ is a function such that $g(x) \leq 1$ for all x and $g(x) \leq 0$ for $x \leq c$. Demonstrate the inequality

$$\Pr(X \geq c) \geq \mathrm{E}[g(X)] \tag{6.13}$$

for any random variable X [89]. Verify that the polynomial

$$g(x) = \frac{(x-c)(c+2d-x)}{d^2}$$

with $d > 0$ satisfies the stated conditions leading to inequality (6.13). If X is nonnegative with $\mathrm{E}(X) = 1$ and $\mathrm{E}(X^2) = \beta$ and $c \in (0, 1)$, then prove that the choice $d = \beta/(1-c)$ yields

$$\Pr(X \geq c) \geq \frac{(1-c)^2}{\beta}.$$

Finally, if $\mathrm{E}(X^2) = 1$ and $\mathrm{E}(X^4) = \beta$, show that

$$\Pr(|X| \geq c) \quad \geq \quad \frac{(1-c^2)^2}{\beta}.$$

43. Let X be a Poisson random variable with mean λ. Demonstrate that the Chernoff bound

$$\Pr(X \geq c) \quad \leq \quad \inf_{t>0} e^{-ct} \mathrm{E}(e^{tX})$$

amounts to

$$\Pr(X \geq c) \quad \leq \quad \frac{(\lambda e)^c}{c^c} e^{-\lambda}$$

for any integer $c > \lambda$. Recall that $\Pr(X = i) = \lambda^i e^{-\lambda}/i!$ for all nonnegative integers i.

44. Use Jensen's inequality to prove the inequality

$$\prod_{k=1}^n x_k^{\alpha_k} + \prod_{k=1}^n y_k^{\alpha_k} \quad \leq \quad \prod_{k=1}^n (x_k + y_k)^{\alpha_k}$$

for positive numbers x_k and y_k and nonnegative numbers α_k with sum $\sum_{k=1}^n \alpha_k = 1$. Prove the inequality

$$\left(1 + \prod_{k=1}^n x_k^{\alpha_k}\right)^{-1} \quad \leq \quad \sum_{k=1}^n \frac{\alpha_k}{1+x_k}$$

when all $x_k \geq 1$ and the reverse inequality when all $x_k \in (0, 1]$.

45. Let $B_n f(x) = \mathrm{E}[f(S_n/n)]$ denote the Bernstein polynomial of degree n approximating $f(x)$ as discussed in Example 6.6.1. Prove that

(a) $B_n f(x)$ is linear in $f(x)$,
(b) $B_n f(x) \geq 0$ if $f(x) \geq 0$,
(c) $B_n f(x) = f(x)$ if $f(x)$ is linear,
(d) $B_n x(1-x) = \frac{n-1}{n} x(1-x)$,
(e) $\|B_n f\|_\infty \leq \|f\|_\infty$.

46. Suppose the function $f(x)$ has continuous derivative $f'(x)$. For $\delta > 0$ show that Bernstein's polynomial satisfies the bound

$$\left| \mathrm{E}\left[f\left(\frac{S_n}{n}\right)\right] - f(x) \right| \quad \leq \quad \delta \|f'\|_\infty + \frac{\|f\|_\infty}{2n\delta^2}.$$

Conclude from this estimate that $\left\| \mathrm{E}\left[f\left(\frac{S_n}{n}\right)\right] - f \right\|_\infty = O(n^{-\frac{1}{3}})$.

47. Let $f(x)$ be a convex function on $[0,1]$. Prove that the Bernstein polynomial of degree n approximating $f(x)$ is also convex. (Hint: Show that

$$\begin{aligned}\frac{d^2}{dx^2}\operatorname{E}\left[f\left(\frac{S_n}{n}\right)\right] &= n(n-1)\Big\{\operatorname{E}\left[f\left(\frac{S_{n-2}+2}{n}\right)\right] \\ &\quad -2\operatorname{E}\left[f\left(\frac{S_{n-2}+1}{n}\right)\right]+\operatorname{E}\left[f\left(\frac{S_{n-2}}{n}\right)\right]\Big\}\end{aligned}$$

in the notation of Example 6.6.1.)

48. Verify the following special cases

$$\begin{aligned}\sum_{m=1}^{n}\frac{a_m}{[m(m+1)]^{1/5}} &\le \left(\sum_{m=1}^{n}|a_m|^{5/4}\right)^{4/5} \\ \sum_{m=1}^{n}\frac{a_m}{\sqrt{m}} &\le \frac{\sqrt{\pi}}{6^{1/4}}\left(\sum_{m=1}^{n}|a_m|^{4/3}\right)^{3/4} \\ \sum_{m=0}^{\infty}a_m x^m &\le (1-x^3)^{-1/3}\left(\sum_{m=0}^{\infty}|a_m|^{3/2}\right)^{2/3}\end{aligned}$$

of Hölder's inequality [243]. In the last inequality we take $0 \le x < 1$.

49. Suppose $1 \le p < \infty$. For a random variable X with $\operatorname{E}(|X|^p) < \infty$, define the norm $\|X\|_p = \operatorname{E}(X^p)^{\frac{1}{p}}$. Now prove Minkowski's triangle inequality $\|X+Y\|_p \le \|X\|_p + \|Y\|_p$. (Hint: Apply Hölder's inequality to the right-hand side of

$$\operatorname{E}(|X+Y|^p) \le \operatorname{E}(|X|\cdot|X+Y|^{p-1}) + \operatorname{E}(|Y|\cdot|X+Y|^{p-1})$$

and rearrange the result.)

50. Suppose X is a random variable satisfying $0 < a \le X \le b < \infty$. Use L'Hôpital's rule to prove that the weighted mean $\operatorname{M}(p) = \operatorname{E}(X^p)^{\frac{1}{p}}$ is continuous at $p = 0$ if we define $\operatorname{M}(0) = e^{\operatorname{E}(\ln X)}$.

51. Suppose the random variable X is bounded below by the positive constant a and above by the positive constant b. Prove Kantorvich's inequality

$$\operatorname{E}(X)\operatorname{E}(X^{-1}) \le \frac{\mu^2}{\gamma^2}$$

for $\mu = \frac{1}{2}(a+b)$ and $\gamma = \sqrt{ab}$. (Hint: By homogeneity it suffices to consider the case $\gamma = 1$. Apply the arithmetic-geometric mean inequality [243].)

7
Block Relaxation

7.1 Introduction

As a gentle introduction to optimization algorithms, we now consider block relaxation. The more descriptive terms block descent and block ascent suggest either minimization or maximization rather than generic optimization. Regardless of what one terms the strategy, in many problems it pays to update only a subset of the parameters at a time. Block relaxation divides the parameters into disjoint blocks and cycles through the blocks, updating only those parameters within the pertinent block at each stage of a cycle [59]. When each block consists of a single parameter, block relaxation is called cyclic coordinate descent or cyclic coordinate ascent. Block relaxation is best suited to unconstrained problems where the domain of the objective function reduces to a Cartesian product of the subdomains associated with the different blocks. Obviously, exact block updates are a huge advantage. Equality constraints usually present insuperable barriers to coordinate descent and ascent because parameters get locked into position. In some problems it is advantageous to consider overlapping blocks.

The rest of this chapter consists of sequence of examples, most of which are drawn from statistics. Details of statistical inference are downplayed, but familiarity with classical statistics certainly helps in understanding. Block relaxation sometimes converges slowly. In compensation, updates are often very cheap to compute. Judging the performance of optimization algorithms is a complex task. Computational speed is only one factor.

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_7,

Reliability and ease of implementation can be equally important. In many problems block relaxation is trivial to implement.

7.2 Examples of Block Relaxation

Example 7.2.1 *Sinkhorn's Algorithm*

Let $\boldsymbol{M} = (m_{ij})$ be a rectangular matrix with positive entries. Sinkhorn's theorem [237] says that there exist two diagonal matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with positive diagonal entries a_i and b_j such that the matrix $\boldsymbol{AMB}$ has prescribed row and column sums. Let $r_i > 0$ be the ith row sum and $c_j > 0$ the jth column sum. Because $\boldsymbol{AMB}$ has entry $a_i m_{ij} b_j$ at the intersection of row i and column j, the constraints are

$$\sum_i a_i m_{ij} b_j = c_j \quad \text{and} \quad \sum_j a_i m_{ij} b_j = r_i.$$

For these constraints to be consistent, we must have

$$\sum_i r_i = \sum_i \sum_j a_i m_{ij} b_j = \sum_j c_j.$$

Given this assumption, we now sketch a method for finding $\boldsymbol{A}$ and $\boldsymbol{B}$. Consider minimizing the smooth function [156]

$$f(\boldsymbol{A}, \boldsymbol{B}) = -\sum_i r_i \ln a_i - \sum_j c_j \ln b_j + \sum_i \sum_j a_i m_{ij} b_j.$$

If any a_i or b_j approaches 0, then $f(\boldsymbol{A}, \boldsymbol{B})$ tends to ∞. In view of this fact, the minimum occurs in a region where the parameters a_i and b_j are uniformly bounded below by a positive constant. Within this region, it follows that $a_i m_{ij} b_j$ tends to ∞ if either a_i or b_j tends to ∞. Hence, the minimum of $f(\boldsymbol{A}, \boldsymbol{B})$ exists. At the minimum, Fermat's principle requires

$$\begin{aligned} \frac{\partial}{\partial a_i} f(\boldsymbol{A}, \boldsymbol{B}) &= -\frac{r_i}{a_i} + \sum_j m_{ij} b_j = 0 \\ \frac{\partial}{\partial b_j} f(\boldsymbol{A}, \boldsymbol{B}) &= -\frac{c_j}{b_j} + \sum_i a_i m_{ij} = 0. \end{aligned}$$

These equations are just a disguised form of Sinkhorn's constraints.

The direct attempt to solve the stationarity equations is almost immediately thwarted. It is much easier to minimize $f(\boldsymbol{A}, \boldsymbol{B})$ with respect to $\boldsymbol{A}$ for $\boldsymbol{B}$ fixed or vice versa. If we fix $\boldsymbol{B}$, then rearranging the first stationarity equation gives

$$a_i = \frac{r_i}{\sum_j m_{ij} b_j}.$$

Similarly, if we fix $\boldsymbol{A}$, then rearranging the second stationarity equation yields

$$b_j = \frac{c_j}{\sum_i a_i m_{ij}}.$$

Sinkhorn's block relaxation algorithm [237] alternates the updates of $\boldsymbol{A}$ and $\boldsymbol{B}$. ■

Example 7.2.2 *Poisson Sports Model*

Consider a simplified version of a model proposed by Maher [185] for a sports contest between two teams in which the number of points scored by team i against team j follows a Poisson process with intensity $e^{o_i - d_j}$, where o_i is an "offensive strength" parameter for team i and d_j is a "defensive strength" parameter for team j. (See Sect. 8.9 for a brief description of Poisson processes.) If t_{ij} is the length of time that i plays j and p_{ij} is the number of points that i scores against j, then the corresponding Poisson loglikelihood function is

$$\ell_{ij}(\boldsymbol{\theta}) = p_{ij}(o_i - d_j) + p_{ij} \ln t_{ij} - t_{ij} e^{o_i - d_j} - \ln p_{ij}!, \tag{7.1}$$

where $\boldsymbol{\theta} = (\boldsymbol{o}, \boldsymbol{d})$ is the parameter vector. Note that the parameters should satisfy a linear constraint such as $d_1 = 0$ in order for the model be identifiable; otherwise, it is clearly possible to add the same constant to each o_i and d_j without altering the likelihood. We make two simplifying assumptions. First, the outcomes of the different games are independent. Second, each team's point total within a single game is independent of its opponent's point total. The second assumption is more suspect than the first since it implies that a team's offensive and defensive performances are somehow unrelated to one another; nonetheless, the model gives an interesting first approximation to reality. Under these assumptions, the full data loglikelihood is obtained by summing $\ell_{ij}(\boldsymbol{\theta})$ over all pairs (i, j). Setting the partial derivatives of the loglikelihood equal to zero leads to the equations

$$e^{-d_j} = \frac{\sum_i p_{ij}}{\sum_i t_{ij} e^{o_i}} \quad \text{and} \quad e^{o_i} = \frac{\sum_j p_{ij}}{\sum_j t_{ij} e^{-d_j}}$$

satisfied by the maximum likelihood estimate $(\hat{\boldsymbol{o}}, \hat{\boldsymbol{d}})$.

These equations do not admit a closed-form solution, so we turn to block relaxation [59]. If we fix the o_i, then we can solve for the d_j and vice versa in the form

$$d_j = -\ln\left(\frac{\sum_i p_{ij}}{\sum_i t_{ij} e^{o_i}}\right) \quad \text{and} \quad o_i = \ln\left(\frac{\sum_j p_{ij}}{\sum_j t_{ij} e^{-d_j}}\right).$$

Block relaxation consists in alternating the updates of the defensive and offensive parameters with the proviso that d_1 is fixed at 0.

TABLE 7.1. Ranking of all 29 NBA teams on the basis of the 2002–2003 regular season according to their estimated offensive plus defensive strengths. Each team played 82 games

Team	$\hat{o}_i + \hat{d}_i$	Wins	Team	$\hat{o}_i + \hat{d}_i$	Wins
Cleveland	−0.0994	17	Phoenix	0.0166	44
Denver	−0.0845	17	New Orleans	0.0169	47
Toronto	−0.0647	24	Philadelphia	0.0187	48
Miami	−0.0581	25	Houston	0.0205	43
Chicago	−0.0544	30	Minnesota	0.0259	51
Atlanta	−0.0402	35	LA Lakers	0.0277	50
LA Clippers	−0.0355	27	Indiana	0.0296	48
Memphis	−0.0255	28	Utah	0.0299	47
New York	−0.0164	37	Portland	0.0320	50
Washington	−0.0153	37	Detroit	0.0336	50
Boston	−0.0077	44	New Jersey	0.0481	49
Golden State	−0.0051	38	San Antonio	0.0611	60
Orlando	−0.0039	42	Sacramento	0.0686	59
Milwaukee	−0.0027	42	Dallas	0.0804	60
Seattle	0.0039	40			

Table 7.1 summarizes our application of the Poisson sports model to the results of the 2002–2003 regular season of the National Basketball Association. In these data, t_{ij} is measured in minutes. A regular game lasts 48 min, and each overtime period, if necessary, adds 5 min. Thus, team i is expected to score $48e^{\hat{o}_i - \hat{d}_j}$ points against team j when the two teams meet and do not tie. Team i is ranked higher than team j if $\hat{o}_i - \hat{d}_j > \hat{o}_j - \hat{d}_i$, which is equivalent to the condition $\hat{o}_i + \hat{d}_i > \hat{o}_j + \hat{d}_j$.

It is worth emphasizing some of the virtues of the model. First, the ranking of the 29 NBA teams on the basis of the estimated sums $\hat{o}_i + \hat{d}_i$ for the 2002–2003 regular season is not perfectly consistent with their cumulative wins; strength of schedule and margins of victory are reflected in the model. Second, the model gives the point-spread function for a particular game as the difference of two independent Poisson random variables. Third, one can easily amend the model to rank individual players rather than teams by assigning to each player an offensive and defensive intensity parameter. If each game is divided into time segments punctuated by substitutions, then the block relaxation algorithm can be adapted to estimate the assigned player intensities. This might provide a rational basis for salary negotiations that takes into account subtle differences between players not reflected in traditional sports statistics. ■

Example 7.2.3 *K-Means Clustering*

In k-means clustering we must divide n points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in R^m into k clusters. Each cluster C_j is characterized by a cluster center $\boldsymbol{\mu}_j$. The best

clustering of the points minimizes the criterion

$$f(\boldsymbol{\mu}, C) = \sum_{j=1}^{k} \sum_{\boldsymbol{x}_i \in C_j} \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^2,$$

where $\boldsymbol{\mu}$ is the matrix whose columns are the $\boldsymbol{\mu}_j$ and C is the collection of clusters. Because this mixed continuous-discrete optimization problem has no obvious analytic solution, block relaxation is attractive. If we hold the clusters fixed, then it is clear from Example 6.5.7 that we should set

$$\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\boldsymbol{x}_i \in C_j} \boldsymbol{x}_i.$$

Similarly, it is clear that if we hold the cluster centers fixed, then we should assign point $\boldsymbol{x}_i$ to the cluster C_j minimizing $\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|$. Block relaxation, known as Lloyd's algorithm in this context, alternates cluster center redefinition and cluster membership reassignment. It is simple and effective. The initial cluster centers can chosen randomly from the n data points. The evidence suggests that this should be done in a biased manner that spreads the centers out [5]. Changing the objective function to

$$g(\boldsymbol{\mu}, C) = \sum_{j=1}^{k} \sum_{\boldsymbol{x}_i \in C_j} \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|_1$$

makes it more resistant to outliers. The recentering step is now solved by replacing means by medians in each coordinate. This takes a little more computation but is usually worth the effort. ■

Example 7.2.4 *Canonical Correlations*

Consider a random vector $\boldsymbol{Z}$ partitioned into a subvector $\boldsymbol{X}$ of predictors and a subvector $\boldsymbol{Y}$ of responses. (See Sect. 9.7 for a brief discussion of random vectors, expectations, and variances.) The most elementary form of canonical correlation analysis seeks two linear combinations $\boldsymbol{a}^*\boldsymbol{X}$ and $\boldsymbol{b}^*\boldsymbol{Y}$ that are maximally correlated [187]. If we partition the variance matrix of $\boldsymbol{Z}$ into blocks

$$\text{Var}(\boldsymbol{Z}) = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

consistent with $\boldsymbol{X}$ and $\boldsymbol{Y}$, then the two linear combinations maximize the covariance $\boldsymbol{a}^*\boldsymbol{\Sigma}_{12}\boldsymbol{b}$ subject to the variance constraints

$$\boldsymbol{a}^*\boldsymbol{\Sigma}_{11}\boldsymbol{a} = \boldsymbol{b}^*\boldsymbol{\Sigma}_{22}\boldsymbol{b} = 1.$$

This constrained maximization problem is an ideal candidate for block relaxation. Problems 8 and 9 relate the best vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ to the singular value decomposition of a matrix.

TABLE 7.2. Iterates in canonical correlation estimation

n	a_{n1}	a_{n2}	b_{n1}	b_{n2}
0	1.000000	1.000000	1.000000	1.000000
1	0.553047	0.520658	0.504588	0.538164
2	0.552159	0.521554	0.504509	0.538242
3	0.552155	0.521558	0.504509	0.538242
4	0.552155	0.521558	0.504509	0.538242

For fixed $\boldsymbol{b}$ we can easily find the best $\boldsymbol{a}$. Introduce the Lagrangian

$$\mathcal{L}(\boldsymbol{a}) \;=\; \boldsymbol{a}^*\boldsymbol{\Sigma}_{12}\boldsymbol{b} - \frac{\lambda}{2}\Big(\boldsymbol{a}^*\boldsymbol{\Sigma}_{11}\boldsymbol{a} - 1\Big),$$

and equate its gradient

$$\nabla\mathcal{L}(\boldsymbol{a}) \;=\; \Sigma_{12}\boldsymbol{b} - \lambda\Sigma_{11}\boldsymbol{a}$$

to $\mathbf{0}$. This gives the maximum point

$$\boldsymbol{a} \;=\; \frac{1}{\lambda}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{b},$$

assuming the submatrix $\boldsymbol{\Sigma}_{11}$ is positive definite. Inserting this value into the constraint $\boldsymbol{a}^*\boldsymbol{\Sigma}_{11}\boldsymbol{a} = 1$ allows us to solve for the Lagrange multiplier λ and hence pin down $\boldsymbol{a}$ as

$$\boldsymbol{a} \;=\; \frac{1}{\sqrt{\boldsymbol{b}^*\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{b}}}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{b}.$$

Because the second differential $d^2\mathcal{L} = -\lambda\boldsymbol{\Sigma}_{11}$ is negative definite, $\boldsymbol{a}$ represents the maximum. Likewise, fixing $\boldsymbol{a}$ and optimizing over $\boldsymbol{b}$ gives the update

$$\boldsymbol{b} \;=\; \frac{1}{\sqrt{\boldsymbol{a}^*\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{a}}}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{a}.$$

As a toy example consider the correlation matrix

$$\mathrm{Var}(\boldsymbol{Z}) \;=\; \begin{pmatrix} 1 & 0.7346 & 0.7108 & 0.7040 \\ 0.7346 & 1 & 0.6932 & 0.7086 \\ 0.7108 & 0.6932 & 1 & 0.8392 \\ 0.7040 & 0.7086 & 0.8392 & 1 \end{pmatrix}$$

with unit variances on its diagonal. Table 7.2 shows the first few iterates of block relaxation starting from $\boldsymbol{a} = \boldsymbol{b} = \mathbf{1}$. Convergence is exceptionally quick; more complex examples exhibit slower convergence. ■

Example 7.2.5 *Iterative Proportional Fitting*

Our next example of block relaxation is taken from the contingency table literature [14, 86]. Consider a three-way contingency table with two-way interactions. If the three factors are indexed by i, j, and k and have r, s, and t levels, respectively, then a loglinear model for the observed count data y_{ijk} is defined by an exponentially parameterized mean

$$\mu_{ijk} = e^{\lambda+\lambda_i^1+\lambda_j^2+\lambda_k^3+\lambda_{ij}^{12}+\lambda_{ik}^{13}+\lambda_{jk}^{23}}$$

for each cell ijk. To ensure that all parameters are identifiable, we make the usual assumption that a parameter set summed over one of its indices yields 0. For instance, $\lambda_\cdot^1 = \sum_i \lambda_i^1 = 0$ and $\lambda_{i\cdot}^{12} = \sum_j \lambda_{ij}^{12} = 0$. The overall effect λ is permitted to be nonzero.

If we postulate independent Poisson distributions for the random variables Y_{ijk} underlying the observed values y_{ijk}, then the loglikelihood is

$$L = \sum_i \sum_j \sum_k (y_{ijk} \ln \mu_{ijk} - \mu_{ijk}). \tag{7.2}$$

Maximizing L with respect to λ can be accomplished by setting

$$\frac{\partial}{\partial \lambda} L = \sum_i \sum_j \sum_k (y_{ijk} - \mu_{ijk}) = 0.$$

This tells us that whatever the other parameters are, λ should be adjusted so that $\mu_{\cdots} = y_{\cdots} = m$ is the total sample size. (Here again the dot convention signifies summation over a lost index.) In other words, if $\mu_{ijk} = e^\lambda \omega_{ijk}$, then λ is chosen so that $e^\lambda = m/\omega_{\cdots}$. With this proviso, the loglikelihood becomes

$$\begin{aligned} L &= \sum_i \sum_j \sum_k y_{ijk} \ln \frac{m\omega_{ijk}}{\omega_{\cdots}} - m \\ &= \sum_i \sum_j \sum_k y_{ijk} \ln \frac{\omega_{ijk}}{\omega_{\cdots}} + m \ln m - m, \end{aligned}$$

which is up to an irrelevant constant just the loglikelihood of a multinomial distribution with probability $\omega_{ijk}/\omega_{\cdots}$ attached to cell ijk. Thus, for purposes of maximum likelihood estimation, we might as well stick with the Poisson sampling model.

Unfortunately, no closed-form solution to the Poisson likelihood equations exists satisfying the complicated linear constraints. The resolution of this dilemma lies in refusing to update all of the parameters simultaneously. Suppose that we consider only the parameters λ, λ_i^1, λ_j^2, and λ_{ij}^{12} pertinent to the first two factors. If in equation (7.2) we let

$$\begin{aligned} \mu_{ij} &= e^{\lambda+\lambda_i^1+\lambda_j^2+\lambda_{ij}^{12}} \\ \alpha_{ijk} &= e^{\lambda_k^3+\lambda_{ik}^{13}+\lambda_{jk}^{23}}, \end{aligned}$$

then setting

$$\begin{aligned}\frac{\partial}{\partial\lambda_{ij}^{12}}L &= \sum_k (y_{ijk}-\mu_{ijk}) \\ &= y_{ij\cdot}-\mu_{ij\cdot} \\ &= y_{ij\cdot}-\mu_{ij}\alpha_{ij\cdot} \\ &= 0\end{aligned}$$

leads to $\mu_{ij} = y_{ij\cdot}/\alpha_{ij\cdot}$. The constraint $\sum_k (y_{ijk}-\mu_{ijk}) = 0$ implies that the other partial derivatives

$$\begin{aligned}\frac{\partial}{\partial\lambda}L &= y_{\cdots}-\mu_{\cdots} \\ \frac{\partial}{\partial\lambda_i^1}L &= y_{i\cdot\cdot}-\mu_{i\cdot\cdot} \\ \frac{\partial}{\partial\lambda_j^2}L &= y_{\cdot j\cdot}-\mu_{\cdot j\cdot}\end{aligned}$$

vanish as well. This stationary point of the loglikelihood is also a stationary point of the Lagrangian with all Lagrange multipliers equal to 0.

Of course, we still must nail down λ, λ_i^1, λ_j^2, and λ_{ij}^{12}. In view of the definition of μ_{ij}, the choice

$$\lambda_{ij}^{12} = \ln\left(\frac{y_{ij\cdot}}{\alpha_{ij\cdot}}\right)-\lambda-\lambda_i^1-\lambda_j^2$$

guarantees that $\mu_{ij} = y_{ij\cdot}/\alpha_{ij\cdot}$. One can check that the further choices

$$\begin{aligned}\lambda &= \frac{1}{rs}\sum_i\sum_j \ln\mu_{ij} \\ \lambda_i^1 &= \frac{1}{s}\sum_j \ln\mu_{ij}-\lambda \\ \lambda_j^2 &= \frac{1}{r}\sum_i \ln\mu_{ij}-\lambda\end{aligned}$$

satisfy the relevant equality constraints $\lambda_\cdot^1 = 0$, $\lambda_\cdot^2 = 0$, $\lambda_{\cdot j}^{12} = 0$, and $\lambda_{i\cdot}^{12} = 0$. The identity $\mu_{ij} = y_{ij\cdot}/\alpha_{ij\cdot}$ is crucial in this regard.

At the second stage, the parameter set $\{\lambda, \lambda_i^1, \lambda_k^3, \lambda_{ik}^{13}\}$ is updated, holding the remaining parameters fixed. At the third stage, the parameter set $\{\lambda, \lambda_j^2, \lambda_k^3, \lambda_{jk}^{23}\}$ is updated, holding the remaining parameters fixed. These three successive stages constitute one iteration of the iterative proportional fitting algorithm. Each stage either leaves all parameters unchanged or increases the loglikelihood. In this example, the parameter blocks are not disjoint. ■

Example 7.2.6 *Matrix Factorization by Alternating Least Squares*

Least squares, the most venerable of the statistical fitting procedures, was initiated by Gauss and Legendre. As explained in Example 1.3.3, the basic setup involves n independent responses that individually take the form

$$y_i = \sum_{j=1}^{p} x_{ij}\theta_j + u_i. \tag{7.3}$$

Here y_i depends linearly on the unknown regression coefficients θ_j through the known predictors x_{ij}. The error u_i is assumed to be normally distributed with mean 0 and variance σ^2. If we collect the y_i into a $n \times 1$ response vector $\boldsymbol{y}$, the x_{ij} into a $n \times p$ design matrix $\boldsymbol{X}$, the θ_j into a $p \times 1$ parameter vector $\boldsymbol{\theta}$, and the u_j into a $n \times 1$ error vector $\boldsymbol{u}$, then the linear regression model can be rewritten in vector notation as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{u}$. Provided the design matrix $\boldsymbol{X}$ has full rank, the least squares estimate of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^*\boldsymbol{X})^{-1}\boldsymbol{X}^*\boldsymbol{y}$. Example 5.2.6 shows how to amend this solution to take into account affine constraints. In weighted least squares, one minimizes the criterion

$$f(\boldsymbol{\theta}) = \sum_{i=1}^{n} c_i \Big(y_i - \sum_{j=1}^{p} x_{ij}\theta_j\Big)^2,$$

where the c_i are positive weights. This reduces to ordinary least squares if one substitutes $\sqrt{c_i}y_i$ for y_i and $\sqrt{c_i}x_{ij}$ for x_{ij}. It is clear that any method for solving an ordinary least squares problem can be immediately adapted to solving a weighted least squares problem.

The history of alternating least squares is summarized by Gifi [104]. Very early on Kruskal [158] applied the method to factorial ANOVA. Here we briefly survey its use in nonnegative matrix factorization [174, 175]. Suppose $\boldsymbol{U}$ is a $n \times q$ matrix whose columns $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q$ represent data vectors. In many applications one wants to explain the data by postulating a reduced number of prototypes $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ and writing

$$\boldsymbol{u}_j \approx \sum_{k=1}^{p} \boldsymbol{v}_k w_{kj}$$

for certain nonnegative weights w_{kj}. The matrix $\boldsymbol{W} = (w_{kj})$ is $p \times q$. If p is small compared to q, then the representation $\boldsymbol{U} \approx \boldsymbol{V}\boldsymbol{W}$ compresses the data for easier storage and retrieval. Depending on the circumstances, further constraints may be advisable [72]. For instance, if the entries of $\boldsymbol{U}$ are nonnegative, then it is often reasonable to demand that the entries of $\boldsymbol{V}$ be nonnegative as well. If we want each $\boldsymbol{u}_j$ to equal a convex combination of the prototypes, then constraining the column sums of $\boldsymbol{W}$ to equal 1 is indicated.

One way of estimating $\boldsymbol{V}$ and $\boldsymbol{W}$ is to minimize the objective function

$$\|\boldsymbol{U} - \boldsymbol{V}\boldsymbol{W}\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{q}\left(u_{ij} - \sum_{k=1}^{p} v_{ik}w_{kj}\right)^2.$$

No explicit solution is known, but alternating least squares offers an iterative attack. If $\boldsymbol{W}$ is fixed, then we can update the ith row of $\boldsymbol{V}$ by minimizing the sum of squares

$$\sum_{j=1}^{q}\left(u_{ij} - \sum_{k=1}^{p} v_{ik}w_{kj}\right)^2.$$

Similarly, if $\boldsymbol{V}$ is fixed, then we can update the jth column of $\boldsymbol{W}$ by minimizing the sum of squares

$$\sum_{i=1}^{n}\left(u_{ij} - \sum_{k=1}^{p} v_{ik}w_{kj}\right)^2.$$

In either case we are faced with solving a sequence of least squares problems. The introduction of nonnegativity constraints and convexity constraints complicates matters. Problem 12 suggests coordinate descent methods for solving these two constrained least squares problems. Coordinate descent is trivial to implement but potentially very slow. We will revisit nonnegative least squares later from a different perspective. ■

7.3 Problems

1. Program and test any one of the six examples in this chapter.

2. Demonstrate that cyclic coordinate descent either diverges or converges to a saddle point of the function $f : \mathsf{R}^2 \to \mathsf{R}$ defined by

$$f(\boldsymbol{x}) = (x_1 - x_2)^2 - 2x_1x_2.$$

   This function of de Leeuw [59] has no minimum.

3. Consider the function $f(\boldsymbol{x}) = (x_1^2 + x_2^2)^{-1} + \ln(x_1^2 + x_2^2)$ for $\boldsymbol{x} \neq \boldsymbol{0}$. Explicitly find the minimum value of $f(\boldsymbol{x})$. Specify the coordinate descent algorithm for finding the minimum. Note any ambiguities in the implementation of coordinate descent, and describe the possible cluster points of the algorithm as a function of the initial point. (Hint: Coordinate descent, properly defined, converges in a finite number of iterations.)

4. Implement cyclic coordinate descent to minimize the posynomial

$$f(\boldsymbol{x}) = \frac{1}{x_1^3} + \frac{3}{x_1 x_2^2} + x_1 x_2$$

over the region $\{\boldsymbol{x} \in \mathsf{R}^2 : x_1 > 0,\ x_2 > 0\}$. Derive the updates

$$\begin{aligned} x_{n+1,1} &= \sqrt{\frac{\frac{3}{x_{n2}^2} + \sqrt{\frac{9}{x_{n2}^4} + 12x_{n2}}}{2x_{n2}}} \\ x_{n+1,2} &= \sqrt[3]{\frac{6}{x_{n+1,1}^2}}. \end{aligned}$$

Compare your numerical results to those displayed in Table 8.2.

5. In Sinkhorn's theorem, suppose the matrix $\boldsymbol{M}$ is square. Show that some entries of $\boldsymbol{M}$ can be 0 as long as some positive power $\boldsymbol{M}^p$ of $\boldsymbol{M}$ has all entries positive.

6. Consider cluster analysis on the real line. Show that Lloyd's algorithm cannot improve on the initial partition $\pi_1 = \{0, 2\}$ and $\pi_2 = \{3.5\}$ despite that the fact that the partition $\pi_1 = \{0\}$ and $\pi_2 = \{2, 3.5\}$ is better. Also demonstrate that Lloyd's algorithm transforms the initial partition $\pi_1 = \{-7, -5\}$, $\pi_2 = \{-4, 4\}$, and $\pi_3 = \{5, 7\}$ into the partition $\pi_1 = \{-7, -5, -4\}$, $\pi_2 = \emptyset$, and $\pi_3 = \{4, 5, 7\}$.

7. Let $\boldsymbol{M} = (m_{ij})$ be a nontrivial $m \times n$ matrix. The dominant part of the singular value decomposition (svd) of $\boldsymbol{M}$ is an outer product matrix $\lambda \boldsymbol{u}\boldsymbol{v}^*$ with $\lambda > 0$ and $\boldsymbol{u}$ and $\boldsymbol{v}$ unit vectors. This outer product minimizes

$$\|\boldsymbol{M} - \lambda \boldsymbol{u}\boldsymbol{v}^*\|_F^2 = \sum_i \sum_j (m_{ij} - \lambda u_i v_j)^2.$$

One can use alternating least squares to find $\lambda \boldsymbol{u}\boldsymbol{v}^*$ [101]. In the first step of the algorithm, one fixes $\boldsymbol{v}$ and estimates $\boldsymbol{w} = \lambda \boldsymbol{u}$ by least squares. Show that $\boldsymbol{w}$ has components $w_i = \sum_j m_{ij} v_j$. Once $\boldsymbol{w}$ is available, we set $\lambda = \|\boldsymbol{w}\|$ and $\boldsymbol{u} = \|\boldsymbol{w}\|^{-1}\boldsymbol{w}$. What are the corresponding updates for $\boldsymbol{v}$ and λ when you fix $\boldsymbol{u}$? To find the next outer product in the svd, form the deflated matrix $\boldsymbol{M} - \lambda \boldsymbol{u}\boldsymbol{v}^*$ and repeat the process. Program and test this algorithm.

8. Continuing Problem 7, prove that minimizing $\|\boldsymbol{M} - \lambda \boldsymbol{u}\boldsymbol{v}^*\|_F^2$ subject to the constraints $\|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$ is equivalent to maximizing $\boldsymbol{u}^* \boldsymbol{M} \boldsymbol{v}$ subject to the same constraints. The solution vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are called singular vectors. The corresponding scalar λ is the singular value. Problem 7 of Chap. 2 relates λ to the spectral norm of $\boldsymbol{M}$.

9. In Example 7.2.4, make the linear change of variables $\boldsymbol{c} = \boldsymbol{\Sigma}_{11}^{1/2}\boldsymbol{a}$ and $\boldsymbol{d} = \boldsymbol{\Sigma}_{22}^{1/2}\boldsymbol{b}$. In the new variables show that one must maximize $\boldsymbol{c}^*\boldsymbol{\Omega}\boldsymbol{d}$ subject to $\|\boldsymbol{c}\| = \|\boldsymbol{d}\| = 1$, where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$. In the language of Problems 7 and 8, the first singular vectors $\boldsymbol{c} = \boldsymbol{u}$ and $\boldsymbol{d} = \boldsymbol{v}$ of the svd of $\boldsymbol{\Omega}$ solve the transformed problem. Obviously, the vectors $\boldsymbol{b} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{u}$ and $\boldsymbol{b} = \boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{v}$ solve the original problem. The advantage of this approach is that one can now define higher-order canonical correlations from the remaining singular vectors of the svd.

10. Suppose $\boldsymbol{A}$ is a symmetric matrix and $\boldsymbol{B}$ is a positive definite matrix of the same dimension. Formulate cyclic coordinate descent and ascent algorithms for minimizing and maximizing the Rayleigh quotient

$$R(\boldsymbol{x}) \;=\; \frac{\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{B}\boldsymbol{x}} \tag{7.4}$$

over the set $\boldsymbol{x} \neq \boldsymbol{0}$. Program and test this algorithm.

11. Continuing Problem 10, demonstrate that the maximum and minimum values of the Rayleigh quotient (7.4) coincide with the maximum and minimum eigenvalues of the matrix $\boldsymbol{B}^{-1}\boldsymbol{A}$.

12. For a positive definite matrix $\boldsymbol{A}$, consider minimizing the quadratic function $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x} + c$ subject to the constraints $x_i \geq 0$ for all i. Show that the cyclic coordinate descent updates are

$$\hat{x}_i \;=\; \max\Big\{0, x_i - a_{ii}^{-1}\Big[\sum_j a_{ij}x_j + b_i\Big]\Big\}.$$

If we impose the additional constraint $\sum_i x_i = 1$, the problem is harder. One line of attack is to minimize the penalized function

$$f_\mu(\boldsymbol{x}) \;=\; f(\boldsymbol{x}) + \frac{\mu}{2}\Big(\sum_i x_i - 1\Big)^2$$

for a large positive constant μ. The theory in Chap. 13 shows that the minimum of $f_\mu(\boldsymbol{x})$ tends to the constrained minimum of $f(\boldsymbol{x})$ as μ tends to ∞. Accepting this result, demonstrate that cyclic coordinate descent for $f_\mu(\boldsymbol{x})$ has updates

$$\hat{x}_i \;=\; \max\Big\{0, x_i - (a_{ii}+\mu)^{-1}\Big[\sum_j a_{ij}x_j + b_i + \mu\Big(\sum_j x_j - 1\Big)\Big]\Big\}.$$

Program this second algorithm and test it for the choices

$$\boldsymbol{A} \;=\; \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \quad \boldsymbol{b} \;=\; \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Start with $\mu = 1$ and double it every time you update the full vector $\boldsymbol{x}$. Do the iterates converge to the minimum of $f(\boldsymbol{x})$ subject to all constraints?

TABLE 7.3. Coronary disease data

Disease status	Cholesterol level	Blood pressure				
		1	2	3	4	Total
Coronary	1	2	3	3	4	12
	2	3	2	1	3	9
	3	8	11	6	6	31
	4	7	12	11	11	41
Total		20	28	21	24	93
No coronary	1	117	121	47	22	307
	2	85	98	43	20	246
	3	119	209	68	43	439
	4	67	99	46	33	245
Total		388	527	204	118	1,237

13. Program and test a k-medians clustering algorithm and concoct an example where it differs from k-means clustering.

14. In fitting splines to data, the problem arises of minimizing the criterion $\|\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\alpha} - \boldsymbol{V}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^*\boldsymbol{W}\boldsymbol{\beta}$ with respect $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ for $\lambda \geq 0$ and $\boldsymbol{W}$ positive semidefinite [166]. Derive the block descent updates

$$\begin{aligned} \boldsymbol{\alpha} &= (\boldsymbol{U}^*\boldsymbol{U})^{-1}\boldsymbol{U}^*(\boldsymbol{y} - \boldsymbol{V}\boldsymbol{\beta}) \\ \boldsymbol{\beta} &= (\boldsymbol{V}^*\boldsymbol{V} + \lambda\boldsymbol{W})^{-1}\boldsymbol{V}^*(\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\alpha}). \end{aligned}$$

15. Consider the coronary disease data [86, 159] displayed in the three-way contingency Table 7.3. Using iterative proportional fitting, find the maximum likelihood estimates for the loglinear model with first-order interactions. Perform a chi-square test to decide whether this model fits the data better than the model postulating independence of the three factors.

16. As noted in the text, the loglinear model for categorical data can be interpreted as assuming independent Poisson distributions for the various categories with category i having mean $\mu_i(\boldsymbol{\theta}) = e^{\boldsymbol{l}_i^*\boldsymbol{\theta}}$, where $\boldsymbol{l}_i$ is a vector whose entries are 0's or 1's. Calculate the observed information $-d^2L(\boldsymbol{\theta}) = \sum_i e^{\boldsymbol{l}_i^*\boldsymbol{\theta}}\boldsymbol{l}_i\boldsymbol{l}_i^*$ in this circumstance, and deduce that it is positive semidefinite. In the presence of affine constraints $\boldsymbol{V}\boldsymbol{\theta} = \boldsymbol{d}$ on $\boldsymbol{\theta}$, show that any maximum likelihood estimate of $\boldsymbol{\theta}$ is necessarily unique provided the null space (kernel) of $\boldsymbol{V}$ is contained in the linear span of the $\boldsymbol{l}_i$.

8
The MM Algorithm

8.1 Introduction

Most practical optimization problems defy exact solution. In the current chapter we discuss an optimization method that relies heavily on convexity arguments and is particularly useful in high-dimensional problems such as image reconstruction [171]. This iterative method is called the MM algorithm. One of the virtues of this acronym is that it does double duty. In minimization problems, the first M of MM stands for majorize and the second M for minimize. In maximization problems, the first M stands for minorize and the second M for maximize. When it is successful, the MM algorithm substitutes a simple optimization problem for a difficult optimization problem. Simplicity can be attained by: (a) separating the variables of an optimization problem, (b) avoiding large matrix inversions, (c) linearizing an optimization problem, (d) restoring symmetry, (e) dealing with equality and inequality constraints gracefully, and (f) turning a nondifferentiable problem into a smooth problem. In simplifying the original problem, we must pay the price of iteration or iteration with a slower rate of convergence.

Statisticians have vigorously developed a special case of the MM algorithm called the EM algorithm, which revolves around notions of missing data [65, 166, 191]. We present the EM algorithm in the next chapter. We prefer to present the MM algorithm first because of its greater generality, its more obvious connection to convexity, and its weaker reliance on difficult statistical principles.

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_8,

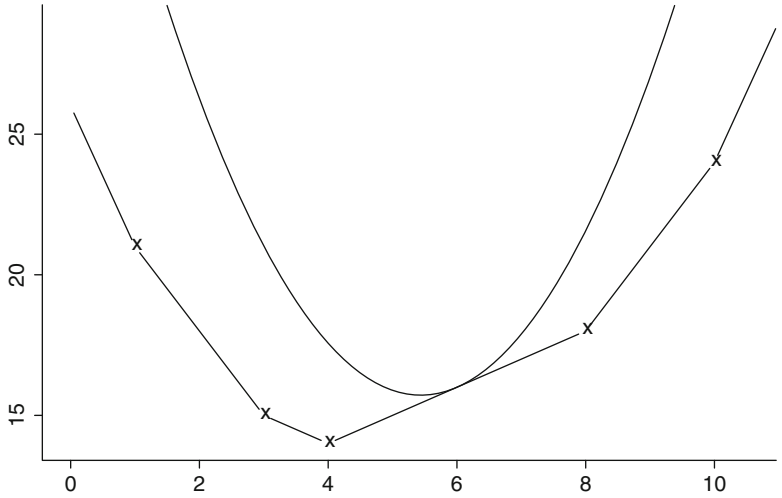

FIGURE 8.1. A quadratic majorizing function for the piecewise linear function $f(x) = |x-1| + |x-3| + |x-4| + |x-8| + |x-10|$ at the point $x_m = 6$

8.2 Philosophy of the MM Algorithm

A function $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ is said to majorize a function $f(\boldsymbol{x})$ at $\boldsymbol{x}_m$ provided

$$f(\boldsymbol{x}_m) = g(\boldsymbol{x}_m \mid \boldsymbol{x}_m) \tag{8.1}$$

$$f(\boldsymbol{x}) \leq g(\boldsymbol{x} \mid \boldsymbol{x}_m), \qquad \boldsymbol{x} \neq \boldsymbol{x}_m. \tag{8.2}$$

In other words, the surface $\boldsymbol{x} \mapsto g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ lies above the surface $f(\boldsymbol{x})$ and is tangent to it at the point $\boldsymbol{x} = \boldsymbol{x}_m$. Here $\boldsymbol{x}_m$ represents the current iterate in a search of the surface $f(\boldsymbol{x})$. Figure 8.1 provides a simple one-dimensional example.

In the minimization version of the MM algorithm, we minimize the surrogate majorizing function $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ rather than the actual function $f(\boldsymbol{x})$. If $\boldsymbol{x}_{m+1}$ denotes the minimum of the surrogate $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$, then we can show that the MM procedure forces $f(\boldsymbol{x})$ downhill. Indeed, the inequalities

$$f(\boldsymbol{x}_{m+1}) \leq g(\boldsymbol{x}_{m+1} \mid \boldsymbol{x}_m) \leq g(\boldsymbol{x}_m \mid \boldsymbol{x}_m) = f(\boldsymbol{x}_m) \tag{8.3}$$

follow directly from the definition of $\boldsymbol{x}_{m+1}$ and the majorization conditions (8.1) and (8.2). The descent property (8.3) lends the MM algorithm remarkable numerical stability. Strictly speaking, it depends only on decreasing $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$, not on minimizing $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$. This fact has practical consequences when the minimum of $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ cannot be found exactly. When $f(\boldsymbol{x})$ is strictly convex, one can show with a few additional mild hypotheses that the iterates $\boldsymbol{x}_m$ converge to the global minimum of $f(\boldsymbol{x})$ regardless of the initial point $\boldsymbol{x}_0$.

If $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ majorizes $f(\boldsymbol{x})$ at an interior point $\boldsymbol{x}_m$ of the domain of $f(\boldsymbol{x})$, then $\boldsymbol{x}_m$ is a stationary point of the difference $g(\boldsymbol{x} \mid \boldsymbol{x}_m) - f(\boldsymbol{x})$, and the gradient identity

$$\nabla g(\boldsymbol{x}_m \mid \boldsymbol{x}_m) \quad = \quad \nabla f(\boldsymbol{x}_m) \tag{8.4}$$

holds. Furthermore, the second differential $d^2 g(\boldsymbol{x}_m \mid \boldsymbol{x}_m) - d^2 f(\boldsymbol{x}_m)$ is positive semidefinite. Problem 1 makes the point that the majorization relation between functions is closed under the formation of sums, nonnegative products, limits, and composition with an increasing function. These rules permit us to work piecemeal in simplifying complicated objective functions. With obvious changes, the MM algorithm also applies to maximization rather than to minimization. To maximize a function $f(\boldsymbol{x})$, we minorize it by a surrogate function $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ and maximize $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ to produce the next iterate $\boldsymbol{x}_{m+1}$.

The reader might well object that the MM algorithm is not so much an algorithm as a vague philosophy for deriving an algorithm. The same objection applies to the EM algorithm. As we proceed through the current chapter, we hope the various examples will convince the reader of the value of a unifying principle and a framework for attacking concrete problems. The strong connection of the MM algorithm to convexity and inequalities has the natural pedagogical advantage of building on the material presented in previous chapters.

8.3 Majorization and Minorization

We will feature five methods for constructing majorizing functions. Two of these simply adapt Jensen's inequality

$$f\Big(\sum_i \alpha_i t_i\Big) \quad \leq \quad \sum_i \alpha_i f(t_i)$$

defining a convex function $f(t)$. It is easy to identify convex functions on the real line, so the first method composes such a function with a linear function $\boldsymbol{c}^* \boldsymbol{x}$ to create a new convex function of the vector $\boldsymbol{x}$. Invoking the definition of convexity with $\alpha_i = c_i y_i / \boldsymbol{c}^* \boldsymbol{y}$ and $t_i = \boldsymbol{c}^* \boldsymbol{y}\, x_i / y_i$ then yields

$$f(\boldsymbol{c}^* \boldsymbol{x}) \quad \leq \quad \sum_i \frac{c_i y_i}{\boldsymbol{c}^* \boldsymbol{y}} f\Big(\frac{\boldsymbol{c}^* \boldsymbol{y}}{y_i} x_i\Big) \quad = \quad g(\boldsymbol{x} \mid \boldsymbol{y}), \tag{8.5}$$

provided all of the components of the vectors $\boldsymbol{c}$, $\boldsymbol{x}$, and $\boldsymbol{y}$ are positive. The surrogate function $g(\boldsymbol{x} \mid \boldsymbol{y})$ equals $f(\boldsymbol{c}^* \boldsymbol{y})$ when $\boldsymbol{x} = \boldsymbol{y}$. One of the virtues of applying inequality (8.5) in defining a surrogate function is that it separates parameters in the surrogate function. This feature is critically important in high-dimensional problems because it reduces optimization

over $\boldsymbol{x}$ to a sequence of one-dimensional optimizations over each component x_i. The argument establishing inequality (8.5) is equally valid if we replace the parameter vector $\boldsymbol{x}$ throughout by a vector-valued function $h(\boldsymbol{x})$ of $\boldsymbol{x}$. The genetics problem in the next section illustrates this variant of the technique.

To relax the positivity restrictions on the vectors $\boldsymbol{c}$, $\boldsymbol{x}$, and $\boldsymbol{y}$, De Pierro [67] suggested in a medical imaging context the alternative majorization

$$f(\boldsymbol{c}^*\boldsymbol{x}) \quad \leq \quad \sum_i \alpha_i f\left\{\frac{c_i}{\alpha_i}(x_i - y_i) + \boldsymbol{c}^*\boldsymbol{y}\right\} \quad = \quad g(\boldsymbol{x} \mid \boldsymbol{y}) \tag{8.6}$$

for a convex function $f(t)$. Here all $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$, and $\alpha_i > 0$ whenever $c_i \neq 0$. In practice, we must somehow tailor the α_i to the problem at hand. Among the obvious candidates for the α_i are

$$\alpha_i \quad = \quad \frac{|c_i|^p}{\sum_j |c_j|^p}$$

for $p \geq 0$. When $p = 0$, we interpret α_i as 0 if $c_i = 0$ and as $1/q$ if c_i is one among q nonzero coefficients.

Our third method involves the linear majorization

$$f(\boldsymbol{x}) \quad \leq \quad f(\boldsymbol{y}) + df(\boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y}) \quad = \quad g(\boldsymbol{x} \mid \boldsymbol{y}) \tag{8.7}$$

satisfied by any concave function $f(\boldsymbol{x})$. Once again we can replace the argument $\boldsymbol{x}$ by a vector-valued function $h(\boldsymbol{x})$.

Our fourth method applies to functions $f(\boldsymbol{x})$ with bounded curvature [16, 59]. Assuming that $f(\boldsymbol{x})$ is twice differentiable, we look for a matrix $\boldsymbol{B}$ satisfying $\boldsymbol{B} \succeq d^2 f(\boldsymbol{x})$ and $\boldsymbol{B} \succ \boldsymbol{0}$ in the sense that $\boldsymbol{B} - d^2 f(\boldsymbol{x})$ is positive semidefinite for all $\boldsymbol{x}$ and $\boldsymbol{B}$ is positive definite. The quadratic bound principle then amounts to the majorization

$$\begin{aligned} f(\boldsymbol{x}) \quad &= \quad f(\boldsymbol{y}) + df(\boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y}) \\ &\qquad + (\boldsymbol{x} - \boldsymbol{y})^* \int_0^1 d^2 f[\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})](1 - t)\, dt\, (\boldsymbol{x} - \boldsymbol{y}) \\ &\leq \quad f(\boldsymbol{y}) + df(\boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{y})^* \boldsymbol{B}(\boldsymbol{x} - \boldsymbol{y}) \\ &= \quad g(\boldsymbol{x} \mid \boldsymbol{y}). \end{aligned} \tag{8.8}$$

Our fifth and final method exploits the generalized arithmetic-geometric mean inequality (6.5) of Chap. 6. With this result in mind, Problem 9 asks the reader to prove the majorization

$$\prod_{i=1}^n x_i^{\alpha_i} \quad \leq \quad \left(\prod_{i=1}^n y_i^{\alpha_i}\right) \sum_{i=1}^n \frac{\alpha_i}{\alpha} \left(\frac{x_i}{y_i}\right)^{\alpha} \quad = \quad g(\boldsymbol{x} \mid \boldsymbol{y}) \tag{8.9}$$

for positive numbers x_i, y_i, and α_i and sum $\alpha = \sum_{i=1}^{n} \alpha_i$. Inequality (8.9) is the key to separating parameters with posynomials. We will use it in sketching an MM algorithm for unconstrained geometric programming.

Any of the first four majorizations can be turned into minorizations by interchanging the adjectives convex and concave and positive definite and negative definite, respectively. Of course, there is an art to applying these methods just as there is an art to applying any mathematical principle. The methods hardly exhaust the possibilities for majorization and minorization. The problems at the end of the chapter sketch other helpful techniques. Readers are also urged to consult the survey papers [59, 122, 142, 171] and the literature on the EM algorithm for a fuller discussion.

8.4 Allele Frequency Estimation

The ABO and Rh genetic loci are usually typed in matching blood donors to blood recipients. The ABO locus incorporates the three alleles A, B, and O and exhibits the four observable phenotypes A, B, AB, and O. These phenotypes arise because each person inherits two alleles, one from his mother and one from his father, and the alleles A and B are genetically dominant to allele O. Dominance amounts to a masking of the O allele by the presence of an A or B allele. For instance, a person inheriting an A allele from one parent and an O allele from the other parent is said to have genotype A/O and is indistinguishable from a person inheriting an A allele from both parents. This second person has genotype A/A.

The MM algorithm for estimating the population frequencies or proportions of the three alleles involves an interplay between observed phenotypes and underlying unobserved genotypes. As just noted, both genotypes A/O and A/A generate the same phenotype A. Likewise, both genotypes B/O and B/B generate the same phenotype B. Phenotypes AB and O correspond to the single genotypes A/B and O/O, respectively.

As a concrete example, Clarke et al. [50] noted that among their population sample of $n = 521$ duodenal ulcer patients, a total of $n_A = 186$ had phenotype A, $n_B = 38$ had phenotype B, $n_{AB} = 13$ had phenotype AB, and $n_O = 284$ had phenotype O. If we want to estimate the frequencies p_A, p_B, and p_O of the three different alleles from this sample, then we can employ the MM algorithm with the four phenotype counts as the observed data.

The likelihood of the data is given by the multinomial distribution in conjunction with the Hardy-Weinberg law of population genetics. This law specifies that each genotype frequency equals the product of the corresponding allele frequencies with an extra factor of 2 included to account for ambiguity in parental source when the two alleles differ. For example, genotype A/A has frequency p_A^2, and genotype A/O has frequency $2p_Ap_O$.

These assumptions are summarized in the multinomial loglikelihood

$$\begin{aligned} f(\boldsymbol{p}) &= n_A \ln\left(p_A^2 + 2p_Ap_O\right) + n_B \ln\left(p_B^2 + 2p_Bp_O\right) + n_{AB}\ln(2p_Ap_B) \\ &\quad + n_O \ln p_O^2 + \ln \binom{n}{n_A,\ n_B,\ n_{AB},\ n_O}. \end{aligned}$$

In maximum likelihood estimation we maximize this function of the allele frequencies subject to the equality constraint $p_A + p_B + p_O = 1$ and the nonnegativity constraints $p_A \geq 0$, $p_B \geq 0$, and $p_O \geq 0$.

The loglikelihood function $f(\boldsymbol{p})$ would be easy to maximize if it were not for the terms $\ln\left(p_A^2 + 2p_Ap_O\right)$ and $\ln\left(p_B^2 + 2p_Bp_O\right)$. In the MM algorithm we attack these functions using the convexity of the function $-\ln x$ and the majorization (8.5). This yields the minorization

$$\begin{aligned} \ln\left(p_A^2 + 2p_Ap_O\right) &\geq \frac{p_{mA}^2}{p_{mA}^2 + 2p_{mA}p_{mO}} \ln\left(\frac{p_{mA}^2 + 2p_{mA}p_{mO}}{p_{mA}^2} p_A^2\right) \\ &\quad + \frac{2p_{mA}p_{mO}}{p_{mA}^2 + 2p_{mA}p_{mO}} \ln\left(\frac{p_{mA}^2 + 2p_{mA}p_{mO}}{2p_{mA}p_{mO}} 2p_Ap_O\right). \end{aligned}$$

A similar minorization applies to $\ln\left(p_B^2 + 2p_Bp_O\right)$. These maneuvers have the virtue of separating parameters because logarithms turn products into sums.

Notationally, things become clearer if we introduce the abbreviations

$$\begin{aligned} n_{mA/A} &= n_A \frac{p_{mA}^2}{p_{mA}^2 + 2p_{mA}p_{mO}} \\ n_{mA/O} &= n_A \frac{2p_{mA}p_{mO}}{p_{mA}^2 + 2p_{mA}p_{mO}} \end{aligned}$$

and likewise for $n_{mB/B}$ and $n_{mB/O}$. We are now faced with maximizing the surrogate function

$$\begin{aligned} g(\boldsymbol{p} \mid \boldsymbol{p}_m) &= n_{mA/A} \ln p_A^2 + n_{mA/O} \ln(2p_Ap_O) + n_{mB/B} \ln p_B^2 \\ &\quad + n_{mB/O} \ln(2p_Bp_O) + n_{AB} \ln(2p_Ap_B) + n_O \ln p_O^2 + c, \end{aligned}$$

where c is an irrelevant constant that depends on the current iterate $\boldsymbol{p}_m$ but not on the potential value $\boldsymbol{p}$ of the next iterate. This completes the minorization step of the algorithm.

The maximization step can be accomplished by introducing a Lagrange multiplier and finding a stationary point of the Lagrangian

$$\mathcal{L}(\boldsymbol{p}, \lambda) = g(\boldsymbol{p} \mid \boldsymbol{p}_m) + \lambda(p_A + p_B + p_O - 1).$$

Here we ignore the nonnegativity constraints under the assumption that they are inactive at the solution. Setting the partial derivatives of $\mathcal{L}(\boldsymbol{p}, \lambda)$,

$$\frac{\partial}{\partial p_A}\mathcal{L}(\boldsymbol{p}, \lambda) = \frac{2n_{mA/A}}{p_A} + \frac{n_{mA/O}}{p_A} + \frac{n_{AB}}{p_A} + \lambda$$

TABLE 8.1. Iterations for ABO duodenal ulcer data

Iteration m	p_{mA}	p_{mB}	p_{mO}
0	0.3000	0.2000	0.5000
1	0.2321	0.0550	0.7129
2	0.2160	0.0503	0.7337
3	0.2139	0.0502	0.7359
4	0.2136	0.0501	0.7363
5	0.2136	0.0501	0.7363

$$\begin{aligned}
\frac{\partial}{\partial p_B}\mathcal{L}(\boldsymbol{p},\lambda) &= \frac{2n_{mB/B}}{p_B}+\frac{n_{mB/O}}{p_B}+\frac{n_{AB}}{p_B}+\lambda \\
\frac{\partial}{\partial p_O}\mathcal{L}(\boldsymbol{p},\lambda) &= \frac{n_{mA/O}}{p_O}+\frac{n_{mB/O}}{p_O}+\frac{2n_O}{p_O}+\lambda \\
\frac{\partial}{\partial \lambda}\mathcal{L}(\boldsymbol{p},\lambda) &= p_A+p_B+p_O-1,
\end{aligned}$$

equal to 0 provides the unique stationary point of $\mathcal{L}(\boldsymbol{p},\lambda)$. The solution of the resulting equations is

$$\begin{aligned}
p_{m+1,A} &= \frac{2n_{mA/A}+n_{mA/O}+n_{AB}}{2n} \\
p_{m+1,B} &= \frac{2n_{mB/B}+n_{mB/O}+n_{AB}}{2n} \\
p_{m+1,O} &= \frac{n_{mA/O}+n_{mB/O}+2n_O}{2n}.
\end{aligned}$$

In other words, the MM update is identical to a form of gene counting in which the unknown genotype counts are imputed based on the current allele frequency estimates [239]. In these updates, the denominator $2n$ is the total number of genes; the numerators are the current best guesses of the number of alleles of each type contained in the hidden and manifest genotypes.

Table 8.1 shows the progress of the MM iterates starting from the initial estimates $p_{0A}=0.3$, $p_{0B}=0.2$, and $p_{0O}=0.5$. The MM updates are simple enough to carry out on a pocket calculator. Convergence occurs quickly in this example.

8.5 Linear Regression

Because t^2 is a convex function, we can majorize each summand of the sum of squares criterion $\sum_{i=1}^n (y_i-\boldsymbol{x}_i^*\boldsymbol{\theta})^2$ using inequality (8.6). The overall

surrogate

$$g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m) \;=\; \sum_{i=1}^{n} \sum_{j} \alpha_{ij} \left[y_i - \frac{x_{ij}}{\alpha_{ij}} (\theta_j - \theta_{mj}) - \boldsymbol{x}_i^* \boldsymbol{\theta}_m \right]^2,$$

achieves equality with the sum of squares when $\boldsymbol{\theta} = \boldsymbol{\theta}_m$. Minimization of $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$ yields the updates

$$\theta_{m+1,j} \;=\; \theta_{mj} + \frac{\sum_{i=1}^{n} x_{ij}(y_i - \boldsymbol{x}_i^* \boldsymbol{\theta}_m)}{\sum_{i=1}^{n} \frac{x_{ij}^2}{\alpha_{ij}}} \tag{8.10}$$

and avoids matrix inversion [171]. Although it seems intuitively reasonable to take $p = 1$ in choosing

$$\alpha_{ij} \;=\; \frac{|x_{ij}|^p}{(\sum_k |x_{ik}|^p)},$$

conceivably other values of p might perform better. In fact, it might accelerate convergence to alternate different values of p as the iterations proceed. For problems involving just a few parameters, this iterative scheme is clearly inferior to the usual single-step solution via matrix inversion. Cyclic coordinate descent also avoids matrix operations, and Problem 11 suggests that it will converge faster than the MM update (8.10).

Least squares estimation suffers from the fact that it is strongly influenced by observations far removed from their predicted values. In least absolute deviation regression, we replace $\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^* \boldsymbol{\theta})^2$ by

$$h(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{n} \left| y_i - \boldsymbol{x}_i^* \boldsymbol{\theta} \right| \;=\; \sum_{i=1}^{n} |r_i(\boldsymbol{\theta})|, \tag{8.11}$$

where $r_i(\boldsymbol{\theta}) = y_i - \boldsymbol{x}_i^* \boldsymbol{\theta}$ is the ith residual.We are now faced with minimizing a nondifferentiable function. Fortunately, the MM algorithm can be implemented by exploiting the concavity of the function $\sqrt{u}$ in inequality (8.7). Because

$$\sqrt{u} \;\leq\; \sqrt{u_m} + \frac{u - u_m}{2\sqrt{u_m}}, \tag{8.12}$$

we find that

$$\begin{aligned} h(\boldsymbol{\theta}) \;&=\; \sum_{i=1}^{n} \sqrt{r_i^2(\boldsymbol{\theta})} \\ &\leq\; h(\boldsymbol{\theta}_m) + \frac{1}{2} \sum_{i=1}^{n} \frac{r_i^2(\boldsymbol{\theta}) - r_i^2(\boldsymbol{\theta}_m)}{\sqrt{r_i^2(\boldsymbol{\theta}_m)}} \\ &=\; g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m). \end{aligned}$$

Minimizing $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$ is accomplished by minimizing the weighted sum of squares

$$\sum_{i=1}^{n} w_i(\boldsymbol{\theta}_m) r_i(\boldsymbol{\theta})^2$$

with ith weight $w_i(\boldsymbol{\theta}_m) = |r_i(\boldsymbol{\theta}_m)|^{-1}$. A slight variation of the usual argument for minimizing a sum of squares leads to the update

$$\boldsymbol{\theta}_{m+1} = [\boldsymbol{X}^* \boldsymbol{W}(\boldsymbol{\theta}_m) \boldsymbol{X}]^{-1} \boldsymbol{X}^* \boldsymbol{W}(\boldsymbol{\theta}_m) \boldsymbol{y},$$

where $\boldsymbol{W}(\boldsymbol{\theta}_m)$ is the diagonal matrix with ith diagonal entry $w_i(\boldsymbol{\theta}_m)$. Unfortunately, the possibility that some weight $w_i(\boldsymbol{\theta}_m)$ is infinite cannot be ruled out. Problem 14 suggests a simple remedy.

8.6 Bradley-Terry Model of Ranking

In the sports version of the Bradley and Terry model [23, 140, 150], each team i in a league of teams is assigned a rank parameter $r_i > 0$. Assuming ties are impossible, team i beats team j with probability $r_i/(r_i + r_j)$. If this outcome occurs y_{ij} times during a season of play, then the probability of the whole season is

$$L(\boldsymbol{r}) = \prod_{i,j} \left(\frac{r_i}{r_i + r_j}\right)^{y_{ij}},$$

assuming the games are independent. To rank the teams, we find the values $\hat{r}_i$ that maximize $f(\boldsymbol{r}) = \ln L(\boldsymbol{r})$. The team with largest $\hat{r}_i$ is considered best, the team with smallest $\hat{r}_i$ is considered worst, and so forth. In view of the fact that $\ln u$ is concave, inequality (8.7) implies

$$\begin{aligned} f(\boldsymbol{r}) &= \sum_{i,j} y_{ij} \Big[\ln r_i - \ln(r_i + r_j) \Big] \\ &\geq \sum_{i,j} y_{ij} \Big[\ln r_i - \ln(r_{mi} + r_{mj}) - \frac{r_i + r_j - r_{mi} - r_{mj}}{r_{mi} + r_{mj}} \Big] \\ &= g(\boldsymbol{r} \mid \boldsymbol{r}_m) \end{aligned}$$

with equality when $\boldsymbol{r} = \boldsymbol{r}_m$. Differentiating $g(\boldsymbol{r} \mid \boldsymbol{r}_m)$ with respect to the ith component r_i of $\boldsymbol{r}$ and setting the result equal to 0 produces the next iterate

$$r_{m+1,i} = \frac{\sum_{j \neq i} y_{ij}}{\sum_{j \neq i} (y_{ij} + y_{ji})/(r_{mi} + r_{mj})}.$$

Because $L(\boldsymbol{r}) = L(\beta \boldsymbol{r})$ for any $\beta > 0$, we constrain $r_1 = 1$ and omit the update $r_{m+1,1}$. In this example, the MM algorithm separates parameters and allows us to maximize $g(\boldsymbol{r} \mid \boldsymbol{r}_m)$ parameter by parameter.

8.7 Linear Logistic Regression

In linear logistic regression, we observe a sequence of independent Bernoulli trials, each resulting in success or failure. The success probability of the ith trial

$$\pi_i(\boldsymbol{\theta}) = \frac{e^{\boldsymbol{x}_i^*\boldsymbol{\theta}}}{1+e^{\boldsymbol{x}_i^*\boldsymbol{\theta}}}$$

depends on a predictor (covariate) vector $\boldsymbol{x}_i$ and parameter vector $\boldsymbol{\theta}$ by analogy with linear regression. The observation y_i at trial i equals 1 for a success and 0 for a failure. In this notation, the likelihood of the data is

$$L(\boldsymbol{\theta}) = \prod_i \pi_i(\boldsymbol{\theta})^{y_i}[1-\pi_i(\boldsymbol{\theta})]^{1-y_i}.$$

As usual in maximum likelihood estimation, we pass to the loglikelihood

$$f(\boldsymbol{\theta}) = \sum_i [y_i \ln \pi_i(\boldsymbol{\theta}) + (1-y_i)\ln[1-\pi_i(\boldsymbol{\theta})].$$

Straightforward calculations show

$$\begin{aligned} df(\boldsymbol{\theta}) &= \sum_i [y_i - \pi_i(\boldsymbol{\theta})]\boldsymbol{x}_i^* \\ d^2 f(\boldsymbol{\theta}) &= -\sum_i \pi_i(\boldsymbol{\theta})[1-\pi_i(\boldsymbol{\theta})]\boldsymbol{x}_i\boldsymbol{x}_i^*. \end{aligned}$$

The loglikelihood $f(\boldsymbol{\theta})$ is therefore concave, and we seek to minorize it by a quadratic rather than majorize it by a quadratic as suggested in inequality (8.8). Hence, we must identify a matrix $\boldsymbol{B}$ such that $\boldsymbol{B}$ is negative definite and $\boldsymbol{B} - d^2 f(\boldsymbol{\theta})$ is negative semidefinite for all $\boldsymbol{\theta}$. In view of the scalar inequality $\pi(1-\pi) \le \frac{1}{4}$, we take $\boldsymbol{B} = -\frac{1}{4}\sum_i \boldsymbol{x}_i\boldsymbol{x}_i^*$. Maximization of the minorizing quadratic

$$f(\boldsymbol{\theta}_m) + df(\boldsymbol{\theta}_m)(\boldsymbol{\theta}-\boldsymbol{\theta}_m) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_m)^*\boldsymbol{B}(\boldsymbol{\theta}-\boldsymbol{\theta}_m)$$

is a problem we have met before. It does involve inversion of the matrix $\boldsymbol{B}$, but once we have computed $\boldsymbol{B}^{-1}$, we can store and reuse it at every iteration.

8.8 Geometric and Signomial Programs

The idea behind these minimization algorithms is best understood in a concrete setting. Consider the posynomial

$$f(\boldsymbol{x}) = \frac{1}{x_1^3} + \frac{3}{x_1 x_2^2} + x_1 x_2 \tag{8.13}$$

with the implied constraints $x_1 > 0$ and $x_2 > 0$. The majorization (8.9) applied to the third term of $f(\boldsymbol{x})$ yields

$$\begin{aligned} x_1 x_2 &\leq x_{m1} x_{m2} \left[\frac{1}{2} \left(\frac{x_1}{x_{m1}} \right)^2 + \frac{1}{2} \left(\frac{x_2}{x_{m2}} \right)^2 \right] \\ &= \frac{x_{m2}}{2x_{m1}} x_1^2 + \frac{x_{m1}}{2x_{m2}} x_2^2. \end{aligned}$$

Applied to the second term of $f(\boldsymbol{x})$, it gives with x_1^{-1} replacing x_1 and x_2^{-1} replacing x_2,

$$\begin{aligned} \frac{3}{x_1 x_2^2} &\leq \frac{3}{x_{m1} x_{m2}^2} \left[\frac{1}{3} \left(\frac{x_{m1}}{x_1} \right)^3 + \frac{2}{3} \left(\frac{x_{m2}}{x_2} \right)^3 \right] \\ &= \frac{x_{m1}^2}{x_{m2}^2} \frac{1}{x_1^3} + \frac{2x_{m2}}{x_{m1}} \frac{1}{x_2^3}. \end{aligned}$$

The second step of the MM algorithm for minimizing $f(\boldsymbol{x})$ therefore splits into minimizing the two surrogate functions

$$\begin{aligned} g_1(x_1 \mid \boldsymbol{x}_m) &= \frac{1}{x_1^3} + \frac{x_{m1}^2}{x_{m2}^2} \frac{1}{x_1^3} + \frac{x_{m2}}{2x_{m1}} x_1^2 \\ g_2(x_2 \mid \boldsymbol{x}_m) &= \frac{2x_{m2}}{x_{m1}} \frac{1}{x_2^3} + \frac{x_{m1}}{2x_{m2}} x_2^2. \end{aligned}$$

If we set the derivatives of each of these equal to 0, then we find the solutions

$$\begin{aligned} x_{m+1,1} &= \sqrt[5]{3 \left(\frac{x_{m1}^2}{x_{m2}^2} + 1 \right) \frac{x_{m1}}{x_{m2}}} \\ x_{m+1,2} &= \sqrt[5]{6 \frac{x_{m2}^2}{x_{m1}^2}}. \end{aligned}$$

It is obvious that the point $\boldsymbol{x} = (\sqrt[5]{6}, \sqrt[5]{6})^*$ is a fixed point of these equations and minimizes $f(\boldsymbol{x})$. Ignoring this fact, Table 8.2 records the iterates of both the MM algorithm and cyclic coordinate descent. Although the MM updates are slower to converge, they are less complicated. See Problem 4 of Chap. 7 for the form of the cyclic coordinate descent updates.

This MM analysis carries over to general posynomials except that we cannot expect to derive explicit solutions of the minimization step. (See Problem 28.) Each separated surrogate function is a posynomial in a single variable. If the powers appearing in one of these posynomials are integers, then the derivative of the posynomial is a rational function, and once we equate it to 0, we are faced with solving a polynomial equation. This can be accomplished by bisection or by Newton's method as discussed in Chap. 10. Introducing posynomial constraints is another matter. Box constraints in

TABLE 8.2. MM and coordinate descent iterates for a geometric program

	MM algorithm			Coordinate descent		
m	x_{m1}	x_{m2}	$f(\boldsymbol{x}_m)$	x_{m1}	x_{m2}	$f(\boldsymbol{x}_m)$
0	1.00000	2.00000	3.75000	1.00000	2.00000	3.75000
1	1.13397	1.88818	3.56899	1.19437	1.61420	3.47886
2	1.19643	1.75472	3.49766	1.32882	1.50339	3.42280
3	1.24544	1.66786	3.46079	1.38616	1.46165	3.41457
4	1.28395	1.60829	3.44074	1.41117	1.44432	3.41312
5	1.31428	1.56587	3.42942	1.42219	1.43685	3.41285
10	1.39358	1.47003	3.41427	1.43082	1.43107	3.41279
20	1.42699	1.43496	3.41280	1.43097	1.43097	3.41279
30	1.43054	1.43140	3.41279	1.43097	1.43097	3.41279
40	1.43092	1.43101	3.41279	1.43097	1.43097	3.41279
50	1.43096	1.43097	3.41279	1.43097	1.43097	3.41279
51	1.43097	1.43097	3.41279	1.43097	1.43097	3.41279

the form $a_i \leq x_i \leq b_i$ are consistent with parameter separation as developed here, but more complicated posynomial constraints are not.

The perturbation

$$f(\boldsymbol{x}) \;=\; \frac{1}{x_1^3} + \frac{3}{x_1 x_2^2} + x_1 x_2 - \sqrt{x_1 x_2}$$

of the function (8.13) is called a signomial rather than a posynomial [20]. If we want to minimize this new function subject to the constraints $x_1 > 0$ and $x_2 > 0$, then a different tactic is needed for the terms with negative coefficients [170]. Consider the minorization $z \geq 1 + \ln z$ around the point $z = 1$ derived from the convexity of $-\ln z$. If we let $z = \sqrt{x_1 x_2}/\sqrt{x_{m1} x_{m2}}$, then the more elaborate minorization

$$\sqrt{x_1 x_2} \;\geq\; \frac{1}{2}\sqrt{x_{m1} x_{m2}}(2 + \ln x_1 + \ln x_2 - \ln x_{m1} - \ln x_{m2})$$

follows. Multiplication of this by -1 now gives the operative majorization. Up to an irrelevant additive constant, the overall surrogate function equals the sum of the two parameter separated surrogates

$$\begin{aligned} g_1(x_1 \mid \boldsymbol{x}_m) &= \frac{1}{x_1^3} + \frac{x_{m1}^2}{x_{m2}^2}\frac{1}{x_1^3} + \frac{x_{m2}}{2x_{m1}}x_1^2 - \frac{1}{2}\sqrt{x_{m1}x_{m2}}\ln x_1 \\ g_2(x_2 \mid \boldsymbol{x}_m) &= \frac{2x_{m2}}{x_{m1}}\frac{1}{x_2^3} + \frac{x_{m1}}{2x_{m2}}x_2^2 - \frac{1}{2}\sqrt{x_{m1}x_{m2}}\ln x_2. \end{aligned}$$

The corresponding minima are roots of the polynomial equations

$$\begin{aligned} 0 &= \frac{x_{m2}}{x_{m1}}x_1^5 - \frac{1}{2}\sqrt{x_{m1}x_{m2}}x_1^3 - 3\left(1 + \frac{x_{m1}^2}{x_{m2}^2}\right) \\ 0 &= \frac{x_{m1}}{x_{m2}}x_2^5 - \frac{1}{2}\sqrt{x_{m1}x_{m2}}x_2^3 - \frac{6x_{m2}}{x_{m1}}. \end{aligned}$$

Table 8.3 displays the MM iterates for this signomial program. Each update relies on Newton's method to solve the two preceding quintic equations. Although there is no convexity guarantee that the converged point is optimal, random sampling of the objective function does not produce a better point.

TABLE 8.3. MM iterates for a signomial program

m	x_{m1}	x_{m2}	$f(\boldsymbol{x}_m)$
0	1.00000	2.00000	2.33579
1	1.19973	2.04970	2.06522
5	1.39153	1.73281	1.94756
10	1.48648	1.61186	1.92935
15	1.52318	1.57173	1.92702
20	1.53763	1.55678	1.92668
25	1.54336	1.55097	1.92663
30	1.54565	1.54867	1.92662
35	1.54656	1.54776	1.92662
40	1.54692	1.54740	1.92662
45	1.54706	1.54725	1.92662
50	1.54712	1.54720	1.92662
55	1.54714	1.54717	1.92662
60	1.54715	1.54716	1.92662
65	1.54716	1.54716	1.92662

8.9 Poisson Processes

In preparation for our exposition of transmission tomography in the next section, let us briefly review the theory of Poisson processes, a topic from probability of considerable interest in its own right. A Poisson process involves points randomly scattered in a region S of R^n [113, 133, 148, 154]. The notion that the points are concentrated on average more in some regions than in others is captured by postulating an intensity function $\lambda(\boldsymbol{x}) \geq 0$ on S. The expected number of points in a subregion T is given by the integral $\omega = \int_T \lambda(\boldsymbol{x}) d\boldsymbol{x}$. If $\omega = \infty$, then an infinite number of random points occur in T. If $\omega < \infty$, then a finite number of random points occur in T, and the probability that this number equals k is given by the Poisson probability

$$p_k(\omega) = \frac{\omega^k}{k!} e^{-\omega}.$$

Derivation of this formula depends critically on the assumption that the numbers N_{T_i} of random points in disjoint regions T_i are independent

random variables. This basically means that knowing the values of some of the N_{T_i} tells one nothing about the values of the remaining N_{T_i}. The model also presupposes that random points never coincide.

The Poisson distribution has a peculiar relationship to the multinomial distribution. Suppose a Poisson random variable Z with mean ω represents the number of outcomes from some experiment, say an experiment involving a Poisson process. Let each outcome be independently classified in one of l categories, the kth of which occurs with probability p_k. Then the number of outcomes Z_k falling in category k is Poisson distributed with mean $\omega_k = p_k\omega$. Furthermore, the random variables $Z_1, \ldots, Z_l$ are independent. Conversely, if $Z = \sum_{k=1}^{l} Z_k$ is a sum of independent Poisson random variables Z_k with means $\omega_k = p_k\omega$, then conditional on $Z = n$, the vector $(Z_1, \ldots, Z_l)^*$ follows a multinomial distribution with n trials and cell probabilities $p_1, \ldots, p_l$. To prove the first two of these assertions, let $n = n_1 + \cdots + n_l$. Then

$$\begin{aligned}
\Pr(Z_1 = n_1, \ldots, Z_l = n_l) &= \frac{\omega^n}{n!} e^{-\omega} \binom{n}{n_1, \ldots, n_l} \prod_{k=1}^{l} p_k^{n_k} \\
&= \prod_{k=1}^{l} \frac{\omega_k^{n_k}}{n_k!} e^{-\omega_k} \\
&= \prod_{k=1}^{l} \Pr(Z_k = n_k).
\end{aligned}$$

To prove the converse, divide the last string of equalities by the probability $\Pr(Z = n) = \omega^n e^{-\omega}/n!$.

The random process of assigning points to categories is termed coloring in the stochastic process literature. When there are just two colors, and only random points of one of the colors are tracked, then the process is termed random thinning. We will see examples of both coloring and thinning in the next section.

8.10 Transmission Tomography

Problems in medical imaging often involve thousands of parameters. As an illustration of the MM algorithm, we treat maximum likelihood estimation in transmission tomography. Traditionally, transmission tomography images have been reconstructed by the methods of Fourier analysis. Fourier methods are fast but do not take into account the uncertainties of photon counts. Statistically based methods give better reconstructions with less patient exposure to harmful radiation.

The purpose of transmission tomography is to reconstruct the local attenuation properties of the object being imaged [124]. Attenuation is to

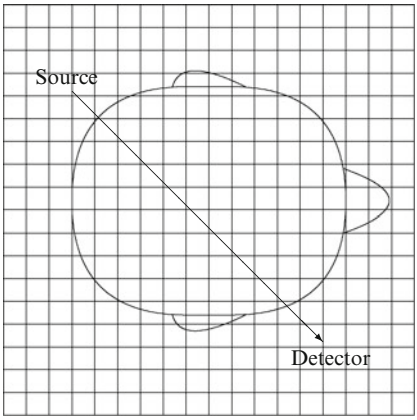

FIGURE 8.2. Cartoon of transmission tomography

be roughly equated with density. In medical applications, material such as bone is dense and stops or deflects X-rays (high-energy photons) better than soft tissue. With enough photons, even small gradations in soft tissue can be detected. A two-dimensional image is constructed from a sequence of photon counts. Each count corresponds to a projection line L drawn from an X-ray source through the imaged object to an X-ray detector. The average number of photons sent from the source along L to the detector is known in advance. The random number of photons actually detected is determined by the probability of a single photon escaping deflection or capture along L. Figure 8.2 shows one such projection line beamed through a cartoon of the human head.

To calculate this probability, we let $\mu(\boldsymbol{x})$ be the intensity (or attenuation coefficient) of photon deflection or capture per unit length at the point $\boldsymbol{x} = (x_1, x_2)$ in the plane. We can imagine that deflection or capture events occur completely randomly along L according to a Poisson process. The first such event effectively prevents the photon from being detected. Thus, the photon is detected with the Poisson probability $p_0(\omega) = e^{-\omega}$ of no such events, where

$$\omega \quad = \quad \int_L \mu(\boldsymbol{x}) ds$$

is the line integral of $\mu(\boldsymbol{x})$ along L. In actual practice, X-rays are beamed through the object along a large number of different projection lines. We therefore face the inverse problem of reconstructing a function $\mu(\boldsymbol{x})$ in the plane from a large number of its measured line integrals. Imposing enough smoothness on $\mu(\boldsymbol{x})$, one can solve this classical deterministic problem by applying Radon transform techniques from Fourier analysis [124].

An alternative to the Fourier method is to pose an explicitly stochastic model and estimate its parameters by maximum likelihood [167, 168].

The MM algorithm suggests itself in this context. The stochastic model depends on dividing the object of interest into small nonoverlapping regions of constant attenuation called pixels. Typically the pixels are squares on a regular grid as depicted in Fig. 8.2. The attenuation attributed to pixel j constitutes parameter θ_j of the model. Since there may be thousands of pixels, implementation of maximum likelihood algorithms such as scoring or Newton's method as discussed in Chap. 10 is out of the question.

To summarize our discussion, each observation Y_i is generated by beaming a stream of X-rays or high-energy photons from an X-ray source toward some detector on the opposite side of the object. The observation (or projection) Y_i counts the number of photons detected along the ith line of flight. Naturally, only a fraction of the photons are successfully transmitted from source to detector. If l_{ij} is the length of the segment of projection line i intersecting pixel j, then the probability of a photon escaping attenuation along projection line i is the exponentiated line integral $\exp(-\sum_j l_{ij}\theta_j)$.

In the absence of the intervening object, the number of photons generated and ultimately detected follows a Poisson distribution. We assume that the mean d_i of this distribution for projection line i is known. Ideally, detectors are long tubes aimed at the source. If a photon is deflected, then it is detected neither by the tube toward which it is initially headed nor by any other tube. In practice, many different detectors collect photons simultaneously from a single source. If we imagine coloring the tubes, then each photon is colored by the tube toward which it is directed. Each stream of colored photons is then thinned by capture or deflection. These considerations imply that the counts Y_i are independent and Poisson distributed with means $d_i \exp(-\sum_j l_{ij}\theta_j)$. It follows that we can express the loglikelihood of the observed data $Y_i = y_i$ as the finite sum

$$\sum_i \Big[-d_i e^{-\sum_j l_{ij}\theta_j} - y_i \sum_j l_{ij}\theta_j + y_i \ln d_i - \ln y_i! \Big]. \tag{8.14}$$

Omitting irrelevant constants, we can rewrite the loglikelihood (8.14) more succinctly as

$$L(\boldsymbol{\theta}) \;=\; -\sum_i f_i(\boldsymbol{l}_i^* \boldsymbol{\theta}),$$

where $f_i(t)$ is the convex function $d_i e^{-t} + y_i t$ and $\boldsymbol{l}_i^* \boldsymbol{\theta} = \sum_j l_{ij}\theta_j$ is the inner product of the attenuation parameter vector $\boldsymbol{\theta}$ and the vector of intersection lengths $\boldsymbol{l}_i$ for projection i.

To generate a surrogate function, we majorize each $f_i(\boldsymbol{l}_i^* \boldsymbol{\theta})$ according to the recipe (8.5). This gives the surrogate function

$$g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m) \;=\; -\sum_i \sum_j \frac{l_{ij}\theta_{mj}}{\boldsymbol{l}_i^* \boldsymbol{\theta}_m} f_i\Big(\frac{\boldsymbol{l}_i^* \boldsymbol{\theta}_m}{\theta_{mj}} \theta_j\Big) \tag{8.15}$$

minorizing $L(\boldsymbol{\theta})$. By construction, maximization of $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$ separates into a sequence of one-dimensional problems, each of which can be solved approximately by one step of Newton's method. We will take up the details of this in Chap. 10.

The images produced by maximum likelihood estimation in transmission tomography look grainy. The cure is to enforce image smoothness by penalizing large differences between estimated attenuation parameters of neighboring pixels. Geman and McClure [103] recommend multiplying the likelihood of the data by a Gibbs prior $\pi(\boldsymbol{\theta})$. Equivalently we add the log prior

$$\ln \pi(\boldsymbol{\theta}) \quad = \quad -\gamma \sum_{\{j,k\} \epsilon N} w_{jk} \psi(\theta_j - \theta_k)$$

to the loglikelihood, where γ and the weights w_{jk} are positive constants, N is a set of unordered pairs $\{j, k\}$ defining a neighborhood system, and $\psi(r)$ is called a potential function. This function should be large whenever $|r|$ is large. Neighborhoods have limited extent. For instance, if the pixels are squares, we might define the weights by $w_{jk} = 1$ for orthogonal nearest neighbors sharing a side and $w_{jk} = 1/\sqrt{2}$ for diagonal nearest neighbors sharing only a corner. The constant γ scales the overall strength assigned to the prior. The sum $L(\boldsymbol{\theta}) + \ln \pi(\boldsymbol{\theta})$ is called the log posterior function; its maximum is the posterior mode.

Choice of the potential function $\psi(r)$ is the most crucial feature of the Gibbs prior. It is convenient to assume that $\psi(r)$ is even and strictly convex. Strict convexity leads to the strict concavity of the log posterior function $L(\boldsymbol{\theta}) + \ln \pi(\boldsymbol{\theta})$ and permits simple modification of the MM algorithm based on the surrogate function $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$ defined by equation (8.15). Many potential functions exist satisfying these conditions. One natural example is $\psi(r) = r^2$. This choice unfortunately tends to deter the formation of boundaries. The gentler alternatives $\psi(r) = \sqrt{r^2 + \epsilon}$ for a small positive ϵ and $\psi(r) = \ln[\cosh(r)]$ are preferred in practice [111]. Problem 35 asks the reader to verify some of the properties of these two potential functions.

One adverse consequence of introducing a prior is that it couples pairs of parameters in the maximization step of the MM algorithm for finding the posterior mode. One can decouple the parameters by exploiting the convexity and evenness of the potential function $\psi(r)$ through the inequality

$$\begin{aligned} \psi(\theta_j - \theta_k) &= \psi\Big(\frac{1}{2}\Big[2\theta_j - \theta_{mj} - \theta_{mk}\Big] + \frac{1}{2}\Big[-2\theta_k + \theta_{mj} + \theta_{mk}\Big]\Big) \\ &\leq \frac{1}{2}\psi(2\theta_j - \theta_{mj} - \theta_{mk}) + \frac{1}{2}\psi(2\theta_k - \theta_{mj} - \theta_{mk}), \end{aligned}$$

which is strict unless $\theta_j + \theta_k = \theta_{mj} + \theta_{mk}$. This inequality allows us to redefine the surrogate function as

$$g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$$

$$\begin{aligned} &= -\sum_i \sum_j \frac{l_{ij}\theta_{mj}}{\boldsymbol{l}_i^*\boldsymbol{\theta}_m} f_i\Big(\frac{\boldsymbol{l}_i^*\boldsymbol{\theta}_m}{\theta_{mj}}\theta_j\Big) \\ &\quad -\frac{\gamma}{2}\sum_{\{j,k\}\epsilon N} w_{jk}[\psi(2\theta_j-\theta_{mj}-\theta_{mk})+\psi(2\theta_k-\theta_{mj}-\theta_{mk})]. \end{aligned}$$

Once again the parameters are separated, and the maximization step reduces to a sequence of one-dimensional problems. Maximizing $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$ drives the log posterior uphill and eventually leads to the posterior mode.

8.11 Poisson Multigraphs

In a graph the number of edges between any two nodes is 0 or 1. A multigraph allows an arbitrary number of edges between any two nodes. Multigraphs are natural structures for modeling the internet and gene and protein networks. Here we consider a multigraph with a random number of edges X_{ij} connecting every pair of nodes $\{i,j\}$. In particular, we assume that the X_{ij} are independent Poisson random variables with means μ_{ij}. As a plausible model for ranking nodes, we take $\mu_{ij} = p_i p_j$, where p_i and p_j are nonnegative propensities [217].

The loglikelihood of the observed edge counts $x_{ij} = x_{ji}$ amounts to

$$\begin{aligned} L(\boldsymbol{p}) &= \sum_{\{i,j\}} (x_{ij}\ln\mu_{ij} - \mu_{ij} - \ln x_{ij}!) \\ &= \sum_{\{i,j\}} [x_{ij}(\ln p_i + \ln p_j) - p_i p_j - \ln x_{ij}!]. \end{aligned}$$

Inspection of $L(\boldsymbol{p})$ shows that the parameters are separated except for the products $p_i p_j$. To achieve full separation of parameters in maximum likelihood estimation, we employ the majorization

$$p_i p_j \leq \frac{p_{mj}}{2p_{mi}}p_i^2 + \frac{p_{mi}}{2p_{mj}}p_j^2.$$

Equality prevails here when $\boldsymbol{p} = \boldsymbol{p}_m$. This majorization leads to the minorization

$$\begin{aligned} L(\boldsymbol{p}) &\geq \sum_{\{i,j\}} [x_{ij}(\ln p_i + \ln p_j) - \frac{p_{mj}}{2p_{mi}}p_i^2 - \frac{p_{mi}}{2p_{mj}}p_j^2 - \ln x_{ij}!] \\ &= g(\boldsymbol{p} \mid \boldsymbol{p}_m). \end{aligned}$$

Maximization of $g(\boldsymbol{p} \mid \boldsymbol{p}_m)$ can be accomplished by setting

$$\frac{\partial}{\partial p_i} g(\boldsymbol{p} \mid \boldsymbol{p}_m) = \sum_{j\neq i}\frac{x_{ij}}{p_i} - \sum_{j\neq i}\frac{p_{mj}}{p_{mi}} p_i = 0$$

The solution

$$p_{m+1,i} = \sqrt{\frac{p_{mi}\sum_{j\neq i} x_{ij}}{\sum_{j\neq i} p_{mj}}} \tag{8.16}$$

is straightforward to implement and maps positive parameters to positive parameters. When edges are sparse, the range of summation in $\sum_{j\neq i} x_{ij}$ can be limited to those nodes j with $x_{ij} > 0$. Observe that these sums need only be computed once. The partial sums $\sum_{j\neq i} p_{mj} = \sum_j p_{mj} - p_{mi}$ require updating the full sum $\sum_j p_{mj}$ once per iteration.

A similar MM algorithm can be derived for a Poisson model of arc formation in a directed multigraph. We now postulate a donor propensity p_i and a recipient propensity q_j for arcs extending from node i to node j. If the number of such arcs X_{ij} is Poisson distributed with mean $p_i q_j$, then under independence we have the loglikelihood

$$L(\boldsymbol{p},\boldsymbol{q}) = \sum_i \sum_{j\neq i} [x_{ij}(\ln p_i + \ln q_j) - p_i q_j - \ln x_{ij}!]$$

With directed arcs the observed numbers x_{ij} and x_{ji} may differ. The minorization

$$L(\boldsymbol{p},\boldsymbol{q}) \geq \sum_i \sum_{j\neq i} [x_{ij}(\ln p_i + \ln q_j) - \frac{q_{mj}}{2p_{mi}} p_i^2 - \frac{p_{mi}}{2q_{mj}} q_j^2 - \ln x_{ij}!]$$

now yields the MM updates

$$p_{m+1,i} = \sqrt{\frac{p_{mi}\sum_{j\neq i} x_{ij}}{\sum_{j\neq i} q_{mj}}}, \quad q_{m+1,j} = \sqrt{\frac{q_{mj}\sum_{i\neq j} x_{ij}}{\sum_{i\neq j} p_{mi}}}. \tag{8.17}$$

Again these are computationally simple to implement and map positive parameters to positive parameters. It is important to observe that the loglikelihood $L(\boldsymbol{p},\boldsymbol{q})$ is invariant under the rescaling cp_i and $c^{-1}q_j$ for some positive constant c and all i and j. This fact suggests that we fix one propensity, say $p_1 = 1$, and omit its update.

There are interesting examples of propensity estimation in literary analysis and attribution. Nodes are words in a text. An arc is drawn between two consecutive words, from the first word to the second word, provided the words are not separated by a punctuation mark. Here we examine Charlotte Bronte's novel *Jane Eyre*. Based on the directed multigraph model, it is possible to calculate the donor and recipient propensities of each word. Given these propensities, one can assign a p-value under the Poisson distribution indicating whether two words have an excess number of connections. Table 8.4 lists the most significant ordered word pairs in *Jane Eyre*.

TABLE 8.4. Most significantly connected word pairs in *Jane Eyre*

Rank	−Log p-value	Observed	Expected	Pair
1	−304.77	323	14.30	I am
2	−293.43	420	33.82	It was
3	−271.98	510	62.60	I had
4	−258.13	306	17.28	It is
5	−256.50	251	9.24	You are
6	−239.92	811	192.31	Of the
7	−233.92	155	1.82	Do not
8	−219.41	609	119.73	In the
9	−208.02	173	4.17	Could not
10	−196.42	320	32.02	To be
11	−191.60	154	3.37	Had been
12	−179.88	259	21.18	I could
13	−168.41	138	3.20	Did not
14	−162.75	337	47.45	On the
15	−153.04	317	44.68	I have
16	−152.62	480	106.93	I was
17	−132.87	108	2.46	Have been
18	−118.87	162	12.10	I should
19	−117.57	112	3.91	As if
20	−115.97	130	6.61	There was
21	−114.57	110	3.92	Would be
22	−113.61	123	5.80	Do you
23	−106.61	171	16.85	You have
24	−104.87	62	0.49	At least
25	−103.23	132	8.78	A little

8.12 Problems

1. Prove that the majorization relation between functions is closed under the formation of sums, nonnegative products, limits, and composition with an increasing function. In what sense is the relation also transitive?

2. Demonstrate the majorizing and minorizing inequalities

$$\begin{aligned} x^q &\leq q x_m^{q-1} x + (1-q) x_m^q \\ \ln x &\leq \frac{x}{x_m} + \ln x_m - 1 \end{aligned}$$

$$\begin{aligned}
x \ln x &\leq \frac{x^2}{x_m} + x \ln x_m - x \\
\|\boldsymbol{x}\| &\geq \frac{\boldsymbol{x}_m^* \boldsymbol{x}}{\|\boldsymbol{x}_m\|} \\
xy &\leq \frac{y_m}{2x_m} x^2 + \frac{x_m}{2y_m} y^2 \\
-xy &\leq -x_m y_m \left[1 + \ln\left(\frac{x}{x_m}\right) + \ln\left(\frac{y}{y_m}\right)\right] \\
\frac{1}{x} &\leq \frac{1}{x_m} - \frac{x - x_m}{x_m^2} + \frac{(x - x_m)^2}{c^3} \\
\frac{1}{x+y} &\leq \left(\frac{x_m}{x_m + y_m}\right)^2 \frac{1}{x} + \left(\frac{y_m}{x_m + y_m}\right)^2 \frac{1}{y}
\end{aligned}$$

Determine the relevant domains of each variable q, x, x_m, y, y_m, and c, and check that equality occurs in each of the inequalities when $x = x_m$ and $y = y_m$ [122].

3. As alternatives to the fifth and sixth examples of Problem 2, demonstrate the majorizations

$$\begin{aligned}
xy &\leq \frac{1}{2}(x^2 + y^2) + \frac{1}{2}(x_m - y_m)^2 - (x_m - y_m)(x - y) \\
-xy &\leq \frac{1}{2}(x^2 + y^2) + \frac{1}{2}(x_m + y_m)^2 - (x_m + y_m)(x + y)
\end{aligned}$$

valid for all values of x, y, x_m, and y_m.

4. Based on Problem 3, devise an MM algorithm to minimize Rosenbrock's function

$$f(\boldsymbol{x}) = 100(x_1^2 - x_2)^2 + (x_1 - 1)^2.$$

Show that up to an irrelevant constant $f(\boldsymbol{x})$ is majorized by the sum of the two functions

$$\begin{aligned}
g_1(x_1 \mid \boldsymbol{x}_m) &= 200x_1^4 - [200(x_{m1}^2 + x_{m2}) - 1]x_1^2 - 2x_1 \\
g_2(x_2 \mid \boldsymbol{x}_m) &= 200x_2^2 - 200(x_{m1}^2 + x_{m2})x_2.
\end{aligned}$$

Hence at each iteration one must minimize a quartic in x_1 and a quadratic in x_2. Implement this MM algorithm, and check whether it converges to the global minimum of $f(\boldsymbol{x})$ at $\boldsymbol{x} = \mathbf{1}$.

5. Devise an MM algorithm to minimize Snell's objective function (1.1).

6. Prove van Ruitenburg's [263] minorization

$$\ln x \geq -\frac{3x_m}{2x} - \frac{x}{2x_m} + \ln x_m + 2\,.$$

Deduce the further minorization

$$x \ln x \quad \geq \quad -\frac{3x_m}{2} - \frac{x^2}{2x_m} + x \ln x_m + 2x \, .$$

7. Suppose $p \in [1, 2]$ and $x_m \neq 0$. Verify the majorizing inequality

$$|x|^p \quad \leq \quad \frac{p}{2}|x_m|^{p-2}x^2 + \left(1 - \frac{p}{2}\right)|x_m|^p.$$

This is important in least ℓ_p regression.

8. The majorization

$$|x| \quad \leq \quad \frac{1}{2|x_m|}\left(x^2 + x_m^2\right) \tag{8.18}$$

for $x_m \neq 0$ is a special case of Problem 7. Use the majorization (8.18) and the identity

$$\max\{x, y\} \quad = \quad \frac{1}{2}|x - y| + \frac{1}{2}x + \frac{1}{2}y$$

to majorize $\max\{x, y\}$ when $x_m \neq y_m$. Note that your majorization contains the product xy up to a negative factor. Describe how one can invoke Problem 2 or Problem 3 to separate x and y.

9. Prove the majorization (8.9) of the text.

10. Consider the function

$$f(x) \quad = \quad \frac{1}{4}x^4 - \frac{1}{2}x^2.$$

This function has global minima at $x = \pm 1$ and a local maximum at $x = 0$. Show that the function

$$g(x \mid x_m) \quad = \quad \frac{1}{4}x^4 + \frac{1}{2}x_m^2 - xx_m$$

majorizes $f(x)$ at x_m and leads to the MM update $x_{m+1} = \sqrt[3]{x_m}$. Prove that the alternative update $x_{m+1} = -\sqrt[3]{x_m}$ leads to the same value of $f(x)$, but the first update always converges while the second oscillates in sign and has two converging subsequences [60].

11. In the regression algorithm (8.10), let p tend to 0. If there are q predictors and all x_{ij} are nonzero, then show that $\alpha_{ij} = 1/q$. This leads to the update

$$\theta_{m+1,j} \quad = \quad \theta_{mj} + \frac{\sum_{i=1}^n x_{ij}(y_i - \boldsymbol{x}_i^* \boldsymbol{\theta}_m)}{q \sum_{i=1}^n x_{ij}^2}. \tag{8.19}$$

On the other hand, argue that cyclic coordinate descent yields the update

$$\theta_{m+1,j} \;=\; \theta_{mj} + \frac{\sum_{i=1}^n x_{ij}(y_i - \boldsymbol{x}_i^* \boldsymbol{\theta}_m)}{\sum_{i=1}^n x_{ij}^2},$$

which definitely takes larger steps.

12. A number μ is said to be a q quantile of the n numbers $x_1, \ldots, x_n$ if it satisfies

$$\frac{1}{n} \sum_{x_i \leq \mu_q} 1 \;\geq\; q \quad \text{and} \quad \frac{1}{n} \sum_{x_i \geq \mu_q} 1 \;\geq\; 1 - q.$$

If we define

$$\rho_q(r) \;=\; \begin{cases} qr & r \geq 0 \\ -(1-q)r & r < 0 \end{cases},$$

then demonstrate that μ is a q quantile if and only if μ minimizes the function $f_q(\mu) = \sum_{i=1}^n \rho_q(x_i - \mu)$. Medians correspond to the case $q = 1/2$.

13. Continuing Problem 12, show that the function $\rho_q(r)$ is majorized by the quadratic

$$\zeta_q(r \mid r_m) \;=\; \frac{1}{4}\left[\frac{r^2}{|r_m|} + (4q-2)r + |r_m|\right].$$

Deduce from this majorization the MM algorithm

$$\begin{aligned} \mu_{m+1} &= \frac{n(2q-1) + \sum_{i=1}^n w_{mi} x_i}{\sum_{i=1}^n w_{mi}} \\ w_{mi} &= \frac{1}{|x_i - \mu_m|} \end{aligned}$$

for finding a q quantile. This interesting algorithm involves no sorting, only arithmetic operations.

14. Suppose we minimize the function

$$h_\epsilon(\boldsymbol{\theta}) \;=\; \sum_{i=1}^p \left\{ \left[y_i - \sum_{j=1}^q x_{ij}\theta_j \right]^2 + \epsilon \right\}^{1/2} \tag{8.20}$$

instead of the function $h(\boldsymbol{\theta})$ in equation (8.11) for a small, positive number ϵ. Show that the same MM algorithm applies with revised weights $w_i(\boldsymbol{\theta}_m) = 1/\sqrt{r_i(\boldsymbol{\theta}_m)^2 + \epsilon}$.

15. At the point $\boldsymbol{y}$ suppose the affine function $\boldsymbol{v}_j^*\boldsymbol{x} + a_j$ is the only contribution to the convex function

$$f(\boldsymbol{x}) = \max_{1 \le i \le n} (\boldsymbol{v}_i^*\boldsymbol{x} + a_i).$$

achieving the maximum. Show that the quadratic function

$$\begin{aligned} g(\boldsymbol{x} \mid \boldsymbol{y}) &= \frac{c}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 + \boldsymbol{v}_j^*\boldsymbol{x} + a_j \\ &= \frac{c}{2}\left\|\boldsymbol{x} - \boldsymbol{y} + \frac{1}{c}\boldsymbol{v}_j\right\|^2 + \boldsymbol{v}_j^*\boldsymbol{y} - \frac{1}{2c}\|\boldsymbol{v}_j\|^2 + a_j \end{aligned}$$

majorizes $f(\boldsymbol{x})$ with anchor $\boldsymbol{y}$ for $c > 0$ sufficiently large. Argue that the best choice of $g(\boldsymbol{x} \mid \boldsymbol{y})$ takes $c = \max_{i \neq j} c_i$, where

$$c_i = \frac{\|\boldsymbol{v}_i - \boldsymbol{v}_j\|^2}{2[a_j - a_i + (\boldsymbol{v}_j - \boldsymbol{v}_i)^*\boldsymbol{y}]},$$

provided none of the denominators vanish. (Hint: Investigate the tangency conditions between $g(\boldsymbol{x} \mid \boldsymbol{y})$ and $\boldsymbol{v}_i^*\boldsymbol{x} + a_i$.)

16. Based on Problem 15, consider majorization of the ℓ_∞ norm $\|\boldsymbol{x}\|_\infty$ around $\boldsymbol{y} \neq \boldsymbol{0}$. If j is the sole index with $|y_j| = \|\boldsymbol{y}\|_\infty$, then prove that

$$\begin{aligned} g(\boldsymbol{x} \mid \boldsymbol{y}) &= \frac{c}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 + \operatorname{sgn}(y_j)x_j \\ c &= \frac{1}{|y_j| - \max_{i \neq j} |y_i|} \end{aligned}$$

majorizes $\|\boldsymbol{x}\|_\infty$ around $\boldsymbol{y}$.

17. Suppose the differentiable function $f(\boldsymbol{x})$ satisfies

$$\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| \le L\|\boldsymbol{y} - \boldsymbol{x}\|$$

for all $\boldsymbol{y}$ and $\boldsymbol{x}$. Deduce the majorization

$$f(\boldsymbol{x}) \le f(\boldsymbol{x}_m) + df(\boldsymbol{x}_m)(\boldsymbol{x} - \boldsymbol{x}_m) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{x}_m\|^2$$

from the expansion $f(\boldsymbol{x}) = f(\boldsymbol{x}_m) + \int_0^1 df\,[\boldsymbol{x}_m + t(\boldsymbol{x} - \boldsymbol{x}_m)](\boldsymbol{x} - \boldsymbol{x}_m)dt$. Verify the special case

$$\begin{aligned} \|M\boldsymbol{x} - \boldsymbol{y}\|^2 &\le \|M\boldsymbol{x}_m - \boldsymbol{y}\|^2 + \|M\|^2\left[\|\boldsymbol{x} - \boldsymbol{x}_m + \boldsymbol{z}_m\|^2 - \|\boldsymbol{z}_m\|^2\right] \\ \boldsymbol{z}_m &= \|M\|^{-2}M^*(M\boldsymbol{x}_m - \boldsymbol{y}) \end{aligned}$$

for the Euclidean and spectral norms.

18. Validate the Euclidean norm majorization

$$\|\boldsymbol{y}-\boldsymbol{x}\|_\dagger \quad \leq \quad \|\boldsymbol{y}-\boldsymbol{x}_m\|_\dagger + c\|\boldsymbol{x}_m-\boldsymbol{x}\|.$$

Here $\|\cdot\|_\dagger$ an arbitrary norm on R^n, and c is the positive constant guaranteed by Example 2.5.6.

19. Let $\boldsymbol{P}$ be an orthogonal projection onto a subspace of R^n. Demonstrate the majorization

$$\|\boldsymbol{P}\boldsymbol{x}\|^2 \quad \leq \quad \|\boldsymbol{x}-\boldsymbol{x}_m\|^2 + 2(\boldsymbol{x}-\boldsymbol{x}_m)^*\boldsymbol{P}\boldsymbol{x}_m + \|\boldsymbol{P}\boldsymbol{x}_m\|^2.$$

20. Consider minimizing the function

$$f(\boldsymbol{\mu}) \quad = \quad \sum_{i=1}^p \|\boldsymbol{x}_i-\boldsymbol{\mu}\|$$

for p points $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p$ in R^n. Demonstrate that

$$g(\boldsymbol{\mu}\mid\boldsymbol{\mu}_m) \quad = \quad \frac{1}{2}\sum_{i=1}^p \frac{\|\boldsymbol{x}_i-\boldsymbol{\mu}\|^2}{\|\boldsymbol{x}_i-\boldsymbol{\mu}_m\|}$$

majorizes $f(\boldsymbol{\mu})$ at the current iterate $\boldsymbol{\mu}_m$ up to an irrelevant constant. Deduce Weiszfeld's algorithm [271]

$$\boldsymbol{\mu}_{m+1} \quad = \quad \frac{1}{\sum_{i=1}^p w_{mi}} \sum_{i=1}^p w_{mi}\boldsymbol{x}_i$$

with weights $w_{mi} = 1/\|\boldsymbol{x}_i-\boldsymbol{\mu}_m\|$. Comment on the relevance of this algorithm to a robust analogue of k-means clustering.

21. Show that the function $f(\boldsymbol{\mu})$ of Problem 20 is strictly convex whenever the points $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p$ are not collinear.

22. Explain how Problem 31 of Chap. 4 delivers a quadratic majorization. Describe circumstances under which this majorization is apt to be poor.

23. In ℓ_1 regression show that the estimate of the parameter vector satisfies the equality

$$\sum_{i=1}^m \operatorname{sgn}[y_i-\mu_i(\boldsymbol{\theta})]\nabla\mu_i(\boldsymbol{\theta}) \quad = \quad \mathbf{0},$$

provided no residual $y_i-\mu_i(\boldsymbol{\theta})=0$ and the regression functions $\mu_i(\boldsymbol{\theta})$ are differentiable. What is the corresponding equality for the modified ℓ_1 criterion (8.20)?

24. Problem 12 of Chap. 7 deals with minimizing the quadratic function $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}^*\boldsymbol{x}+c$ subject to the constraints $x_i \geq 0$. If one drops the assumption that $\boldsymbol{A} = (a_{ij})$ is positive definite, it is still possible to devise an MM algorithm. Define matrices $\boldsymbol{A}^+$ and $\boldsymbol{A}^-$ with entries $\max\{a_{ij}, 0\}$ and $-\min\{a_{ij}, 0\}$, respectively. Based on the fifth and sixth majorizations of Problem 2, derive the MM updates

$$x_{m+1,i} \;=\; x_{m,i}\left[\frac{-b_i + \sqrt{b_i^2 + 4(\boldsymbol{A}^+\boldsymbol{x}_m)_i(\boldsymbol{A}^-\boldsymbol{x}_m)_i}}{2(\boldsymbol{A}^+\boldsymbol{x}_m)_i}\right]$$

of Sha et al. [235]. All entries of the initial point $\boldsymbol{x}_0$ should be positive.

25. Show that the Bradley-Terry loglikelihood $f(\boldsymbol{r}) = \ln L(\boldsymbol{r})$ of Sect. 8.6 is concave under the reparameterization $r_i = e^{\theta_i}$.

26. In the Bradley-Terry model of Sect. 8.6, suppose we want to include the possibility of ties. One way of doing this is to write the probabilities of the three outcomes of i versus j as

$$\begin{aligned}
\Pr(\text{i wins}) &= \frac{r_i}{r_i + r_j + \theta\sqrt{r_i r_j}} \\
\Pr(\text{i ties}) &= \frac{\theta\sqrt{r_i r_j}}{r_i + r_j + \theta\sqrt{r_i r_j}} \\
\Pr(\text{i loses}) &= \frac{r_j}{r_i + r_j + \theta\sqrt{r_i r_j}},
\end{aligned}$$

where $\theta > 0$ is an additional parameter to be estimated. Let y_{ij} represent the number of times i beats j and t_{ij} the number of times i ties j. Prove that the loglikelihood of the data is

$$\begin{aligned}
&L(\theta, \boldsymbol{r}) \\
&= \frac{1}{2}\sum_{i,j}\left(2y_{ij}\ln\frac{r_i}{r_i + r_j + \theta\sqrt{r_i r_j}} + t_{ij}\ln\frac{\theta\sqrt{r_i r_j}}{r_i + r_j + \theta\sqrt{r_i r_j}}\right).
\end{aligned}$$

One way of maximizing $L(\theta, \boldsymbol{r})$ is to alternate between updating θ and $\boldsymbol{r}$. Both of these updates can be derived from the perspective of the MM algorithm. Two minorizations are now involved. The first proceeds using the convexity of $-\ln t$ just as in the text. This produces a function involving $-\sqrt{r_i r_j}$ terms. Demonstrate the minorization

$$-\sqrt{r_i r_j} \;\geq\; -\frac{r_i}{2}\sqrt{\frac{r_{mj}}{r_{mi}}} - \frac{r_j}{2}\sqrt{\frac{r_{mi}}{r_{mj}}},$$

and use it to minorize $L(\theta, \boldsymbol{r})$. Finally, determine $\boldsymbol{r}_{m+1}$ for θ fixed at θ_m and θ_{m+1} for $\boldsymbol{r}$ fixed at $\boldsymbol{r}_{m+1}$. The details are messy, but the overall strategy is straightforward.

27. In the linear logistic model of Sect. 8.7, it is possible to separate parameters and avoid matrix inversion altogether. In constructing a minorizing function, first prove the inequality

$$\begin{aligned}\ln[1-\pi(\boldsymbol{\theta})] &= -\ln\left(1+e^{\boldsymbol{x}_i^*\boldsymbol{\theta}}\right)\\ &\geq -\ln\left(1+e^{\boldsymbol{x}_i^*\boldsymbol{\theta}_m}\right)-\frac{e^{\boldsymbol{x}_i^*\boldsymbol{\theta}}-e^{\boldsymbol{x}_i^*\boldsymbol{\theta}_m}}{1+e^{\boldsymbol{x}_i^*\boldsymbol{\theta}_m}},\end{aligned}$$

with equality when $\boldsymbol{\theta}=\boldsymbol{\theta}_m$. This eliminates the log terms. Now apply the arithmetic-geometric mean inequality to the exponential functions $e^{\boldsymbol{x}_i^*\boldsymbol{\theta}}$ to separate parameters. Assuming that $\boldsymbol{\theta}$ has n components and that there are k observations, show that these maneuvers lead to the minorizing function

$$g(\boldsymbol{\theta}\mid\boldsymbol{\theta}_m) = -\frac{1}{n}\sum_{i=1}^k\frac{e^{\boldsymbol{x}_i^*\boldsymbol{\theta}_m}}{1+e^{\boldsymbol{x}_i^*\boldsymbol{\theta}_m}}\sum_{j=1}^n e^{nx_{ij}(\theta_j-\theta_{mj})}+\sum_{i=1}^k y_i\boldsymbol{x}_i^*\boldsymbol{\theta}$$

up to a constant that does not depend on $\boldsymbol{\theta}$. Finally, prove that maximizing $g(\boldsymbol{\theta}\mid\boldsymbol{\theta}_m)$ consists in solving the transcendental equation

$$-\sum_{i=1}^k\frac{e^{\boldsymbol{x}_i^*\boldsymbol{\theta}_m}x_{ij}e^{-nx_{ij}\theta_{mj}}}{1+e^{\boldsymbol{x}_i^*\boldsymbol{\theta}_m}}e^{nx_{ij}\theta_j}+\sum_{i=1}^k y_ix_{ij} = 0$$

for each j. This can be accomplished numerically.

28. Consider the general posynomial of n variables

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}\in S}c_{\boldsymbol{\alpha}}\prod_{i=1}^n x_i^{\alpha_i}$$

subject to the constraints $x_i>0$ for each i. Here the index set $S\subset\mathsf{R}^n$ is finite, and the coefficients $c_{\boldsymbol{\alpha}}$ are positive. We can assume that at least one $\alpha_i>0$ and at least one $\alpha_i<0$ for every i. Otherwise, $f(\boldsymbol{x})$ can be reduced by sending x_i to ∞ or 0. Demonstrate that $f(\boldsymbol{x})$ is majorized by the sum

$$\begin{aligned}g(\boldsymbol{x}\mid\boldsymbol{x}_m) &= \sum_{i=1}^n g_i(x_i\mid\boldsymbol{x}_m)\\ g_i(x_i\mid\boldsymbol{x}_m) &= \sum_{\boldsymbol{\alpha}\in S}c_{\boldsymbol{\alpha}}\left(\prod_{j=1}^n x_{mj}^{\alpha_j}\right)\frac{|\alpha_i|}{\|\boldsymbol{\alpha}\|_1}\left(\frac{x_i}{x_{mi}}\right)^{\|\boldsymbol{\alpha}\|_1\,\mathrm{sgn}(\alpha_i)},\end{aligned}$$

where $\|\boldsymbol{\alpha}\|_1=\sum_{j=1}^n|\alpha_j|$ and $\mathrm{sgn}(\alpha_i)$ is the sign function. To prove that the MM algorithm is well defined and produces iterates with positive entries, demonstrate that

$$\lim_{x_i\to\infty}g_i(x_i\mid\boldsymbol{x}_m) = \lim_{x_i\to 0}g_i(x_i\mid\boldsymbol{x}_m) = \infty.$$

Finally change variables by setting

$$\begin{aligned} y_i &= \ln x_i \\ h_i(y_i \mid \boldsymbol{x}_m) &= g_i(x_i \mid \boldsymbol{x}_m) \end{aligned}$$

for each i. Show that $h_i(y_i \mid \boldsymbol{x}_m)$ is strictly convex in y_i and therefore possesses a unique minimum point. The latter property carries over to the surrogate function $g_i(x_i \mid \boldsymbol{x}_m)$.

29. Devise MM algorithms based on Problem 28 to minimize the posynomials

$$\begin{aligned} f_1(\boldsymbol{x}) &= \frac{1}{x_1 x_2^2} + x_1 x_2^2 \\ f_2(\boldsymbol{x}) &= \frac{1}{x_1 x_2^2} + x_1 x_2. \end{aligned}$$

In the first case, demonstrate that the MM algorithm iterates according to

$$x_{m+1,1} = \sqrt[3]{\frac{x_{m1}^2}{x_{m2}^2}}, \quad x_{m+1,2} = \sqrt[3]{\frac{x_{m2}}{x_{m1}}}.$$

Furthermore, show that (a) $f_1(\boldsymbol{x})$ attains its minimum value of 2 whenever $x_1 x_2^2 = 1$, (b) the MM algorithm converges after a single iteration to the value 2, and (c) the converged point $\boldsymbol{x}_1$ depends on the initial point $\boldsymbol{x}_0$. In the second case, demonstrate that the MM algorithm iterates according to

$$x_{m+1,1} = \sqrt[5]{\frac{x_{m1}^3}{x_{m2}^3}}, \quad x_{m+1,2} = \sqrt[5]{2\frac{x_{m2}^2}{x_{m1}^2}}.$$

Furthermore, show that (a) the infimum of $f_2(\boldsymbol{x})$ is 0, (b) the MM algorithm satisfies the identities

$$x_{m1} x_{m2}^{3/2} = 2^{3/10}, \quad x_{m+1,2} = 2^{2/25} x_{m2}$$

for all $m \geq 2$, and (c) the minimum value 0 is attained asymptotically with x_{m1} tending to 0 and x_{m2} tending to ∞.

30. A general posynomial of n variables can be represented as

$$h(\boldsymbol{y}) = \sum_{\boldsymbol{\alpha} \in S} c_{\boldsymbol{\alpha}} e^{\boldsymbol{\alpha}^* \boldsymbol{y}}$$

in the parameterization $y_i = \ln x_i$. Here the index set $S \subset \mathsf{R}^n$ is finite and the coefficients $c_{\boldsymbol{\alpha}}$ are positive. Show that $h(\boldsymbol{y})$ is strictly convex if and only if the power vectors $\boldsymbol{\alpha} \in S$ span R^n.

31. Demonstrate the minorization

$$\prod_{i=1}^n x_i^{\alpha_i} \geq \prod_{j=1}^n x_{mj}^{\alpha_j} \left(1 + \sum_{i=1}^n \alpha_i \ln x_i - \sum_{i=1}^n \alpha_i \ln x_{mi}\right).$$

How is this helpful in signomial programming [170]? As a concrete example, devise an MM algorithm to minimize the function

$$f(\boldsymbol{x}) = x_1^2 x_2^2 - 2x_1 x_2 x_3 x_4 + x_3^2 x_4^2 = (x_1 x_2 - x_3 x_4)^2.$$

Describe the solution as you vary the initial point.

32. Even more functions can be can be brought under the umbrella of signomial programming. For instance, majorization of the functions $-\ln f(\boldsymbol{x})$ and $\ln f(\boldsymbol{x})$ is possible for any posynomial

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha} \in S} c_{\boldsymbol{\alpha}} \prod_{i=1}^n x_i^{\alpha_i}.$$

In the first case show that

$$-\ln f(\boldsymbol{x}) \leq -\sum_{\boldsymbol{\alpha} \in S} \frac{d_{m\boldsymbol{\alpha}}}{e_m} \Big[\sum_{i=1}^n \alpha_i \ln x_i + \ln\Big(\frac{c_{\boldsymbol{\alpha}} e_m}{d_{m\boldsymbol{\alpha}}}\Big)\Big] \quad (8.21)$$

holds for $d_{m\boldsymbol{\alpha}} = c_{\boldsymbol{\alpha}} \prod_{i=1}^n x_{mi}^{\alpha_i}$ and $e_m = \sum_{\boldsymbol{\alpha}} d_{m\boldsymbol{\alpha}}$. In the second case, show that

$$\ln f(\boldsymbol{x}) \leq \ln f(\boldsymbol{x}_m) + \frac{1}{f(\boldsymbol{x}_m)} \Big[f(\boldsymbol{x}) - f(\boldsymbol{x}_m)\Big].$$

This second majorization yields a posynomial, which can be majorized by the methods already described. Note that the coefficients $c_{\boldsymbol{\alpha}}$ can be negative as well as positive in the second case [170].

33. Show that the loglikelihood (8.14) for the transmission tomography model is concave. State a necessary condition for strict concavity in terms of the number of pixels and the number of projections.

34. In the maximization phase of the MM algorithm for transmission tomography without a smoothing prior, demonstrate that the exact solution of the one-dimensional equation

$$\frac{\partial}{\partial \theta_j} g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m) = 0$$

exists and is positive when $\sum_i l_{ij} d_i > \sum_i l_{ij} y_i$. Why would this condition typically hold in practice?

35. Prove that the functions $\psi(r) = \sqrt{r^2 + \epsilon}$ and $\psi(r) = \ln[\cosh(r)]$ are even, strictly convex, infinitely differentiable, and asymptotic to $|r|$ as $|r| \to \infty$.

36. In positron emission tomography (PET), one seeks to estimate an object's Poisson emission intensities $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^*$. Here p pixels are arranged in a 2-dimensional grid surrounded by an array of photon detectors [167, 265]. The observed data are coincidence counts $(y_1, \ldots y_d)$ along d lines of flight connecting pairs of photon detectors. The loglikelihood under the PET model is

$$L(\boldsymbol{\lambda}) \;=\; \sum_i \Big[y_i \ln \Big(\sum_j e_{ij}\lambda_j \Big) - \sum_j e_{ij}\lambda_j \Big],$$

where the constants e_{ij} are determined from the geometry of the grid and the detectors. One can assume without loss of generality that $\sum_i e_{ij} = 1$. Use Jensen's inequality to derive the minorization

$$L(\boldsymbol{\lambda}) \;\geq\; \sum_i y_i \sum_j w_{nij} \ln \Big(\frac{e_{ij}\lambda_j}{w_{nij}} \Big) - \sum_i \sum_j e_{ij}\lambda_j \;=\; Q(\boldsymbol{\lambda} \mid \boldsymbol{\lambda}_n),$$

where $w_{nij} = e_{ij}\lambda_{nj}/(\sum_k e_{ik}\lambda_{nk})$. Show that the stationarity conditions for the surrogate function $Q(\boldsymbol{\lambda} \mid \boldsymbol{\lambda}_n)$ entail the MM updates

$$\lambda_{n+1,j} \;=\; \frac{\sum_i y_i w_{nij}}{\sum_i e_{ij}}.$$

To smooth the image, one can maximize the penalized loglikelihood

$$f(\boldsymbol{\lambda}) \;=\; L(\boldsymbol{\lambda}) - \frac{\mu}{2} \sum_{\{j,k\} \in \mathcal{N}} (\lambda_j - \lambda_k)^2,$$

where μ is a tuning constant, and $\mathcal{N}$ is a neighborhood system that pairs spatially adjacent pixels. If we adopt the majorization

$$(\lambda_j - \lambda_k)^2 \;\leq\; \frac{1}{2}(2\lambda_j - \lambda_{nj} - \lambda_{nk})^2 + \frac{1}{2}(2\lambda_k - \lambda_{nj} - \lambda_{nk})^2,$$

then show that the revised stationarity conditions are

$$0 \;=\; \sum_i \Big[\frac{y_i w_{nij}}{\lambda_j} - e_{ij} \Big] - \mu \sum_{k:\in\mathcal{N}_j} (2\lambda_j - \lambda_{nj} - \lambda_{nk}).$$

From these equations derive the smoothed MM updates

$$\lambda_{n+1,j} \;=\; \frac{-b_{nj} - \sqrt{b_{nj}^2 - 4a_j c_{nj}}}{2a_j},$$

where

$$\begin{aligned} a_j &= -2\mu \sum_{k \in \mathcal{N}_j} 1 \\ b_{nj} &= \mu \sum_{k \in \mathcal{N}_j} (\lambda_{nj} + \lambda_{nk}) - 1 \\ c_{nj} &= \sum_i y_i w_{nij}. \end{aligned}$$

Note that $a_j < 0$, so we take the negative sign before the square root.

37. Program and test on real data either of the random multigraph algorithms (8.16) or (8.17). The internet is a rich source of random graph data.

38. The Dirichlet-multinomial distribution is used to model multivariate count data $\boldsymbol{x} = (x_1, \ldots, x_d)^*$ that is too over-dispersed to be handled reliably by the multinomial distribution. Recall that the Dirichlet-multinomial distribution represents a mixture of multinomial distribution with a Dirichlet prior on the cell probabilities $p_1, \ldots, p_d$. If $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^*$ denotes the Dirichlet parameter vector, Δ_d denotes the unit simplex in R^d, and $|\boldsymbol{\alpha}| = \sum_{i=1}^d \alpha_i$ and $|\boldsymbol{x}| = \sum_{i=1}^d x_i$, then the discrete density of the Dirichlet-multinomial is

$$\begin{aligned} f(\boldsymbol{x} \mid \boldsymbol{\alpha}) &= \int_{\Delta_d} \binom{|\boldsymbol{x}|}{\boldsymbol{x}} \prod_{j=1}^d p_j^{x_j} \frac{\Gamma(|\boldsymbol{\alpha}|)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_d)} \prod_{j=1}^d p_j^{\alpha_j - 1} dp_1 \cdots dp_d \\ &= \binom{|\boldsymbol{x}|}{\boldsymbol{x}} \frac{\Gamma(\alpha_1 + x_1) \cdots \Gamma(\alpha_d + x_d)}{\Gamma(|\boldsymbol{\alpha}| + |\boldsymbol{x}|)} \frac{\Gamma(|\boldsymbol{\alpha}|)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_d)} \\ &= \binom{|\boldsymbol{x}|}{\boldsymbol{x}} \frac{\prod_{j=1}^d \alpha_j(\alpha_j + 1) \cdots (\alpha_j + x_j - 1)}{|\boldsymbol{\alpha}|(|\boldsymbol{\alpha}| + 1) \cdots (|\boldsymbol{\alpha}| + |\boldsymbol{x}| - 1)}. \end{aligned}$$

Devise an MM algorithm to maximize $\ln f(\boldsymbol{x} \mid \boldsymbol{\alpha})$ by applying the supporting hyperplane inequality to each term $-\ln(|\boldsymbol{\alpha}| + k)$ and Jensen's inequality to each term $\ln(\alpha_j + k)$. These minorizations are designed to separate parameters. Show that the MM updates are

$$\alpha_{m+1,j} = \alpha_{mj} \frac{\sum_k \frac{s_{jk}}{\alpha_{mj} + k}}{\sum_k \frac{r_k}{|\alpha_m| + k}}$$

for appropriate constants r_k and s_{jk}. This problem and similar problems are treated in the reference [282].

39. In the dictionary model of motif finding [227], a DNA sequence is viewed as a concatenation of words independently drawn from a dictionary having the four letters A, C, G, and T. The words of the

dictionary of length k have collective probability q_k. The EM algorithm offers one method of estimating the q_k. Omitting many details, the EM algorithm maximizes the function

$$Q(\boldsymbol{q} \mid \boldsymbol{q}_m) \quad = \quad \sum_{k=1}^{l} c_{mk} \ln q_k - \ln\Big(\sum_{k=1}^{l} k q_k\Big).$$

Here the constants c_{mk} are positive, l is the maximum word length, and maximization is performed subject to the constraints $q_k \geq 0$ for $k = 1, \ldots, l$ and $\sum_{k=1}^{l} q_k = 1$. Because this problem can not be solved in closed form, it is convenient to follow the EM minorization with a second minorization based on the inequality

$$\ln x \quad \leq \quad \ln y + x/y - 1. \tag{8.22}$$

Application of inequality (8.22) produces the minorizing function

$$h(\boldsymbol{q} \mid \boldsymbol{q}_m) \quad = \quad \sum_{k=1}^{l} c_{mk} \ln q_k - \ln\Big(\sum_{k=1}^{l} k q_{mk}\Big) - d_m \sum_{k=1}^{l} k q_k + 1$$

with $d_m = 1/(\sum_{k=1}^{l} k q_{mk})$.

(a) Show that the function $h(\boldsymbol{q} \mid \boldsymbol{q}_m)$ minorizes $Q(\boldsymbol{q} \mid \boldsymbol{q}_m)$.

(b) Maximize $h(\boldsymbol{q} \mid \boldsymbol{q}_m)$ using the method of Lagrange multipliers. At the current iteration, show that the solution has components

$$q_k \quad = \quad \frac{c_{mk}}{d_m k - \lambda}$$

for an unknown Lagrange multiplier λ.

(c) Using the constraints, prove that λ exists and is unique.

(d) Describe a reliable method for computing λ.

(e) As an alternative to the exact method, construct a quadratic approximation to $h(\boldsymbol{q} \mid \boldsymbol{q}_m)$ near $\boldsymbol{q}_m$ of the form

$$\frac{1}{2}(\boldsymbol{q} - \boldsymbol{q}_m)^* \boldsymbol{A}(\boldsymbol{q} - \boldsymbol{q}_m) + \boldsymbol{b}^*(\boldsymbol{q} - \boldsymbol{q}_m) + c.$$

In particular, what are $\boldsymbol{A}$ and $\boldsymbol{b}$?

(f) Show that the quadratic approximation has maximum

$$\boldsymbol{q} \quad = \quad \boldsymbol{q}_m - \boldsymbol{A}^{-1}\Big(\boldsymbol{b} - \frac{\mathbf{1}^* \boldsymbol{A}^{-1} \boldsymbol{b}}{\mathbf{1}^* \boldsymbol{A}^{-1} \mathbf{1}} \mathbf{1}\Big) \tag{8.23}$$

subject to the constraint $\sum_{k=1}^{l} (q_k - q_{mk}) = 0$.

(g) In the dictionary model, demonstrate that the solution (8.23) takes the form

$$q_j = q_{mj} + \frac{(q_{mj})^2}{c_{mj}} \left[\frac{c_{mj}}{q_{mj}} - d_m j - \frac{1 - \sum_{k=1}^{l} d_m k \frac{(q_{mk})^2}{c_{mk}}}{\sum_{k=1}^{l} \frac{(q_{mk})^2}{c_{mk}}} \right]$$

for the jth component of q.

(h) Point out two potential pitfalls of this particular solution in conjunction with maximizing $h(\boldsymbol{q} \mid \boldsymbol{q}_m)$.

40. In the balanced ANOVA model with two factors, we estimate the parameter vector $\boldsymbol{\theta} = (\mu, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)^*$ by minimizing the sum of squares

$$f(\boldsymbol{\theta}) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} w_{ijk}(y_{ijk} - \mu - \alpha_i - \beta_j)^2$$

with all weights $w_{ijk} = 1$. If some of the observations y_{ijk} are missing, then we take the corresponding weights to be 0. The missing observations are now irrelevant, but it is possible to replace each one by its predicted value

$$\hat{y}_{ijk} = \mu + \alpha_i + \beta_j$$

given the current parameter values. If there are missing observations, de Leeuw [59] notes that

$$g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (z_{ijk} - \mu - \alpha_i - \beta_j)^2$$

majorizes $f(\boldsymbol{\theta})$ provided we define

$$z_{ijk} = \begin{cases} y_{ijk} & \text{for a regular observation} \\ \hat{y}_{ijk} & \text{for a missing observation.} \end{cases}$$

Prove this fact and calculate the MM update of $\boldsymbol{\theta}$, assuming the sum to 0 constraints $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0$.

41. Consider the weighted sum of squares criterion

$$f(\boldsymbol{\theta}) = \sum_i w_i [y_i - \mu_i(\boldsymbol{\theta})]^2$$

with weights w_i drawn from the unit interval $[0, 1]$. Show that the function

$$\begin{aligned} g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m) &= \sum_i [w_i y_i + (1 - w_i)\mu_i(\boldsymbol{\theta}_m) - \mu_i(\boldsymbol{\theta})]^2 + c_m \\ c_m &= \sum_i w_i (1 - w_i)[y_i - \mu_i(\boldsymbol{\theta}_m)]^2 \end{aligned}$$

majorizes $f(\boldsymbol{\theta})$ [121, 153, 242]. This majorization converts a weighted least squares problem into an ordinary least squares problem. (Hints: Write $[y_i-\mu_i(\boldsymbol{\theta})]^2 = [y_i-\mu_i(\boldsymbol{\theta}_m)+\mu_i(\boldsymbol{\theta}_m)-\mu_i(\boldsymbol{\theta})]^2$, expand, majorize $w_i[\mu_i(\boldsymbol{\theta}_m) - \mu_i(\boldsymbol{\theta})]^2$, and complete the square.)

42. Luce's model [181, 182] is a convenient scheme for ranking items such as candidates in an election, consumer goods in a certain category, or academic departments in a reputational survey. Some people will be too lazy or uncertain to rank each and every item, preferring to focus on just their top choices. How can we use this form of limited voting to rank the entire set of items? A partial ranking by a person is a sequence of random choices $X_1, \ldots, X_l$, with X_1 the highest ranked item, X_2 the second highest ranked item, and so forth. If there are r items, then the number l of items ranked may be strictly less than r; $l = 1$ is a distinct possibility. The data arrive at our doorstep as a random sample of s independent partial rankings, which we must integrate in some coherent fashion. One possibility is to adopt multinomial sampling without replacement. This differs from ordinary multinomial sampling in that once an item is chosen, it cannot be chosen again. However, remaining items are selected with the conditional probabilities dictated by the original sampling probabilities. Show that the likelihood under Luce's model reduces to

$$\begin{aligned} & \prod_{i=1}^{s} \Pr(X_{i1} = x_{i1}, \ldots, X_{il_i} = x_{il_i}) \\ = & \prod_{i=1}^{s} p_{x_{i1}} \prod_{j=1}^{l_i-1} \frac{p_{x_{i,j+1}}}{\sum_{k \notin \{x_{i1},\ldots,x_{ij}\}} p_k}, \end{aligned}$$

where x_{ij} is the jth choice out of l_i choices for person i and p_k is the multinomial probability assigned to item k. If we can estimate the p_k, then we can rank the items accordingly. The item with largest estimated probability is ranked first and so on.

The model has the added virtue of leading to straightforward estimation by the MM algorithm. Use the supporting hyperplane inequality

$$-\ln t \geq -\ln t_m - \frac{1}{t_m}(t - t_m).$$

to generate the minorization

$$Q(\boldsymbol{p} \mid \boldsymbol{p}_m) = \sum_{i=1}^{s} \sum_{j=1}^{l_i} \ln p_{x_{ij}} - \sum_{i=1}^{s} \sum_{j=1}^{l_i-1} w_{ij} \sum_{k \notin \{x_{i1},\ldots,x_{ij}\}} p_k$$

of the loglikelihood up to an irrelevant constant. Specify the positive weights w_{ij} and derive the maximization step of the MM algorithm.

Show that your update has the intuitive interpretation of equating the expected number of choices of item k to the observed number of choices of item k across all voters. Finally, generalize the model so that person i's choices are limited to a subset S_i of the items. For instance, in rating academic departments, some people may only feel competent to rank those departments in their state or region. What form does the MM algorithm take in this setting?

9
The EM Algorithm

9.1 Introduction

Maximum likelihood is the dominant form of estimation in applied statistics. Because closed-form solutions to likelihood equations are the exception rather than the rule, numerical methods for finding maximum likelihood estimates are of paramount importance. In this chapter we study maximum likelihood estimation by the EM algorithm [65, 179, 191], a special case of the MM algorithm. At the heart of every EM algorithm is some notion of missing data. Data can be missing in the ordinary sense of a failure to record certain observations on certain cases. Data can also be missing in a theoretical sense. We can think of the E (expectation) step of the algorithm as filling in the missing data. This action replaces the log-likelihood of the observed data by a minorizing function. This surrogate function is then maximized in the M step. Because the surrogate function is usually much simpler than the likelihood, we can often solve the M step analytically. The price we pay for this simplification is that the EM algorithm is iterative. Reconstructing the missing data is bound to be slightly wrong if the parameters do not already equal their maximum likelihood estimates.

One of the advantages of the EM algorithm is its numerical stability. As an MM algorithm, any EM algorithm leads to a steady increase in the likelihood of the observed data. Thus, the EM algorithm avoids wildly overshooting or undershooting the maximum of the likelihood along its current direction of search. Besides this desirable feature, the EM handles

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_9,

parameter constraints gracefully. Constraint satisfaction is by definition built into the solution of the M step. In contrast, competing methods of maximization must incorporate special techniques to cope with parameter constraints. The EM shares some of the negative features of the more general MM algorithm. For example, the EM algorithm often converges at an excruciatingly slow rate in a neighborhood of the maximum point. This rate directly reflects the amount of missing data in a problem. In the absence of concavity, there is also no guarantee that the EM algorithm will converge to the global maximum. The global maximum can usually be reached by starting the parameters at good but suboptimal estimates such as method-of-moments estimates or by choosing multiple random starting points.

9.2 Definition of the EM Algorithm

A sharp distinction is drawn in the EM algorithm between the observed, incomplete data $\boldsymbol{y}$ and the unobserved, complete data $\boldsymbol{x}$ of a statistical experiment [65, 179, 251]. Some function $t(\boldsymbol{x}) = \boldsymbol{y}$ collapses $\boldsymbol{x}$ onto $\boldsymbol{y}$. For instance, if we represent $\boldsymbol{x}$ as $(\boldsymbol{y}, \boldsymbol{z})$, with $\boldsymbol{z}$ as the missing data, then t is simply projection onto the $\boldsymbol{y}$-component of $\boldsymbol{x}$. It should be stressed that the missing data can consist of more than just observations missing in the ordinary sense. In fact, the definition of $\boldsymbol{x}$ is left up to the intuition and cleverness of the statistician. The general idea is to choose $\boldsymbol{x}$ so that maximum likelihood estimation becomes trivial for the complete data.

The complete data are assumed to have a probability density $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ that is a function of a parameter vector $\boldsymbol{\theta}$ as well as of $\boldsymbol{x}$. In the E step of the EM algorithm, we calculate the conditional expectation

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n) \quad = \quad \mathrm{E}[\ln f(\boldsymbol{X} \mid \boldsymbol{\theta}) \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{\theta}_n].$$

Here $\boldsymbol{\theta}_n$ is the current estimated value of $\boldsymbol{\theta}$, upper case letters indicate random vectors, and lower case letters indicate corresponding realizations of these random vectors. In the M step, we maximize $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n)$ with respect to $\boldsymbol{\theta}$. This yields the new parameter estimate $\boldsymbol{\theta}_{n+1}$, and we repeat this two-step process until convergence occurs. Note that $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_n$ play fundamentally different roles in $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n)$.

If $\ln g(\boldsymbol{y} \mid \boldsymbol{\theta})$ denotes the loglikelihood of the observed data, then the EM algorithm enjoys the ascent property

$$\ln g(\boldsymbol{y} \mid \boldsymbol{\theta}_{n+1}) \quad \geq \quad \ln g(\boldsymbol{y} \mid \theta_n).$$

Proof of this assertion unfortunately involves measure theory, so some readers may want to take it on faith and skip the rest of this section. A necessary preliminary is the following well-known inequality from statistics.

Proposition 9.2.1 (Information Inequality) *Let h and k be probability densities with respect to a measure μ. Suppose $h > 0$ and $k > 0$ almost everywhere relative to μ. If E_h denotes expectation with respect to the probability measure $h d\mu$, then $\mathrm{E}_h(\ln h) \geq \mathrm{E}_h(\ln k)$, with equality if and only if $h = k$ almost everywhere relative to μ.*

Proof: Because $-\ln(w)$ is a strictly convex function on $(0, \infty)$, Proposition 6.6.1 applied to the random variable k/h yields

$$\begin{aligned}
\mathrm{E}_h(\ln h) - \mathrm{E}_h(\ln k) &= \mathrm{E}_h\left(-\ln \frac{k}{h}\right) \\
&\geq -\ln \mathrm{E}_h\left(\frac{k}{h}\right) \\
&= -\ln \int \frac{k}{h} h \, d\mu \\
&= -\ln \int k \, d\mu \\
&= 0.
\end{aligned}$$

Equality holds if and only if $k/h = \mathrm{E}_h(k/h)$ almost everywhere relative to μ. This necessary and sufficient condition is equivalent to $h = k$ since $\mathrm{E}_h(k/h) = 1$. ■

To prove the ascent property of the EM algorithm, it suffices to demonstrate the minorization inequality

$$\ln g(\boldsymbol{y} \mid \boldsymbol{\theta}) \geq Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n) + \ln g(\boldsymbol{y} \mid \boldsymbol{\theta}_n) - Q(\boldsymbol{\theta}_n \mid \boldsymbol{\theta}_n),$$

where $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n) = \mathrm{E}[\ln f(\boldsymbol{X} \mid \boldsymbol{\theta}) \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{\theta}_n]$. With this end in mind, note that both $f(\boldsymbol{x} \mid \boldsymbol{\theta})/g(\boldsymbol{y} \mid \boldsymbol{\theta})$ and $f(\boldsymbol{x} \mid \boldsymbol{\theta}_n)/g(\boldsymbol{y} \mid \boldsymbol{\theta}_n)$ are conditional densities of $\boldsymbol{X}$ on $\{\boldsymbol{x} : t(\boldsymbol{x}) = \boldsymbol{y}\}$ with respect to some measure $\mu_{\boldsymbol{y}}$. The information inequality now indicates that

$$\begin{aligned}
Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n) - \ln g(\boldsymbol{y} \mid \boldsymbol{\theta}) &= \mathrm{E}\left(\ln \left[\frac{f(\boldsymbol{X} \mid \boldsymbol{\theta})}{g(\boldsymbol{Y} \mid \boldsymbol{\theta})}\right] \Big| \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{\theta}_n\right) \\
&\leq \mathrm{E}\left(\ln \left[\frac{f(\boldsymbol{X} \mid \boldsymbol{\theta}_n)}{g(\boldsymbol{Y} \mid \boldsymbol{\theta}_n)}\right] \Big| \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{\theta}_n\right) \\
&= Q(\boldsymbol{\theta}_n \mid \boldsymbol{\theta}_n) - \ln g(\boldsymbol{y} \mid \boldsymbol{\theta}_n).
\end{aligned}$$

Maximizing $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n)$ therefore drives $\ln g(\boldsymbol{y} \mid \boldsymbol{\theta})$ uphill. The ascent inequality is strict whenever the conditional density $f(\boldsymbol{x} \mid \boldsymbol{\theta})/g(\boldsymbol{y} \mid \boldsymbol{\theta})$ differs at the parameter points $\boldsymbol{\theta}_n$ and $\boldsymbol{\theta}_{n+1}$ or

$$Q(\boldsymbol{\theta}_{n+1} \mid \boldsymbol{\theta}_n) > Q(\boldsymbol{\theta}_n \mid \boldsymbol{\theta}_n).$$

The preceding proof is a little vague as to the meaning of the conditional density $f(\boldsymbol{x} \mid \boldsymbol{\theta})/g(\boldsymbol{y} \mid \boldsymbol{\theta})$ and its associated measure $\mu_{\boldsymbol{y}}$. Commonly the

complete data decomposes as $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{z})$, where $\boldsymbol{z}$ is considered the missing data and $t(\boldsymbol{y}, \boldsymbol{z}) = \boldsymbol{y}$ is projection onto the observed data. Suppose $(\boldsymbol{y}, \boldsymbol{z})$ has joint density $f(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\theta})$ relative to a product measure $\omega \times \mu\,(\boldsymbol{y}, \boldsymbol{z})$; ω and μ are typically Lebesgue measure or counting measure. In this framework, we define $g(\boldsymbol{y} \mid \boldsymbol{\theta}) = \int f(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\theta}) d\mu(\boldsymbol{z})$ and set $\mu_{\boldsymbol{y}} = \mu$. The function $g(\boldsymbol{y} \mid \boldsymbol{\theta})$ serves as a density relative to ω. To check that these definitions make sense, it suffices to prove that $\int h(\boldsymbol{y}, \boldsymbol{z}) f(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\theta}) / g(\boldsymbol{y} \mid \boldsymbol{\theta}) d\mu(\boldsymbol{z})$ is a version of the conditional expectation $\mathrm{E}[h(\boldsymbol{Y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}]$ for every well-behaved function $h(\boldsymbol{y}, \boldsymbol{z})$. This assertion can be verified by showing

$$\mathrm{E}\{1_S(\boldsymbol{Y})\,\mathrm{E}[h(\boldsymbol{Y}, \boldsymbol{Z}) \mid \boldsymbol{Y}]\} \quad = \quad \mathrm{E}[1_S(\boldsymbol{Y}) h(\boldsymbol{Y}, \boldsymbol{Z})]$$

for every measurable set S. With

$$\mathrm{E}[h(\boldsymbol{Y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}] \quad = \quad \int h(\boldsymbol{y}, \boldsymbol{z}) \frac{f(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\theta})}{g(\boldsymbol{y} \mid \boldsymbol{\theta})}\, d\mu(\boldsymbol{z}),$$

we calculate

$$\begin{aligned} & \mathrm{E}\{1_S(\boldsymbol{Y})\,\mathrm{E}[h(\boldsymbol{Y}, \boldsymbol{Z}) \mid \boldsymbol{Y} = \boldsymbol{y}]\} \\ = & \int_S \int h(\boldsymbol{y}, \boldsymbol{z}) \frac{f(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\theta})}{g(\boldsymbol{y} \mid \boldsymbol{\theta})}\, d\mu(\boldsymbol{z}) g(\boldsymbol{y} \mid \boldsymbol{\theta})\, d\omega(\boldsymbol{y}) \\ = & \int_S \int h(\boldsymbol{y}, \boldsymbol{z}) f(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\theta})\, d\mu(\boldsymbol{z})\, d\omega(\boldsymbol{y}) \\ = & \mathrm{E}[1_S(\boldsymbol{Y}) h(\boldsymbol{Y}, \boldsymbol{Z})]. \end{aligned}$$

Hence in this situation, $f(\boldsymbol{x} \mid \boldsymbol{\theta}) / g(\boldsymbol{y} \mid \boldsymbol{\theta})$ is indeed the conditional density of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$.

9.3 Missing Data in the Ordinary Sense

The most common application of the EM algorithm is to data missing in the ordinary sense. For example, Problem 40 of Chap. 8 considers a balanced ANOVA model with two factors. Missing observations in this setting break the symmetry that permits explicit solution of the likelihood equations. Thus, there is ample incentive for filling in the missing observations. If the observations follow an exponential model, and missing data are missing completely at random, then the EM algorithm replaces the sufficient statistic of each missing observation by its expected value.

The density of a random variable Y from an exponential family can be written as

$$f(y \mid \boldsymbol{\theta}) \quad = \quad g(y) e^{\beta(\boldsymbol{\theta}) + h(y)^* \boldsymbol{\gamma}(\boldsymbol{\theta})} \tag{9.1}$$

relative to some measure ν [73, 218]. The normal, Poisson, binomial, negative binomial, gamma, beta, and multinomial families are prime examples

of exponential families. The function $h(y)$ in equation (9.1) is the sufficient statistic. The maximum likelihood estimate of the parameter vector $\boldsymbol{\theta}$ depends on an observation y only through $h(y)$. Predictors of y are incorporated into the functions $\beta(\boldsymbol{\theta})$ and $\gamma(\boldsymbol{\theta})$.

To fill in a missing observation y, we take the ordinary expectation

$$\mathrm{E}[\ln f(Y \mid \boldsymbol{\theta}) \mid \boldsymbol{\theta}_n] = \mathrm{E}[\ln g(Y) \mid \boldsymbol{\theta}_n] + \beta(\boldsymbol{\theta}) + \mathrm{E}[h(Y) \mid \boldsymbol{\theta}_n]^* \gamma(\theta)$$

of the complete data loglikelihood. This function is added to the loglikelihood of the regular observations $y_1, \ldots, y_m$ to generate the surrogate function $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n)$. For example, if a typical observation is normally distributed with mean $\mu(\boldsymbol{\alpha})$ and variance σ^2, then $\boldsymbol{\theta}$ is the vector $(\boldsymbol{\alpha}^*, \sigma^2)^*$ and

$$\begin{aligned}\mathrm{E}[\ln f(Y \mid \boldsymbol{\theta}) \mid \boldsymbol{\theta}_n] &= \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \mathrm{E}\{[Y - \mu(\boldsymbol{\alpha})]^2 \mid \boldsymbol{\theta}_n\} \\ &= \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\{\sigma_n^2 + [\mu(\boldsymbol{\alpha}_n) - \mu(\boldsymbol{\alpha})]^2\}.\end{aligned}$$

Once we have filled in the missing data, we can estimate $\boldsymbol{\alpha}$ without reference to σ^2. This is accomplished by adding each square $[\mu_i(\boldsymbol{\alpha}_n) - \mu_i(\boldsymbol{\alpha})]^2$ corresponding to a missing observation y_i to the sum of squares for the actual observations and then minimizing the entire sum over $\boldsymbol{\alpha}$. In classical models such as balanced ANOVA, the M step is exact. Once the iterative limit $\lim_{n\to\infty} \boldsymbol{\alpha}_n = \hat{\boldsymbol{\alpha}}$ is reached, we can estimate σ^2 in one step by the formula

$$\widehat{\sigma^2} = \frac{1}{m} \sum_{i=1}^{m} [y_i - \mu_i(\hat{\boldsymbol{\alpha}})]^2$$

using only the observed y_i. The reader is urged to work Problem 40 of Chap. 8 to see the whole process in action.

9.4 Allele Frequency Estimation

It is instructive to compare the EM and MM algorithms on identical problems. Even when the two algorithms specify the same iteration scheme, the differences in deriving the algorithms are illuminating. Consider the ABO allele frequency estimation problem of Sect. 8.4. From the EM perspective, the complete data $\boldsymbol{x}$ are genotype counts rather than phenotype counts $\boldsymbol{y}$. In passing from the complete data to the observed data, nature collapses genotypes A/A and A/O into phenotype A and genotypes B/B and B/O into phenotype B. In view of the Hardy-Weinberg equilibrium law, the complete data multinomial loglikelihood becomes

$$\ln f(X \mid \boldsymbol{p}) = n_{A/A} \ln p_A^2 + n_{A/O} \ln(2p_A p_O) + n_{B/B} \ln p_B^2$$

$$+ n_{B/O} \ln(2p_B p_O) + n_{AB} \ln(2p_A p_B) + n_O \ln p_O^2 + \ln \binom{n}{n_{A/A},\, n_{A/O},\, n_{B/B},\, n_{B/O},\, n_{AB},\, n_O}. \quad (9.2)$$

In the E step of the EM algorithm we take the expectation of $\ln f(X \mid \boldsymbol{p})$ conditional on the observed counts n_A, n_B, n_{AB}, and n_O and the current parameter vector $\boldsymbol{p}_m = (p_{mA}, p_{mB}, p_{mO})^*$. This action yields the surrogate function

$$\begin{aligned} Q(\boldsymbol{p} \mid \boldsymbol{p}_m) &= \mathrm{E}(n_{A/A} \mid Y, \boldsymbol{p}_m) \ln p_A^2 + \mathrm{E}(n_{A/O} \mid Y, \boldsymbol{p}_m) \ln(2p_A p_O) \\ &\quad + \mathrm{E}(n_{B/B} \mid Y, \boldsymbol{p}_m) \ln p_B^2 + \mathrm{E}(n_{B/O} \mid Y, \boldsymbol{p}_m) \ln(2p_B p_O) \\ &\quad + \mathrm{E}(n_{AB} \mid Y, \boldsymbol{p}_m) \ln(2p_A p_B) + \mathrm{E}(n_O \mid Y, \boldsymbol{p}_m) \ln p_O^2 \\ &\quad + \mathrm{E}\left[\ln \binom{n}{n_{A/A},\, n_{A/O},\, n_{B/B},\, n_{B/O},\, n_{AB},\, n_O} \Big| Y, \boldsymbol{p}_m \right]. \end{aligned}$$

It is obvious that

$$\begin{aligned} \mathrm{E}(n_{AB} \mid Y, \boldsymbol{p}_m) &= n_{AB} \\ \mathrm{E}(n_O \mid Y, \boldsymbol{p}_m) &= n_O. \end{aligned}$$

Application of Bayes' rule gives

$$\begin{aligned} n_{mA/A} &= \mathrm{E}(n_{A/A} \mid Y, \boldsymbol{p}_m) \\ &= n_A \frac{p_{mA}^2}{p_{mA}^2 + 2p_{mA}p_{mO}} \\ n_{mA/O} &= \mathrm{E}(n_{A/O} \mid Y, \boldsymbol{p}_m) \\ &= n_A \frac{2p_{mA}p_{mO}}{p_{mA}^2 + 2p_{mA}p_{mO}}. \end{aligned}$$

The conditional expectations $n_{mB/B}$ and $n_{mB/O}$ reduce to similar expressions. Hence, the surrogate function $Q(\boldsymbol{p} \mid \boldsymbol{p}_m)$ derived from the complete data likelihood matches the surrogate function of the MM algorithm up to a constant, and the maximization step proceeds as described earlier. One of the advantages of the EM derivation is that it explicitly reveals the nature of the conditional expectations $n_{mA/A}$, $n_{mA/O}$, $n_{mB/B}$, and $n_{mB/O}$.

9.5 Clustering by EM

The k-means clustering algorithm discussed in Example 7.2.3 makes hard choices in cluster assignment. The alternative of soft choices is possible with admixture models [192, 259]. An admixture probability density $h(\boldsymbol{y})$ can be written as a convex combination

$$h(\boldsymbol{y}) = \sum_{j=1}^{k} \pi_j h_j(\boldsymbol{y}), \quad (9.3)$$

where the π_j are nonnegative probabilities that sum to 1 and $h_j(\boldsymbol{y})$ is the probability density of group j. According to Bayes' rule, the posterior probability that an observation $\boldsymbol{y}$ belongs to group j equals the ratio

$$\frac{\pi_j h_j(\boldsymbol{y})}{\sum_{i=1}^k \pi_i h_i(\boldsymbol{y})}. \tag{9.4}$$

If hard assignment is necessary, then the rational procedure is to assign $\boldsymbol{y}$ to the group with highest posterior probability.

Suppose the observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ represent a random sample from the admixture density (9.3). In practice we want to estimate the admixture proportions and whatever further parameters $\boldsymbol{\theta}$ characterize the densities $h_j(\boldsymbol{y} \mid \boldsymbol{\theta})$. The EM algorithm is natural in this context with group membership as the missing data. If we let z_{ij} be an indicator specifying whether observation $\boldsymbol{y}_i$ comes from group j, then the complete data loglikelihood amounts to

$$\sum_{i=1}^m \sum_{j=1}^k z_{ij} \left[\ln \pi_j + \ln h_j(\boldsymbol{y}_i \mid \boldsymbol{\theta})\right].$$

To find the surrogate function, we must find the conditional expectation w_{ij} of z_{ij}. But this reduces to the Bayes' rule (9.4) with $\boldsymbol{\theta}$ fixed at $\boldsymbol{\theta}_n$ and $\boldsymbol{\pi}$ fixed at $\boldsymbol{\pi}_n$, where as usual n indicates iteration number. Note that the property $\sum_{j=1}^k z_{ij} = 1$ entails the property $\sum_{j=1}^k w_{ij} = 1$.

Fortunately, the E step of the EM algorithm separates the $\boldsymbol{\pi}$ parameters from the $\boldsymbol{\theta}$ parameters. The problem of maximizing

$$\sum_{j=1}^k c_j \ln \pi_j$$

with $c_j = \sum_{i=1}^m w_{ij}$ should be familiar by now. Since $\sum_{j=1}^k c_j = m$, Example (1.4.2) shows that $\pi_{n+1,j} = \frac{c_j}{m}$.

We now undertake estimation of the remaining parameters assuming the groups are normally distributed with a common variance matrix $\boldsymbol{\Omega}$ but different mean vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$. The pertinent part of the surrogate function is

$$\begin{aligned}
&\sum_{i=1}^m \sum_{j=1}^k w_{ij} \Big[-\frac{1}{2} \ln \det \boldsymbol{\Omega} - \frac{1}{2} (\boldsymbol{y}_i - \boldsymbol{\mu}_j)^* \boldsymbol{\Omega}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_j) \Big] \\
= \; & -\frac{m}{2} \ln \det \boldsymbol{\Omega} - \frac{1}{2} \sum_{j=1}^k \sum_{i=1}^m w_{ij} (\boldsymbol{y}_i - \boldsymbol{\mu}_j)^* \boldsymbol{\Omega}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_j) \\
= \; & -\frac{m}{2} \ln \det \boldsymbol{\Omega} - \frac{1}{2} \operatorname{tr} \Big[\boldsymbol{\Omega}^{-1} \sum_{j=1}^k \sum_{i=1}^m w_{ij} (\boldsymbol{y}_i - \boldsymbol{\mu}_j)(\boldsymbol{y}_i - \boldsymbol{\mu}_j)^* \Big]. \qquad (9.5)
\end{aligned}$$

Differentiating the surrogate (9.5) with respect to $\boldsymbol{\mu}_j$ gives the equation

$$\sum_{i=1}^{m} w_{ij}\Omega^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_j) \;=\; \mathbf{0}$$

with solution

$$\boldsymbol{\mu}_{n+1,j} \;=\; \frac{1}{\sum_{i=1}^{m} w_{ij}} \sum_{i=1}^{m} w_{ij}\boldsymbol{y}_i.$$

Maximization of the surrogate (9.5) with respect to $\boldsymbol{\Omega}$ can be rephrased as maximization of

$$-\frac{m}{2}\ln\det\boldsymbol{\Omega} - \frac{1}{2}\operatorname{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{M})$$

for the choice

$$\boldsymbol{M} \;=\; \sum_{j=1}^{k}\sum_{i=1}^{m} w_{ij}(\boldsymbol{y}_i - \boldsymbol{\mu}_{n+1,j})(\boldsymbol{y}_i - \boldsymbol{\mu}_{n+1,j})^*.$$

Abstractly this is just the problem we faced in Example 6.5.7. Inspection of the arguments there shows that

$$\boldsymbol{\Omega}_{n+1} \;=\; \frac{1}{m}\boldsymbol{M}. \tag{9.6}$$

There is no guarantee of a unique mode in this model. Fortunately, k-means clustering generates good starting values for the parameters. The cluster centers provide the group means. If we set w_{ij} equal to 1 or 0 depending on whether observation i belongs to cluster j or not, then the matrix (9.6) serves as an initial guess of the common variance matrix. The initial admixture proportion π_j can be taken to be the proportion of the observations assigned to cluster j.

9.6 Transmission Tomography

The EM and MM algorithms for transmission tomography differ. The MM algorithm is easier to derive and computationally more efficient. In other examples, the opposite is true.

In the transmission tomography example of Sect. 8.10, it is natural to view the missing data as the number of photons X_{ij} entering each pixel j along each projection line i. These random variables supplemented by the observations Y_i constitute the complete data. If projection line i does not intersect pixel j, then $X_{ij} = 0$. Although X_{ij} and $X_{ij'}$ are not independent,

the collection $\{X_{ij}\}_j$ indexed by projection i is independent of the collection $\{X_{i'j}\}_j$ indexed by another projection i'. This allows us to work projection by projection in writing the complete data likelihood. We will therefore temporarily drop the projection subscript i and relabel pixels, starting with pixel 1 adjacent to the source and ending with pixel $m-1$ adjacent to the detector. In this notation X_1 is the number of photons leaving the source, X_j is the number of photons entering pixel j, and $X_m = Y$ is the number of photons detected.

By assumption X_1 follows a Poisson distribution with mean d. Conditional on $X_1, \ldots, X_j$, the random variable X_{j+1} is binomially distributed with X_j trials and success probability $e^{-l_j\theta_j}$. In other words, each of the X_j photons entering pixel j behaves independently and has a chance $e^{-l_j\theta_j}$ of avoiding attenuation in pixel j. It follows that the complete data loglikelihood for the current projection is

$$\begin{aligned} &-d + X_1 \ln d - \ln X_1! \\ &+ \sum_{j=1}^{m-1} \left[\ln \binom{X_j}{X_{j+1}} + X_{j+1} \ln e^{-l_j\theta_j} + (X_j - X_{j+1}) \ln(1 - e^{-l_j\theta_j}) \right]. \end{aligned} \tag{9.7}$$

To perform the E step of the EM algorithm, we merely need to compute the conditional expectations $\mathrm{E}(X_j \mid X_m = y, \boldsymbol{\theta})$ for $1 \leq j \leq m$. The conditional expectations of other terms such as $\ln \binom{X_j}{X_{j+1}}$ appearing in (9.7) are irrelevant in the subsequent M step.

Reasoning as above, we infer that the unconditional mean of X_j is

$$\mu_j \;=\; \mathrm{E}(X_j) \;=\; de^{-\sum_{k=1}^{j-1} l_k\theta_k}$$

and that the distribution of X_m conditional on X_j is binomial with X_j trials and success probability

$$\frac{\mu_m}{\mu_j} \;=\; e^{-\sum_{k=j}^{m-1} l_k\theta_k}.$$

In view of our remarks about random thinning in Chap. 8, the joint probability density of X_j and X_m therefore reduces to

$$\Pr(X_j = x_j, X_m = x_m) \;=\; e^{-\mu_j} \frac{\mu_j^{x_j}}{x_j!} \binom{x_j}{x_m} \left(\frac{\mu_m}{\mu_j}\right)^{x_m} \left(1 - \frac{\mu_m}{\mu_j}\right)^{x_j - x_m},$$

and the conditional probability density of X_j given X_m becomes

$$\begin{aligned} \Pr(X_j = x_j \mid X_m = x_m) &= \frac{e^{-\mu_j} \frac{\mu_j^{x_j}}{x_j!} \binom{x_j}{x_m} (\frac{\mu_m}{\mu_j})^{x_m} (1 - \frac{\mu_m}{\mu_j})^{x_j - x_m}}{e^{-\mu_m} \frac{\mu_m^{x_m}}{x_m!}} \\ &= e^{-(\mu_j - \mu_m)} \frac{(\mu_j - \mu_m)^{x_j - x_m}}{(x_j - x_m)!}. \end{aligned}$$

In other words, conditional on X_m, the difference $X_j - X_m$ follows a Poisson distribution with mean $\mu_j - \mu_m$. This implies in particular that

$$\begin{aligned} \mathrm{E}(X_j \mid X_m) &= \mathrm{E}(X_j - X_m \mid X_m) + X_m \\ &= \mu_j - \mu_m + X_m. \end{aligned}$$

Reverting to our previous notation, it is now possible to assemble the surrogate function $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n)$ of the E step. Define

$$\begin{aligned} M_{ij} &= d_i(e^{-\sum_{k \in S_{ij}} l_{ik}\theta_{nk}} - e^{-\sum_k l_{ik}\theta_{nk}}) + y_i \\ N_{ij} &= d_i(e^{-\sum_{k \in S_{ij} \cup \{j\}} l_{ik}\theta_{nk}} - e^{-\sum_k l_{ik}\theta_{nk}}) + y_i, \end{aligned}$$

where S_{ij} is the set of pixels between the source and pixel j along projection i. If j' is the next pixel after pixel j along projection i, then

$$\begin{aligned} M_{ij} &= \mathrm{E}(X_{ij} \mid Y_i = y_i, \boldsymbol{\theta}_n) \\ N_{ij} &= \mathrm{E}(X_{ij'} \mid Y_i = y_i, \boldsymbol{\theta}_n). \end{aligned}$$

In view of expression (9.7), we find that

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n) = \sum_i \sum_j \Big[-N_{ij} l_{ij} \theta_j + (M_{ij} - N_{ij}) \ln(1 - e^{-l_{ij}\theta_j}) \Big]$$

up to an irrelevant constant.

If we try to maximize $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n)$ by setting its partial derivatives equal to 0, we get for pixel j the equation

$$-\sum_i N_{ij} l_{ij} + \sum_i \frac{(M_{ij} - N_{ij}) l_{ij}}{e^{l_{ij}\theta_j} - 1} = 0. \tag{9.8}$$

This is an intractable transcendental equation in the single variable θ_j, and the M step must be solved numerically, say by Newton's method. It is straightforward to check that the left-hand side of equation (9.8) is strictly decreasing in θ_j and has exactly one positive solution. Thus, the EM algorithm like the MM algorithm has the advantages of decoupling the parameters in the likelihood equations and of satisfying the natural boundary constraints $\theta_j \geq 0$. The MM algorithm is preferable to the EM algorithm because the MM algorithm involves far fewer exponentiations in defining its surrogate function.

9.7 Factor Analysis

In some instances, the missing data framework of the EM algorithm offers the easiest way to exploit convexity in deriving an MM algorithm. The complete data for a given problem is often fairly natural, and the difficulty

in deriving an EM algorithm shifts toward specifying the E step. Statisticians are particularly adept at calculating complicated conditional expectations connected with sampling distributions. We now illustrate these truths for estimation in factor analysis. Factor analysis explains the covariation among the components of a random vector by approximating the vector by a linear transformation of a small number of uncorrelated factors. Because factor analysis models usually involve normally distributed random vectors, Appendix A.2 reviews some basic facts about the multivariate normal distribution.

For the sake of notational convenience, we now extend the expectation and variance operators to random vectors. The expectation of a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)^*$ is defined componentwise by

$$\mathrm{E}(\boldsymbol{X}) \;=\; \begin{pmatrix} \mathrm{E}[X_1] \\ \vdots \\ \mathrm{E}[X_n] \end{pmatrix}.$$

Linearity carries over from the scalar case in the sense that

$$\begin{aligned} \mathrm{E}(\boldsymbol{X} + \boldsymbol{Y}) &= \mathrm{E}(\boldsymbol{X}) + \mathrm{E}(\boldsymbol{Y}) \\ \mathrm{E}(\boldsymbol{M}\boldsymbol{X}) &= \boldsymbol{M}\,\mathrm{E}(\boldsymbol{X}) \end{aligned}$$

for a compatible random vector $\boldsymbol{Y}$ and a compatible matrix $\boldsymbol{M}$. The same componentwise conventions hold for the expectation of a random matrix and the variances and covariances of a random vector. Thus, we can express the variance matrix of a random vector $\boldsymbol{X}$ as

$$\mathrm{Var}(\boldsymbol{X}) \;=\; \mathrm{E}\{[\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})][\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})]^*\} \;=\; \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^*) - \mathrm{E}(\boldsymbol{X})\,\mathrm{E}(\boldsymbol{X})^*.$$

These notational choices produce many other compact formulas. For instance, the random quadratic form $\boldsymbol{X}^*\boldsymbol{M}\boldsymbol{X}$ has expectation

$$\mathrm{E}(\boldsymbol{X}^*\boldsymbol{M}\boldsymbol{X}) \;=\; \mathrm{tr}[\boldsymbol{M}\,\mathrm{Var}(\boldsymbol{X})] + \mathrm{E}(\boldsymbol{X})^*\boldsymbol{M}\,\mathrm{E}(\boldsymbol{X}). \tag{9.9}$$

To verify this assertion, observe that

$$\begin{aligned} \mathrm{E}(\boldsymbol{X}^*\boldsymbol{M}\boldsymbol{X}) &= \mathrm{E}\Big(\sum_i \sum_j X_i m_{ij} X_j\Big) \\ &= \sum_i \sum_j m_{ij}\,\mathrm{E}(X_i X_j) \\ &= \sum_i \sum_j m_{ij}[\mathrm{Cov}(X_i, X_j) + \mathrm{E}(X_i)\,\mathrm{E}(X_j)] \\ &= \mathrm{tr}[\boldsymbol{M}\,\mathrm{Var}(\boldsymbol{X})] + \mathrm{E}(\boldsymbol{X})^*\boldsymbol{M}\,\mathrm{E}(\boldsymbol{X}). \end{aligned}$$

The classical factor analysis model deals with l independent multivariate observations of the form

$$\boldsymbol{Y}_k \;=\; \boldsymbol{\mu} + \boldsymbol{F}\boldsymbol{X}_k + \boldsymbol{U}_k.$$

Here the $p \times q$ factor loading matrix $\boldsymbol{F}$ transforms the unobserved factor score $\boldsymbol{X}_k$ into the observed $\boldsymbol{Y}_k$. The random vector $\boldsymbol{U}_k$ represents random measurement error. Typically, q is much smaller than p. The random vectors $\boldsymbol{X}_k$ and $\boldsymbol{U}_k$ are independent and normally distributed with means and variances

$$\begin{aligned} \mathrm{E}(\boldsymbol{X}_k) &= \boldsymbol{0}, & \mathrm{Var}(\boldsymbol{X}_k) &= \boldsymbol{I} \\ \mathrm{E}(\boldsymbol{U}_k) &= \boldsymbol{0}, & \mathrm{Var}(\boldsymbol{U}_k) &= \boldsymbol{D}, \end{aligned}$$

where $\boldsymbol{I}$ is the $q \times q$ identity matrix and $\boldsymbol{D}$ is a $p \times p$ diagonal matrix with ith diagonal entry d_i. The entries of the mean vector $\boldsymbol{\mu}$, the factor loading matrix $\boldsymbol{F}$, and the diagonal matrix $\boldsymbol{D}$ constitute the parameters of the model. For a particular realization $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_l$ of the model, the maximum likelihood estimation of $\boldsymbol{\mu}$ is simply the sample mean $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$. This fact is a consequence of the reasoning given in Example 6.5.7. Therefore, we replace each $\boldsymbol{y}_k$ by $\boldsymbol{y}_k - \bar{\boldsymbol{y}}$, assume $\boldsymbol{\mu} = \boldsymbol{0}$, and focus on estimating $\boldsymbol{F}$ and $\boldsymbol{D}$.

The random vector $(\boldsymbol{X}_k^*, \boldsymbol{Y}_k^*)^*$ is the obvious choice of the complete data for case k. If $f(\boldsymbol{x}_k)$ is the density of $\boldsymbol{X}_k$ and $g(\boldsymbol{y}_k \mid \boldsymbol{x}_k)$ is the conditional density of $\boldsymbol{Y}_k$ given $\boldsymbol{X}_k = \boldsymbol{x}_k$, then the complete data loglikelihood can be expressed as

$$\begin{aligned} & \sum_{k=1}^{l} \ln f(\boldsymbol{x}_k) + \sum_{k=1}^{l} \ln g(\boldsymbol{y}_k \mid \boldsymbol{x}_k) \\ = \; & -\frac{l}{2} \ln \det \boldsymbol{I} - \frac{1}{2} \sum_{k=1}^{l} \boldsymbol{x}_k^* \boldsymbol{x}_k - \frac{l}{2} \ln \det \boldsymbol{D} \\ & - \frac{1}{2} \sum_{k=1}^{l} (\boldsymbol{y}_k - \boldsymbol{F}\boldsymbol{x}_k)^* \boldsymbol{D}^{-1} (\boldsymbol{y}_k - \boldsymbol{F}\boldsymbol{x}_k). \end{aligned} \tag{9.10}$$

We can simplify this by noting that $\ln \det \boldsymbol{I} = 0$ and $\ln \det \boldsymbol{D} = \sum_{i=1}^{p} \ln d_i$.

The key to performing the E step is to note that $(\boldsymbol{X}_k^*, \boldsymbol{Y}_k^*)^*$ follows a multivariate normal distribution with variance matrix

$$\mathrm{Var} \begin{pmatrix} \boldsymbol{X}_k \\ \boldsymbol{Y}_k \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{F}^* \\ \boldsymbol{F} & \boldsymbol{F}\boldsymbol{F}^* + \boldsymbol{D} \end{pmatrix}.$$

Equation (A.1) of Appendix A.2 then permits us to calculate the conditional expectation

$$\begin{aligned} \boldsymbol{v}_k &= \mathrm{E}(\boldsymbol{X}_k \mid \boldsymbol{Y}_k = \boldsymbol{y}_k, \boldsymbol{F}_n, \boldsymbol{D}_n) \\ &= \boldsymbol{F}_n^* (\boldsymbol{F}_n \boldsymbol{F}_n^* + \boldsymbol{D}_n)^{-1} \boldsymbol{y}_k \end{aligned}$$

and the conditional variance

$$\begin{aligned} \boldsymbol{A}_k &= \mathrm{Var}(\boldsymbol{X}_k \mid \boldsymbol{Y}_k = \boldsymbol{y}_k, \boldsymbol{F}_n, \boldsymbol{D}_n) \\ &= \boldsymbol{I} - \boldsymbol{F}_n^* (\boldsymbol{F}_n \boldsymbol{F}_n^* + \boldsymbol{D}_n)^{-1} \boldsymbol{F}_n, \end{aligned}$$

given the observed data and the current values of the matrices $\boldsymbol{F}$ and $\boldsymbol{D}$. Combining these results with equation (9.9) yields

$$\begin{aligned} & \mathrm{E}[(\boldsymbol{Y}_k - \boldsymbol{F}\boldsymbol{X}_k)^* \boldsymbol{D}^{-1}(\boldsymbol{Y}_k - \boldsymbol{F}\boldsymbol{X}_k) \mid \boldsymbol{Y}_k = \boldsymbol{y}_k] \\ = \; & \mathrm{tr}(\boldsymbol{D}^{-1}\boldsymbol{F}\boldsymbol{A}_k\boldsymbol{F}^*) + (\boldsymbol{y}_k - \boldsymbol{F}\boldsymbol{v}_k)^* \boldsymbol{D}^{-1}(\boldsymbol{y}_k - \boldsymbol{F}\boldsymbol{v}_k) \\ = \; & \mathrm{tr}\{\boldsymbol{D}^{-1}[\boldsymbol{F}\boldsymbol{A}_k\boldsymbol{F}^* + (\boldsymbol{y}_k - \boldsymbol{F}\boldsymbol{v}_k)(\boldsymbol{y}_k - \boldsymbol{F}\boldsymbol{v}_k)^*]\}. \end{aligned}$$

If we define

$$\boldsymbol{\Lambda} = \sum_{k=1}^{l} [\boldsymbol{A}_k + \boldsymbol{v}_k \boldsymbol{v}_k^*], \quad \boldsymbol{\Gamma} = \sum_{k=1}^{l} \boldsymbol{v}_k \boldsymbol{y}_k^*, \quad \boldsymbol{\Omega} = \sum_{k=1}^{l} \boldsymbol{y}_k \boldsymbol{y}_k^*$$

and take conditional expectations in equation (9.10), then we can write the surrogate function of the E step as

$$\begin{aligned} & Q(\boldsymbol{F}, \boldsymbol{D} \mid \boldsymbol{F}_n, \boldsymbol{D}_n) \\ = \; & -\frac{l}{2} \sum_{i=1}^{p} \ln d_i - \frac{1}{2} \mathrm{tr}[\boldsymbol{D}^{-1}(\boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{F}^* - \boldsymbol{F}\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^*\boldsymbol{F}^* + \boldsymbol{\Omega})], \end{aligned}$$

omitting the additive constant

$$-\frac{1}{2} \sum_{k=1}^{l} \mathrm{E}(\boldsymbol{X}_k^* \boldsymbol{X}_k \mid \boldsymbol{Y}_k = \boldsymbol{y}_k, \boldsymbol{F}_n, \boldsymbol{D}_n),$$

which depends on neither $\boldsymbol{F}$ nor $\boldsymbol{D}$.

To perform the M step, we first maximize $Q(\boldsymbol{F}, \boldsymbol{D} \mid \boldsymbol{F}_n, \boldsymbol{D}_n)$ with respect to $\boldsymbol{F}$, holding $\boldsymbol{D}$ fixed. We can do so by permuting factors and completing the square in the trace

$$\begin{aligned} & \mathrm{tr}[\boldsymbol{D}^{-1}(\boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{F}^* - \boldsymbol{F}\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^*\boldsymbol{F}^* + \boldsymbol{\Omega})] \\ = \; & \mathrm{tr}[\boldsymbol{D}^{-1}(\boldsymbol{F} - \boldsymbol{\Gamma}^*\boldsymbol{\Lambda}^{-1})\boldsymbol{\Lambda}(\boldsymbol{F} - \boldsymbol{\Gamma}^*\boldsymbol{\Lambda}^{-1})^*] + \mathrm{tr}[\boldsymbol{D}^{-1}(\boldsymbol{\Omega} - \boldsymbol{\Gamma}^*\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma})] \\ = \; & \mathrm{tr}[\boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{F} - \boldsymbol{\Gamma}^*\boldsymbol{\Lambda}^{-1})\boldsymbol{\Lambda}(\boldsymbol{F} - \boldsymbol{\Gamma}^*\boldsymbol{\Lambda}^{-1})^*\boldsymbol{D}^{-\frac{1}{2}}] + \mathrm{tr}[\boldsymbol{D}^{-1}(\boldsymbol{\Omega} - \boldsymbol{\Gamma}^*\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma})]. \end{aligned}$$

This calculation depends on the existence of the inverse matrix $\boldsymbol{\Lambda}^{-1}$. Now $\boldsymbol{\Lambda}$ is certainly positive definite if $\boldsymbol{A}_k$ is positive definite, and Problem 22 asserts that $\boldsymbol{A}_k$ is positive definite. It follows that $\boldsymbol{\Lambda}^{-1}$ not only exists but is positive definite as well. Furthermore, the matrix

$$\boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{F} - \boldsymbol{\Gamma}^*\boldsymbol{\Lambda}^{-1})\boldsymbol{\Lambda}(\boldsymbol{F} - \boldsymbol{\Gamma}^*\boldsymbol{\Lambda}^{-1})^*\boldsymbol{D}^{-\frac{1}{2}}$$

is positive semidefinite and has a nonnegative trace. Hence, the maximum value of the surrogate function $Q(\boldsymbol{F}, \boldsymbol{D} \mid \boldsymbol{F}_n, \boldsymbol{D}_n)$ with respect to $\boldsymbol{F}$ is attained at the point $\boldsymbol{F} = \boldsymbol{\Gamma}^*\boldsymbol{\Lambda}^{-1}$, regardless of the value of $\boldsymbol{D}$. In other words, the EM update of $\boldsymbol{F}$ is $\boldsymbol{F}_{n+1} = \boldsymbol{\Gamma}^*\boldsymbol{\Lambda}^{-1}$. It should be stressed that

$\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$ implicitly depend on the previous values $\boldsymbol{F}_n$ and $\boldsymbol{D}_n$. Once $\boldsymbol{F}_{n+1}$ is determined, the equation

$$\begin{aligned} 0 &= \frac{\partial}{\partial d_i} Q(\boldsymbol{F}, \boldsymbol{D} \mid \boldsymbol{F}_n, \boldsymbol{D}_n) \\ &= -\frac{l}{2d_i} + \frac{1}{2d_i^2}(\boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{F}^* - \boldsymbol{F}\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^*\boldsymbol{F}^* + \boldsymbol{\Omega})_{ii} \end{aligned}$$

provides the update

$$d_{n+1,i} = \frac{1}{l}(\boldsymbol{F}_{n+1}\boldsymbol{\Lambda}\boldsymbol{F}_{n+1}^* - \boldsymbol{F}_{n+1}\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^*\boldsymbol{F}_{n+1}^* + \boldsymbol{\Omega})_{ii}.$$

One of the frustrating features of factor analysis is that the factor loading matrix $\boldsymbol{F}$ is not uniquely determined. To understand the source of the ambiguity, consider replacing $\boldsymbol{F}$ by $\boldsymbol{F}\boldsymbol{O}$, where $\boldsymbol{O}$ is a $q \times q$ orthogonal matrix. The distribution of each random vector $\boldsymbol{Y}_k$ is normal with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{F}\boldsymbol{F}^* + \boldsymbol{D}$. If we substitute $\boldsymbol{F}\boldsymbol{O}$ for $\boldsymbol{F}$, then the variance $\boldsymbol{F}\boldsymbol{O}\boldsymbol{O}^*\boldsymbol{F}^* + \boldsymbol{D} = \boldsymbol{F}\boldsymbol{F}^* + \boldsymbol{D}$ remains the same. Another problem in factor analysis is the existence of more than one local maximum. Which one of these the EM algorithm converges to depends on its starting value [76]. For a suggestion of how to improve the chances of converging to the dominant mode, see the article [281].

9.8 Hidden Markov Chains

A hidden Markov chain incorporates both observed data and missing data. The missing data are the sequence of states visited by the chain; the observed data provide partial information about this sequence of states. Denote the sequence of visited states by $Z_1, \ldots, Z_n$ and the observation taken at epoch i when the chain is in state Z_i by $Y_i = y_i$. Baum's algorithms [8, 71] recursively compute the likelihood of the observed data

$$P = \Pr(Y_1 = y_1, \ldots, Y_n = y_n) \tag{9.11}$$

without actually enumerating all possible realizations $Z_1, \ldots, Z_n$. Baum's algorithms can be adapted to perform an EM search. The references [78, 165, 216] discuss several concrete examples of hidden Markov chains.

The likelihood (9.11) is constructed from three ingredients: (a) the initial distribution $\boldsymbol{\pi}$ at the first epoch of the chain, (b) the epoch-dependent transition probabilities $p_{ijk} = \Pr(Z_{i+1} = k \mid Z_i = j)$, and (c) the conditional densities $\phi_i(y_i \mid j) = \Pr(Y_i = y_i \mid Z_i = j)$. The dependence of the transition probability p_{ijk} on i allows the chain to be inhomogeneous over time and promotes greater flexibility in modeling. Implicit in the definition of $\phi_i(y_i \mid j)$ are the assumptions that $Y_1, \ldots, Y_n$ are independent given

$Z_1, \ldots, Z_n$ and that Y_i depends only on Z_i. For simplicity, we will assume that the Y_i are discretely distributed.

Baum's forward algorithm is based on recursively evaluating the joint probabilities

$$\alpha_i(j) \;=\; \Pr(Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}, Z_i = j).$$

At the first epoch, $\alpha_1(j) = \pi_j$ by definition. The obvious update to $\alpha_i(j)$ is

$$\alpha_{i+1}(k) \;=\; \sum_j \alpha_i(j)\phi_i(y_i \mid j)p_{ijk}. \tag{9.12}$$

The likelihood (9.11) can be recovered by computing the sum

$$P \;=\; \sum_j \alpha_n(j)\phi_n(y_n \mid j)$$

at the final epoch n.

In Baum's backward algorithm, we recursively evaluate the conditional probabilities

$$\beta_i(k) \;=\; \Pr(Y_{i+1} = y_{i+1}, \ldots, Y_n = y_n \mid Z_i = k),$$

starting by convention at $\beta_n(k) = 1$ for all k. The required update is clearly

$$\beta_i(j) \;=\; \sum_k p_{ijk}\phi_{i+1}(y_{i+1} \mid k)\beta_{i+1}(k). \tag{9.13}$$

In this instance, the likelihood is recovered at the first epoch by forming the sum $P = \sum_j \pi_j \phi_1(y_1 \mid j)\beta_1(j)$.

Baum's algorithms also interdigitate beautifully with the E step of the EM algorithm. It is natural to summarize the missing data by a collection of indicator random variables X_{ij}. If the chain occupies state j at epoch i, then we take $X_{ij} = 1$. Otherwise, we take $X_{ij} = 0$. In this notation, the complete data loglikelihood can be written as

$$\begin{aligned} L_{\text{com}}(\boldsymbol{\theta}) \;&=\; \sum_j X_{1j} \ln \pi_j + \sum_{i=1}^{n} \sum_j X_{ij} \ln \phi_i(Y_i \mid j) \\ &\quad + \sum_{i=1}^{n-1} \sum_j \sum_k X_{ij} X_{i+1,k} \ln p_{ijk}. \end{aligned}$$

Execution of the E step amounts to calculation of the conditional expectations

$$\begin{aligned} \mathrm{E}(X_{ij}X_{i+1,k} \mid \boldsymbol{Y}, \boldsymbol{\theta}_m) \;&=\; \frac{\alpha_i(j)\phi_i(y_i \mid j)p_{ijk}\phi_{i+1}(y_{i+1} \mid k)\beta_{i+1}(k)}{P}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_m} \\ \mathrm{E}(X_{ij} \mid \boldsymbol{Y}, \boldsymbol{\theta}_m) \;&=\; \frac{\alpha_i(j)\phi_i(y_i \mid j)\beta_i(j)}{P}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_m}, \end{aligned}$$

where $\boldsymbol{Y} = \boldsymbol{y}$ is the observed data, P is the likelihood of the observed data, and $\boldsymbol{\theta}_m$ is the current parameter vector.

The M step may or may not be exactly solvable. If it is not, then one can always revert to the MM gradient algorithm discussed in Sect. 10.4. In the case of hidden multinomial trials, it is possible to carry out the M step analytically. Hidden multinomial trials may govern (a) the choice of the initial state j, (b) the choice of an observed outcome Y_i at the ith epoch given the hidden state j of the chain at that epoch, or (c) the choice of the next state k given the current state j in a time-homogeneous chain. In the first case, the multinomial parameters are the π_j; in the last case, they are the common transition probabilities p_{jk}.

As a concrete example, consider estimation of the initial distribution $\boldsymbol{\pi}$ at the first epoch of the chain. For estimation to be accurate, there must be multiple independent runs of the chain. Let the superscript r index the various runs. The surrogate function delivered by the E step equals

$$Q(\boldsymbol{\pi} \mid \boldsymbol{\pi}_m) = \sum_r \sum_j \mathrm{E}(X_{1j}^r \mid \boldsymbol{Y}^r = \boldsymbol{y}^r, \boldsymbol{\pi}_m) \ln \pi_j$$

up to an additive constant. Maximizing $Q(\boldsymbol{\pi} \mid \boldsymbol{\pi}_m)$ subject to the constraints $\sum_j \pi_j = 1$ and $\pi_j \geq 0$ for all j is done as in Example 1.4.2. The resulting EM updates

$$\pi_{m+1,j} = \frac{\sum_r \mathrm{E}(X_{1j}^r \mid \boldsymbol{Y}^r = \boldsymbol{y}^r, \boldsymbol{\pi}_m)}{R}$$

for R runs can be interpreted as multinomial proportions with fractional category counts. Problem 24 asks the reader to derive the EM algorithm for estimating time homogeneous transition probabilities. Problem 25 covers estimation of the parameters of the conditional densities $\phi_i(y_i \mid j)$ for some common densities.

9.9 Problems

1. Code and test any of the algorithms discussed in the text or problems of this chapter.

2. The entropy of a probability density $p(\boldsymbol{x})$ on R^n is defined by

$$-\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x}. \tag{9.14}$$

Among all densities with a fixed mean vector $\boldsymbol{\mu} = \int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x}$ and variance matrix $\boldsymbol{\Omega} = \int (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^* p(\boldsymbol{x}) d\boldsymbol{x}$, prove that the multivariate normal has maximum entropy. (Hints: Apply Proposition 9.2.1 and formula (9.9).)

3. In statistical mechanics, entropy is employed to characterize the stationary distribution of many independently behaving particles. Let $p(\boldsymbol{x})$ be the probability density that a particle is found at position $\boldsymbol{x}$ in phase space R^n, and suppose that each position $\boldsymbol{x}$ is assigned an energy $u(\boldsymbol{x})$. If the average energy $U = \int u(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$ per particle is fixed, then Nature chooses $p(\boldsymbol{x})$ to maximize entropy as defined in equation (9.14). Show that if constants α and β exist satisfying

$$\int \alpha e^{\beta u(\boldsymbol{x})} d\boldsymbol{x} \;=\; 1 \quad \text{and} \quad \int u(\boldsymbol{x}) \alpha e^{\beta u(\boldsymbol{x})} d\boldsymbol{x} \;=\; U,$$

then $p(\boldsymbol{x}) = \alpha e^{\beta u(\boldsymbol{x})}$ does indeed maximize entropy subject to the average energy constraint. The density $p(\boldsymbol{x})$ is the celebrated Maxwell-Boltzmann density.

4. Show that the normal, Poisson, binomial, negative binomial, gamma, beta, and multinomial families are exponential by writing their densities in the form (9.1). What are the corresponding measure and sufficient statistic in each case?

5. In the EM algorithm [65], suppose that the complete data X possesses a regular exponential density

$$f(x \mid \boldsymbol{\theta}) \;=\; g(x) e^{\beta(\boldsymbol{\theta}) + h(x)^* \boldsymbol{\theta}}$$

relative to some measure ν. Prove that the unconditional mean of the sufficient statistic $h(X)$ is given by the negative gradient $-\nabla\beta(\boldsymbol{\theta})$ and that the EM update is characterized by the condition

$$\mathrm{E}[h(X) \mid Y, \theta_n] \;=\; -\nabla\beta(\boldsymbol{\theta}_{n+1}).$$

6. Suppose the phenotypic counts in the ABO allele frequency estimation example satisfy $n_A + n_{AB} > 0$, $n_B + n_{AB} > 0$, and $n_O > 0$. Show that the loglikelihood is strictly concave and possesses a single global maximum on the interior of the feasible region.

7. In a genetic linkage experiment, 197 animals are randomly assigned to four categories according to the multinomial distribution with cell probabilities $\pi_1 = \frac{1}{2} + \frac{\theta}{4}$, $\pi_2 = \frac{1-\theta}{4}$, $\pi_3 = \frac{1-\theta}{4}$ and $\pi_4 = \frac{\theta}{4}$. If the corresponding observations are

$$y \;=\; (y_1, y_2, y_3, y_4)^* \;=\; (125, 18, 20, 34)^*,$$

then devise an EM algorithm and use it to estimate $\hat{\theta} = .6268$ [218]. (Hint: Split the first category into two so that there are five categories for the complete data.)

8. Derive the EM algorithm solving Problem 7 as an MM algorithm. No mention of missing data is necessary.

9. Consider the data from *The London Times* [259] during the years 1910–1912 given in Table 9.1. The two columns labeled "Deaths i" refer to the number of deaths to women 80 years and older reported by day. The columns labeled "Frequency n_i" refer to the number of days with i deaths. A Poisson distribution gives a poor fit to these data, possibly because of different patterns of deaths in winter and summer. A mixture of two Poissons provides a much better fit. Under the Poisson admixture model, the likelihood of the observed data is

$$\prod_{i=0}^{9}\left[\alpha e^{-\mu_1}\frac{\mu_1^i}{i!}+(1-\alpha)e^{-\mu_2}\frac{\mu_2^i}{i!}\right]^{n_i},$$

where α is the admixture parameter and μ_1 and μ_2 are the means of the two Poisson distributions.

TABLE 9.1. Death notices from *The London Times*

Deaths i	**Frequency** n_i	**Deaths** i	**Frequency** n_i
0	162	5	61
1	267	6	27
2	271	7	8
3	185	8	3
4	111	9	1

Formulate an EM algorithm for this model. Let $\boldsymbol{\theta} = (\alpha, \mu_1, \mu_2)^*$ and

$$z_i(\boldsymbol{\theta}) = \frac{\alpha e^{-\mu_1}\mu_1^i}{\alpha e^{-\mu_1}\mu_1^i + (1-\alpha)e^{-\mu_2}\mu_2^i}$$

be the posterior probability that a day with i deaths belongs to Poisson population 1. Show that the EM algorithm is given by

$$\begin{aligned}
\alpha_{m+1} &= \frac{\sum_i n_i z_i(\boldsymbol{\theta}_m)}{\sum_i n_i} \\
\mu_{m+1,1} &= \frac{\sum_i n_i i z_i(\boldsymbol{\theta}_m)}{\sum_i n_i z_i(\boldsymbol{\theta}_m)} \\
\mu_{m+1,2} &= \frac{\sum_i n_i i[1-z_i(\boldsymbol{\theta}_m)]}{\sum_i n_i[1-z_i(\boldsymbol{\theta}_m)]}.
\end{aligned}$$

From the initial estimates $\alpha_0 = 0.3$, $\mu_{01} = 1.$ and $\mu_{02} = 2.5$, compute via the EM algorithm the maximum likelihood estimates $\hat{\alpha} = 0.3599$, $\hat{\mu}_1 = 1.2561$, and $\hat{\mu}_2 = 2.6634$. Note how slowly the EM algorithm converges in this example.

10. Derive the least squares algorithm (8.19) as an EM algorithm [112]. (Hint: Decompose y_i as the sum $\sum_{j=1}^{q} y_{ij}$ of realizations from independent normal deviates with means $x_{ij}\theta_j$ and variances $1/q$.)

11. Let $x_1, \ldots, x_m$ be an i.i.d. sample from a normal density with mean μ and variance σ^2. Suppose for each x_i we observe $y_i = |x_i|$ rather than x_i. Formulate an EM algorithm for estimating μ and σ^2, and show that its updates are

$$\begin{aligned} \mu_{n+1} &= \frac{1}{m}\sum_{i=1}^{m}(w_{ni1}y_i - w_{ni2}y_i) \\ \sigma^2_{n+1} &= \frac{1}{m}\sum_{i=1}^{m}[w_{ni1}(y_i - \mu_{n+1})^2 + w_{ni2}(-y_i - \mu_{n+1})^2] \end{aligned}$$

with weights

$$\begin{aligned} w_{ni1} &= \frac{f(y_i \mid \boldsymbol{\theta}_n)}{f(y_i \mid \boldsymbol{\theta}_n) + f(-y_i \mid \boldsymbol{\theta}_n)} \\ w_{ni2} &= \frac{f(-y_i \mid \boldsymbol{\theta}_n)}{f(y_i \mid \boldsymbol{\theta}_n) + f(-y_i \mid \boldsymbol{\theta}_n)}, \end{aligned}$$

where $f(x \mid \boldsymbol{\theta})$ is the normal density with $\boldsymbol{\theta} = (\mu, \sigma^2)^*$. Demonstrate that the modes of the likelihood of the observed data come in symmetric pairs differing only in the sign of μ. This fact does not prevent accurate estimation of $|\mu|$ and σ^2.

12. Consider an i.i.d. sample drawn from a bivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)^*$ and variance matrix

$$\boldsymbol{\Omega} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

Suppose through some random accident that the first p observations are missing their first component, the next q observations are missing their second component, and the last r observations are complete. Design an EM algorithm to estimate the five mean and variance parameters, taking as complete data the original data before the accidental loss.

13. The standard linear regression model can be written in matrix notation as $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{U}$. Here $\boldsymbol{X}$ is the $r \times 1$ vector of responses, $\boldsymbol{A}$ is the $r \times s$ design matrix, $\boldsymbol{\beta}$ is the $s \times 1$ vector of regression coefficients, and $\boldsymbol{U}$ is the $r \times 1$ normally distributed error vector with mean $\boldsymbol{0}$ and variance $\sigma^2\boldsymbol{I}$. The responses are right censored if for each i there is a constant c_i such that only $Y_i = \min\{c_i, X_i\}$ is observed.

The EM algorithm offers a vehicle for estimating the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^*, \sigma^2)^*$ in the presence of right censoring [65, 251]. Show that

$$\begin{aligned}\boldsymbol{\beta}_{n+1} &= (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^* \operatorname{E}(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{\theta}_n) \\ \sigma^2_{n+1} &= \frac{1}{r} \operatorname{E}[(\boldsymbol{X} - \boldsymbol{A}\boldsymbol{\beta}_{n+1})^*(\boldsymbol{X} - \boldsymbol{A}\boldsymbol{\beta}_{n+1}) \mid \boldsymbol{Y}, \boldsymbol{\theta}_n].\end{aligned}$$

To compute the conditional expectations appearing in these formulas, let $\boldsymbol{a}_i$ be the ith row of $\boldsymbol{A}$ and define

$$H(v) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}}}{\frac{1}{\sqrt{2\pi}} \int_v^\infty e^{-\frac{w^2}{2}} dw}.$$

For a censored observation $y_i = c_i < \infty$, prove that

$$\begin{aligned}\operatorname{E}(X_i \mid Y_i = c_i, \boldsymbol{\theta}_n) &= \boldsymbol{a}_i\boldsymbol{\beta}_n + \sigma_n H\Big(\frac{c_i - \boldsymbol{a}_i\boldsymbol{\beta}_n}{\sigma_n}\Big) \\ \operatorname{E}(X_i^2 \mid Y_i = c_i, \boldsymbol{\theta}_n) &= (\boldsymbol{a}_i\boldsymbol{\beta}_n)^2 + \sigma_n^2 \\ &\quad + \sigma_n(c_i + \boldsymbol{a}_i\boldsymbol{\beta}_n) H\Big(\frac{c_i - \boldsymbol{a}_i\boldsymbol{\beta}_n}{\sigma_n}\Big).\end{aligned}$$

Use these formulas to complete the specification of the EM algorithm.

14. In the transmission tomography model it is possible to approximate the solution of equation (9.8) to good accuracy in certain situations. Verify the expansion

$$\frac{1}{e^s - 1} = \frac{1}{s} - \frac{1}{2} + \frac{s}{12} + O(s^2).$$

Using the approximation $1/(e^s - 1) \approx 1/s - 1/2$ for $s = l_{ij}\theta_j$, show that

$$\theta_{n+1,j} = \frac{\sum_i (M_{ij} - N_{ij})}{\frac{1}{2}\sum_i (M_{ij} + N_{ij}) l_{ij}}$$

results. Can you motivate this result heuristically?

15. Suppose that the complete data in the EM algorithm involve N binomial trials with success probability θ per trial. Here N can be random or fixed. If M trials result in success, then the complete data likelihood can be written as $\theta^M(1-\theta)^{N-M}c$, where c is an irrelevant constant. The E step of the EM algorithm amounts to forming

$$Q(\theta \mid \theta_n) = \operatorname{E}(M \mid \boldsymbol{Y}, \theta_n) \ln\theta + \operatorname{E}(N - M \mid \boldsymbol{Y}, \theta_n) \ln(1-\theta) + \ln c.$$

The binomial trials are hidden because only a function $\boldsymbol{Y}$ of them is directly observed. The brief derivation in Sect. 9.8 shows that the EM update amounts to

$$\theta_{n+1} = \frac{\mathrm{E}(M \mid \boldsymbol{Y}, \theta_n)}{\mathrm{E}(N \mid \boldsymbol{Y}, \theta_n)}.$$

Prove that this is equivalent to the update

$$\theta_{n+1} = \theta_n + \frac{\theta_n(1-\theta_n)}{\mathrm{E}(N \mid \boldsymbol{Y}, \theta_n)} \frac{d}{d\theta} L(\theta_n),$$

where $L(\theta)$ is the loglikelihood of the observed data $\boldsymbol{Y}$ [270]. (Hint: Apply identity (8.4) of Chap. 8.)

16. As an example of hidden binomial trials, consider a random sample of twin pairs. Let u of these pairs consist of male pairs, v consist of female pairs, and w consist of opposite sex pairs. A simple model to explain these data involves a random Bernoulli choice for each pair dictating whether it consists of identical or nonidentical twins. Suppose that identical twins occur with probability p and nonidentical twins with probability $1-p$. Once the decision is made as to whether the twins are identical, then sexes are assigned to the twins. If the twins are identical, one assignment of sex is made. If the twins are nonidentical, then two independent assignments of sex are made. Suppose boys are chosen with probability q and girls with probability $1-q$. Model these data as hidden binomial trials. Derive the EM algorithm for estimating p and q.

17. Chun Li has derived an EM update for hidden multinomial trials. Let N denote the number of hidden trials, θ_i the probability of outcome i of k possible outcomes, and $L(\boldsymbol{\theta})$ the loglikelihood of the observed data $\boldsymbol{Y}$. Derive the EM update

$$\theta_{n+1,i} = \theta_{ni} + \frac{\theta_{ni}}{\mathrm{E}(N \mid \boldsymbol{Y}, \boldsymbol{\theta}_n)} \left[\frac{\partial}{\partial \theta_i} L(\boldsymbol{\theta}_n) - \sum_{j=1}^{k} \theta_{nj} \frac{\partial}{\partial \theta_j} L(\boldsymbol{\theta}_n) \right]$$

following the reasoning of Problem 15.

18. In this problem you are asked to formulate models for hidden Poisson and exponential trials [270]. If the number of trials is N and the mean per trial is θ, then show that the EM update in the Poisson case is

$$\theta_{n+1} = \theta_n + \frac{\theta_n}{\mathrm{E}(N \mid \boldsymbol{Y}, \theta_n)} \frac{d}{d\theta} L(\theta_n)$$

and in the exponential case is

$$\theta_{n+1} = \theta_n + \frac{\theta_n^2}{\mathrm{E}(N \mid \boldsymbol{Y}, \theta_n)} \frac{d}{d\theta} L(\theta_n),$$

where $L(\theta)$ is the loglikelihood of the observed data $\boldsymbol{Y}$.

19. Suppose light bulbs have an exponential lifetime with mean θ. Two experiments are conducted. In the first, the lifetimes $y_1, \ldots, y_m$ of m independent bulbs are observed. In the second, p independent bulbs are observed to burn out before time t, and q independent bulbs are observed to burn out after time t. In other words, the lifetimes in the second experiment are both left and right censored. Construct an EM algorithm for finding the maximum likelihood estimate of θ [95].

20. In many discrete probability models, only data with positive counts are observed. Counts that are 0 are missing. Show that the likelihoods for the binomial, Poisson, and negative binomial models truncated at 0 amount to

$$\begin{aligned} L_1(p) &= \prod_i \frac{\binom{m_i}{x_i} p^{x_i}(1-p)^{m_i-x_i}}{1-(1-p)^{m_i}} \\ L_2(\lambda) &= \prod_i \frac{\lambda^{x_i} e^{-\lambda}}{x_i!(1-e^{-\lambda})} \\ L_3(p) &= \prod_i \frac{\binom{m_i+x_i-1}{x_i}(1-p)^{x_i} p^{m_i}}{1-p^{m_i}}. \end{aligned}$$

For observation i of the binomial model, there are x_i successes out of m_i trials with success probability p per trial. For observation i of the negative binomial model, there are x_i failures before m_i required successes. For each model, devise an EM algorithm that fills in the missing observations by imputing a geometrically distributed number of truncated observations for every real observation. Show that the EM updates reduce to

$$\begin{aligned} p_{n+1} &= \frac{\sum_i x_i}{\sum_i \frac{m_i}{1-(1-p_n)^{m_i}}} \\ \lambda_{n+1} &= \frac{\sum_i x_i}{\sum_i \frac{1}{1-e^{-\lambda_n}}} \\ p_{n+1} &= \frac{\sum_i \frac{m_i}{1-p_n^{m_i}}}{\sum_i (x_i + \frac{m_i}{1-p_n^{m_i}})} \end{aligned}$$

for the three models.

21. Demonstrate that the EM updates of the previous problem can be derived as MM updates based on the minorization

$$-\ln(1-u) \geq -\ln(1-u_n) + \frac{u_n}{1-u_n} \ln \frac{u}{u_n}$$

for u and u_n in the interval $(0, 1)$. Prove this minorization first. (Hint: If you rearrange the minorization, then Proposition 9.2.1 applies.)

22. Suppose that $\boldsymbol{\Sigma}$ is a positive definite matrix. Prove that the matrix $\boldsymbol{I} - \boldsymbol{F}^*(\boldsymbol{F}\boldsymbol{F}^* + \boldsymbol{\Sigma})^{-1}\boldsymbol{F}$ is also positive definite. This result is used in the derivation of the EM algorithm in Sect. 9.7. (Hints: For readers familiar with the sweep operator of computational statistics, the simplest proof relies on applying Propositions 7.5.2 and 7.5.3 of the reference [166].)

23. A certain company asks consumers to rate movies on an integer scale from 1 to 5. Let M_i be the set of movies rated by person i. Denote the cardinality of M_i by $|M_i|$. Each rater does so in one of two modes that we will call "quirky" and "consensus". In quirky mode, i has a private rating distribution $(q_{i1}, q_{i2}, q_{i3}, q_{i4}, q_{i5})$ that applies to every movie regardless of its intrinsic merit. In consensus mode, rater i rates movie j according to the distribution $(c_{j1}, c_{j2}, c_{j3}, c_{j4}, c_{j5})$ shared with all other raters in consensus mode. For every movie i rates, he or she makes a quirky decision with probability π_i and a consensus decision with probability $1 - \pi_i$. These decisions are made independently across raters and movies. If x_{ij} is the rating given to movie j by rater i, then prove that the likelihood of the data is

$$L = \prod_i \prod_{j \in M_i} [\pi_i q_{ix_{ij}} + (1 - \pi_i) c_{jx_{ij}}].$$

Once we estimate the parameters, we can rank the reliability of rater i by the estimate $\hat{\pi}_i$ and the popularity of movie j by its estimated average rating $\sum_k k\hat{c}_{jk}$.

If we choose the natural course of estimating the parameters by maximum likelihood, then it is possible to derive an EM or MM algorithm. From the right perspectives, these two algorithms coincide. Let n denote iteration number and w_{nij} the weight

$$w_{nij} = \frac{\pi_{ni} q_{nix_{ij}}}{\pi_{ni} q_{nix_{ij}} + (1 - \pi_{ni}) c_{njx_{ij}}}.$$

Derive either algorithm and show that it updates the parameters by

$$\begin{aligned}
\pi_{n+1,i} &= \frac{1}{|M_i|} \sum_{j \in M_i} w_{nij} \\
q_{n+1,ix} &= \frac{\sum_{j \in M_i} 1_{\{x_{ij} = x\}} w_{nij}}{\sum_{j \in M_i} w_{nij}} \\
c_{n+1,jx} &= \frac{\sum_i 1_{\{x_{ij} = x\}} (1 - w_{nij})}{\sum_i (1 - w_{nij})}.
\end{aligned}$$

These updates are easy to implement. Can you motivate them as ratios of expected counts?

24. In the hidden Markov chain model, suppose that the chain is time homogeneous with transition probabilities p_{jk}. Derive an EM algorithm for estimating the p_{jk} from one or more independent runs of the chain.

25. In the hidden Markov chain model, consider estimation of the parameters of the conditional densities $\phi_i(y_i \mid j)$ of the observed data $y_1, \ldots, y_n$. When Y_i given $Z_i = j$ is Poisson distributed with mean μ_j, show that the EM algorithm updates μ_j by

$$\mu_{m+1,j} = \frac{\sum_{i=1}^n w_{mij} y_i}{\sum_{i=1}^n w_{mij}},$$

where the weight $w_{mij} = \mathrm{E}(X_{ij} \mid Y, \boldsymbol{\mu}_m)$. Show that the same update applies when Y_i given $Z_i = i$ is exponentially distributed with mean μ_j or normally distributed with mean μ_j and common variance σ^2. In the latter setting, demonstrate that the EM update of σ^2 is

$$\sigma^2_{m+1} = \frac{\sum_{i=1}^n \sum_j w_{mij}(y_i - \mu_{m+1,j})^2}{\sum_{i=1}^n \sum_j w_{mij}}.$$

10
Newton's Method and Scoring

10.1 Introduction

Block relaxation and the MM algorithm are hardly the only methods of optimization. Newton's method is better known and more widely applied. Despite its defects, Newton's method is the gold standard for speed of convergence and forms the basis of most modern optimization algorithms in low dimensions. Its many variants seek to retain its fast convergence while taming its defects. The variants all revolve around the core idea of locally approximating the objective function by a strictly convex quadratic function. At each iteration the quadratic approximation is optimized. Safeguards are introduced to keep the iterates from veering toward irrelevant stationary points.

Statisticians are among the most avid consumers of optimization techniques. Statistics, like other scientific disciplines, has a special vocabulary. We will meet some of that vocabulary in this chapter as we discuss optimization methods important in computational statistics. Thus, we will take up Fisher's scoring algorithm and the Gauss-Newton method of nonlinear least squares. We have already encountered likelihood functions and the device of passing to loglikelihoods. In statistics, the gradient of the loglikelihood is called the score, and the negative of the second differential is called the observed information. One major advantage of maximizing the loglikelihood rather than the likelihood is that the loglikelihood, score, and observed information are all additive functions of independent observations.

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_10,

10.2 Newton's Method and Root Finding

One of the virtues of Newton's method is that it is a root-finding technique as well as an optimization technique. Consider a function $f(\boldsymbol{x})$ mapping R^n into R^n, and suppose a root of $f(\boldsymbol{x}) = \mathbf{0}$ occurs at $\boldsymbol{y}$. If the slope matrix in the expansion

$$\begin{aligned} \mathbf{0} - f(\boldsymbol{x}) &= f(\boldsymbol{y}) - f(\boldsymbol{x}) \\ &= s(\boldsymbol{y}, \boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \end{aligned}$$

is invertible, then we can solve for $\boldsymbol{y}$ as

$$\boldsymbol{y} = \boldsymbol{x} - s(\boldsymbol{y}, \boldsymbol{x})^{-1} f(\boldsymbol{x}).$$

In practice, $\boldsymbol{y}$ is unknown and the slope $s(\boldsymbol{y}, \boldsymbol{x})$ is unavailable. However, if $\boldsymbol{y}$ is close to $\boldsymbol{x}$, then $s(\boldsymbol{y}, \boldsymbol{x})$ should be close to $df(\boldsymbol{x})$. Thus, Newton's method iterates according to

$$\boldsymbol{x}_{m+1} = \boldsymbol{x}_m - df(\boldsymbol{x}_m)^{-1} f(\boldsymbol{x}_m). \tag{10.1}$$

Example 10.2.1 *Division without Dividing*

Forming the reciprocal of a number $a > 0$ is equivalent to solving for a root of the equation $f(x) = a - x^{-1}$. Newton's method (10.1) iterates according to

$$x_{m+1} = x_m - \frac{a - x_m^{-1}}{x_m^{-2}} = x_m(2 - a x_m),$$

which involves multiplication and subtraction but no division. If x_{m+1} is to be positive, then x_m must lie on the interval $(0, 2/a)$. If x_m does indeed reside there, then x_{m+1} will reside on the shorter interval $(0, 1/a)$ because the quadratic $x(2 - ax)$ attains its maximum of $1/a$ at $x = 1/a$. Furthermore, $x_{m+1} > x_m$ if and only if $2 - a x_m > 1$, and this latter inequality holds if and only if $x_m < 1/a$. Thus, starting on $(0, 1/a)$, the iterates x_m monotonically increase to their limit $1/a$. Starting on $[1/a, 2/a)$, the first iterate satisfies $x_1 \leq 1/a$, and subsequent iterates monotonically increase to $1/a$. Finally, starting outside $(0, 2/a)$ leads either to fixation at 0 or divergence to $-\infty$. ■

Example 10.2.2 *Extraction of nth Roots*

Newton's method can be used to extract square roots, cube roots, and so forth. Consider the function $f(x) = x^n - a$ for some integer $n > 1$ and $a > 0$. Newton's method amounts to the iteration scheme

$$x_{m+1} = x_m - \frac{x_m^n - a}{n x_m^{n-1}} = \frac{1}{n}\left[(n-1)x_m + \frac{a}{x_m^{n-1}}\right]. \tag{10.2}$$

TABLE 10.1. Newton's iterates for $x^4 - x^2$

m	x_m	x_m	x_m
0	−0.74710	−0.66710	−0.500000
1	−2.16581	1.01669	−0.125000
2	−1.68896	1.00066	−0.061491
3	−1.35646	1.00000	−0.030628
4	−1.14388	1.00000	−0.015300
5	−1.03477	1.00000	−0.007648
6	−1.00270	1.00000	−0.003823
7	−1.00002	1.00000	−0.001911
8	−1.00000	1.00000	−0.000955
9	−1.00000	1.00000	−0.000477
10	−1.00000	1.00000	−0.000238

This sequence converges to $\sqrt[n]{a}$ regardless of the starting point $x_0 > 0$. To demonstrate this fact, we first note that the right-hand side of equation (10.2) is the arithmetic mean of $n-1$ copies of the number x_m and a/x_m^{n-1}. Because the arithmetic mean exceeds the geometric mean $\sqrt[n]{a}$, it follows that $x_m \geq \sqrt[n]{a}$ for all $m \geq 1$. Given this inequality, we have $a/x_m^{n-1} \leq x_m$. Again viewing equation (10.2) as a weighted average of x_m and the ratio $a/x_m^{n-1} \leq x_m$, it follows that $x_{m+1} \leq x_m$ for all $m \geq 1$. Hence, the sequence $x_1, x_2, \ldots$ is bounded below and is monotonically decreasing. By continuity, its limit is $\sqrt[n]{a}$. ■

Example 10.2.3 *Sensitivity to Initial Conditions*

In contrast to the previous two well-behaved examples, finding a root of the polynomial $f(x) = x^4 - x^2$ is more problematic. These roots are clearly -1, 0, and 1. We anticipate trouble when $f'(x) = 4x^3 - 2x = 0$ at the points $-1/\sqrt{2}$, 0, and $1/\sqrt{2}$. Consider initial points near $-1/\sqrt{2}$. Just to the left of $-1/\sqrt{2}$, Newton's method converges to -1. For a narrow zone just to the right of $-1/\sqrt{2}$, it converges to 1, and beyond this zone but to the left of $1/\sqrt{2}$, it converges to 0. Table 10.1 gives three typical examples of this extreme sensitivity to initial conditions. The slower convergence to the middle root 0 is hardly surprising given that $f'(0) = 0$. It appears that the discrepancy $x_m - 0$ roughly halves at each iteration. Problem 2 clarifies this behavior. ■

Example 10.2.4 *Secant Method*

There are several ways of estimating the differential $df(\boldsymbol{x}_m)$ appearing in Newton's formula (10.1). In one dimension the secant method approximates $f'(x_m)$ by the slope $s(x_{m-1}, x_m)$ using the canonical slope function

$$s(y, x) = \frac{f(y) - f(x)}{y - x}.$$

This produces the secant update

$$x_{m+1} = x_m - \frac{f(x_m)(x_{m-1} - x_m)}{f(x_{m-1}) - f(x_m)}, \tag{10.3}$$

which is the prototype for the quasi-Newton updates treated in Chap. 11. While the secant method avoids computation of derivatives, it typically takes more iterations to converge than Newton's method. Safeguards must also be put into place to ensure its reliability. ■

Example 10.2.5 *Newton's Method of Matrix Inversion*

Newton's method for finding the reciprocal of a number can be generalized to compute the inverse of a matrix [138]. Consider the matrix-valued function $f(\boldsymbol{B}) = \boldsymbol{A} - \boldsymbol{B}^{-1}$ for some invertible $n \times n$ matrix $\boldsymbol{A}$. Example 5.4.5 provides the first-order differential approximation

$$f(\boldsymbol{A}^{-1}) - f(\boldsymbol{B}) = \boldsymbol{0} - (\boldsymbol{A} - \boldsymbol{B}^{-1}) \approx \boldsymbol{B}^{-1}(\boldsymbol{A}^{-1} - \boldsymbol{B})\boldsymbol{B}^{-1}.$$

Multiplying this equation on the left and right by $\boldsymbol{B}$ and rearranging give

$$\boldsymbol{A}^{-1} \approx 2\boldsymbol{B} - \boldsymbol{B}\boldsymbol{A}\boldsymbol{B}$$

and hence Newton's scheme

$$\boldsymbol{B}_{m+1} = 2\boldsymbol{B}_m - \boldsymbol{B}_m\boldsymbol{A}\boldsymbol{B}_m.$$

Further rearrangement yields

$$\boldsymbol{A}^{-1} - \boldsymbol{B}_{m+1} = (\boldsymbol{A}^{-1} - \boldsymbol{B}_m)\boldsymbol{A}(\boldsymbol{A}^{-1} - \boldsymbol{B}_m),$$

which entails

$$\|\boldsymbol{A}^{-1} - \boldsymbol{B}_{m+1}\| \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{A}^{-1} - \boldsymbol{B}_m\|^2$$

for every matrix norm. It follows that the sequence $\boldsymbol{B}_m$ converges at a quadratic rate to $\boldsymbol{A}^{-1}$ if $\boldsymbol{B}_0$ is sufficiently close to $\boldsymbol{A}^{-1}$. ■

10.3 Newton's Method and Optimization

Suppose we want to minimize the real-valued function $f(\boldsymbol{x})$ defined on an open set $S \subset \mathsf{R}^n$. Assuming that $f(\boldsymbol{x})$ is twice differentiable, the expansion

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + df(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^* s^2(\boldsymbol{y}, \boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})$$

suggests that we substitute $d^2 f(\boldsymbol{x})$ for the second slope $s^2(\boldsymbol{y}, \boldsymbol{x})$ and approximate $f(\boldsymbol{y})$ by the resulting quadratic. If we take this approximation seriously, then we can solve for its minimum point $\boldsymbol{y}$ as

$$\boldsymbol{y} = \boldsymbol{x} - d^2 f(\boldsymbol{x})^{-1} \nabla f(\boldsymbol{x}).$$

In Newton's method we iterate according to

$$\boldsymbol{x}_{m+1} = \boldsymbol{x}_m - d^2 f(\boldsymbol{x}_m)^{-1} \nabla f(\boldsymbol{x}_m). \tag{10.4}$$

It should come as no surprise that algorithm (10.4) coincides with the earlier version of Newton's method seeking a root of $\nabla f(\boldsymbol{x})$. For this reason, any stationary point $\boldsymbol{x}$ of $f(\boldsymbol{x})$ is a fixed point of algorithm (10.4).

Example 10.3.1 *Newton's Method for the Poisson Multigraph Model*

Newton's method can be applied to the Poisson multigraph model introduced in Sect. 8.11. The score vector has entries

$$\frac{\partial}{\partial p_i} L(\boldsymbol{p}) = \sum_{j \neq i} \left(\frac{x_{ij}}{p_i} - p_j \right),$$

and the observed information matrix has entries

$$-\frac{\partial^2}{\partial p_i \partial p_j} L(\boldsymbol{p}) = \begin{cases} 1 & i \neq j \\ \frac{1}{p_i^2} \sum_{k \neq i} x_{ik} & i = j. \end{cases}$$

For n nodes the matrix $-d^2 L(\boldsymbol{p})$ is $n \times n$, and inverting it seems out of the question when n is large. Fortunately, the Sherman-Morrison formula comes to the rescue. If we write $-d^2 L(\boldsymbol{p})$ as $\boldsymbol{D} + \boldsymbol{1}\boldsymbol{1}^*$ with $\boldsymbol{D}$ diagonal, then the explicit inverse

$$(\boldsymbol{D} + \boldsymbol{1}\boldsymbol{1}^*)^{-1} = \boldsymbol{D}^{-1} - \frac{1}{1 + \boldsymbol{1}^* \boldsymbol{D}^{-1} \boldsymbol{1}} \boldsymbol{D}^{-1} \boldsymbol{1}\boldsymbol{1}^* \boldsymbol{D}^{-1}$$

is available. This makes Newton's method trivial to implement as long as one respects the bounds $p_i \geq 0$. More generally, it is always cheap to invert a low-rank perturbation of an explicitly invertible matrix. See Problem 10 of Chap. 11 for Woodbury's generalization of the Sherman-Morrison formula. ■

There are two potential problems with Newton's method. First, it may be expensive computationally to evaluate or invert $d^2 f(\boldsymbol{x})$. Second, far from the minimum, Newton's method is equally happy to head uphill or down. In other words, Newton's method is not a descent algorithm in the sense that $f(\boldsymbol{x}_{m+1}) < f(\boldsymbol{x}_m)$. This second defect can be remedied by modifying the Newton increment so that it is a partial step in a descent direction. A descent direction $\boldsymbol{v}$ at the point $\boldsymbol{x}$ satisfies the inequality $df(\boldsymbol{x})\boldsymbol{v} < 0$. The formula

$$\lim_{t \downarrow 0} \frac{f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x})}{t} = df(\boldsymbol{x})\boldsymbol{v}$$

for the forward directional derivative shows that $f(\boldsymbol{x} + t\boldsymbol{v}) < f(\boldsymbol{x})$ for $t > 0$ sufficiently small. The key to generating a descent direction is to define

$\boldsymbol{v} = -\boldsymbol{H}^{-1}\nabla f(\boldsymbol{x})$ using a positive definite matrix $\boldsymbol{H}$. Here we assume that $\boldsymbol{x}$ is not a stationary point and recall the fact that the inverse of a positive definite matrix is positive definite.

The necessary modifications of Newton's method to achieve a descent algorithm are now clear. We simply replace $d^2 f(\boldsymbol{x}_m)$ by a positive definite approximating matrix $\boldsymbol{H}_m$ and take a sufficiently short step in the direction $\Delta\boldsymbol{x}_m = -\boldsymbol{H}_m^{-1}\nabla f(\boldsymbol{x}_m)$. If we believe that the proposed increment is reasonable, then we will be reluctant to shrink $\Delta\boldsymbol{x}_m$ too much. This suggests backtracking, the simplest form of which is step halving. In step halving, if the initial increment $\Delta\boldsymbol{x}_m$ does not produce a decrease in $f(\boldsymbol{x})$, then try $\Delta\boldsymbol{x}_m/2$. If $\Delta\boldsymbol{x}_m/2$ fails, then try $\Delta\boldsymbol{x}_m/4$, and so forth. We will meet more sophisticated backtracking schemes later. Note at this juncture we have said nothing about how well $\boldsymbol{H}_m$ approximates $d^2 f(\boldsymbol{x}_m)$. The quality of this approximation obviously affects the rate of convergence toward any local minimum.

If we minimize $f(\boldsymbol{x})$ subject to the linear equality constraints $\boldsymbol{V}\boldsymbol{x} = \boldsymbol{d}$, then minimization of the approximating quadratic can be accomplished as indicated in Example 5.2.6 of Chap. 5. Because

$$\boldsymbol{V}(\boldsymbol{x}_{m+1} - \boldsymbol{x}_m) \;=\; \boldsymbol{0},$$

the revised increment $\Delta\boldsymbol{x}_m = \boldsymbol{x}_{m+1} - \boldsymbol{x}_m$ is

$$\Delta\boldsymbol{x}_m \;=\; -\left[\boldsymbol{H}_m^{-1} - \boldsymbol{H}_m^{-1}\boldsymbol{V}^*(\boldsymbol{V}\boldsymbol{H}_m^{-1}\boldsymbol{V}^*)^{-1}\boldsymbol{V}\boldsymbol{H}_m^{-1}\right]\nabla f(\boldsymbol{x}_m). \qquad (10.5)$$

This can be viewed as the projection of the unconstrained increment onto the null space of $\boldsymbol{V}$. Problem 12 shows that step halving also works for the projected increment.

10.4 MM Gradient Algorithm

Often it is impossible to solve the optimization step of the MM algorithm exactly. If $f(\boldsymbol{x})$ is the objective function and $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ minorizes or majorizes $f(\boldsymbol{x})$ at $\boldsymbol{x}_m$, then Newton's method can be applied to optimize $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$. As we shall see later, one step of Newton's method preserves the overall rate of convergence of the MM algorithm. Thus, the MM gradient algorithm iterates according to

$$\begin{aligned}\boldsymbol{x}_{m+1} &= \boldsymbol{x}_m - d^2 g(\boldsymbol{x}_m \mid \boldsymbol{x}_m)^{-1}\nabla g(\boldsymbol{x}_m \mid \boldsymbol{x}_m) \\ &= \boldsymbol{x}_m - d^2 g(\boldsymbol{x}_m \mid \boldsymbol{x}_m)^{-1}\nabla f(\boldsymbol{x}_m).\end{aligned}$$

Here derivatives are taken with respect to the left argument of $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$. Substitution of $\nabla f(\boldsymbol{x}_m)$ for $\nabla g(\boldsymbol{x}_m \mid \boldsymbol{x}_m)$ is justified by the comments in Sect. 8.2. In practice the surrogate function $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ is either convex or concave, and its second differential $d^2 g(\boldsymbol{x}_m \mid \boldsymbol{x}_m)$ needs no adjustment to give a descent or ascent algorithm.

Example 10.4.1 *Newton's Method in Transmission Tomography*

In the transmission tomography model of Chap. 8, the surrogate function $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$ of equation (8.15) minorizes the loglikelihood $L(\boldsymbol{\theta})$ in the absence of a smoothing prior. Differentiating $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$ with respect to θ_j gives the transcendental equation

$$0 = \sum_i l_{ij}\left[d_i e^{-\boldsymbol{l}_i^* \boldsymbol{\theta}_m \theta_j/\theta_{mj}} - y_i\right].$$

One step of Newton's method starting at $\theta_j = \theta_{mj}$ produces the next iterate

$$\begin{aligned}\theta_{m+1,j} &= \theta_{mj} + \frac{\theta_{mj}\sum_i l_{ij}(d_i e^{-\boldsymbol{l}_i^*\boldsymbol{\theta}_m} - y_i)}{\sum_i l_{ij}\boldsymbol{l}_i^*\boldsymbol{\theta}_m d_i e^{-\boldsymbol{l}_i^*\boldsymbol{\theta}_m}} \\ &= \theta_{mj}\frac{\sum_i l_{ij}[d_i e^{-\boldsymbol{l}_i^*\boldsymbol{\theta}_m}(1+\boldsymbol{l}_i^*\boldsymbol{\theta}_m) - y_i]}{\sum_i l_{ij}\boldsymbol{l}_i^*\boldsymbol{\theta}_m d_i e^{-\boldsymbol{l}_i^*\boldsymbol{\theta}_m}}.\end{aligned}$$

This step typically increases $L(\boldsymbol{\theta})$. The comparable EM gradient update involves solving the transcendental equation (9.8). ■

Example 10.4.2 *Estimation with the Dirichlet Distribution*

As another example, consider parameter estimation for the Dirichlet distribution [154]. This distribution has probability density

$$\frac{\Gamma(\sum_{i=1}^n \theta_i)}{\prod_{i=1}^n \Gamma(\theta_i)} \prod_{i=1}^n y_i^{\theta_i - 1} \tag{10.6}$$

on the unit simplex $\{\boldsymbol{y} = (y_1, \ldots, y_n)^* : y_1 > 0, \ldots, y_n > 0, \sum_{i=1}^n y_i = 1\}$ endowed with the uniform measure. The Dirichlet distribution is used to represent random proportions. The beta distribution is the special case $n = 2$.

If $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_l$ are randomly sampled vectors from the Dirichlet distribution, then their loglikelihood is

$$L(\boldsymbol{\theta}) = l \ln \Gamma\Big(\sum_{i=1}^n \theta_i\Big) - l\sum_{i=1}^n \ln \Gamma(\theta_i) + \sum_{k=1}^l \sum_{i=1}^n (\theta_i - 1)\ln y_{ki}.$$

Except for the first term on the right, the parameters are separated. Fortunately as demonstrated in Example 6.3.11, the function $\ln \Gamma(t)$ is convex. Denoting its derivative by $\psi(t)$, we exploit the minorization

$$\ln \Gamma\Big(\sum_{i=1}^n \theta_i\Big) \geq \ln \Gamma\Big(\sum_{i=1}^n \theta_{mi}\Big) + \psi\Big(\sum_{i=1}^n \theta_{mi}\Big)\sum_{i=1}^n (\theta_i - \theta_{mi})$$

and create the surrogate function

$$\begin{aligned} g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m) &= l \ln \Gamma\Big(\sum_{i=1}^n \theta_{mi}\Big) + l\psi\Big(\sum_{i=1}^n \theta_{mi}\Big) \sum_{i=1}^n (\theta_i - \theta_{mi}) \\ &\quad - l \sum_{i=1}^n \ln \Gamma(\theta_i) + \sum_{k=1}^l \sum_{i=1}^n (\theta_i - 1) \ln y_{ki}. \end{aligned}$$

Owing to the presence of the terms $\ln \Gamma(\theta_i)$, the maximization step is intractable. However, the MM gradient algorithm can be readily implemented because the parameters are now separated and the functions $\psi(t)$ and $\psi'(t)$ are easily computed as suggested in Problem 14. The whole process is carried out in the references [163, 199] on actual data. ■

10.5 Ad Hoc Approximations of $d^2 f(\theta)$

In minimization problems, we have emphasized the importance of approximating $d^2 f(\boldsymbol{\theta})$ by a positive definite matrix. Three key ideas drive the process of approximation. One is the recognition that outer product matrices are positive semidefinite. Another is a feel for when terms are small on average. Usually this involves comparison of random variables and their means. Finally, it is almost always advantageous to avoid the explicit calculation of complicated second derivatives.

For example, consider the problem of least squares estimation with nonlinear regression functions. Let us formulate the problem slightly more generally as one of minimizing the sum of squares

$$f(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^n w_i [y_i - \mu_i(\boldsymbol{\theta})]^2$$

involving a weight $w_i > 0$ and response y_i for each case i. Here y_i is a realization of a random variable Y_i with mean $\mu_i(\boldsymbol{\theta})$. In linear regression, $\mu_i(\boldsymbol{\theta}) = \sum_k x_{ik}\theta_k$. To implement Newton's method, we need

$$\begin{aligned} \nabla f(\theta) &= -\sum_{i=1}^n w_i [y_i - \mu_i(\boldsymbol{\theta})] \nabla \mu_i(\boldsymbol{\theta}) \\ d^2 f(\boldsymbol{\theta}) &= \sum_{i=1}^n w_i \nabla \mu_i(\boldsymbol{\theta}) d\mu_i(\boldsymbol{\theta}) - \sum_{i=1}^n w_i [y_i - \mu_i(\boldsymbol{\theta})] d^2 \mu_i(\boldsymbol{\theta}). \quad (10.7) \end{aligned}$$

In the Gauss-Newton algorithm, we approximate

$$d^2 f(\boldsymbol{\theta}) \approx \sum_{i=1}^n w_i \nabla \mu_i(\boldsymbol{\theta}) d\mu_i(\boldsymbol{\theta})$$

on the rationale that either the weighted residuals $w_i[y_i - \mu_i(\boldsymbol{\theta})]$ are small or the regression functions $\mu_i(\boldsymbol{\theta})$ are nearly linear. In both instances, the Gauss-Newton algorithm shares the fast convergence of Newton's method.

Maximum likelihood estimation with the Poisson distribution furnishes another example. Here the count data $y_1, \ldots, y_n$ have loglikelihood, score, and negative observed information

$$\begin{aligned} L(\boldsymbol{\theta}) &= \sum_{i=1}^n [y_i \ln \lambda_i(\boldsymbol{\theta}) - \lambda_i(\boldsymbol{\theta}) - \ln y_i!] \\ \nabla L(\boldsymbol{\theta}) &= \sum_{i=1}^n \Big[\frac{y_i}{\lambda_i(\boldsymbol{\theta})} \nabla \lambda_i(\boldsymbol{\theta}) - \nabla \lambda_i(\boldsymbol{\theta})\Big] \\ d^2 L(\boldsymbol{\theta}) &= \sum_{i=1}^n \Big[- \frac{y_i}{\lambda_i(\boldsymbol{\theta})^2} \nabla \lambda_i(\boldsymbol{\theta}) d\lambda_i(\boldsymbol{\theta}) + \frac{y_i}{\lambda_i(\boldsymbol{\theta})} d^2 \lambda_i(\boldsymbol{\theta}) - d^2 \lambda_i(\boldsymbol{\theta})\Big], \end{aligned}$$

where $\mathrm{E}(y_i) = \lambda_i(\boldsymbol{\theta})$. Given that the ratio $y_i/\lambda_i(\boldsymbol{\theta})$ has average value 1, the negative semidefinite approximations

$$\begin{aligned} d^2 L(\boldsymbol{\theta}) &\approx - \sum_{i=1}^n \frac{y_i}{\lambda_i(\boldsymbol{\theta})^2} \nabla \lambda_i(\boldsymbol{\theta}) d\lambda_i(\boldsymbol{\theta}) \\ &\approx - \sum_{i=1}^n \frac{1}{\lambda_i(\boldsymbol{\theta})} \nabla \lambda_i(\boldsymbol{\theta}) d\lambda_i(\boldsymbol{\theta}) \end{aligned}$$

are reasonable. The second of these leads to the scoring algorithm discussed in the next section.

The exponential distribution offers a third illustration. Now the data have means $\mathrm{E}(y_i) = 1/\lambda_i(\boldsymbol{\theta})$. The loglikelihood

$$L(\boldsymbol{\theta}) = \sum_{i=1}^n [\ln \lambda_i(\boldsymbol{\theta}) - y_i \lambda_i(\boldsymbol{\theta})]$$

yields the score and negative observed information

$$\begin{aligned} \nabla L(\boldsymbol{\theta}) &= \sum_{i=1}^n \Big[\frac{1}{\lambda_i(\boldsymbol{\theta})} \nabla \lambda_i(\boldsymbol{\theta}) - y_i \nabla \lambda_i(\boldsymbol{\theta})\Big] \\ d^2 L(\boldsymbol{\theta}) &= \sum_{i=1}^n \Big[- \frac{1}{\lambda_i(\boldsymbol{\theta})^2} \nabla \lambda_i(\boldsymbol{\theta}) d\lambda_i(\boldsymbol{\theta}) + \frac{1}{\lambda_i(\boldsymbol{\theta})} d^2 \lambda_i(\boldsymbol{\theta}) - y_i d^2 \lambda_i(\boldsymbol{\theta})\Big]. \end{aligned}$$

Replacing observations by their means suggests the approximation

$$d^2 L(\boldsymbol{\theta}) \approx - \sum_{i=1}^n \frac{1}{\lambda_i(\boldsymbol{\theta})^2} \nabla \lambda_i(\boldsymbol{\theta}) d\lambda_i(\boldsymbol{\theta})$$

made in the scoring algorithm. Table 10.2 summarizes the scoring algorithm with means $\mu_i(\boldsymbol{\theta})$ replacing intensities $\lambda_i(\boldsymbol{\theta})$.

Our final example involves maximum likelihood estimation with the multinomial distribution. The observations $y_1, \ldots, y_j$ are now cell counts over n independent trials. Cell i is assigned probability $p_i(\boldsymbol{\theta})$ and averages a total of $np_i(\boldsymbol{\theta})$ counts. The loglikelihood, score, and negative observed information amount to

$$\begin{aligned} L(\boldsymbol{\theta}) &= \sum_{i=1}^{j} y_i \ln p_i(\boldsymbol{\theta}) \\ \nabla L(\boldsymbol{\theta}) &= \sum_{i=1}^{j} \frac{y_i}{p_i(\boldsymbol{\theta})} \nabla p_i(\boldsymbol{\theta}) \\ d^2 L(\boldsymbol{\theta}) &= \sum_{i=1}^{j} \Big[-\frac{y_i}{p_i(\boldsymbol{\theta})^2} \nabla p_i(\boldsymbol{\theta}) dp_i(\boldsymbol{\theta}) + \frac{y_i}{p_i(\boldsymbol{\theta})} d^2 p_i(\boldsymbol{\theta}) \Big]. \end{aligned}$$

In light of the identity $\mathrm{E}(y_i) = np_i(\boldsymbol{\theta})$, the approximation

$$\sum_{i=1}^{j} \frac{y_i}{p_i(\boldsymbol{\theta})} d^2 p_i(\boldsymbol{\theta}) \approx n \sum_{i=1}^{j} d^2 p_i(\boldsymbol{\theta}) = nd^2 1 = \mathbf{0}$$

is reasonable. This suggests the further negative semidefinite approximations

$$\begin{aligned} d^2 L(\boldsymbol{\theta}) &\approx -\sum_{i=1}^{j} \frac{y_i}{p_i(\boldsymbol{\theta})^2} \nabla p_i(\boldsymbol{\theta}) dp_i(\boldsymbol{\theta}) \\ &\approx -n \sum_{i=1}^{j} \frac{1}{p_i(\boldsymbol{\theta})} \nabla p_i(\boldsymbol{\theta}) dp_i(\boldsymbol{\theta}), \end{aligned}$$

the second of which coincides with the scoring algorithm.

10.6 Scoring and Exponential Families

As we have just witnessed, one can approximate the observed information in a variety of ways. The method of steepest ascent replaces the observed information by the identity matrix $\boldsymbol{I}$. The usually more efficient scoring algorithm replaces the observed information by the expected information $J(\boldsymbol{\theta}) = \mathrm{E}[-d^2 L(\boldsymbol{\theta})]$, where $L(\boldsymbol{\theta})$ is the loglikelihood. The alternative representation $J(\boldsymbol{\theta}) = \mathrm{Var}[\nabla L(\boldsymbol{\theta})]$ of $J(\boldsymbol{\theta})$ as a variance matrix shows that it is positive semidefinite and hence a good replacement for $-d^2 L(\boldsymbol{\theta})$ in Newton's method. An extra dividend of scoring is that the inverse matrix $J(\hat{\boldsymbol{\theta}})^{-1}$ immediately supplies the asymptotic variances and covariances of

the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ [218]. Scoring shares this benefit with Newton's method since the observed information is under natural assumptions asymptotically equivalent to the expected information.

To prove that $J(\boldsymbol{\theta}) = \text{Var}[\nabla L(\boldsymbol{\theta})]$, suppose the data has density $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ relative to some measure ν, which is usually ordinary volume measure or a discrete counting measure. We first note that the score conveniently has vanishing expectation because

$$\begin{aligned} \text{E}[\nabla L(\boldsymbol{\theta})] &= \int \frac{\nabla f(\boldsymbol{y} \mid \boldsymbol{\theta})}{f(\boldsymbol{y} \mid \boldsymbol{\theta})} f(\boldsymbol{y} \mid \boldsymbol{\theta}) \, d\nu(\boldsymbol{y}) \\ &= \nabla \int f(\boldsymbol{y} \mid \boldsymbol{\theta}) \, d\nu(\boldsymbol{y}) \end{aligned}$$

and $\int f(\boldsymbol{y} \mid \boldsymbol{\theta}) \, d\nu(\boldsymbol{y}) = 1$. Here the interchange of differentiation and expectation must be proved, but we will not stop to do so. See the references [176, 218]. The formal calculation

$$\begin{aligned} \text{E}[-d^2 L(\boldsymbol{\theta})] &= -\int \left[\frac{d^2 f(\boldsymbol{y} \mid \boldsymbol{\theta})}{f(\boldsymbol{y} \mid \boldsymbol{\theta})} - \frac{\nabla f(\boldsymbol{y} \mid \boldsymbol{\theta}) df(\boldsymbol{y} \mid \boldsymbol{\theta})}{f(\boldsymbol{y} \mid \boldsymbol{\theta})^2} \right] f(\boldsymbol{y} \mid \boldsymbol{\theta}) \, d\nu(\boldsymbol{y}) \\ &= -d^2 \int f(\boldsymbol{y} \mid \boldsymbol{\theta}) \, d\nu(\boldsymbol{y}) \\ &\quad + \int \nabla L(\boldsymbol{\theta}) dL(\boldsymbol{\theta}) f(\boldsymbol{y} \mid \boldsymbol{\theta}) \, d\nu(\boldsymbol{y}) \\ &= -\mathbf{0} + \text{E}[\nabla L(\boldsymbol{\theta}) dL(\boldsymbol{\theta})] \end{aligned}$$

then completes the verification.

The score and expected information simplify considerably for exponential families of densities [22, 43, 110, 146, 202]. Based on equation (9.1), the score and expected information can be expressed succinctly in terms of the mean vector $\mu(\boldsymbol{\theta}) = \text{E}[h(\boldsymbol{y})]$ and the variance matrix $\Sigma(\boldsymbol{\theta}) = \text{Var}[h(\boldsymbol{y})]$ of the sufficient statistic $h(\boldsymbol{y})$. Our point of departure in deriving these quantities is the identity

$$dL(\boldsymbol{\theta}) = d\beta(\boldsymbol{\theta}) + h(\boldsymbol{y})^* d\gamma(\boldsymbol{\theta}). \tag{10.8}$$

If $\gamma(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$, then $J(\boldsymbol{\theta}) = -d^2 L(\boldsymbol{\theta}) = -d^2 \beta(\boldsymbol{\theta})$, and scoring coincides with Newton's method. If in addition $J(\boldsymbol{\theta})$ is positive definite, then $L(\boldsymbol{\theta})$ is strictly concave and possesses at most a single local maximum, which is necessarily the global maximum.

For an exponential family, the fact that $\text{E}[\nabla L(\theta)] = \mathbf{0}$ can be restated as

$$d\beta(\boldsymbol{\theta}) + \mu(\boldsymbol{\theta})^* d\gamma(\boldsymbol{\theta}) = \mathbf{0}^*. \tag{10.9}$$

Subtracting equation (10.9) from equation (10.8) yields the alternative representation

$$dL(\boldsymbol{\theta}) = [h(\boldsymbol{y}) - \mu(\boldsymbol{\theta})]^* d\gamma(\boldsymbol{\theta}) \tag{10.10}$$

of the first differential. This representation implies that the expected information is

$$J(\boldsymbol{\theta}) = \operatorname{Var}[\nabla L(\boldsymbol{\theta})] = d\gamma(\boldsymbol{\theta})^* \Sigma(\boldsymbol{\theta}) d\gamma(\boldsymbol{\theta}). \tag{10.11}$$

To eliminate $d\gamma(\boldsymbol{\theta})$ in equations (10.10) and (10.11), note that

$$\begin{aligned} d\mu(\boldsymbol{\theta}) &= \int h(\boldsymbol{y}) df(\boldsymbol{y} \mid \boldsymbol{\theta})\, d\nu(\boldsymbol{y}) \\ &= \int h(\boldsymbol{y}) dL(\boldsymbol{\theta}) f(\boldsymbol{y} \mid \boldsymbol{\theta})\, d\nu(\boldsymbol{y}) \\ &= \int h(\boldsymbol{y})[h(\boldsymbol{y}) - \mu(\boldsymbol{\theta})]^* \, d\gamma(\boldsymbol{\theta}) f(\boldsymbol{y} \mid \boldsymbol{\theta})\, d\nu(\boldsymbol{y}) \\ &= \int [h(\boldsymbol{y}) - \mu(\boldsymbol{\theta})][h(\boldsymbol{y}) - \mu(\boldsymbol{\theta})]^* f(\boldsymbol{y} \mid \boldsymbol{\theta})\, d\nu(\boldsymbol{y})\, d\gamma(\boldsymbol{\theta}) \\ &= \Sigma(\boldsymbol{\theta}) d\gamma(\boldsymbol{\theta}). \end{aligned}$$

When $\Sigma(\boldsymbol{\theta})$ is invertible, this calculation implies $d\gamma(\boldsymbol{\theta}) = \Sigma(\boldsymbol{\theta})^{-1} d\mu(\boldsymbol{\theta})$, which in view of equations (10.10) and (10.11) yields

$$dL(\boldsymbol{\theta}) = [h(\boldsymbol{y}) - \mu(\boldsymbol{\theta})]^* \Sigma(\boldsymbol{\theta})^{-1} d\mu(\boldsymbol{\theta}) \tag{10.12}$$

$$J(\boldsymbol{\theta}) = d\mu(\boldsymbol{\theta})^* \Sigma(\boldsymbol{\theta})^{-1} d\mu(\boldsymbol{\theta}). \tag{10.13}$$

One can verify these formulas directly for the normal, Poisson, exponential, and multinomial distributions studied in the previous section. In each instance the sufficient statistic for case i is just y_i.

Based on equations (9.1), (10.12), and (10.13), Table 10.2 displays the loglikelihood, score vector, and expected information matrix for some commonly applied exponential families. In this table, x represents a single observation from the binomial, Poisson, and exponential families. For the multinomial family with m categories, $\boldsymbol{x} = (x_1, \ldots, x_m)^*$ gives the category counts. The quantity μ denotes the mean of x for the Poisson and exponential families. For the binomial family, we express the mean np as the product of the number of trials n and the success probability p per trial. A similar convention holds for the multinomial family.

The multinomial family deserves further comment. Straightforward calculation shows that the variance matrix $\Sigma(\boldsymbol{\theta})$ has entries

$$n[1_{\{i=j\}} p_i(\boldsymbol{\theta}) - p_i(\boldsymbol{\theta}) p_j(\boldsymbol{\theta})].$$

Here the matrix $\Sigma(\boldsymbol{\theta})$ is singular, so the generalized inverse applies in formulas (10.12) and (10.13). In this case it is easier to derive the expected information by taking the expectation of the observed information given in Sect. 10.5.

In the ABO allele frequency estimation problem studied in Chaps. 8 and 9, scoring can be implemented by taking as basic parameters p_A and p_B

TABLE 10.2. Score and information for some exponential families

Family	$L(\boldsymbol{\theta})$	$\nabla L(\boldsymbol{\theta})$	$J(\boldsymbol{\theta})$
Binomial	$x \ln \frac{p}{1-p} + n \ln(1-p)$	$\frac{x-np}{p(1-p)} \nabla p$	$\frac{n}{p(1-p)} \nabla p dp$
Multinomial	$\sum_i x_i \ln p_i$	$\sum_i \frac{x_i}{p_i} \nabla p_i$	$\sum_i \frac{n}{p_i} \nabla p_i dp_i$
Poisson	$-\mu + x \ln \mu$	$-\nabla \mu + \frac{x}{\mu} \nabla \mu$	$\frac{1}{\mu} \nabla \mu d\mu$
Exponential	$-\ln \mu - \frac{x}{\mu}$	$-\frac{1}{\mu} \nabla \mu + \frac{x}{\mu^2} \nabla \mu$	$\frac{1}{\mu^2} \nabla \mu d\mu$

and expressing $p_O = 1 - p_A - p_B$. Scoring then leads to the same maximum likelihood point $(\hat{p}_A, \hat{p}_B, \hat{p}_O) = (.2136, .0501, .7363)$ as the EM algorithm. The quicker convergence of scoring here—four iterations as opposed to five starting from $(.3, .2, .5)$—is often more dramatic in other problems. Scoring also has the advantage over EM of immediately providing asymptotic standard deviations of the parameter estimates. These are $(.0135, .0068, .0145)$ for the estimates $(\hat{p}_A, \hat{p}_B, \hat{p}_O)$.

10.7 The Gauss-Newton Algorithm

Armed with our better understanding of scoring, let us revisit nonlinear regression. Suppose that the n independent responses $y_1, \ldots, y_n$ are normally distributed with means $\mu_i(\boldsymbol{\theta})$ and variances σ^2/w_i, where the w_i are known constants. To estimate the mean parameter vector $\boldsymbol{\theta}$ and the variance parameter σ^2 by scoring, we first write the loglikelihood up to a constant as the function

$$L(\boldsymbol{\phi}) \;=\; -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} w_i [y_i - \mu_i(\boldsymbol{\theta})]^2 \;=\; -\frac{n}{2} \ln \sigma^2 - \frac{f(\boldsymbol{\theta})}{\sigma^2}$$

of the parameters $\boldsymbol{\phi} = (\boldsymbol{\theta}^*, \sigma^2)^*$.

Straightforward differentiation yields the score

$$\nabla L(\phi) \;=\; \begin{pmatrix} \frac{1}{\sigma^2} \sum_{i=1}^{n} w_i [y_i - \mu_i(\boldsymbol{\theta})] \nabla \mu_i(\boldsymbol{\theta}) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} w_i [y_i - \mu_i(\boldsymbol{\theta})]^2 \end{pmatrix}.$$

To derive the expected information

$$J(\boldsymbol{\phi}) \;=\; \begin{pmatrix} \frac{1}{\sigma^2} \sum_{i=1}^{n} w_i \nabla \mu_i(\boldsymbol{\theta}) d\mu_i(\boldsymbol{\theta}) & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix},$$

we note that the observed information matrix can be written as the symmetric block matrix

$$-d^2L(\phi) \;=\; \begin{pmatrix} \boldsymbol{H}_{\boldsymbol{\theta\theta}} & \boldsymbol{H}_{\boldsymbol{\theta}\sigma^2} \\ \boldsymbol{H}_{\sigma^2\boldsymbol{\theta}} & H_{\sigma^2\sigma^2} \end{pmatrix}.$$

The upper-left block $H_{\boldsymbol{\theta\theta}}$ equals $d^2f(\boldsymbol{\theta})/\sigma^2$ with $d^2f(\boldsymbol{\theta})$ given by equation (10.7). The displayed value of the expectation $\mathrm{E}(\boldsymbol{H}_{\boldsymbol{\theta\theta}})$ follows directly from the identity $\mathrm{E}[y_i - \mu_i(\boldsymbol{\theta})] = 0$. The upper-right block $\boldsymbol{H}_{\boldsymbol{\theta}\sigma^2}$ amounts to $-\nabla f(\boldsymbol{\theta})/\sigma^4$, and its expectation vanishes because again $\mathrm{E}[y_i - \mu_i(\boldsymbol{\theta})] = 0$. Finally, the lower-right block $H_{\sigma^2\sigma^2}$ equals

$$-\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{i=1}^n w_i[y_i - \mu_i(\boldsymbol{\theta})]^2.$$

Its expectation

$$\mathrm{E}(H_{\sigma^2\sigma^2}) \;=\; -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{i=1}^n w_i\,\mathrm{E}\{[y_i - \mu_i(\boldsymbol{\theta})]^2\} \;=\; \frac{n}{2\sigma^4}$$

because $\mathrm{Var}(y_i) = \sigma^2/w_i$. Readers experienced in calculating variances and covariances can verify the blocks of $J(\boldsymbol{\theta})$ by forming $\mathrm{Var}[\nabla L(\boldsymbol{\theta})]$.

In any event, scoring updates θ by

$$\begin{aligned} &\boldsymbol{\theta}_{m+1} \\ =\;& \boldsymbol{\theta}_m + \Big[\sum_{i=1}^n w_i \nabla\mu_i(\boldsymbol{\theta}_m) d\mu(\boldsymbol{\theta}_m)\Big]^{-1} \sum_{i=1}^n w_i[y_i - \mu_i(\boldsymbol{\theta}_m)]\nabla\mu_i(\boldsymbol{\theta}_m) \end{aligned} \tag{10.14}$$

and σ^2 by

$$\sigma^2_{m+1} \;=\; \frac{1}{n}\sum_{i=1}^n w_i[y_i - \mu_i(\boldsymbol{\theta}_m)]^2.$$

The scoring algorithm (10.14) for $\boldsymbol{\theta}$ amounts to nothing more than the Gauss-Newton algorithm. The Gauss-Newton updates can be carried out blithely neglecting the updates of σ^2.

10.8 Generalized Linear Models

The generalized linear model [202] deals with exponential families (9.1) in which the sufficient statistic $h(y)$ is y and the mean $\mu(\boldsymbol{\theta})$ of y completely determines the distribution of y. In many applications it is natural to postulate that $\mu(\boldsymbol{\theta}) = q(\boldsymbol{x}^*\boldsymbol{\theta})$ is a monotone function $q(s)$ of some linear combination of known predictors $\boldsymbol{x}$. The inverse of $q(s)$ is called the

TABLE 10.3. AIDS data from Australia during 1983–1986

Quarter	Deaths	Quarter	Deaths	Quarter	Deaths
1	0	6	4	11	20
2	1	7	9	12	25
3	2	8	18	13	37
4	3	9	23	14	45
5	1	10	31		

link function. In this setting, $d\mu(\boldsymbol{\theta}) = q'(\boldsymbol{x}^*\boldsymbol{\theta})\boldsymbol{x}^*$. It follows from equations (10.12) and (10.13) that if $y_1, \ldots, y_j$ are independent responses with corresponding predictor vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_j$, then the score and expected information can be written as

$$\begin{aligned} \nabla L(\boldsymbol{\theta}) &= \sum_{i=1}^{j} \frac{y_i - \mu_i(\boldsymbol{\theta})}{\sigma_i^2(\boldsymbol{\theta})} q'(\boldsymbol{x}_i^*\boldsymbol{\theta})\boldsymbol{x}_i \\ J(\boldsymbol{\theta}) &= \sum_{i=1}^{j} \frac{1}{\sigma_i^2(\boldsymbol{\theta})} q'(\boldsymbol{x}_i^*\boldsymbol{\theta})^2 \boldsymbol{x}_i \boldsymbol{x}_i^*, \end{aligned}$$

where $\sigma_i^2(\boldsymbol{\theta}) = \text{Var}(y_i)$.

Table 10.3 contains quarterly data on AIDS deaths in Australia that illustrate the application of a generalized linear model [73, 273]. A simple plot of the data suggests exponential growth. A plausible model therefore involves Poisson distributed observations y_i with means $\mu_i(\boldsymbol{\theta}) = e^{\theta_1 + i\theta_2}$. Because this parameterization renders scoring equivalent to Newton's method, scoring gives the quick convergence noted in Table 10.4.

10.9 Accelerated MM

We now consider the question of how to accelerate the often excruciatingly slow convergence of the MM algorithm. The simplest device is to just double each MM step [61, 163]. Thus, if $F(\boldsymbol{x}_m)$ is the MM algorithm map from R^p to R^p, then we move to $\boldsymbol{x}_m + 2[F(\boldsymbol{x}_m) - \boldsymbol{x}_m]$ rather than to $F(\boldsymbol{x}_m)$. Step doubling is a standard tactic that usually halves the number of iterations until convergence. However, in many problems something more radical is necessary. Because Newton's method enjoys exceptionally quick convergence in a neighborhood of the optimal point, an attractive strategy is to amend the MM algorithm so that it resembles Newton's method. The papers [144, 145, 164] take up this theme from the perspective of optimizing the objective function by Newton's method. It is also possible to apply Newton's method to find a root of the equation $\mathbf{0} = \boldsymbol{x} - F(\boldsymbol{x})$. This alternative perspective has the advantage of dealing directly with the iterates of

TABLE 10.4. Scoring iterates for the AIDS model

Iteration	Step halves	θ_1	θ_2
1	0	0.0000	0.0000
2	3	−1.3077	0.4184
3	0	0.6456	0.2401
4	0	0.3744	0.2542
5	0	0.3400	0.2565
6	0	0.3396	0.2565

the MM algorithm. Let $G(\boldsymbol{x})$ denote the difference $G(\boldsymbol{x}) = \boldsymbol{x} - F(\boldsymbol{x})$. Because $G(\boldsymbol{x})$ has differential $dG(\boldsymbol{x}) = \boldsymbol{I} - dF(\boldsymbol{x})$, Newton's method iterates according to

$$\begin{aligned} \boldsymbol{x}_{m+1} &= \boldsymbol{x}_m - dG(\boldsymbol{x}_m)^{-1} G(\boldsymbol{x}_m) \\ &= \boldsymbol{x}_m - [\boldsymbol{I} - dF(\boldsymbol{x}_m)]^{-1} G(\boldsymbol{x}_m). \end{aligned} \tag{10.15}$$

If we can approximate $dF(\boldsymbol{x}_m)$ by a low-rank matrix $\boldsymbol{M}$, then we can replace $\boldsymbol{I} - dF(\boldsymbol{x}_m)$ by $\boldsymbol{I} - \boldsymbol{M}$ and explicitly form the inverse $(\boldsymbol{I} - \boldsymbol{M})^{-1}$. Let us see where this strategy leads.

Quasi-Newton methods operate by secant approximations [27, 32]. It is easy to generate a secant condition by taking two MM iterates starting from the current point $\boldsymbol{x}_m$. Close to the optimal point $\boldsymbol{y}$, the linear approximation

$$F \circ F(\boldsymbol{x}_m) - F(\boldsymbol{x}_m) \approx \boldsymbol{M}[F(\boldsymbol{x}_m) - \boldsymbol{x}_m]$$

holds, where $\boldsymbol{M} = dF(\boldsymbol{y})$. If $\boldsymbol{v}$ is the vector $F \circ F(\boldsymbol{x}_m) - F(\boldsymbol{x}_m)$ and $\boldsymbol{u}$ is the vector $F(\boldsymbol{x}_m) - \boldsymbol{x}_m$, then the secant condition is $\boldsymbol{M}\boldsymbol{u} = \boldsymbol{v}$. In fact, the best results may require several secant conditions $\boldsymbol{M}\boldsymbol{u}_i = \boldsymbol{v}_i$ for $i = 1, \ldots, q$, where $q \leq p$. These can be generated at the current iterate $\boldsymbol{x}_m$ and the previous $q - 1$ iterates. For convenience represent the secant conditions in the matrix form $\boldsymbol{M}\boldsymbol{U} = \boldsymbol{V}$ for $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q)$ and $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q)$. Example 5.2.7 shows that the choice $\boldsymbol{M} = \boldsymbol{V}(\boldsymbol{U}^*\boldsymbol{U})^{-1}\boldsymbol{U}^*$ minimizes the Frobenius norm of $\boldsymbol{M}$ subject to the secant constraint $\boldsymbol{M}\boldsymbol{U} = \boldsymbol{V}$. In practice, it is better to make a controlled approximation to $dF(\boldsymbol{y})$ than a wild guess.

To apply the approximation, we must invert the matrix $\boldsymbol{I}$–$\boldsymbol{V}(\boldsymbol{U}^*\boldsymbol{U})^{-1}\boldsymbol{U}^*$. Fortunately, we have the explicit inverse

$$[\boldsymbol{I} - \boldsymbol{V}(\boldsymbol{U}^*\boldsymbol{U})^{-1}\boldsymbol{U}^*]^{-1} = \boldsymbol{I} + \boldsymbol{V}[\boldsymbol{U}^*\boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{V}]^{-1}\boldsymbol{U}^*. \tag{10.16}$$

The reader can readily check this variant of the Sherman-Morrison formula. It is noteworthy that the $q \times q$ matrix $\boldsymbol{U}^*\boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{V}$ is trivial to invert for q small even when p is large. With these results in hand, the Newton update

(10.15) can be replaced by the quasi-Newton update

$$\begin{aligned}\boldsymbol{x}_{m+1} &= \boldsymbol{x}_m - [\boldsymbol{I} - \boldsymbol{V}(\boldsymbol{U}^*\boldsymbol{U})^{-1}\boldsymbol{U}^*]^{-1}[\boldsymbol{x}_m - F(\boldsymbol{x}_m)] \\ &= \boldsymbol{x}_m - [\boldsymbol{I} + \boldsymbol{V}(\boldsymbol{U}^*\boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{V})^{-1}\boldsymbol{U}^*][\boldsymbol{x}_m - F(\boldsymbol{x}_m)] \\ &= F(\boldsymbol{x}_m) - \boldsymbol{V}(\boldsymbol{U}^*\boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{V})^{-1}\boldsymbol{U}^*[\boldsymbol{x}_m - F(\boldsymbol{x}_m)].\end{aligned}$$

The special case $q = 1$ is interesting in its own right. A brief calculation shows that the quasi-Newton update for $q = 1$ is

$$\begin{aligned}\boldsymbol{x}_{m+1} &= (1 - c_m)F(\boldsymbol{x}_m) + c_m F \circ F(\boldsymbol{x}_m) \\ c_m &= -\frac{\|F(\boldsymbol{x}_m) - \boldsymbol{x}_m\|^2}{[F \circ F(\boldsymbol{x}_m) - 2F(\boldsymbol{x}_m) + \boldsymbol{x}_m]^*[F(\boldsymbol{x}_m) - \boldsymbol{x}_m]}.\end{aligned} \tag{10.17}$$

This quasi-Newton acceleration enjoys several desirable properties in high-dimensional problems. First, the computational effort per iteration is relatively light: two MM updates and a few matrix times vector multiplications. Second, memory demands are also light. If we fix q in advance, the most onerous requirement is storage of the secant matrices $\boldsymbol{U}$ and $\boldsymbol{V}$. These two matrices can be updated by replacing the earliest retained secant pair by the latest secant pair generated. Third, the whole scheme is consistent with linear constraints. Thus, if the parameter space satisfies a linear constraint $\boldsymbol{w}^*\boldsymbol{x} = a$ for all feasible $\boldsymbol{x}$, then the quasi-Newton iterates also satisfy $\boldsymbol{w}^*\boldsymbol{x}_m = a$ for all m. This claim follows from the equalities $\boldsymbol{w}^*F(\boldsymbol{x}) = a$ and $\boldsymbol{w}^*\boldsymbol{V} = \boldsymbol{0}$. Finally, if the quasi-Newton update at $\boldsymbol{x}_m$ fails the ascent or descent test, then one can always revert to the second MM update $F \circ F(\boldsymbol{x}_m)$. Balanced against these advantages is the failure of the quasi-Newton acceleration to respect parameter lower and upper bounds.

Example 10.9.1 *A Mixture of Poissons*

Problem 9 of Chap. 9 describes a Poisson mixture model for mortality data from *The London Times*. Starting from the method of moments estimates $(\mu_{01}, \mu_{02}, \pi_0) = (1.101, 2.582, .2870)$, the EM algorithm takes an excruciating 535 iterations for the loglikelihood $L(\theta)$ to attain its maximum of -1989.946. Even worse, it takes 1,749 iterations for the parameters to reach the maximum likelihood estimates $(\hat{\mu_1}, \hat{\mu_2}, \hat{\pi}) = (1.256, 2.663, .3599)$. The sizable difference in convergence rates to the maximum loglikelihood and the maximum likelihood estimates indicates that the likelihood surface is quite flat. In contrast, the accelerated EM algorithm converges to the maximum loglikelihood in about 10–150 iterations, depending on the value of q. Figure 10.1 plots the progress of the EM algorithm and the different versions of the quasi-Newton acceleration. Titterington et al. [259] report that Newton's method typically takes 8–11 iterations to converge when it converges for these data. For about a third of their initial points, Newton's method fails. ■

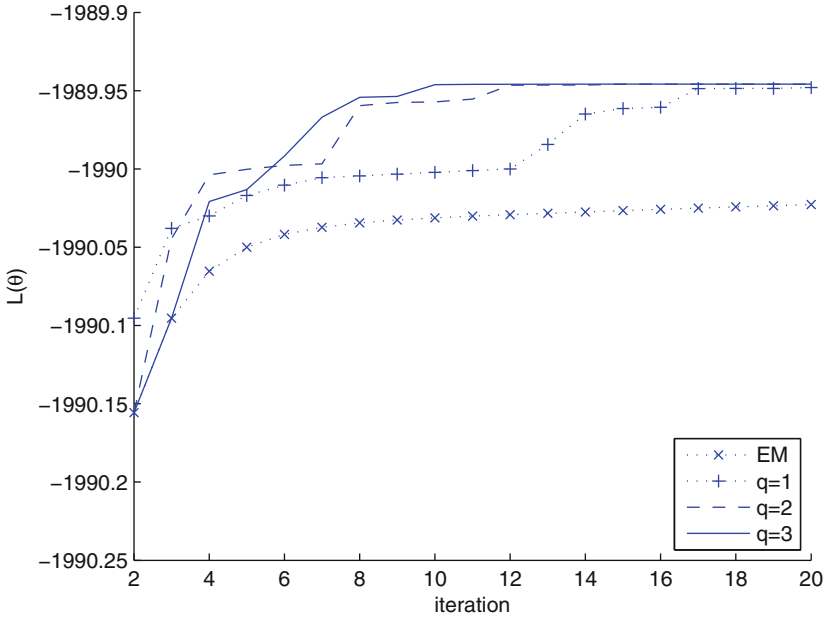

FIGURE 10.1. MM acceleration for the mixture of Poissons example

Although we have couched quasi-Newton acceleration in terms of the MM algorithm, it applies to any optimization algorithm with reasonably smooth parameter updates. Block relaxation is another potential beneficiary. Block relaxation shares the ascent-descent property of the MM algorithm, so if acceleration fails to improve the objective function, then one can still make progress by reverting to the original double step executed in constructing a new secant.

10.10 Problems

1. What happens when you apply Newton's method to the functions

$$f(x) = \begin{cases} \sqrt{x} & x \geq 0 \\ -\sqrt{-x} & x < 0 \end{cases}$$

and $g(x) = \sqrt[3]{x}$?

2. Consider a function $f(x) = (x-r)^k g(x)$ with a root r of multiplicity k. If $g'(x)$ is continuous at r, and the Newton iterates x_m converge to r, then show that the iterates satisfy

$$\lim_{m\to\infty} \frac{|x_{m+1} - r|}{|x_m - r|} = 1 - \frac{1}{k}.$$

3. As an illustration of Problem 2, use Newton's method to extract a root of the polynomials $p_1(x) = x^2-1$ and $p_2(x) = x^2-2x+1$ starting from $x_0 = 2$. Notice how much more slowly convergence occurs for $p_2(x)$ than for $p_1(x)$.

4. Suppose the real-valued $f(x)$ satisfies $f'(x) > 0$ and $f''(x) > 0$ for all x in its domain (d, ∞). If the equation $f(x) = 0$ has a root r, then demonstrate that r is unique and that Newton's method converges to r regardless of its starting point. Further, prove that x_m converges monotonically to r from above when $x_0 > r$ and that $x_1 > r$ when $x_0 < r$. How are these results pertinent to Example 10.2.2?

5. Problem 4 applies to polynomials $p(x)$ having only real roots. Suppose $p(x)$ is a polynomial of degree d with roots $r_1 < r_2 < \cdots < r_d$ and leading coefficient $c_d > 0$. Show that on the interval (r_d, ∞) the functions $p(x)$, $p'(x)$, and $p''(x)$ are all positive. Hence, if we seek r_d by Newton's method starting at $x_0 > r_d$, then the iterates x_m decrease monotonically to r_d. (Hint: According to Rolle's theorem, what can we say about the roots of $p'(x)$ and $p''(x)$?)

6. Suppose that the polynomial $p(x)$ has the known roots $r_1, \ldots, r_d$. Maehly's algorithm [246] attempts to extract one additional root r_{d+1} by iterating via

$$x_{m+1} \;=\; x_m - \frac{p(x_m)}{p'(x_m) - \sum_{k=1}^{d} \frac{p(x_m)}{(x_m - r_k)}}.$$

   Show that this is just a disguised version of Newton's method. It has the virtue of being more numerically accurate than Newton's method applied to the deflated polynomial calculated from $p(x)$ by synthetic division. (Hint: Consider the polynomial $q(x) = p(x) \prod_{k=1}^{d} (x-r_k)^{-1}$.)

7. Apply Maehly's algorithm as sketched in Problem 6 to find the roots of the polynomial $p(x) = x^4 - 12x^3 + 47x^2 - 60x$.

8. Consider the map

$$f(\boldsymbol{x}) \;=\; \begin{pmatrix} x_1^2 + x_2^2 - 2 \\ x_1 - x_2 \end{pmatrix}$$

   of the plane into itself. Show that $f(\boldsymbol{x}) = \boldsymbol{0}$ has the roots $-\boldsymbol{1}$ and $\boldsymbol{1}$ and no other roots. Prove that Newton's method iterates according to

$$x_{m+1,1} \;=\; x_{m+1,2} \;=\; \frac{x_{m1}^2 + x_{m2}^2 + 2}{2(x_{m1} + x_{m2})}$$

   and that these iterates converge to the root $-\boldsymbol{1}$ if $x_{01} + x_{02}$ is negative and to the root $\boldsymbol{1}$ if $x_{01} + x_{02}$ is positive. If $x_{01} + x_{02} = 0$, then the

first iterate is undefined. Finally, prove that

$$\lim_{m\to\infty} \frac{|x_{m+1,1} - y_1|}{|x_{m1} - y_1|^2} \quad = \quad \lim_{m\to\infty} \frac{|x_{m+1,2} - y_2|}{|x_{m2} - y_2|^2} \quad = \quad \frac{1}{2},$$

where $\boldsymbol{y}$ is the root relevant to the initial point $\boldsymbol{x}_0$.

9. Continuing Example 10.2.5, consider iterating according to

$$\boldsymbol{B}_{m+1} \quad = \quad \boldsymbol{B}_m \sum_{i=0}^{j} (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{B}_m)^i \tag{10.18}$$

to find $\boldsymbol{A}^{-1}$ [138]. Example 10.2.5 covers the special case $j = 1$. Verify the alternative representation

$$\boldsymbol{B}_{m+1} \quad = \quad \sum_{i=0}^{j} (\boldsymbol{I} - \boldsymbol{B}_m\boldsymbol{A})^i \boldsymbol{B}_m,$$

and use it to prove that $\boldsymbol{B}_{m+1}$ is symmetric whenever $\boldsymbol{A}$ and $\boldsymbol{B}_m$ are. Also show that

$$\boldsymbol{A}^{-1} - \boldsymbol{B}_{m+1} \quad = \quad (\boldsymbol{A}^{-1} - \boldsymbol{B}_m)[\boldsymbol{A}(\boldsymbol{A}^{-1} - \boldsymbol{B}_m)]^j.$$

From this last identity deduce the norm inequality

$$\|\boldsymbol{A}^{-1} - \boldsymbol{B}_{m+1}\| \quad \leq \quad \|\boldsymbol{A}\|^j \|\boldsymbol{A}^{-1} - \boldsymbol{B}_m\|^{j+1}.$$

Thus, the algorithm converges at a cubic rate when $j = 2$, at a quartic rate when $j = 3$, and so forth.

10. Example 10.2.2 can be adapted to extract the nth root of a positive semidefinite matrix $\boldsymbol{A}$ [107]. Consider the iteration scheme

$$\boldsymbol{B}_{m+1} \quad = \quad \frac{n-1}{n}\boldsymbol{B}_m + \frac{1}{n}\boldsymbol{B}_m^{-n+1}\boldsymbol{A}$$

starting with $\boldsymbol{B}_0 = c\boldsymbol{I}$ for some positive constant c. Show by induction that (a) $\boldsymbol{B}_m$ commutes with $\boldsymbol{A}$, (b) $\boldsymbol{B}_m$ is symmetric, and (c) $\boldsymbol{B}_m$ is positive definite. To prove that $\boldsymbol{B}_m$ converges to $\boldsymbol{A}^{1/n}$, consider the spectral decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^*$ of $\boldsymbol{A}$ with $\boldsymbol{D}$ diagonal and $\boldsymbol{U}$ orthogonal. Show that $\boldsymbol{B}_m$ has a similar spectral decomposition $\boldsymbol{B}_m = \boldsymbol{U}\boldsymbol{D}_m\boldsymbol{U}^*$ and that the ith diagonal entries of $\boldsymbol{D}_m$ and $\boldsymbol{D}$ satisfy

$$d_{m+1,i} \quad = \quad \frac{n-1}{n} d_{mi} + \frac{1}{n} d_{mi}^{-n+1} d_i.$$

Example 10.2.2 implies that d_{mi} converges to $\sqrt[n]{d_i}$ when $d_i > 0$. This convergence occurs at a fast quadratic rate as explained in Proposition 12.2.2. If $d_i = 0$, then d_{mi} converges to 0 at the linear rate $\frac{n-1}{n}$.

11. Program the algorithm of Problem 10 and extract the square roots of the two matrices

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Describe the apparent rate of convergence in each case and any difficulties you encounter with roundoff error.

12. Prove that the increment (10.5) can be expressed as

$$\begin{aligned} & \Delta \boldsymbol{x}_m \\ = \; & -\boldsymbol{H}_m^{-1/2} \left[\boldsymbol{I} - \boldsymbol{H}_m^{-1/2} \boldsymbol{V}^* (\boldsymbol{V} \boldsymbol{H}_m^{-1} \boldsymbol{V}^*)^{-1} \boldsymbol{V} \boldsymbol{H}_m^{-1/2} \right] \boldsymbol{H}_m^{-1/2} \nabla f(\boldsymbol{x}_m) \\ = \; & -\boldsymbol{H}_m^{-1/2} (\boldsymbol{I} - \boldsymbol{P}_m) \boldsymbol{H}_m^{-1/2} \nabla f(\boldsymbol{x}_m) \end{aligned}$$

using the symmetric square root $\boldsymbol{H}_m^{-1/2}$ of $\boldsymbol{H}_m^{-1}$. Check that the matrix $\boldsymbol{P}_m$ is a projection in the sense that $\boldsymbol{P}_m^* = \boldsymbol{P}_m$ and $\boldsymbol{P}_m^2 = \boldsymbol{P}_m$ and that these properties carry over to $\boldsymbol{I} - \boldsymbol{P}_m$. Now argue that

$$-df(\boldsymbol{x}_m) \Delta \boldsymbol{x}_m \;=\; \|(\boldsymbol{I} - \boldsymbol{P}_m) \boldsymbol{H}_m^{-1/2} \nabla f(\boldsymbol{x}_m)\|^2$$

and consequently that step halving is bound to produce a decrease in $f(\boldsymbol{x})$ if $(\boldsymbol{I} - \boldsymbol{P}_m) \boldsymbol{H}_m^{-1/2} \nabla f(\boldsymbol{x}_m) \neq \boldsymbol{0}$.

13. Show that Newton's method converges in one iteration to the minimum of

$$f(\boldsymbol{\theta}) \;=\; \frac{1}{2} \boldsymbol{\theta}^* \boldsymbol{A} \boldsymbol{\theta} + \boldsymbol{b}^* \boldsymbol{\theta} + c$$

when the symmetric matrix $\boldsymbol{A}$ is positive definite. Note that this implies that the Gauss-Newton algorithm (10.14) converges in a single step when the regression functions $\mu_i(\boldsymbol{\theta})$ are linear.

14. In Example 10.4.2, digamma and trigamma functions $\psi(t)$ and $\psi'(t)$ must be evaluated. Show that these functions satisfy the recurrence relations

$$\begin{aligned} \psi(t) &= -t^{-1} + \psi(t+1) \\ \psi'(t) &= t^{-2} + \psi'(t+1). \end{aligned}$$

Thus, if $\psi(t)$ and $\psi'(t)$ can be accurately evaluated via asymptotic expansions for large t, then they can be accurately evaluated for small t. For example, it is known that $\psi(t) = \ln t - (2t)^{-1} + O(t^{-2})$ and $\psi'(t) = t^{-1} + (\sqrt{2}t)^{-2} + O(t^{-3})$ as $t \to \infty$.

15. Compute the score vector and the observed and expected information matrices for the Dirichlet distribution (10.6). Explicitly invert the expected information using the Sherman-Morrison formula.

16. Verify the score and information entries in Table 10.2.

17. Let $g(x)$ and $h(x)$ be probability densities defined on the real line. Show that the admixture density $f(x) = \theta g(x) + (1-\theta)h(x)$ for $\theta \in [0,1]$ has score and expected information

$$\begin{aligned} L'(\theta) &= \frac{g(x)-h(x)}{\theta g(x)+(1-\theta)h(x)} \\ J(\theta) &= \int \frac{[g(x)-h(x)]^2}{\theta g(x)+(1-\theta)h(x)}dx \\ &= \frac{1}{\theta(1-\theta)}\Big[1-\int \frac{g(x)h(x)}{\theta g(x)+(1-\theta)h(x)}dx\Big]. \end{aligned}$$

What happens to $J(\theta)$ when $g(x)$ and $h(x)$ coincide? What does $J(\theta)$ equal when $g(x)$ and $h(x)$ have nonoverlapping domains? (Hint: The identities

$$h-g = \frac{\theta g+(1-\theta)h-g}{1-\theta}, \quad g-h = \frac{\theta g+(1-\theta)h-h}{\theta}$$

will help.)

18. A quantal response model involves independent binomial observations $y_1, \ldots, y_j$ with n_i trials and success probability $\pi_i(\boldsymbol{\theta})$ per trial for the ith observation. If $\boldsymbol{x}_i$ is a predictor vector and $\boldsymbol{\theta}$ a parameter vector, then the specification

$$\pi_i(\boldsymbol{\theta}) = \frac{e^{\boldsymbol{x}_i^*\boldsymbol{\theta}}}{1+e^{\boldsymbol{x}_i^*\boldsymbol{\theta}}}$$

gives a generalized linear model. Use the scoring algorithm to estimate $\hat{\boldsymbol{\theta}} = (-5.1316, 0.0677)^*$ for the ingot data of Cox [53] displayed in Table 10.5.

TABLE 10.5. Ingot data for a quantal response model

Trials n_i	**Observation** y_i	**Covariate** x_{i1}	**Covariate** x_{i2}
55	0	1	7
157	2	1	14
159	7	1	27
16	3	1	57

19. In robust regression it is useful to consider location-scale families with densities of the form

$$\frac{c}{\sigma}e^{-\rho(\frac{x-\mu}{\sigma})}, \qquad x \in (-\infty, \infty). \tag{10.19}$$

Here $\rho(r)$ is a strictly convex even function, decreasing to the left of 0 and symmetrically increasing to the right of 0. Without loss of generality, one can take $\rho(0) = 0$. The normalizing constant c is determined by $c\int_{-\infty}^{\infty} e^{-\rho(r)}dr = 1$. Show that a random variable X with density (10.19) has mean μ and variance

$$\text{Var}(X) \;=\; c\sigma^2 \int_{-\infty}^{\infty} r^2 e^{-\rho(r)} dr.$$

If μ depends on a parameter vector $\boldsymbol{\theta}$, demonstrate that the score corresponding to a single observation $X = x$ amounts to

$$\nabla L(\boldsymbol{\phi}) \;=\; \begin{pmatrix} \frac{1}{\sigma}\rho'(\frac{x-\mu}{\sigma})\nabla\mu \\ -\frac{1}{\sigma} + \rho'(\frac{x-\mu}{\sigma})\frac{x-\mu}{\sigma^2} \end{pmatrix}$$

for $\boldsymbol{\phi} = (\boldsymbol{\theta}^*, \sigma)^*$. Finally, prove that the expected information $J(\boldsymbol{\phi})$ is block diagonal with upper-left block

$$\frac{c}{\sigma^2}\int_{-\infty}^{\infty} \rho''(r)e^{-\rho(r)}dr\nabla\mu(\theta)d\mu(\theta)$$

and lower-right block

$$\frac{c}{\sigma^2}\int_{-\infty}^{\infty} \rho''(r)r^2e^{-\rho(r)}dr + \frac{1}{\sigma^2}.$$

20. In the context of Problem 19, take $\rho(r) = \ln\cosh^2(\frac{r}{2})$. Show that this corresponds to the logistic distribution with density

$$f(x) \;=\; \frac{e^{-x}}{(1+e^{-x})^2}.$$

Compute the integrals

$$\begin{aligned} \frac{\pi^2}{3} &= c\int_{-\infty}^{\infty} r^2 e^{-\rho(r)}dr \\ \frac{1}{3} &= c\int_{-\infty}^{\infty} \rho''(r)e^{-\rho(r)}dr \\ \frac{1}{3}+\frac{\pi^2}{9} &= c\int_{-\infty}^{\infty} \rho''(r)r^2e^{-\rho(r)}dr + 1 \end{aligned}$$

determining the variance and expected information of the density (10.19) for this choice of $\rho(r)$.

21. Continuing Problems 19 and 20, compute the normalizing constant c and the three integrals determining the variance and expected information for Huber's function

$$\rho(r) \;=\; \begin{cases} \frac{r^2}{2} & |r| \le k \\ k|r| - \frac{k^2}{2} & |r| > k\,. \end{cases}$$

22. A family of discrete density functions $p_n(\theta)$ defined on $\{0, 1, \ldots\}$ and indexed by a parameter $\theta > 0$ is said to be a power series family if for all n

$$p_n(\theta) = \frac{c_n \theta^n}{g(\theta)}, \tag{10.20}$$

where $c_n \geq 0$, and where $g(\theta) = \sum_{n=0}^{\infty} c_n \theta^n$ is the appropriate normalizing constant. If $y_1, \ldots, y_j$ are independent observations from the discrete density (10.20), then show that the maximum likelihood estimate of θ is a root of the equation

$$\frac{1}{j} \sum_{i=1}^{j} y_i = \frac{\theta g'(\theta)}{g(\theta)}. \tag{10.21}$$

Prove that the expected information in a single observation is

$$J(\theta) = \frac{\sigma^2(\theta)}{\theta^2},$$

where $\sigma^2(\theta)$ is the variance of the density (10.20).

23. Continuing problem 22, equation (10.21) suggests that one can find the maximum likelihood estimate $\hat{\theta}$ by iterating via

$$\theta_{m+1} = \frac{\bar{x} g(\theta_m)}{g'(\theta_m)} = f(\theta_m),$$

where $\bar{x}$ is the sample mean. The question now arises whether this iteration scheme is likely to converge to $\hat{\theta}$. Local convergence hinges on the condition $|f'(\hat{\theta})| < 1$. When this condition is true, the map $\theta_{m+1} = f(\theta_m)$ is locally contractive near the fixed point $\hat{\theta}$. Prove that

$$f'(\hat{\theta}) = 1 - \frac{\sigma^2(\hat{\theta})}{\mu(\hat{\theta})},$$

where

$$\mu(\theta) = \frac{\theta g'(\theta)}{g(\theta)}$$

is the mean of a single realization. Thus, convergence depends on the ratio of the variance to the mean. (Hints: By differentiating $g(\theta)$ it is easy to compute the mean and the second factorial moment

$$\mathrm{E}[X(X-1)] = \frac{\theta^2 g''(\theta)}{g(\theta)}.$$

Substitute this in $f'(\hat{\theta})$, recall $\mathrm{Var}(X) = \mathrm{E}[X(X-1)] + \mathrm{E}(X) - \mathrm{E}(X)^2$, and invoke equality (10.21).)

24. In the Gauss-Newton algorithm (10.14), the matrix

$$d^2 f(\boldsymbol{\theta}_m) = \sum_{i=1}^{j} w_i \nabla \mu_i(\boldsymbol{\theta}_m) d\mu(\boldsymbol{\theta}_m)$$

can be singular or nearly so. To cure this ill, Marquardt suggested choosing $\lambda > 0$, substituting

$$\boldsymbol{H}_m = \sum_{i=1}^{j} w_i \nabla \mu_i(\boldsymbol{\theta}_m) d\mu(\boldsymbol{\theta}_m) + \lambda I$$

for $d^2 f(\boldsymbol{\theta}_m)$, and iterating according to

$$\boldsymbol{\theta}_{m+1} = \boldsymbol{\theta}_m + \boldsymbol{H}_m^{-1} \sum_{i=1}^{j} w_i [x_i - \mu_i(\boldsymbol{\theta}_m)] \nabla \mu_i(\boldsymbol{\theta}_m). \tag{10.22}$$

Prove that the increment $\Delta\boldsymbol{\theta}_m = \boldsymbol{\theta}_{m+1} - \boldsymbol{\theta}_m$ proposed in equation (10.22) minimizes the criterion

$$\frac{1}{2} \sum_{i=1}^{j} w_i [x_i - \mu_i(\boldsymbol{\theta}_m) - d\mu_i(\boldsymbol{\theta}_m) \Delta\boldsymbol{\theta}_m]^2 + \frac{\lambda}{2} \|\Delta\boldsymbol{\theta}_m\|^2.$$

25. Survival analysis deals with nonnegative random variables T modeling random lifetimes. Let such a random variable $T \geq 0$ have density function $f(t)$ and distribution function $F(t)$. The hazard function

$$\begin{aligned} h(t) &= \lim_{s \downarrow 0} \frac{\Pr(t < T \leq t + s \mid T > t)}{s} \\ &= \frac{f(t)}{1 - F(t)} \end{aligned}$$

represents the instantaneous rate of death under lifetime T. Statisticians call the right-tail probability $1 - F(t) = S(t)$ the survival function and view $h(t)$ as the derivative

$$h(t) = -\frac{d}{dt} \ln S(t).$$

The cumulative hazard function $H(t) = \int_0^t h(s) ds$ obviously satisfies the identity

$$S(t) = e^{-H(t)}.$$

In Cox's proportional hazards model, longevity depends not only on time but also predictors. This is formalized by taking

$$h(t) = \lambda(t) e^{\boldsymbol{x}^* \boldsymbol{\alpha}},$$

where $\boldsymbol{x}$ and $\boldsymbol{\alpha}$ are column vectors of predictors and regression coefficients, respectively. For instance, $\boldsymbol{x}$ might be $(1, d)^*$, where d indicates dosage of a life-prolonging drug.

Many clinical trials involve right censoring. In other words, instead of observing a lifetime $T = t$, we observe $T > t$. Censored and ordinary data can be mixed in the same study. Generally, each observation T comes with a censoring indicator W. If T is censored, then $W = 1$; otherwise, $W = 0$.

(a) Show that

$$H(t) = \Lambda(t)e^{\boldsymbol{x}^*\boldsymbol{\alpha}},$$

where

$$\Lambda(t) = \int_0^t \lambda(s)ds.$$

In the Weibull proportional hazards model, $\lambda(t) = \beta t^{\beta-1}$. Show that this translates into the survival and density functions

$$\begin{aligned} S(t) &= e^{-t^\beta e^{\boldsymbol{x}^*\boldsymbol{\alpha}}} \\ f(t) &= \beta t^{\beta-1} e^{\boldsymbol{x}^*\boldsymbol{\alpha} - t^\beta e^{\boldsymbol{x}^*\boldsymbol{\alpha}}}. \end{aligned}$$

(b) Consider n independent possibly censored observations $t_1, \ldots, t_n$ with corresponding predictor vectors $\boldsymbol{x}_1 \ldots, \boldsymbol{x}_n$ and censoring indicators $w_1, \ldots, w_n$. Prove that the loglikelihood of the data is

$$L(\boldsymbol{\alpha}, \beta) = \sum_{i=1}^n w_i \ln S_i(t_i) + \sum_{i=1}^n (1 - w_i) \ln f_i(t_i),$$

where $S_i(t)$ and $f_i(t)$ are the survival and density functions of the ith case.

(c) Calculate the score and observed information for the Weibull model as posed. The observed information is

$$\begin{aligned} -d^2 L(\boldsymbol{\alpha}, \beta) &= \sum_{i=1}^n t_i^\beta e^{\boldsymbol{x}_i^*\boldsymbol{\alpha}} \begin{pmatrix} \boldsymbol{x}_i \\ \ln t_i \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_i \\ \ln t_i \end{pmatrix}^* \\ &\quad + \sum_{i=1}^n (1 - w_i) \begin{pmatrix} 0 & 0 \\ 0 & \beta^{-2} \end{pmatrix}. \end{aligned}$$

(d) Show that the loglikelihood $L(\boldsymbol{\alpha}, \beta)$ for the Weibull model is concave. Demonstrate that it is strictly concave if and only if the n vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ span R^m, where $\boldsymbol{\alpha}$ has m components.

(e) Describe in detail how you would implement Newton's method for finding the maximum likelihood estimate of the parameter vector $(\boldsymbol{\alpha}, \beta)$. What difficulties might you encounter? Why is concavity of the loglikelihood helpful?

26. Write a computer program and reproduce the iterates displayed in Table 10.4.

27. Let $x_1, \ldots, x_m$ be a random sample from the gamma density

$$f(x) = \Gamma(\alpha)^{-1}\beta^{\alpha}x^{\alpha-1}e^{-\beta x}$$

on $(0, \infty)$. Find the score, observed information, and expected information for the parameters α and β, and demonstrate that Newton's method and scoring coincide.

28. Continuing Problem 27, derive the method of moments estimators

$$\hat{\alpha} = \frac{\overline{x}^2}{s^2}, \qquad \hat{\beta} = \frac{\overline{x}}{s^2},$$

where $\overline{x} = \frac{1}{m}\sum_{i=1}^m x_i$ and $s^2 = \frac{1}{m}\sum_{i=1}^m (x_i - \overline{x})^2$ are the sample mean and variance, respectively. These are not necessarily the best explicit estimators of the two parameters. Show that setting the score function equal to $\mathbf{0}$ implies that $\beta = \alpha/\overline{x}$ is a stationary point of the loglikelihood $L(\alpha, \beta)$ of the sample $x_1, \ldots, x_m$ for α fixed. Why does $\beta = \alpha/\overline{x}$ furnish the maximum? Now argue that substituting this value of β in the loglikelihood reduces maximum likelihood estimation to optimization of the profile loglikelihood

$$L(\alpha) = m\alpha \ln \alpha - m\alpha \ln \overline{x} - m \ln \Gamma(\alpha) + m(\alpha - 1)\overline{\ln x} - m\alpha.$$

Here $\overline{\ln x} = \frac{1}{m}\sum_{i=1}^m \ln x_i$. There are two nasty terms in $L(\alpha)$. One is $\alpha \ln \alpha$, and the other is $\ln \Gamma(\alpha)$. We can eliminate both by appealing to a version of Stirling's formula. Ordinarily Stirling's formula is only applied for large factorials. This limitation is inconsistent with small α. However, Gosper's version of Stirling's formula is accurate for all arguments. This little-known version of Stirling's formula says that

$$\Gamma(\alpha + 1) \approx \sqrt{(\alpha + 1/6)2\pi}\, \alpha^{\alpha} e^{-\alpha}.$$

Given that $\Gamma(\alpha) = \Gamma(\alpha+1)/\alpha$, show that the application of Gosper's formula leads to the approximate maximum likelihood estimate

$$\hat{\alpha} = \frac{3 - d + \sqrt{(3-d)^2 + 24d}}{12d},$$

where $d = \ln \overline{x} - \overline{\ln x}$ [47]. Why is this estimate of α positive? Why does one take the larger root of the defining quadratic?

29. In the multilogit model, items are drawn from m categories. Let y_i denote the ith outcome of l independent draws and $\boldsymbol{x}_i$ a corresponding predictor vector. The probability π_{ij} that $y_i = j$ is given by

$$\pi_{ij}(\boldsymbol{\theta}) = \begin{cases} \frac{e^{\boldsymbol{x}_i^* \boldsymbol{\theta}_j}}{1+\sum_{k=1}^{m-1} e^{\boldsymbol{x}_i^* \boldsymbol{\theta}_k}} & 1 \leq j < m \\ \frac{1}{1+\sum_{k=1}^{m-1} e^{\boldsymbol{x}_i^* \boldsymbol{\theta}_k}} & j = m. \end{cases}$$

Find the loglikelihood, score, observed information, and expected information. Demonstrate that Newton's method and scoring coincide. (Hint: You can achieve compact expressions by stacking vectors and using matrix Kronecker products.)

30. Derive formulas (10.16) and (10.17).

11
Conjugate Gradient and Quasi-Newton

11.1 Introduction

Our discussion of Newton's method has highlighted both its strengths and its weaknesses. Related algorithms such as scoring and Gauss-Newton exploit special features of the objective function $f(\boldsymbol{x})$ in overcoming the defects of Newton's method. We now consider algorithms that apply to generic functions $f(\boldsymbol{x})$. These algorithms also operate by locally approximating $f(\boldsymbol{x})$ by a strictly convex quadratic function. Indeed, the guiding philosophy behind many modern optimization algorithms is to see what techniques work well with quadratic functions and then to modify the best techniques to accommodate generic functions.

The conjugate gradient algorithm [94, 127] is noteworthy for three properties: (a) it minimizes a quadratic function $f(\boldsymbol{x})$ from R^n to R in n steps, (b) it does not require evaluation of $d^2 f(\boldsymbol{x})$, and (c) it does not involve storage or inversion of any $n \times n$ matrices. Property (c) makes the method particularly suitable for optimization in high-dimensional settings. One of the drawbacks of the conjugate gradient method is that it requires exact line searches.

Quasi-Newton algorithms [10, 56, 91, 93] enjoy properties (a) and (b) but not property (c). In compensation for the failure of (c), inexact line searches are usually adequate with quasi-Newton algorithms. Furthermore, quasi-Newton methods adapt more readily to parameter constraints. Except for a discussion of trust regions, the current chapter considers only unconstrained optimization problems.

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_11,

11.2 Centers of Spheres and Centers of Ellipsoids

As an introduction to many of the central ideas of the chapter, it is instructive to explore a simple algorithm for finding the center of a sphere. The fact that we already know the answer should not deter us from considering algorithmic issues. If the center is the origin, then obviously we can find it by minimizing the scaled distance function

$$g(\boldsymbol{x}) \;=\; \tfrac{1}{2}\|\boldsymbol{x}\|^2 \;=\; \tfrac{1}{2}\textstyle\sum_{i=1}^n x_i^2$$

with gradient $\nabla g(\boldsymbol{x}) = \boldsymbol{x}$. In cyclic coordinate descent, we minimize $g(\boldsymbol{x})$ along each coordinate direction in turn, starting from a point $\boldsymbol{x}_1$ and generating successive points $\boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n+1}$. The search along coordinate direction $\boldsymbol{e}_i$ at iteration i amounts to minimizing the function

$$g(\boldsymbol{x}_i + t\boldsymbol{e}_i) \;=\; \frac{1}{2}(x_{ii}+t)^2 + \frac{1}{2}\sum_{j\neq i} x_{ij}^2$$

of the scalar t. The minimum occurs at $t = -x_{ii}$ and yields $\boldsymbol{x}_{i+1}$. It is trivial to check that this procedure achieves the minimum in n iterations and satisfies at iteration i the identities

$$\boldsymbol{e}_i^* \boldsymbol{e}_j \;=\; 0 \quad \text{and} \quad dg(\boldsymbol{x}_i)\boldsymbol{e}_j \;=\; 0 \tag{11.1}$$

for all $j < i$. Furthermore, $\boldsymbol{x}_{i+1}$ minimizes the function

$$h(t_1, \ldots, t_i) \;=\; g\Big(\boldsymbol{x}_1 + \sum_{j=1}^{i} t_j \boldsymbol{e}_j\Big)$$

defined on the i-dimensional plane $\boldsymbol{x}_1 + t_1\boldsymbol{e}_1 + \cdots + t_i\boldsymbol{e}_i$ formed from all linear combinations of the first i search directions. Because of the spherical symmetry of the function $g(\boldsymbol{x})$, we can substitute any set of nontrivial orthogonal vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ and reach the same conclusions.

If we consider an arbitrary strictly convex quadratic function

$$\begin{aligned} f(\boldsymbol{y}) \;&=\; \frac{1}{2}\boldsymbol{y}^*\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}^*\boldsymbol{y} + c \qquad (11.2)\\ &=\; \frac{1}{2}(\boldsymbol{y} + \boldsymbol{A}^{-1}\boldsymbol{b})^*\boldsymbol{A}(y + \boldsymbol{A}^{-1}\boldsymbol{b}) - \frac{1}{2}\boldsymbol{b}^*\boldsymbol{A}^{-1}\boldsymbol{b} + c, \end{aligned}$$

then the situation becomes more interesting. Here the matrix $\boldsymbol{A}$ is positive definite, so there is no doubt that its inverse $\boldsymbol{A}^{-1}$ exists. Because the minimum of $f(\boldsymbol{y})$ occurs at $\boldsymbol{y} = -\boldsymbol{A}^{-1}\boldsymbol{b}$, any method of minimizing $f(\boldsymbol{y})$ gives in effect a method for solving the linear equation $\boldsymbol{A}\boldsymbol{y} = -\boldsymbol{b}$. The solution of such equations in high dimensions is one of the primary applications of the conjugate gradient method.

We can reduce the problem of minimizing the quadratic function (11.2) to the previous spherical minimization problem by making the change of variables $\boldsymbol{y} = \boldsymbol{A}^{-1/2}\boldsymbol{x} - \boldsymbol{A}^{-1}\boldsymbol{b}$ involving the symmetric square root $\boldsymbol{A}^{-1/2}$ of $\boldsymbol{A}^{-1}$. If $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^*$ is the spectral decomposition of the positive definite matrix $\boldsymbol{A}$, then $\boldsymbol{A}^{-1/2} = \boldsymbol{U}\boldsymbol{D}^{-1/2}\boldsymbol{U}^*$. The invertible transformation $\boldsymbol{x} \mapsto \boldsymbol{y}$ sends lines into lines and planes into planes. It also sends the function $f(\boldsymbol{y})$ into the function

$$\begin{aligned} g(\boldsymbol{x}) &= f(\boldsymbol{y}) \\ &= f(\boldsymbol{A}^{-1/2}\boldsymbol{x} - \boldsymbol{A}^{-1}\boldsymbol{b}) \\ &= \frac{1}{2}\|\boldsymbol{x}\|^2 - \frac{1}{2}\boldsymbol{b}^*\boldsymbol{A}^{-1}\boldsymbol{b} + c \end{aligned}$$

and puts us back where we started, minimizing $\frac{1}{2}\|\boldsymbol{x}\|^2$. If we have a set of nontrivial orthogonal vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ in $\boldsymbol{x}$ space, then we can search along each of these directions in turn and achieve the global minima of $g(\boldsymbol{x})$ and $f(\boldsymbol{y})$ in n iterations.

For later use, it is important to identify the analogs of the orthogonality conditions (11.1). The direction $\boldsymbol{u}_i$ in $\boldsymbol{x}$ space corresponds to the direction $\boldsymbol{v}_i$ in $\boldsymbol{y}$ space defined by $\boldsymbol{A}^{1/2}\boldsymbol{v}_i = \boldsymbol{u}_i$. Thus, the condition $\boldsymbol{u}_i^*\boldsymbol{u}_j = 0$ for all $j < i$ translates into the condition

$$\boldsymbol{v}_i^*\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2}\boldsymbol{v}_j = \boldsymbol{v}_i^*\boldsymbol{A}\boldsymbol{v}_j = 0$$

for all $j < i$. Two vectors $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ satisfying such an orthogonality relation are said to be conjugate. Conjugacy is equivalent to orthogonality under the nonstandard inner product $\boldsymbol{v}_i^*\boldsymbol{A}\boldsymbol{v}_j$. A finite set of conjugate vectors is necessarily linearly independent.

In view of the chain rule, we have $dg(\boldsymbol{x}_i) = df(\boldsymbol{y}_i)\boldsymbol{A}^{-1/2}$ for the point $\boldsymbol{y}_i = \boldsymbol{A}^{-1/2}\boldsymbol{x}_i - \boldsymbol{A}^{-1}\boldsymbol{b}$. Thus, the condition $dg(\boldsymbol{x}_i)\boldsymbol{u}_j = 0$ for all $j < i$ translates into the condition

$$df(\boldsymbol{y}_i)\boldsymbol{A}^{-1/2}\boldsymbol{A}^{1/2}\boldsymbol{v}_j = df(\boldsymbol{y}_i)\boldsymbol{v}_j = 0 \tag{11.3}$$

for all $j < i$. Alternatively, the condition $df(\boldsymbol{y}_i)\boldsymbol{v}_j = 0$ for all $j < i$ is an immediate consequence of the fact that $\boldsymbol{y}_i$ minimizes the function

$$h(t_1, \ldots, t_{i-1}) = f\Big(\boldsymbol{y}_1 + \sum_{j=1}^{i-1} t_j\boldsymbol{v}_j\Big) \tag{11.4}$$

defined on the plane $\boldsymbol{y}_1 + t_1\boldsymbol{v}_1 + \cdots + t_{i-1}\boldsymbol{v}_{i-1}$ formed from all linear combinations of the first $i-1$ search directions.

11.3 The Conjugate Gradient Algorithm

The flaw with this analysis is that it omits any description of how the initial point $\boldsymbol{y}_1$ and conjugate directions $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are chosen. Choice of $\boldsymbol{y}_1$ is

more or less arbitrary, depending on the particular problem and relevant external information. The obvious choice $\boldsymbol{v}_1 = -\nabla f(\boldsymbol{y}_1)$ is consistent with an initial search along the direction of steepest descent. At iteration $i > 1$ the conjugate gradient algorithm inductively chooses the search direction

$$\boldsymbol{v}_i \;=\; -\nabla f(\boldsymbol{y}_i) + \alpha_i \boldsymbol{v}_{i-1}, \tag{11.5}$$

where

$$\alpha_i \;=\; \frac{df(\boldsymbol{y}_i)\boldsymbol{A}\boldsymbol{v}_{i-1}}{\boldsymbol{v}_{i-1}^* \boldsymbol{A}\boldsymbol{v}_{i-1}} \tag{11.6}$$

is defined so that $\boldsymbol{v}_i^* \boldsymbol{A}\boldsymbol{v}_{i-1} = 0$. For $1 \le j < i-1$, the conjugacy condition $\boldsymbol{v}_i^* \boldsymbol{A}\boldsymbol{v}_j = 0$ requires

$$0 \;=\; -df(\boldsymbol{y}_i)\boldsymbol{A}\boldsymbol{v}_j + \alpha_i \boldsymbol{v}_{i-1}^* \boldsymbol{A}\boldsymbol{v}_j \;=\; -df(\boldsymbol{y}_i)\boldsymbol{A}\boldsymbol{v}_j. \tag{11.7}$$

Equality (11.7) is hardly obvious, but we can attack it by noting that in view of definition (11.5) the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{i-1}$ and $\nabla f(\boldsymbol{y}_1), \ldots, \nabla f(\boldsymbol{y}_{i-1})$ span the same vector subspace. Hence, the condition $df(\boldsymbol{y}_i)\boldsymbol{v}_j = 0$ for all $j < i$ is equivalent to the condition $df(\boldsymbol{y}_i)\nabla f(\boldsymbol{y}_j) = 0$ for all $j < i$. However, $df(\boldsymbol{y}_i)\boldsymbol{v}_j = 0$ for all $j < i$ because $\boldsymbol{y}_i$ minimizes the function (11.4). Since $\nabla f(\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}$ and $\boldsymbol{y}_{j+1} = \boldsymbol{y}_j + t_j \boldsymbol{v}_j$ for some optimal constant t_j, we can also write

$$\nabla f(\boldsymbol{y}_{j+1}) \;=\; \nabla f(\boldsymbol{y}_j) + t_j \boldsymbol{A}\boldsymbol{v}_j. \tag{11.8}$$

It follows that

$$df(\boldsymbol{y}_i)\boldsymbol{A}\boldsymbol{v}_j \;=\; \frac{1}{t_j} df(\boldsymbol{y}_i)[\nabla f(\boldsymbol{y}_{j+1}) - \nabla f(\boldsymbol{y}_j)] \;=\; 0$$

for $1 \le j < i-1$. Except for the details addressed in the next paragraph, this demonstrates equality (11.7) and completes the proof that the search directions $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are conjugate.

If at any iteration we have $\nabla f(\boldsymbol{y}_i) = \boldsymbol{0}$, then the algorithm terminates with the global minimum. Otherwise, equations (11.3) and (11.5) show that all search directions $\boldsymbol{v}_i$ satisfy

$$\begin{aligned} df(\boldsymbol{y}_i)\boldsymbol{v}_i &= -\|\nabla f(\boldsymbol{y}_i)\|^2 + \alpha_i df(\boldsymbol{y}_i)\boldsymbol{v}_{i-1} \\ &= -\|\nabla f(\boldsymbol{y}_i)\|^2 \\ &< 0. \end{aligned}$$

As a consequence, $\boldsymbol{v}_i \neq \boldsymbol{0}$, $\boldsymbol{v}_i^* \boldsymbol{A}\boldsymbol{v}_i > 0$, and α_{i+1} is well defined. Finally, the inequality $df(\boldsymbol{y}_i)\boldsymbol{v}_i < 0$ implies that the search direction $\boldsymbol{v}_i$ leads downhill from $\boldsymbol{y}_i$ and that the search constant $t_i > 0$. In fact, the stationarity condition $0 = df(\boldsymbol{y}_{i+1})\boldsymbol{v}_i$ and equation (11.8) lead to the conclusion

$$t_i \;=\; -\frac{df(\boldsymbol{y}_i)\boldsymbol{v}_i}{\boldsymbol{v}_i^* \boldsymbol{A}\boldsymbol{v}_i} \;=\; -\frac{(\boldsymbol{A}\boldsymbol{y}_i + \boldsymbol{b})^* \boldsymbol{v}_i}{\boldsymbol{v}_i^* \boldsymbol{A}\boldsymbol{v}_i}. \tag{11.9}$$

In generalizing the conjugate gradient algorithm to non-quadratic functions, we preserve most of the structure of the algorithm. Thus, the revised algorithm first searches along the negative gradient $\boldsymbol{v}_1 = -\nabla f(\boldsymbol{y}_1)$ emanating from the initial point $\boldsymbol{y}_1$. At iteration i it searches along the direction $\boldsymbol{v}_i$ defined by equality (11.5), avoiding the explicit formula (11.9) for t_i. The formula (11.6) for α_i is problematic because it appears to require $\boldsymbol{A}$. Hestenes and Stiefel recommend the alternative formula

$$\alpha_i = \frac{df(\boldsymbol{y}_i)[\nabla f(\boldsymbol{y}_i) - \nabla f(\boldsymbol{y}_{i-1})]}{\boldsymbol{v}_{i-1}^*[\nabla f(\boldsymbol{y}_i) - \nabla f(\boldsymbol{y}_{i-1})]} \tag{11.10}$$

based on the substitution

$$\boldsymbol{A}\boldsymbol{v}_{i-1} = \frac{1}{t_{i-1}}[\nabla f(\boldsymbol{y}_i) - \nabla f(\boldsymbol{y}_{i-1})].$$

Polak and Ribière suggest the further substitutions

$$\begin{aligned} \boldsymbol{v}_{i-1}^* \nabla f(\boldsymbol{y}_i) &= 0 \\ \boldsymbol{v}_{i-1}^* \nabla f(\boldsymbol{y}_{i-1}) &= [-df(\boldsymbol{y}_{i-1}) + \alpha_{i-1}\boldsymbol{v}_{i-2})]\nabla f(\boldsymbol{y}_{i-1}) \\ &= -\|\nabla f(\boldsymbol{y}_{i-1})\|^2. \end{aligned}$$

These produce the second alternative

$$\alpha_i = \frac{df(\boldsymbol{y}_i)[\nabla f(\boldsymbol{y}_i) - \nabla f(\boldsymbol{y}_{i-1})]}{\|\nabla f(\boldsymbol{y}_{i-1})\|^2}. \tag{11.11}$$

Finally, Fletcher and Reeves note that the identity

$$\begin{aligned} df(\boldsymbol{y}_i)\nabla f(\boldsymbol{y}_{i-1}) &= df(\boldsymbol{y}_i)[-\boldsymbol{v}_{i-1} + \alpha_{i-1}\boldsymbol{v}_{i-2})] \\ &= 0 \end{aligned}$$

yields the third alternative

$$\alpha_i = \frac{\|\nabla f(\boldsymbol{y}_i)\|^2}{\|\nabla f(\boldsymbol{y}_{i-1})\|^2}. \tag{11.12}$$

Almost no one now uses the Hestenes-Stiefel update (11.10). Current opinion is divided between the Polak-Ribière update (11.11) and the Fletcher-Reeves update (11.12). *Numerical Recipes* [215] codes both formulas but leans toward the Polak-Ribière formula. In addition to this issue, there are other practical concerns in implementing the conjugate gradient algorithm. For example, if we fail to stop once the gradient $\nabla f(\boldsymbol{y})$ vanishes, then the Polak-Ribière and Fletcher-Reeves updates are undefined. This suggests stopping when $\|\nabla f(\boldsymbol{y}_i)\| \leq \epsilon\|\nabla f(\boldsymbol{y}_1)\|$ for some small $\epsilon > 0$. There is also the problem of loss of conjugacy. Assuming $f(\boldsymbol{y})$ is defined on R^n, it is common practice to restart the conjugate gradient algorithm with

the steepest descent direction every n iterations. This is also a good idea whenever the descent condition $df(\boldsymbol{y}_i)\boldsymbol{v}_i < 0$ fails. Finally, the algorithm is incomplete without specifying a line search algorithm. The next section discusses some of the ways of conducting a line search. The references [2, 92, 215] provide a fuller account and appropriate computer code.

11.4 Line Search Methods

The secant method of Example 10.2.4 can obviously be adapted to minimize the objective function $f(\boldsymbol{y})$ along a line $r \mapsto \boldsymbol{x} + r\boldsymbol{v}$ in R^n. If we define $g(r) = f(\boldsymbol{x} + r\boldsymbol{v})$, then we proceed by searching for a zero of the derivative $g'(r) = df(\boldsymbol{x} + r\boldsymbol{v})\boldsymbol{v}$. In this guise, the secant method is known as the method of false position. It iterates according to

$$r_{m+1} = r_m - \frac{g'(r_m)(r_{m-1} - r_m)}{g'(r_{m-1}) - g'(r_m)}.$$

Two criticisms of the method of false position immediately come to mind. One is that it indiscriminately heads for maxima as well as minima. Another is that it does not make full use of the available information.

A better alternative to the method of false position is to approximate $g(r)$ by a cubic polynomial matching the values of $g(r)$ and $g'(r)$ at r_m and r_{m-1}. Minimizing the cubic should lead to an improved estimate r_{m+1} of the minimum of $g(r)$. It simplifies matters notationally to rescale the interval by setting $h(s) = g(r_{m-1} + sd_m)$ and $d_m = r_m - r_{m-1}$. Now $s = 0$ corresponds to r_{m-1} and $s = 1$ corresponds to r_m. Furthermore, the chain rule implies $h'(s) = g'(r_{m-1} + sd_m)d_m$. Given these conventions, the theory of Hermite interpolation [123] suggests approximating $h(s)$ by the cubic polynomial

$$\begin{aligned} & p(s) \\ = \; & (s-1)^2 h_0 + s^2 h_1 + s(s-1)[(s-1)(h_0' + 2h_0) + s(h_1' - 2h_1)] \\ = \; & (2h_0 + h_0' - 2h_1 + h_1')s^3 + (-3h_0 - 2h_0' + 3h_1 - h_1')s^2 + h_0' s + h_0, \end{aligned}$$

where $h_0 = h(0)$, $h_0' = h'(0)$, $h_1 = h(1)$, and $h_1' = h'(1)$. One can readily verify that $p(0) = h_0$, $p'(0) = h_0'$, $p(1) = h_1$, and $p'(1) = h_1'$.

The conjugate gradient method is locally descending in the sense that $p'(0) = h_0' < 0$. To be on the cautious side, $p'(1) = h_1' > 0$ should hold and $p(s)$ should be convex throughout the interval $[0, 1]$. To check convexity, it suffices to check the conditions $p''(0) \geq 0$ and $p''(1) \geq 0$ since $p''(s)$ is linear. Straightforward calculation shows that

$$\begin{aligned} p''(0) &= -6h_0 + 6h_1 - 4h_0' - 2h_1' \\ p''(1) &= 6h_0 - 6h_1 + 2h_0' + 4h_1'. \end{aligned}$$

Thus, $p(s)$ is convex throughout $[0,1]$ if and only if

$$\tfrac{1}{3}h_1' + \tfrac{2}{3}h_0' \;\le\; h_1 - h_0 \;\le\; \tfrac{2}{3}h_1' + \tfrac{1}{3}h_0'. \tag{11.13}$$

Under these conditions, a local minimum of $p(s)$ occurs on $[0,1]$. The pertinent root of the two possible roots of $p'(s) = 0$ is determined by the sign of the coefficient $2h_0 + h_0' - 2h_1 + h_1'$ of s^3 in $p(s)$. If this coefficient is positive, then the right root furnishes the minimum, and if this coefficient is negative, then the left root furnishes the minimum. The two roots can be calculated simultaneously by solving the quadratic equation

$$\begin{aligned} p'(s) &= 3(2h_0 + h_0' - 2h_1 + h_1')s^2 + 2(-3h_0 - 2h_0' + 3h_1 - h_1')s + h_0' \\ &= 0. \end{aligned}$$

If the condition $p'(1) = h_1' > 0$ or the convexity conditions (11.13) fail, or if the minimum of the cubic leads to an increase in $g(r)$, then one should fall back on more conservative search methods. Golden search involves recursively bracketing a minimum by three points $a < b < c$ satisfying $g(b) < \min\{g(a), g(c)\}$. The analogous method of bisection brackets a zero of $g(r)$ by two points $a < b$ satisfying $g(a)g(b) < 0$. For the moment we ignore the question of how the initial three points a, b, and c are chosen in golden search.

To replace the bracketing interval (a, c) by a shorter interval, we choose $d \in (a, c)$ so that d belongs to the longer of the two intervals (a, b) and (b, c). Without loss of generality, suppose $b < d < c$. If $g(d) < g(b)$, then the three points $b < d < c$ bracket a minimum. If $g(d) > g(b)$, then the three points $a < b < d$ bracket a minimum. In the case of a tie $g(d) = g(b)$, we choose $b < d < c$ when $g(c) < g(a)$ and $a < b < d$ when $g(a) < g(c)$.

These sensible rules do not address the problem of choosing d. Consider the fractional distances

$$\beta \;=\; \tfrac{b-a}{c-a}, \qquad \delta \;=\; \tfrac{d-b}{c-a}$$

along the interval (a, c). The next bracketing interval will have a fractional length of either $1 - \beta$ or $\beta + \delta$. To guard against the worst case, we should take $1 - \beta = \beta + \delta$. This determines $\delta = 1 - 2\beta$ and hence d. One could leave matters as they now stand, but the argument is taken one step farther in golden search. If we imagine repeatedly performing golden search, then scale similarity is expected to set in eventually so that

$$\beta \;=\; \frac{b-a}{c-a} \;=\; \frac{d-b}{c-b} \;=\; \frac{\delta}{1-\beta}.$$

Substituting $\delta = 1 - 2\beta$ in this identity and cross multiplying give the quadratic $\beta^2 - 3\beta + 1 = 0$ with solution

$$\beta \;=\; \frac{3 - \sqrt{5}}{2}$$

equal to the golden mean of ancient Greek mathematics. Following this reasoning, we should take $\delta = \sqrt{5} - 2 = 0.2361$.

There is little theory to guide us in finding an initial bracketing triple $a < b < c$. It is clear that $a = 0$ is one natural choice. In view of the condition $g'(0) < 0$, the point b can be chosen close to 0 as well. This leaves c, which is usually selected based on specific knowledge of $f(y)$, parabolic extrapolation, or repeated doubling of some small arbitrary distance.

11.5 Stopping Criteria

Deciding when to terminate an iterative method is more subtle than it might seem. In solving a one-dimensional nonlinear equation $g(x) = 0$, there are basically two tests. One can declare convergence when $|g(x_n)|$ is small or when x_n does not change much from one iteration to the next. Ideally, both tests should be satisfied. However, there are questions of scale. Our notion of small depends on the typical magnitudes of $g(x)$ and x, and stopping criteria should reflect these magnitudes [66]. Suppose $a > 0$ represents the typical magnitude of $g(x)$. Then a sensible criterion of the first kind is to stop when $|g(x_n)| < \epsilon a$ for $\epsilon > 0$ small. If $b > 0$ represents the typical magnitude of x, then a sensible criterion of the second kind is to stop when

$$|x_n - x_{n-1}| \quad \leq \quad \epsilon \max\{|x_n|, b\}. \tag{11.14}$$

To achieve p significant digits in the solution x_∞, take $\epsilon = 10^{-p}$.

When we optimize a function $f(x)$ with derivative $g(x) = f'(x)$, a third test comes into play. Now it is desirable for $f(x)$ to remain relatively constant near convergence. If $c > 0$ represents the typical magnitude of $f(x)$, then our final stopping criterion is

$$|f(x_n) - f(x_{n-1})| \quad \leq \quad \epsilon \max\{|f(x_n)|, c\}.$$

The second and third criteria generalize better than the first criterion to higher-dimensional problems because solutions often occur on boundaries or manifolds where the gradient $\nabla f(\boldsymbol{x})$ is not required to vanish. The Karush-Kuhn-Tucker conditions are acceptable substitute for Fermat's condition provided the Lagrange multipliers are known. In higher dimensions, one should apply the criterion (11.14) to each coordinate of $\boldsymbol{x}$. Choice of the typical magnitudes a, b, and c is problem specific, and some optimization programs leave this up to the discretion of the user. Often problems can be rescaled by an appropriate choice of units so that the choice $a = b = c = 1$ is reasonable. When in doubt about typical magnitudes, take this default and check whether the output of a preliminary computer run justifies the assumption.

11.6 Quasi-Newton Methods

Quasi-Newton methods of minimization update the current approximation $\boldsymbol{H}_i$ to the second differential $d^2 f(\boldsymbol{x}_i)$ of the objective function $f(\boldsymbol{x})$ by a low-rank perturbation satisfying a secant condition. The secant condition originates from the first-order Taylor approximation

$$\nabla f(\boldsymbol{x}_{i+1}) - \nabla f(\boldsymbol{x}_i) \quad \approx \quad d^2 f(\boldsymbol{x}_i)(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i).$$

If we set

$$\begin{aligned} \boldsymbol{g}_i &= \nabla f(\boldsymbol{x}_{i+1}) - \nabla f(\boldsymbol{x}_i) \\ \boldsymbol{d}_i &= \boldsymbol{x}_{i+1} - \boldsymbol{x}_i, \end{aligned}$$

then the secant condition reads $\boldsymbol{H}_{i+1}\boldsymbol{d}_i = \boldsymbol{g}_i$. The unique, symmetric, rank-one update to $\boldsymbol{H}_i$ satisfying the secant condition is furnished by Davidon's formula [56]

$$\boldsymbol{H}_{i+1} = \boldsymbol{H}_i + c_i \boldsymbol{v}_i \boldsymbol{v}_i^* \tag{11.15}$$

with the constant c_i and the vector $\boldsymbol{v}_i$ specified by

$$\begin{aligned} c_i &= -\frac{1}{(\boldsymbol{H}_i\boldsymbol{d}_i - \boldsymbol{g}_i)^*\boldsymbol{d}_i} \\ \boldsymbol{v}_i &= \boldsymbol{H}_i\boldsymbol{d}_i - \boldsymbol{g}_i. \end{aligned} \tag{11.16}$$

An immediate concern is that the constant c_i is undefined when the inner product $(\boldsymbol{H}_i\boldsymbol{d}_i - \boldsymbol{g}_i)^*\boldsymbol{d}_i = 0$. In such situations or when

$$|(\boldsymbol{H}_i\boldsymbol{d}_i - \boldsymbol{g}_i)^*\boldsymbol{d}_i| \ll \|\boldsymbol{H}_i\boldsymbol{d}_i - \boldsymbol{g}_i\| \cdot \|\boldsymbol{d}_i\|,$$

then the secant adjustment is ignored, and the value $\boldsymbol{H}_i$ is retained for $\boldsymbol{H}_{i+1}$.

We have stressed the desirability of maintaining a positive definite approximation $\boldsymbol{H}_i$ to the second differential $d^2 f(\boldsymbol{x}_i)$. Because this is not always possible with the rank-one update, numerical analysts have investigated rank-two updates. The involvement of the vectors $\boldsymbol{g}_i$ and $\boldsymbol{H}_i\boldsymbol{d}_i$ in the rank-one update suggests trying a rank-two update of the form

$$\boldsymbol{H}_{i+1} = \boldsymbol{H}_i + b_i \boldsymbol{g}_i \boldsymbol{g}_i^* + c_i \boldsymbol{H}_i \boldsymbol{d}_i \boldsymbol{d}_i^* \boldsymbol{H}_i. \tag{11.17}$$

Taking the product of both sides of this equation with d_i gives

$$\boldsymbol{H}_{i+1}\boldsymbol{d}_i = \boldsymbol{H}_i\boldsymbol{d}_i + b_i \boldsymbol{g}_i \boldsymbol{g}_i^* \boldsymbol{d}_i + c_i \boldsymbol{H}_i \boldsymbol{d}_i \boldsymbol{d}_i^* \boldsymbol{H}_i \boldsymbol{d}_i.$$

To achieve consistency with the secant condition $\boldsymbol{H}_{i+1}\boldsymbol{d}_i = \boldsymbol{g}_i$, we set

$$b_i = \frac{1}{\boldsymbol{g}_i^*\boldsymbol{d}_i}, \qquad c_i = -\frac{1}{\boldsymbol{d}_i^*\boldsymbol{H}_i\boldsymbol{d}_i}.$$

The resulting rank-two update was proposed by Broyden, Fletcher, Goldfarb, and Shanno and is consequently known as the BFGS update.

The symmetric rank-one update (11.15) certainly preserves positive definiteness when $c_i \geq 0$. If $c_i < 0$, then $\boldsymbol{H}_{i+1}$ is positive definite only if

$$\begin{aligned} \boldsymbol{v}_i^* \boldsymbol{H}_i^{-1}[\boldsymbol{H}_i + c_i \boldsymbol{v}_i \boldsymbol{v}_i^*]\boldsymbol{H}_i^{-1}\boldsymbol{v}_i &= \boldsymbol{v}_i^* \boldsymbol{H}_i^{-1}\boldsymbol{v}_i[1 + c_i \boldsymbol{v}_i^* \boldsymbol{H}_i^{-1}\boldsymbol{v}_i] \\ &> 0. \end{aligned}$$

In other words, the condition

$$1 + c_i \boldsymbol{v}_i^* \boldsymbol{H}_i^{-1}\boldsymbol{v}_i > 0 \tag{11.18}$$

must hold. Conversely, condition (11.18) is sufficient to guarantee positive definiteness of $\boldsymbol{H}_{i+1}$. This fact can be most easily demonstrated by noting the Sherman-Morrison inversion formula [197]

$$[\boldsymbol{H}_i + c_i \boldsymbol{v}_i \boldsymbol{v}_i^*]^{-1} = \boldsymbol{H}_i^{-1} - \frac{c_i}{1 + c_i \boldsymbol{v}_i^* \boldsymbol{H}_i^{-1}\boldsymbol{v}_i} \boldsymbol{H}_i^{-1}\boldsymbol{v}_i[\boldsymbol{H}_i^{-1}\boldsymbol{v}_i]^*. \tag{11.19}$$

Formula (11.19) shows that $[\boldsymbol{H}_i + c_i \boldsymbol{v}_i \boldsymbol{v}_i^*]^{-1}$ exists and is positive definite under condition (11.18). Since the inverse of a positive definite matrix is positive definite, it follows that $\boldsymbol{H}_i + c_i \boldsymbol{v}_i \boldsymbol{v}_i^*$ is positive definite as well.

If $c_i < 0$ in the rank-one update but condition (11.18) fails, then there are various options. The preferred is implementation of the trust region strategy discussed in Sect. 11.7. Alternatively, one can shrink c_i to maintain positive definiteness. Unfortunately, condition (11.18) gives too little guidance. Problem 12 shows how to control the size of $\det \boldsymbol{H}_{i+1}$ while simultaneously forcing positive definiteness. An even better strategy that monitors the condition number of $\boldsymbol{H}_{i+1}$ rather than $\det \boldsymbol{H}_{i+1}$ is sketched in Problem 14. Finally, there is the option of using c_i as defined but perturbing $\boldsymbol{H}_{i+1}$ by adding a constant multiple $\mu \boldsymbol{I}$ of the identity matrix. This tactic is similar in spirit to the trust region method. If λ_1 is the smallest eigenvalue of $\boldsymbol{H}_{i+1}$, then $\boldsymbol{H}_{i+1} + \mu I$ is positive definite whenever $\lambda_1 + \mu > 0$. Problem 15 discusses a fast algorithm for finding λ_1. With appropriate safeguards, some numerical analysts [51, 152] consider the rank-one update superior to the BFGS update.

Positive definiteness is almost automatic with the BFGS update (11.17). The key turns out to be the inequality

$$0 < \boldsymbol{g}_i^* \boldsymbol{d}_i = df(\boldsymbol{x}_{i+1})\boldsymbol{d}_i - df(\boldsymbol{x}_i)\boldsymbol{d}_i. \tag{11.20}$$

This is ordinarily true for two reasons. First, because $\boldsymbol{d}_i$ is proportional to the current search direction $\boldsymbol{v}_i = -\boldsymbol{H}_i^{-1}\nabla f(\boldsymbol{x}_i)$, positive definiteness of $\boldsymbol{H}_i^{-1}$ implies $-df(\boldsymbol{x}_i)\boldsymbol{d}_i > 0$. Second, when a full search is conducted, the identity $df(\boldsymbol{x}_{i+1})\boldsymbol{d}_i = 0$ holds. Even a partial search typically entails condition (11.20). Section 12.6 takes up the issue of partial line searches.

The speed of partial line searches compared to that of full line searches makes quasi-Newton methods superior to the conjugate gradient method on small-scale problems.

To show that the BFGS update $\boldsymbol{H}_{i+1}$ is positive definite when condition (11.20) holds, we examine the quadratic form

$$\boldsymbol{u}^*\boldsymbol{H}_{i+1}\boldsymbol{u} \;=\; \boldsymbol{u}^*\boldsymbol{H}_i\boldsymbol{u} + \frac{(\boldsymbol{g}_i^*\boldsymbol{u})^2}{\boldsymbol{g}_i^*\boldsymbol{d}_i} - \frac{(\boldsymbol{u}^*\boldsymbol{H}_i\boldsymbol{d}_i)^2}{\boldsymbol{d}_i^*\boldsymbol{H}_i\boldsymbol{d}_i} \tag{11.21}$$

for $\boldsymbol{u} \neq \boldsymbol{0}$. Applying Cauchy's inequality to the vectors $\boldsymbol{a} = \boldsymbol{H}_i^{1/2}\boldsymbol{u}$ and $\boldsymbol{b} = \boldsymbol{H}_i^{1/2}\boldsymbol{d}_i$ gives

$$(\boldsymbol{u}^*\boldsymbol{H}_i\boldsymbol{d}_i)^2 \;\leq\; (\boldsymbol{u}^*\boldsymbol{H}_i\boldsymbol{u})(\boldsymbol{d}_i^*\boldsymbol{H}_i\boldsymbol{d}_i),$$

with equality if and only if $\boldsymbol{u}$ is proportional to $\boldsymbol{d}_i$. Hence, the sum of the first and third terms on the right of equality (11.21) is nonnegative. In the event that $\boldsymbol{u}$ is proportional to $\boldsymbol{d}_i$, the second term on the right of equality (11.21) is positive by assumption. It follows that $\boldsymbol{u}^*\boldsymbol{H}_{i+1}\boldsymbol{u} > 0$ and therefore that $\boldsymbol{H}_{i+1}$ is positive definite.

In successful applications of quasi-Newton methods, choice of the initial matrix $\boldsymbol{H}_1$ is critical. Setting $\boldsymbol{H}_1 = \boldsymbol{I}$ is convenient but often poorly scaled for a particular problem. In maximum likelihood estimation, the expected information matrix $J(\boldsymbol{x}_1)$, if available, is preferable to the identity matrix. In some problems, $J(\boldsymbol{x})$ is cheap to compute and manipulate for special values of $\boldsymbol{x}$. For instance, $J(\boldsymbol{x})$ may be diagonal in certain circumstances. These special $\boldsymbol{x}$ should be considered as starting points for a quasi-Newton search.

It is possible to carry forward approximations $\boldsymbol{K}_i$ of $d^2 f(\boldsymbol{x}_i)^{-1}$ rather than of $d^2 f(\boldsymbol{x}_i)$. This tactic has the advantage of avoiding matrix inversion in computing the quasi-Newton search direction $\boldsymbol{v}_i = -\boldsymbol{K}_i \nabla f(\boldsymbol{x}_i)$. The basic idea is to restate the secant condition $\boldsymbol{H}_{i+1}\boldsymbol{d}_i = \boldsymbol{g}_i$ as the inverse secant condition $\boldsymbol{K}_{i+1}\boldsymbol{g}_i = \boldsymbol{d}_i$. This substitution leads to the symmetric rank-one update

$$\boldsymbol{K}_{i+1} \;=\; \boldsymbol{K}_i + c_i \boldsymbol{w}_i \boldsymbol{w}_i^*, \tag{11.22}$$

where $c_i = -[(\boldsymbol{K}_i\boldsymbol{g}_i - \boldsymbol{d}_i)^*\boldsymbol{g}_i]^{-1}$ and $\boldsymbol{w}_i = \boldsymbol{K}_i\boldsymbol{g}_i - \boldsymbol{d}_i$. Note that monitoring positive definiteness of $\boldsymbol{K}_i$ is still an issue.

For a rank-two update, our earlier arguments apply provided we interchange the roles of $\boldsymbol{d}_i$ and $\boldsymbol{g}_i$. The Davidon-Fletcher-Powell (DFP) update

$$\boldsymbol{K}_{i+1} \;=\; \boldsymbol{K}_i + b_i \boldsymbol{d}_i \boldsymbol{d}_i^* + c_i \boldsymbol{K}_i \boldsymbol{g}_i \boldsymbol{g}_i^* \boldsymbol{K}_i \tag{11.23}$$

with

$$b_i \;=\; \frac{1}{\boldsymbol{g}_i^*\boldsymbol{d}_i}, \qquad c_i \;=\; -\frac{1}{\boldsymbol{g}_i^*\boldsymbol{K}_i\boldsymbol{g}_i}$$

is a competitor to the BFGS update, but the consensus seems to be that the BFGS update is superior to the DFP update in practice [66].

In closing this section, we would like to prove that the BFGS algorithm with an exact line search converges in n or fewer iterations for the strictly convex quadratic function (11.2) defined on R^n. Recall that at iteration i we search along the direction $\boldsymbol{v}_i = -\boldsymbol{H}_i^{-1}\nabla f(\boldsymbol{x}_i)$ and then in preparation for the next iteration construct $\boldsymbol{H}_{i+1}$ according to the BFGS formula (11.17). Unless $\nabla f(\boldsymbol{x}_i) = \boldsymbol{0}$ and the iterates converge prematurely, the current increment $\boldsymbol{d}_i$ is a positive multiple $t_i\boldsymbol{v}_i$ of the search direction $\boldsymbol{v}_i$. Our proof of convergence consists of a subtle inductive argument proving three claims in parallel. These claims amount to the conjugacy condition $\boldsymbol{v}_{i+1}\boldsymbol{A}\boldsymbol{v}_j = 0$, the extended secant condition $\boldsymbol{H}_{i+1}\boldsymbol{d}_j = \boldsymbol{g}_j$, and the gradient perpendicularity condition $df(\boldsymbol{x}_{i+1})\boldsymbol{v}_j = 0$, each for all $1 \leq j \leq i$ and all $i \leq n$. Given the efficacy of successive searches along conjugate directions as demonstrated in Sect. 11.2, the conjugacy condition $\boldsymbol{v}_{i+1}\boldsymbol{A}\boldsymbol{v}_j = 0$ guarantees convergence to the minimum of $f(\boldsymbol{y})$ in n or fewer iterations.

The case $i = 0$, where all three conditions are vacuous, gets the induction on i started. In general, assume that the three conditions are true for $i-1$ and all $1 \leq j \leq i-1$. Equation (11.3) validates the gradient perpendicularity condition for any set of conjugate directions $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_i$, not just the ones determined by the BFGS update (11.17). Given the gradient identity $\boldsymbol{A}\boldsymbol{d}_j = \boldsymbol{g}_j$ and the validity of the extended secant condition $\boldsymbol{H}_{i+1}\boldsymbol{d}_j = \boldsymbol{g}_j$, we calculate

$$\begin{aligned}
\boldsymbol{v}_{i+1}^*\boldsymbol{A}\boldsymbol{v}_j &= -df(\boldsymbol{x}_{i+1})\boldsymbol{H}_{i+1}^{-1}\boldsymbol{A}\boldsymbol{v}_j \\
&= -t_j^{-1}df(\boldsymbol{x}_{i+1})\boldsymbol{H}_{i+1}^{-1}\boldsymbol{A}\boldsymbol{d}_j \\
&= -t_j^{-1}df(\boldsymbol{x}_{i+1})\boldsymbol{H}_{i+1}^{-1}\boldsymbol{g}_j \\
&= -t_j^{-1}df(\boldsymbol{x}_{i+1})\boldsymbol{d}_j \\
&= -df(\boldsymbol{x}_{i+1})\boldsymbol{v}_j \\
&= 0,
\end{aligned} \tag{11.24}$$

which is the required conjugacy condition.

Thus, it suffices to prove $\boldsymbol{H}_{i+1}\boldsymbol{d}_j = \boldsymbol{g}_j$ for $j \leq i$. The case $j = i$ is just the ordinary secant requirement. For $j < i$, we observe that

$$\boldsymbol{H}_{i+1}\boldsymbol{d}_j = \boldsymbol{H}_i\boldsymbol{d}_j + b_i\boldsymbol{g}_i\boldsymbol{g}_i^*\boldsymbol{d}_j + c_i\boldsymbol{H}_i\boldsymbol{d}_i\boldsymbol{d}_i^*\boldsymbol{H}_i\boldsymbol{d}_j.$$

Now the equalities $\boldsymbol{H}_i\boldsymbol{d}_j = \boldsymbol{g}_j$ and $\boldsymbol{g}_i^*\boldsymbol{d}_j = \boldsymbol{d}_i^*\boldsymbol{A}\boldsymbol{d}_j = 0$ hold by the induction hypothesis. Likewise,

$$\begin{aligned}
\boldsymbol{d}_i^*\boldsymbol{H}_i\boldsymbol{d}_j &= \boldsymbol{d}_i^*\boldsymbol{g}_j \\
&= \boldsymbol{d}_i^*\boldsymbol{A}\boldsymbol{d}_j \\
&= 0
\end{aligned}$$

follows from the induction hypothesis. Combining these equalities makes it clear that $\boldsymbol{H}_{i+1}\boldsymbol{d}_j = \boldsymbol{g}_j$ and completes the induction and the proof.

11.7 Trust Regions

If the quadratic approximation

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_i) + df(\boldsymbol{x}_i)(\boldsymbol{x} - \boldsymbol{x}_i) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_i)^* \boldsymbol{H}_i (\boldsymbol{x} - \boldsymbol{x}_i)$$

to the objective function $f(\boldsymbol{x})$ is poor, then the naive step

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \boldsymbol{H}_i^{-1} \nabla f(\boldsymbol{x}_i) \tag{11.25}$$

computed in a quasi-Newton method may be absurdly large. This situation often occurs for early iterates. One remedy is to minimize the quadratic approximation to $f(\boldsymbol{x})$ subject to the spherical constraint $\|\boldsymbol{x} - \boldsymbol{x}_i\|^2 \leq r^2$ for a fixed radius r. This constrained optimization problem has a solution regardless of whether $\boldsymbol{H}_i$ is positive definite, but to simplify matters in the remaining discussion, we assume that $\boldsymbol{H}_i$ is positive definite. According to Proposition 5.2.1, the solution satisfies the multiplier rule

$$\boldsymbol{0} = \nabla f(\boldsymbol{x}_i) + \boldsymbol{H}_i(\boldsymbol{x} - \boldsymbol{x}_i) + \mu(\boldsymbol{x} - \boldsymbol{x}_i) \tag{11.26}$$

for some nonnegative constant μ. If the point $\boldsymbol{x}_{i+1}$ generated by the ordinary step (11.25) occurs within the open ball $\|\boldsymbol{x} - \boldsymbol{x}_i\| < r$, then the multiplier $\mu = 0$. Of course, there is no guarantee that this choice of $\boldsymbol{x}_{i+1}$ will lead to a decrease in $f(\boldsymbol{x})$. If it does not, then one should reduce r, for instance to $\frac{r}{2}$, and try again.

When the point $\boldsymbol{x}_{i+1}$ generated by the ordinary step (11.25) occurs outside the open ball $\|\boldsymbol{x} - \boldsymbol{x}_i\| < r$, we are obliged to look for a minimum point $\boldsymbol{x}$ to the quadratic approximation satisfying $\|\boldsymbol{x} - \boldsymbol{x}_i\| = r$. In this case the multiplier μ may be positive. The fact that μ is unknown makes it impossible to find the minimum in closed form. In principle, one can overcome this difficulty by solving for μ iteratively, say by Newton's method. Hence, we view equation (11.26) as defining $\boldsymbol{x} - \boldsymbol{x}_i$ as a function of μ and ask for the value of μ that yields $\|\boldsymbol{x} - \boldsymbol{x}_i\| = r$. To simplify notation, let $\boldsymbol{H} = \boldsymbol{H}_i$, $\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{x}_i$, and $\boldsymbol{e} = -\nabla f(\boldsymbol{x}_i)$. We now seek a zero of the function

$$\phi(\mu) = \frac{1}{r} - \frac{1}{\|\boldsymbol{y}(\mu)\|} \tag{11.27}$$

with $\boldsymbol{y}(\mu)$ defined by $(\boldsymbol{H} + \mu \boldsymbol{I})\boldsymbol{y}(\mu) = \boldsymbol{e}$. Note that $\phi(0) > 0$ if and only if $\|\boldsymbol{y}(0)\| > r$. To implement Newton's method, we need $\phi'(\mu)$. An easy calculation shows that

$$\phi'(\mu) = \frac{\boldsymbol{y}(\mu)^* \boldsymbol{y}'(\mu)}{\|\boldsymbol{y}(\mu)\|^3}.$$

Unfortunately, this formula contains the unknown derivative $\boldsymbol{y}'(\mu)$. However, differentiation of the equation $(\boldsymbol{H} + \mu \boldsymbol{I})\boldsymbol{y}(\mu) = \boldsymbol{e}$ readily yields

$$\boldsymbol{y}(\mu) + (\boldsymbol{H} + \mu \boldsymbol{I})\boldsymbol{y}'(\mu) = \boldsymbol{0}, \tag{11.28}$$

which implies $\boldsymbol{y}'(\mu) = -(\boldsymbol{H} + \mu\boldsymbol{I})^{-1}\boldsymbol{y}(\mu)$. The complete formula

$$\phi'(\mu) = -\frac{\boldsymbol{y}(\mu)^*(\boldsymbol{H} + \mu\boldsymbol{I})^{-1}\boldsymbol{y}(\mu)}{\|\boldsymbol{y}(\mu)\|^3}$$

shows that $\phi(\mu)$ is strictly decreasing. Problem 16 asks the reader to calculate $\phi''(\mu)$ and verify that it is nonnegative. Problem 17 asserts that $\phi(\mu)$ is negative for large μ. Hence, there is a unique Lagrange multiplier $\mu_i > 0$ solving $\phi(\mu) = 0$ whenever $\phi(0) > 0$. The corresponding $\boldsymbol{y}(\mu_i)$ solves the trust region problem [198, 241].

If one is willing to extract the spectral decomposition of $\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^t$ of $\boldsymbol{H}_i$, then the process can be simplified. Let $\boldsymbol{z} = \boldsymbol{U}^t(\boldsymbol{x} - \boldsymbol{x}_i)$ and $\boldsymbol{b} = \boldsymbol{U}^t\nabla f(\boldsymbol{x}_i)$. Then the trust region problem reduces to minimizing $\frac{1}{2}\boldsymbol{z}^t\boldsymbol{D}\boldsymbol{z} + \boldsymbol{b}^t\boldsymbol{z}$ subject to $\|\boldsymbol{z}\|^2 \leq r^2$. The stationarity conditions for the corresponding Lagrangian

$$\mathcal{L}(\boldsymbol{z}, \mu) = \frac{1}{2}\boldsymbol{z}^t\boldsymbol{D}\boldsymbol{z} + \boldsymbol{b}^t\boldsymbol{z} + \frac{\mu}{2}\left(\|\boldsymbol{z}\|^2 - r^2\right)$$

yield

$$z_j = -\frac{b_j}{d_j + \mu}.$$

where d_j is the jth diagonal entry of $\boldsymbol{D}$. When $\boldsymbol{z} = -\boldsymbol{D}^{-1}\boldsymbol{b}$ satisfies the constraint $\|\boldsymbol{z}\|^2 \leq r^2$, we take $\mu = 0$. Otherwise, we solve the constraint equality

$$r^2 = \sum_j \left(\frac{b_j}{d_j + \mu}\right)^2$$

for μ numerically and determine $\boldsymbol{z}$ and $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{z} + \boldsymbol{x}_i$ accordingly. For more details about trust regions and their practical implementation, see the books [66, 107, 151, 205].

11.8 Problems

1. Suppose you possess n conjugate vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ for the $n \times n$ positive definite matrix $\boldsymbol{A}$. Describe how you can use the expansion $\boldsymbol{x} = \sum_{i=1}^n c_i\boldsymbol{v}_i$ to solve the linear equation $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.

2. Suppose that $\boldsymbol{A}$ is an $n \times n$ positive definite matrix and that the nontrivial vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ satisfy

$$\boldsymbol{u}_i^*\boldsymbol{A}\boldsymbol{u}_j = 0, \qquad \boldsymbol{u}_i^*\boldsymbol{u}_j = 0$$

for all $i \neq j$. Demonstrate that the $\boldsymbol{u}_i$ are eigenvectors of $\boldsymbol{A}$.

3. Suppose that the $n \times n$ symmetric matrix $\boldsymbol{A}$ satisfies $\boldsymbol{v}^* \boldsymbol{A} \boldsymbol{v} \neq 0$ for all $\boldsymbol{v} \neq \boldsymbol{0}$ and that $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ is a basis of R^n. If one defines $\boldsymbol{v}_1 = \boldsymbol{u}_1$ and inductively

$$\boldsymbol{v}_k \quad = \quad \boldsymbol{u}_k - \sum_{j=1}^{k-1} \frac{\boldsymbol{u}_k^* \boldsymbol{A} \boldsymbol{v}_j}{\boldsymbol{v}_j^* \boldsymbol{A} \boldsymbol{v}_j} \boldsymbol{v}_j$$

for $k = 2, \ldots, n$, then show that the vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ are conjugate and provide a basis of R^n. Note that $\boldsymbol{A}$ need not be positive definite.

4. Consider extracting a root of the equation $x^2 = 0$ by Newton's method starting from $x_0 = 1$. Show that it is impossible to satisfy the convergence criterion

$$|x_n - x_{n-1}| \quad \leq \quad \epsilon \max\{|x_n|, |x_{n-1}|\}$$

for $\epsilon = 10^{-7}$ [66]. This example favors the alternative stopping rule (11.14).

5. Let $\boldsymbol{M}$ be an $n \times n$ matrix and $\boldsymbol{d}$ and $\boldsymbol{g}$ be $n \times 1$ vectors. Show that the matrix

$$\boldsymbol{N}_{\text{opt}} \quad = \quad \boldsymbol{M} + \|\boldsymbol{d}\|^{-2} (\boldsymbol{g} - \boldsymbol{M}\boldsymbol{d}) \boldsymbol{d}^*$$

minimizes the distance $\|\boldsymbol{N} - \boldsymbol{M}\|$ between $\boldsymbol{M}$ and an arbitrary $n \times n$ matrix $\boldsymbol{N}$ subject to the secant condition $\boldsymbol{N}\boldsymbol{d} = \boldsymbol{g}$ [66]. Unfortunately, the rank-one update $\boldsymbol{N}_{\text{opt}}$ is not symmetric when $\boldsymbol{M}$ is symmetric. (Hints: Note that $(\boldsymbol{N} - \boldsymbol{M})\boldsymbol{d} = (\boldsymbol{N}_{\text{opt}} - \boldsymbol{M})\boldsymbol{d}$ for every such $\boldsymbol{N}$ and that the outer product $\boldsymbol{d}\boldsymbol{d}^*$ has induced matrix norm $\|\boldsymbol{d}\|^2$.)

6. Let $\boldsymbol{M}$ be an $n \times n$ symmetric matrix and $\boldsymbol{d}$ and $\boldsymbol{g}$ be $n \times 1$ vectors. Powell proposed the rank-two update

$$\boldsymbol{N}_{\text{opt}} \quad = \quad \boldsymbol{M} + \frac{(\boldsymbol{g} - \boldsymbol{M}\boldsymbol{d})\boldsymbol{d}^* + \boldsymbol{d}(\boldsymbol{g} - \boldsymbol{M}\boldsymbol{d})^*}{\|d\|^2} - \frac{(\boldsymbol{g} - \boldsymbol{M}\boldsymbol{d})^* \boldsymbol{d}\, \boldsymbol{d}\boldsymbol{d}^*}{\|\boldsymbol{d}\|^4}$$

to $\boldsymbol{M}$. Show that $\boldsymbol{N}_{\text{opt}}$ is symmetric, has rank two, and satisfies the secant condition $\boldsymbol{N}_{\text{opt}}\boldsymbol{d} = \boldsymbol{g}$ [66].

7. Continuing Problem 6, show that the matrix $\boldsymbol{N}_{\text{opt}}$ minimizes the distance $\|\boldsymbol{N} - \boldsymbol{M}\|_F$ between $\boldsymbol{M}$ and an arbitrary $n \times n$ symmetric matrix $\boldsymbol{N}$ subject to the secant condition $\boldsymbol{N}\boldsymbol{d} = \boldsymbol{g}$. Here $\|\boldsymbol{A}\|_F$ denotes the Frobenius norm of the matrix $\boldsymbol{A}$ viewed as a vector. Unfortunately, Powell's update does not preserve positive definiteness. (Hints: Note that $(\boldsymbol{N} - \boldsymbol{M})\boldsymbol{d} = (\boldsymbol{N}_{\text{opt}} - \boldsymbol{M})\boldsymbol{d}$ for every such $\boldsymbol{N}$ and

$\|(\boldsymbol{N}-\boldsymbol{M})\boldsymbol{v}\| \geq \|(\boldsymbol{N}_{\text{opt}}-\boldsymbol{M})\boldsymbol{v}\|$ for every $\boldsymbol{v}$ with $\boldsymbol{v}^*\boldsymbol{d}=0$. Apply the identities

$$\|\boldsymbol{A}\|_F^2 \;=\; \|\boldsymbol{A}\boldsymbol{O}\|_F^2 \;=\; \sum_{i=1}^{n}\|\boldsymbol{A}\boldsymbol{o}_m\|_F^2$$

for an orthogonal matrix $\boldsymbol{O}$ with columns $\boldsymbol{o}_1,\ldots,\boldsymbol{o}_n$.)

8. Consider the quadratic function

$$Q(\boldsymbol{x}) \;=\; \frac{1}{2}\boldsymbol{x}^*\begin{pmatrix}2 & 1\\ 1 & 1\end{pmatrix}\boldsymbol{x}+(1,1)\boldsymbol{x}$$

defined on R^2. Compute by hand the iterates of the conjugate gradient and BFGS algorithms starting from $\boldsymbol{x}_1=\boldsymbol{0}$. For the BFGS algorithm take $\boldsymbol{H}_1=\boldsymbol{I}$ and use an exact line search. You should find that the two sequences of iterates coincide. This phenomenon holds more generally for any strictly convex quadratic function in the BFGS algorithm given $\boldsymbol{H}_1=\boldsymbol{I}$ [200].

9. Write a program to implement the conjugate gradient algorithm. Apply it to the function

$$f(\boldsymbol{x}) \;=\; \frac{1}{4}x_1^4+\frac{1}{2}x_2^2-x_1x_2+x_1-x_2$$

with two local minima. Demonstrate that your program will converge to either minimum depending on its starting value.

10. Prove Woodbury's generalization

$$(\boldsymbol{A}+\boldsymbol{U}\boldsymbol{B}\boldsymbol{V}^*)^{-1} \;=\; \boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{B}^{-1}+\boldsymbol{V}^*\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}^*\boldsymbol{A}^{-1}$$

of the Sherman-Morrison matrix inversion formula for compatible matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{U}$, and $\boldsymbol{V}$. Apply this formula to the BFGS rank-two update.

11. A quasi-Newton minimization of the strictly convex quadratic function (11.2) generates a sequence of points $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n+1}$ with $\boldsymbol{A}$-conjugate differences $\boldsymbol{d}_i=\boldsymbol{x}_{i+1}-\boldsymbol{x}_i$. At the final iteration we have argued that the approximate Hessian satisfies $\boldsymbol{H}_{n+1}\boldsymbol{d}_i=\boldsymbol{g}_i$ for $1\leq i\leq n$ and $\boldsymbol{g}_i=\nabla f(\boldsymbol{x}_{i+1})-\nabla f(\boldsymbol{x}_i)$. Show that this implies $\boldsymbol{H}_{n+1}=\boldsymbol{A}$.

12. In the rank-one update, suppose we want both $\boldsymbol{H}_{i+1}$ to remain positive definite and $\det\boldsymbol{H}_{i+1}$ to exceed some constant $\epsilon>0$. Explain how these criteria can be simultaneously met by replacing $c_i<0$ by

$$\max\left\{c_i,\left(\frac{\epsilon}{\det\boldsymbol{H}_i}-1\right)\frac{1}{\boldsymbol{v}_i^*\boldsymbol{H}_i^{-1}\boldsymbol{v}_i}\right\}$$

in updating $\boldsymbol{H}_i$. (Hint: In verifying this sufficient condition, you may want to use the one-dimensional version of the identity

$$\det(\boldsymbol{A})\det(\boldsymbol{B}^{-1} - \boldsymbol{U}^*\boldsymbol{A}^{-1}\boldsymbol{U}) \quad = \quad \det(\boldsymbol{A} - \boldsymbol{U}\boldsymbol{B}\boldsymbol{U}^*)\det(\boldsymbol{B}^{-1})$$

for compatible matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{U}$.)

13. Let $\boldsymbol{H}$ be a positive definite matrix. Prove [33] that

$$\operatorname{tr}(\boldsymbol{H}) - \ln\det(\boldsymbol{H}) \quad \geq \quad \ln[\operatorname{cond}_2(\boldsymbol{H})]. \tag{11.29}$$

The condition number $\operatorname{cond}_2(\boldsymbol{H})$ of $\boldsymbol{H}$ equals $\|\boldsymbol{H}\| \cdot \|\boldsymbol{H}^{-1}\|$, that is the ratio of the largest to smallest eigenvalue of $\boldsymbol{H}$. (Hint: Express $\operatorname{tr}(\boldsymbol{H}) - \ln\det(\boldsymbol{H})$ in terms of the eigenvalues of $\boldsymbol{H}$. Then use the inequalities $\lambda - \ln\lambda \geq 1$ and $\lambda > 2\ln\lambda$ for all $\lambda > 0$.)

14. In Davidon's symmetric rank-one update (11.15), it is possible to control the condition number of $\boldsymbol{H}_{i+1}$ by shrinking the constant c_i. Suppose a moderately sized number δ is chosen. Due to inequality (11.29), one can avoid ill-conditioning in the matrices $\boldsymbol{H}_i$ by imposing the constraint $\operatorname{tr}(\boldsymbol{H}_i) - \ln\det(\boldsymbol{H}_i) \leq \delta$. To see how this fits into the updating scheme (11.15), verify that

$$\begin{aligned} \ln\det(\boldsymbol{H}_{i+1}) &= \ln\det(\boldsymbol{H}_i) + \ln(1 + c_i\boldsymbol{v}_i^*\boldsymbol{H}_i^{-1}\boldsymbol{v}_i) \\ \operatorname{tr}(\boldsymbol{H}_{i+1}) &= \operatorname{tr}(\boldsymbol{H}_i) + c_i\|\boldsymbol{v}_i\|^2. \end{aligned}$$

Employing these results, deduce that $\operatorname{tr}(\boldsymbol{H}_{i+1}) - \ln\det(\boldsymbol{H}_{i+1}) \leq \delta$ provided c_i satisfies

$$c_i\|\boldsymbol{v}_i\|^2 - \ln(1 + c_i\boldsymbol{v}_i^*\boldsymbol{H}_i^{-1}\boldsymbol{v}_i) \quad \leq \quad \delta - \operatorname{tr}(\boldsymbol{H}_i) + \ln\det(\boldsymbol{H}_i).$$

15. Suppose the $n \times n$ symmetric matrix $\boldsymbol{A}$ has eigenvalues

$$\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_{n-1} < \lambda_n.$$

The iterative scheme $x_{i+1} = (\boldsymbol{A} - \eta_i\boldsymbol{I})\boldsymbol{x}_i$ can be used to approximate either λ_1 or λ_n. Consider the criterion

$$\sigma_i \quad = \quad \frac{\boldsymbol{x}_{i+1}^*\boldsymbol{A}\boldsymbol{x}_{i+1}}{\boldsymbol{x}_{i+1}^*\boldsymbol{x}_{i+1}}.$$

Choosing η_i to maximize σ_i causes $\lim_{i\to\infty}\sigma_i = \lambda_n$, while choosing η_i to minimize σ_i causes $\lim_{i\to\infty}\sigma_i = \lambda_1$. Show that the extrema of σ_i as a function of η are given by the roots of the quadratic equation

$$0 \quad = \quad \det\begin{pmatrix} 1 & \eta & \eta^2 \\ \tau_0 & \tau_1 & \tau_2 \\ \tau_1 & \tau_2 & \tau_3 \end{pmatrix},$$

where $\tau_k = \boldsymbol{x}_i^* \boldsymbol{A}^k \boldsymbol{x}_i$. Apply this algorithm to find the largest and smallest eigenvalue of the matrix

$$\boldsymbol{A} = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}.$$

You should find $\lambda_1 = 0.01015$ and $\lambda_4 = 30.2887$ [48].

16. Calculate the second derivative $\phi''(\mu)$ of the function defined in equation (11.27). Prove that $\phi''(\mu) \geq 0$.

17. Show that the solution $\boldsymbol{y}(\mu)$ of the equation $(\boldsymbol{H} + \mu \boldsymbol{I})\boldsymbol{y}(\mu) = \boldsymbol{e}$ satisfies $\lim_{\mu \to \infty} \|\boldsymbol{y}(\mu)\| = 0$. How does this justify the conclusion that the function (11.27) has a zero on $(0, \infty)$ when $\phi(0) > 0$?

12
Analysis of Convergence

12.1 Introduction

Proving convergence of the various optimization algorithms is a delicate exercise. In general, it is helpful to consider local and global convergence patterns separately. The local convergence rate of an algorithm provides a useful benchmark for comparing it to other algorithms. On this basis, Newton's method wins hands down. However, the tradeoffs are subtle. Besides the sheer number of iterations until convergence, the computational complexity and numerical stability of an algorithm are critically important. The MM algorithm is often the epitome of numerical stability and computational simplicity. Scoring lies somewhere between Newton's method and the MM algorithm. It tends to converge more quickly than the MM algorithm and to behave more stably than Newton's method. Quasi-Newton methods also occupy this intermediate zone. Because the issues are complex, all of these algorithms survive and prosper in certain computational niches.

The following short overview of convergence manages to cover only highlights. For the sake of simplicity, only unconstrained problems are treated. Quasi-Newton methods are also ignored. The efforts of a generation of numerical analysts in understanding quasi-Newton methods defy easy summary or digestion. Interested readers can consult one of the helpful references [66, 105, 183, 204]. We emphasize MM and gradient algorithms, partially because a fairly coherent theory for them can be reviewed in a few pages.

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_12,

12.2 Local Convergence

Local convergence of many optimization algorithms hinges on the following result [207].

Proposition 12.2.1 (Ostrowski) *Let the differentiable map $h(\boldsymbol{x})$ from an open set $U \subset \mathsf{R}^n$ into R^n have fixed point $\boldsymbol{y}$. If $\|dh(\boldsymbol{y})\|_\dagger < 1$ for some induced matrix norm, and if $\boldsymbol{x}_0$ is sufficiently close to $\boldsymbol{y}$, then the iterates $\boldsymbol{x}_{m+1} = h(\boldsymbol{x}_m)$ converge to $\boldsymbol{y}$.*

Proof: Let $h(\boldsymbol{x})$ have slope function $s(\boldsymbol{x}, \boldsymbol{y})$ near $\boldsymbol{y}$. For any constant r satisfying $\|dh(\boldsymbol{y})\|_\dagger < r < 1$, we have $\|s(\boldsymbol{x}, \boldsymbol{y})\|_\dagger \leq r$ for $\boldsymbol{x}$ sufficiently close to $\boldsymbol{y}$. It therefore follows from the identities

$$\boldsymbol{x}_{m+1} - \boldsymbol{y} \;=\; h(\boldsymbol{x}_m) - h(\boldsymbol{y}) \;=\; s(\boldsymbol{x}_m, \boldsymbol{y})(\boldsymbol{x}_m - \boldsymbol{y})$$

that a proper choice of $\boldsymbol{x}_0$ yields

$$\|\boldsymbol{x}_{m+1} - \boldsymbol{y}\|_\dagger \;\leq\; \|s(\boldsymbol{x}_m, \boldsymbol{y})\|_\dagger \|\boldsymbol{x}_m - \boldsymbol{y}\|_\dagger \;\leq\; r\|\boldsymbol{x}_m - \boldsymbol{y}\|_\dagger.$$

In other words, the distance from $\boldsymbol{x}_m$ to $\boldsymbol{y}$ contracts by a factor of at least r at every iteration. This proves convergence. ■

Two comments are worth making about Proposition 12.2.1. First, the appearance of a general vector norm and its induced matrix norm obscures the fact that the condition $\rho[dh(\boldsymbol{y})] < 1$ on the spectral radius of $dh(\boldsymbol{y})$ is the operative criterion. One can prove that any induced matrix norm exceeds the spectral radius and that some induced matrix norm comes within ϵ of it for any small $\epsilon > 0$ [166]. Later in this section, we will generate a tight matrix norm by taking an $n \times n$ invertible matrix $\boldsymbol{T}$ and forming $\|\boldsymbol{u}\|_{\boldsymbol{T}} = \|\boldsymbol{T}\boldsymbol{u}\|$. It is easy to check that this defines a legitimate vector norm and that the induced matrix norm $\|\boldsymbol{M}\|_{\boldsymbol{T}}$ on $n \times n$ matrices $\boldsymbol{M}$ satisfies

$$\|\boldsymbol{M}\|_{\boldsymbol{T}} \;=\; \sup_{\boldsymbol{u} \neq \boldsymbol{0}} \frac{\|\boldsymbol{T}\boldsymbol{M}\boldsymbol{u}\|}{\|\boldsymbol{T}\boldsymbol{u}\|} \;=\; \sup_{\boldsymbol{v} \neq \boldsymbol{0}} \frac{\|\boldsymbol{T}\boldsymbol{M}\boldsymbol{T}^{-1}\boldsymbol{v}\|}{\|\boldsymbol{v}\|}.$$

In other words, $\|\boldsymbol{M}\|_{\boldsymbol{T}} = \|\boldsymbol{T}\boldsymbol{M}\boldsymbol{T}^{-1}\|$, and we are back in the familiar terrain covered by the spectral norm.

Our second comment involves two definitions. A sequence $\boldsymbol{x}_m$ is said to converge linearly to a point $\boldsymbol{y}$ at rate $r < 1$ provided

$$\lim_{m \to \infty} \frac{\|\boldsymbol{x}_{m+1} - \boldsymbol{y}\|}{\|\boldsymbol{x}_m - \boldsymbol{y}\|} \;=\; r.$$

The sequence converges quadratically if the limit

$$\lim_{m \to \infty} \frac{\|\boldsymbol{x}_{m+1} - \boldsymbol{y}\|}{\|\boldsymbol{x}_m - \boldsymbol{y}\|^2} \;=\; c$$

exists. Ostrowski's result guarantees at least linear convergence; Newton's method improves linear convergence to quadratic convergence.

Our intention is to apply Ostrowski's result to iteration maps of the type

$$h(\boldsymbol{x}) \;=\; \boldsymbol{x} - A(\boldsymbol{x})^{-1} b(\boldsymbol{x}). \tag{12.1}$$

A point $\boldsymbol{y}$ is fixed by the map $h(\boldsymbol{x})$ if and only if $b(\boldsymbol{y}) = \boldsymbol{0}$. In optimization problems, $b(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$ for some real-valued function $f(\boldsymbol{x})$ defined on R^n. Thus, fixed points correspond to stationary points. The matrix $A(\boldsymbol{x})$ is typically $d^2 f(\boldsymbol{x})$ or a positive definite or negative definite approximation to it. For instance, in statistical applications, $-A(\boldsymbol{x})$ could be either the observed or expected information. In the MM gradient algorithm, $A(\boldsymbol{x})$ is the second differential $d^2 g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ of the surrogate function.

Our first order of business is to compute the differential $dh(\boldsymbol{y})$ and an associated slope function $s_h(\boldsymbol{x}, \boldsymbol{y})$ at a fixed point $\boldsymbol{y}$ of $h(\boldsymbol{x})$ in terms of the slope function $s_b(\boldsymbol{x}, \boldsymbol{y})$ of $b(\boldsymbol{x})$. Because $b(\boldsymbol{y}) = \boldsymbol{0}$ at a fixed point, the calculation

$$\begin{aligned} h(\boldsymbol{x}) - h(\boldsymbol{y}) &= \boldsymbol{x} - \boldsymbol{y} - A(\boldsymbol{x})^{-1}[b(\boldsymbol{x}) - b(\boldsymbol{y})] \\ &= [\boldsymbol{I} - A(\boldsymbol{x})^{-1} s_b(\boldsymbol{x}, \boldsymbol{y})](\boldsymbol{x} - \boldsymbol{y}) \end{aligned} \tag{12.2}$$

identifies the slope function

$$s_h(\boldsymbol{x}, \boldsymbol{y}) \;=\; \boldsymbol{I} - A(\boldsymbol{x})^{-1} s_b(\boldsymbol{x}, \boldsymbol{y})$$

and corresponding differential

$$dh(\boldsymbol{y}) \;=\; \boldsymbol{I} - A(\boldsymbol{y})^{-1} db(\boldsymbol{y}). \tag{12.3}$$

In Newton's method, $A(\boldsymbol{y}) = db(\boldsymbol{y})$ and

$$\boldsymbol{I} - A(\boldsymbol{y})^{-1} db(\boldsymbol{y}) \;=\; \boldsymbol{I} - db(\boldsymbol{y})^{-1} db(\boldsymbol{y}) \;=\; \boldsymbol{0}.$$

Proposition 12.2.1 therefore implies that the Newton iterates are locally attracted to a fixed point $\boldsymbol{y}$. Of course, this conclusion is predicated on the suppositions that $db(\boldsymbol{x})$ tends to $db(\boldsymbol{y})$ as $\boldsymbol{x}$ tends to $\boldsymbol{y}$ and that $db(\boldsymbol{y})$ is invertible. To demonstrate quadratic rather than linear convergence, we now assume that $b(\boldsymbol{x})$ is differentiable and that its differential $db(\boldsymbol{x})$ is Lipschitz in a neighborhood of $\boldsymbol{y}$ with Lipschitz constant λ. Given these assumptions and the identities $h(\boldsymbol{y}) = \boldsymbol{y}$ and $A(\boldsymbol{x}) = db(\boldsymbol{x})$, equation (12.2) implies

$$\begin{aligned} h(\boldsymbol{x}) - \boldsymbol{y} &= -\, db(\boldsymbol{x})^{-1} \left[b(\boldsymbol{x}) - b(\boldsymbol{y}) - db(\boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y}) \right] \\ &\quad +\, db(\boldsymbol{x})^{-1} \left[db(\boldsymbol{x}) - db(\boldsymbol{y}) \right] (\boldsymbol{x} - \boldsymbol{y}) \end{aligned}$$

Since Problem 31 of Chap. 4 supplies the bound

$$\|b(\boldsymbol{x}) - b(\boldsymbol{y}) - db(\boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y})\| \;\leq\; \frac{\lambda}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2,$$

it follows that

$$\|h(\boldsymbol{x}) - \boldsymbol{y}\| \quad \leq \quad \left(\frac{\lambda}{2} + \lambda\right) \|db(\boldsymbol{x})^{-1}\| \cdot \|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

The next proposition summarizes our results.

Proposition 12.2.2 *Let $\boldsymbol{y}$ be a $\mathbf{0}$ of the continuously differentiable function $b(\boldsymbol{x})$ from an open set $U \subset \mathsf{R}^n$ into R^n. If $db(\boldsymbol{y})^{-1}$ is invertible and $db(\boldsymbol{x})$ is Lipschitz in U with constant λ, then Newton's method converges to $\boldsymbol{y}$ at a quadratic rate or better whenever $\boldsymbol{x}_0$ is sufficiently close to $\boldsymbol{y}$.*

Proof: The preceding remarks make it clear that

$$\limsup_{m\to\infty} \frac{\|\boldsymbol{x}_{m+1} - \boldsymbol{y}\|}{\|\boldsymbol{x}_m - \boldsymbol{y}\|^2} \quad \leq \quad \frac{3\lambda}{2}\|db(\boldsymbol{y})^{-1}\|,$$

and this suffices for quadratic convergence or better. ■

We now turn to the MM gradient algorithm. Suppose we are minimizing $f(\boldsymbol{x})$ via the surrogate function $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$. If $\boldsymbol{y}$ is a local minimum of $f(\boldsymbol{x})$, it is reasonable to assume that the matrices $\boldsymbol{C} = d^2 f(\boldsymbol{y})$ and $\boldsymbol{D} = d^2 g(\boldsymbol{y} \mid \boldsymbol{y})$ are positive definite. Because $g(\boldsymbol{x} \mid \boldsymbol{y}) - f(\boldsymbol{x})$ attains its minimum at $\boldsymbol{x} = \boldsymbol{y}$, the matrix difference $\boldsymbol{D} - \boldsymbol{C}$ is certainly positive semidefinite. The MM gradient algorithm iterates take $A(\boldsymbol{x}) = d^2 g(\boldsymbol{x} \mid \boldsymbol{x})$. In view of formula (12.3), the iteration map $h(\boldsymbol{x})$ has differential $\boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{C}$ at $\boldsymbol{y}$. If we let $\boldsymbol{T}$ be the symmetric square root $\boldsymbol{D}^{1/2}$ of $\boldsymbol{D}$, then

$$\begin{aligned} \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{C} &= \boldsymbol{D}^{-1}(\boldsymbol{D} - \boldsymbol{C}) \\ &= \boldsymbol{T}^{-1}\boldsymbol{T}^{-1}(\boldsymbol{D} - \boldsymbol{C})\boldsymbol{T}^{-1}\boldsymbol{T}. \end{aligned}$$

Hence, $\boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{C}$ is similar to $\boldsymbol{T}^{-1}(\boldsymbol{D} - \boldsymbol{C})\boldsymbol{T}^{-1}$.

To establish local attraction of the MM gradient algorithm to $\boldsymbol{y}$, we need to choose an appropriate matrix norm. The choice $\|\boldsymbol{M}\|_{\boldsymbol{T}} = \|\boldsymbol{T}\boldsymbol{M}\boldsymbol{T}^{-1}\|$ serves well because Example 1.4.3 and Proposition 2.2.1 imply that

$$\begin{aligned} \|\boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{C}\|_{\boldsymbol{T}} &= \|\boldsymbol{T}\boldsymbol{T}^{-1}\boldsymbol{T}^{-1}(\boldsymbol{D} - \boldsymbol{C})\boldsymbol{T}^{-1}\boldsymbol{T}\boldsymbol{T}^{-1}\| \\ &= \|\boldsymbol{T}^{-1}(\boldsymbol{D} - \boldsymbol{C})\boldsymbol{T}^{-1}\| \\ &= \sup_{\boldsymbol{u}\neq\boldsymbol{0}} \frac{\boldsymbol{u}^*\boldsymbol{T}^{-1}(\boldsymbol{D} - \boldsymbol{C})\boldsymbol{T}^{-1}\boldsymbol{u}}{\boldsymbol{u}^*\boldsymbol{u}} \\ &= \sup_{\boldsymbol{v}\neq\boldsymbol{0}} \frac{\boldsymbol{v}^*(\boldsymbol{D} - \boldsymbol{C})\boldsymbol{v}}{\boldsymbol{v}^*\boldsymbol{T}^*\boldsymbol{T}\boldsymbol{v}} \\ &= \sup_{\boldsymbol{v}\neq\boldsymbol{0}} \frac{\boldsymbol{v}^*(\boldsymbol{D} - \boldsymbol{C})\boldsymbol{v}}{\boldsymbol{v}^*\boldsymbol{D}\boldsymbol{v}} \\ &= 1 - \inf_{\|\boldsymbol{v}\|=1} \frac{\boldsymbol{v}^*\boldsymbol{C}\boldsymbol{v}}{\boldsymbol{v}^*\boldsymbol{D}\boldsymbol{v}}. \end{aligned}$$

The symmetry and positive semidefiniteness of $\boldsymbol{D} - \boldsymbol{C}$ come into play in the third equality in this string of equalities. By virtue of the positive definiteness of $\boldsymbol{C}$ and $\boldsymbol{D}$, the continuous ratio $\boldsymbol{v}^*\boldsymbol{C}\boldsymbol{v}/\boldsymbol{v}^*\boldsymbol{D}\boldsymbol{v}$ is bounded below by a positive constant on the compact sphere $\{\boldsymbol{v} : \|\boldsymbol{v}\| = 1\}$. It follows that $\|\boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{C}\|_{\boldsymbol{T}} < 1$, and Ostrowski's result applies. Hence, the iterates $\boldsymbol{x}_m$ are locally attracted to $\boldsymbol{y}$.

Calculation of the differential $dh(\boldsymbol{y})$ of an MM iteration map $h(\boldsymbol{x})$ is equally interesting. This map satisfies the equation

$$\nabla g[h(\boldsymbol{x}) \mid \boldsymbol{x}] \;=\; \boldsymbol{0}$$

Assuming that the matrix $d^2g(\boldsymbol{y} \mid \boldsymbol{y})$ is invertible, the implicit function theorem, Proposition 4.6.2, shows that $h(\boldsymbol{x})$ is continuously differentiable with differential

$$dh(\boldsymbol{x}) \;=\; -d^2g[h(\boldsymbol{x}) \mid \boldsymbol{x}]^{-1}d^{11}g[h(\boldsymbol{x}) \mid \boldsymbol{x}]. \tag{12.4}$$

Here $d^{11}g(\boldsymbol{u} \mid \boldsymbol{v})$ denotes the differential of $dg(\boldsymbol{u} \mid \boldsymbol{v})$ with respect to $\boldsymbol{v}$. At the fixed point $\boldsymbol{y}$ of $h(\boldsymbol{x})$, equation (12.4) becomes

$$dh(\boldsymbol{y}) \;=\; -d^2g(\boldsymbol{y} \mid \boldsymbol{y})^{-1}d^{11}g(\boldsymbol{y} \mid \boldsymbol{y}). \tag{12.5}$$

Further simplification can be achieved by taking the differential of

$$\nabla f(\boldsymbol{x}) - \nabla g(\boldsymbol{x} \mid \boldsymbol{x}) \;=\; \boldsymbol{0}$$

and setting $\boldsymbol{x} = \boldsymbol{y}$. These actions give

$$d^2f(\boldsymbol{y}) - d^2g(\boldsymbol{y} \mid \boldsymbol{y}) - d^{11}g(\boldsymbol{y} \mid \boldsymbol{y}) \;=\; \boldsymbol{0}.$$

This last equation can be solved for $d^{11}g(\boldsymbol{y} \mid \boldsymbol{y})$, and the result substituted in equation (12.5). It follows that

$$\begin{aligned} dh(\boldsymbol{y}) &= -d^2g(\boldsymbol{y} \mid \boldsymbol{y})^{-1}[d^2f(\boldsymbol{y}) - d^2g(\boldsymbol{y} \mid \boldsymbol{y})] \\ &= \boldsymbol{I} - d^2g(\boldsymbol{y} \mid \boldsymbol{y})^{-1}d^2f(\boldsymbol{y}), \end{aligned} \tag{12.6}$$

which is precisely the differential computed for the MM gradient algorithm. Hence, the MM and MM gradient algorithms display exactly the same behavior in converging to a stationary point of $f(\boldsymbol{x})$.

These apparently esoteric details have considerable practical value. In the setting of the EM algorithm, we replace $\boldsymbol{x}$ by $\boldsymbol{\theta}$, $f(\boldsymbol{x})$ by the observed data loglikelihood $L(\boldsymbol{\theta})$, and $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ by the minorizing function $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$. The rate of convergence of an EM algorithm is determined by how well the observed data information $-d^2L(\boldsymbol{\theta})$ is approximated by the complete data information matrix $-d^2Q(\boldsymbol{\theta} \mid \boldsymbol{\theta})$ at the optimal point. The difference between these two matrices is termed the missing data information [191, 206]. If the amount of missing data is high, then the algorithm will converge

slowly. The art in devising an EM or MM algorithm lies in choosing a tractable surrogate that matches the objective function as closely possible.

Local convergence of the scoring algorithm is not guaranteed by Proposition 12.2.1 because nothing prevents an eigenvalue of

$$dh(\boldsymbol{y}) \quad = \quad \boldsymbol{I} + J(\boldsymbol{y})^{-1} d^2 L(\boldsymbol{y})$$

from falling below -1. Here $L(\boldsymbol{x})$ is the loglikelihood, $J(\boldsymbol{x})$ is the expected information, and $h(\boldsymbol{x})$ is the scoring iteration map. Scoring with a fixed partial step,

$$\boldsymbol{x}_{m+1} \quad = \quad \boldsymbol{x}_m + tJ(\boldsymbol{x}_m)^{-1} \nabla L(\boldsymbol{x}_m),$$

will converge locally for $t > 0$ sufficiently small. In practice, no adjustment is usually necessary. For reasonably large sample sizes, the expected information matrix $J(\boldsymbol{y})$ approximates the observed information matrix $-d^2 L(\boldsymbol{y})$ well, and the spectral radius of $dh(\boldsymbol{y})$ is nearly 0.

Finally, let us consider local convergence of block relaxation. The argument $\boldsymbol{x} = (\boldsymbol{x}_{[1]}, \boldsymbol{x}_{[2]}, \ldots, \boldsymbol{x}_{[b]})$ of the objective function $f(\boldsymbol{x})$ now splits into disjoint blocks, and $f(\boldsymbol{x})$ is minimized along each block of components $\boldsymbol{x}_{[i]}$ in turn. Let $M_i(\boldsymbol{x})$ denote the update to block i. To compute the differential of the full update $M(\boldsymbol{x})$ at a local optimum $\boldsymbol{y}$, we need compact notation. Let $\partial_i f(\boldsymbol{x})$ denote the partial differential of $f(\boldsymbol{x})$ with respect to block i; the transpose of $\partial_i f(\boldsymbol{x})$ is the partial gradient $\nabla_i f(\boldsymbol{x})$. The updates satisfy the partial gradient equations

$$\boldsymbol{0} \quad = \quad \nabla_i f[M_1(\boldsymbol{x}), \ldots, M_i(\boldsymbol{x}), \boldsymbol{x}_{[i+1]}, \ldots, \boldsymbol{x}_{[b]}]. \tag{12.7}$$

Now let $\partial_j \nabla_i f(x)$ denote the partial differential of the partial gradient $\nabla_i f(\boldsymbol{x})$ with respect to block j. Taking the partial differential of equation (12.7) with respect to block j, applying the chain rule, and substituting the optimal point $\boldsymbol{y} = M(\boldsymbol{y})$ for $\boldsymbol{x}$ yield

$$\begin{aligned} \boldsymbol{0} \quad &= \quad \sum_{k=1}^{i} \partial_k \nabla_i f(\boldsymbol{y}) \partial_j M_k(\boldsymbol{y}), \quad j \leq i \\ \boldsymbol{0} \quad &= \quad \sum_{k=1}^{i} \partial_k \nabla_i f(\boldsymbol{y}) \partial_j M_k(\boldsymbol{y}) + \partial_j \nabla_i f(\boldsymbol{y}), \quad j > i. \end{aligned} \tag{12.8}$$

It is helpful to express these equations in block matrix form.

For example in the case of $b = 3$ blocks, the linear system of equations (12.8) can be represented as $\boldsymbol{L}\, dM(\boldsymbol{y}) = \boldsymbol{D} - \boldsymbol{U}$, where $\boldsymbol{U} = \boldsymbol{L}^*$ and

$$dM(\boldsymbol{y}) \quad = \quad \begin{pmatrix} \partial_1 M_1(\boldsymbol{y}) & \partial_2 M_1(\boldsymbol{y}) & \partial_3 M_1(\boldsymbol{y}) \\ \partial_1 M_2(\boldsymbol{y}) & \partial_2 M_2(\boldsymbol{y}) & \partial_3 M_2(\boldsymbol{y}) \\ \partial_1 M_3(\boldsymbol{y}) & \partial_2 M_3(\boldsymbol{y}) & \partial_3 M_3(\boldsymbol{y} \end{pmatrix}$$

$$\begin{aligned}
\boldsymbol{L} &= \begin{pmatrix} \partial_1\nabla_1 f(\boldsymbol{y}) & \mathbf{0} & \mathbf{0} \\ \partial_1\nabla_2 f(\boldsymbol{y}) & \partial_2\nabla_2 f(\boldsymbol{y}) & \mathbf{0} \\ \partial_1\nabla_3 f(\boldsymbol{y}) & \partial_2\nabla_3 f(\boldsymbol{y}) & \partial_3\nabla_3 f(\boldsymbol{y} \end{pmatrix} \\
\boldsymbol{D} &= \begin{pmatrix} \partial_1\nabla_1 f(\boldsymbol{y}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \partial_2\nabla_2 f(\boldsymbol{y}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \partial_3\nabla_3 f(\boldsymbol{y} \end{pmatrix}.
\end{aligned}$$

The identity $[\partial_j\nabla_i f(\boldsymbol{y})]^* = \partial_i\nabla_j f(\boldsymbol{y})$ between two nontrivial blocks of $\boldsymbol{U}$ and $\boldsymbol{L}$ is a consequence of the equality of mixed partials. The matrix equation $\boldsymbol{L}\,dM(\boldsymbol{y}) = \boldsymbol{D} - \boldsymbol{U}$ can be explicitly solved in the form

$$dM(\boldsymbol{y}) = \boldsymbol{L}^{-1}(\boldsymbol{D} - \boldsymbol{U}).$$

Here $\boldsymbol{L}$ is invertible provided its diagonal blocks $\partial_i\nabla_i f(\boldsymbol{y})$ are invertible. At an optimal point $\boldsymbol{y}$, the partial Hessian matrix $\partial_i\nabla_i f(\boldsymbol{y})$ is always positive semidefinite and usually positive definite as well.

Local convergence of block relaxation hinges on whether the spectral radius ρ of the matrix $\boldsymbol{L}^{-1}(\boldsymbol{U} - \boldsymbol{D})$ satisfies $\rho < 1$. Suppose that λ is an eigenvalue of $\boldsymbol{L}^{-1}(\boldsymbol{D} - \boldsymbol{U})$ with eigenvector $\boldsymbol{v}$. These can be complex. The equality $\boldsymbol{L}^{-1}(\boldsymbol{D} - \boldsymbol{U})\boldsymbol{v} = \lambda\boldsymbol{v}$ implies $(1 - \lambda)\boldsymbol{L}\boldsymbol{v} = (\boldsymbol{L} + \boldsymbol{U} - \boldsymbol{D})\boldsymbol{v}$. Premultiplying this by the conjugate transpose $\boldsymbol{v}^*$ gives

$$\frac{1}{1-\lambda} = \frac{\boldsymbol{v}^*\boldsymbol{L}\boldsymbol{v}}{\boldsymbol{v}^*(\boldsymbol{L} + \boldsymbol{U} - \boldsymbol{D})\boldsymbol{v}}.$$

Hence, the real part of $1/(1 - \lambda)$ satisfies

$$\begin{aligned}
\mathsf{Re}\Big(\frac{1}{1-\lambda}\Big) &= \frac{\boldsymbol{v}^*(\boldsymbol{L} + \boldsymbol{U})\boldsymbol{v}}{2\boldsymbol{v}^*(\boldsymbol{L} + \boldsymbol{U} - \boldsymbol{D})\boldsymbol{v}} \\
&= \frac{1}{2}\left[1 + \frac{\boldsymbol{v}^*\boldsymbol{D}\boldsymbol{v}}{\boldsymbol{v}^* d^2 f(\boldsymbol{y})\boldsymbol{v}}\right] \\
&> \frac{1}{2}
\end{aligned}$$

for $d^2 f(\boldsymbol{y})$ positive definite. If $\lambda = \alpha + \beta\sqrt{-1}$, then the last inequality entails

$$\frac{1-\alpha}{(1-\alpha)^2 + \beta^2} > \frac{1}{2},$$

which is equivalent to $|\lambda|^2 = \alpha^2 + \beta^2 < 1$. Hence, the spectral radius $\rho < 1$.

12.3 Coercive Functions

The concept of coerciveness is critical in establishing the existence of minimum points. A function $f(\boldsymbol{x})$ on R^n is said to be coercive if

$$\lim_{\|\boldsymbol{x}\| \to \infty} f(\boldsymbol{x}) = \infty.$$

If $f(\boldsymbol{x})$ is both lower semicontinuous and coercive, then all sublevel sets $\{\boldsymbol{x} : f(\boldsymbol{x}) \leq c\}$ are compact, and the minimum value of $f(\boldsymbol{x})$ is attained. This improvement of Weierstrass' theorem (Proposition 2.5.4) plays a key role in optimization theory.

Two strategies stand out in proving coerciveness. One revolves around comparing one function to another. For example, suppose $f(x) = x^2 + \sin x$ and $g(x) = x^2 - 1$. Then $g(x)$ is clearly coercive and $f(x) \geq g(x)$. Hence, $f(x)$ is also coercive. As explained in the next proposition, the second strategy is restricted to convex functions. In stating the proposition, we allow $f(\boldsymbol{x})$ to have the value ∞.

Proposition 12.3.1 *Suppose $f(\boldsymbol{x})$ is a convex lower semicontinuous function on R^n. Choose any point $\boldsymbol{y}$ with $f(\boldsymbol{y}) < \infty$. Then $f(\boldsymbol{x})$ is coercive if and only $f(\boldsymbol{x})$ is coercive along all nontrivial rays $\{\boldsymbol{x} \in \mathsf{R}^n : \boldsymbol{x} = \boldsymbol{y} + t\boldsymbol{v}, t \geq 0\}$ emanating from $\boldsymbol{y}$.*

Proof: The stated condition is obviously necessary. To prove that it is sufficient, suppose it holds, but $f(\boldsymbol{x})$ is not coercive. It suffices to take $\boldsymbol{y} = \boldsymbol{0}$ because one can always consider the translated function $g(\boldsymbol{x}) = f(\boldsymbol{x} - \boldsymbol{y})$, which retains the properties of convexity and lower semicontinuity. Let $\boldsymbol{x}_n$ be a sequence such that $\lim_{n\to\infty} \|\boldsymbol{x}_n\| = \infty$ and $\limsup_{n\to\infty} f(\boldsymbol{x}_n) < \infty$. By passing to a subsequence if necessary, we can assume that the unit vectors $\boldsymbol{v}_n = \|\boldsymbol{x}_n\|^{-1}\boldsymbol{x}_n$ converge to a unit vector $\boldsymbol{v}$. For $t > 0$ and n large enough, convexity implies

$$f(t\boldsymbol{v}_n) \quad \leq \quad \frac{t}{\|\boldsymbol{x}_n\|} f(\boldsymbol{x}_n) + \Big(1 - \frac{t}{\|\boldsymbol{x}_n\|}\Big) f(\boldsymbol{0}).$$

It follows that $\liminf_{n\to\infty} f(t\boldsymbol{v}_n) \leq f(\boldsymbol{0})$. On the other hand, lower semicontinuity entails

$$f(t\boldsymbol{v}) \quad \leq \quad \liminf_{n\to\infty} f(t\boldsymbol{v}_n) \quad \leq \quad f(\boldsymbol{0}).$$

Hence, $f(\boldsymbol{x})$ does not tend to ∞ along the ray $t\boldsymbol{v}$, contradicting our assumption. ■

For example, consider the quadratic $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x} + c$. If $\boldsymbol{A}$ is positive definite and $\boldsymbol{v} \neq \boldsymbol{0}$, then $f(t\boldsymbol{v})$ is a quadratic in t with positive leading coefficient. Hence, $f(\boldsymbol{x})$ is coercive along the ray $t\boldsymbol{v}$. It follows that $f(\boldsymbol{x})$ is coercive. When $\boldsymbol{A}$ is positive semidefinite and $\boldsymbol{A}\boldsymbol{v} = \boldsymbol{0}$, $f(t\boldsymbol{v})$ is linear in t. If the leading coefficient $\boldsymbol{b}^*\boldsymbol{v}$ is positive, then $f(\boldsymbol{x})$ is coercive along the ray $t\boldsymbol{v}$. However, $f(\boldsymbol{x})$ then fails to be coercive along the ray $-t\boldsymbol{v}$. Hence, $f(\boldsymbol{x})$ is not coercive. This fact does not prevent $f(\boldsymbol{x})$ from attaining its minimum. If $\boldsymbol{x}$ satisfies $\boldsymbol{A}\boldsymbol{x} = -\boldsymbol{b}$, then the convex function $f(\boldsymbol{x})$ has a stationary point, which necessarily furnishes a minimum value.

Posynomials present another interesting test case. In the exponential parameterization $t_k = e^{x_k}$, one can represent a posynomial as

$$f(\boldsymbol{x}) \quad = \quad \sum_{i=1}^{j} c_i e^{\boldsymbol{\beta}_i^* \boldsymbol{x}}.$$

At the preferred point $\mathbf{0}$, it is obvious that $f(t\boldsymbol{v})$ tends to ∞ as t tends to ∞ if and only if at least one $\boldsymbol{\beta}_i$ satisfies $\boldsymbol{\beta}_i^* \boldsymbol{v} > 0$. In other words, $f(\boldsymbol{x})$ is coercive if and only if the polar cone

$$C \quad = \quad \{\boldsymbol{v} : \boldsymbol{\beta}_i^* \boldsymbol{v} \leq 0 \text{ for all } i\}$$

consists of the trivial vector $\mathbf{0}$ alone. Section 14.3.7 treats polar cones in more detail.

In the original posynomial parameterization, t_k tends to 0 as x_k tends to $-\infty$. This suggests the need for a broader definition of coerciveness consistent with Weierstrass' theorem. Suppose the lower semicontinuous function $f(\boldsymbol{x})$ is defined on an open set U. To avoid colliding with the boundary of U, we assume that the set

$$C_{\boldsymbol{y}} \quad = \quad \{\boldsymbol{x} \in U : f(\boldsymbol{x}) \leq f(\boldsymbol{y})\}$$

is compact for every $\boldsymbol{y} \in U$. If this is the case, then $f(\boldsymbol{x})$ attains its minimum somewhere in U. The essence of the expanded definition of coerciveness is that $f(\boldsymbol{x})$ tends to ∞ as $\boldsymbol{x}$ approaches the boundary of U or $\|\boldsymbol{x}\|$ approaches ∞.

12.4 Global Convergence of the MM Algorithm

In this section and the next, we tackle global convergence. We begin with the MM algorithm and consider without loss of generality minimization of the objective function $f(\boldsymbol{x})$ via the majorizing surrogate $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$. In studying global convergence, we must carefully specify the parameter domain U. Let us take U to be any open convex subset of R^n. It is convenient to assume that $f(\boldsymbol{x})$ is coercive on U in the sense just specified and that whenever necessary $f(\boldsymbol{x})$ and $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ and their various first and second differentials are jointly continuous in $\boldsymbol{x}$ and $\boldsymbol{x}_m$.

We also demand that the second differential $d^2 g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ be positive definite. This implies that $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ is strictly convex. Strict convexity in turn implies that the solution $\boldsymbol{x}_{m+1}$ of the minimization step is unique. Existence of a solution fortunately is guaranteed by coerciveness. Indeed, the closed set

$$\{\boldsymbol{x} : g(\boldsymbol{x} \mid \boldsymbol{x}_m) \leq g(\boldsymbol{x}_m \mid \boldsymbol{x}_m) = f(\boldsymbol{x}_m)\}$$

is compact because it is contained within the compact set

$$\{\boldsymbol{x} : f(\boldsymbol{x}) \leq f(\boldsymbol{x}_m)\}.$$

Finally, the implicit function theorem, Proposition 4.6.2, shows that the iteration map $\boldsymbol{x}_{m+1} = M(\boldsymbol{x}_m)$ is continuously differentiable in a neighborhood of every point $\boldsymbol{x}_m$. Local differentiability of $M(\boldsymbol{x})$ clearly extends to global differentiability.

Gradient versions of the algorithm (12.1) have the property that stationary points of the objective function and fixed points of the iteration map coincide. This property also applies to the MM algorithm. Here we recall the two identities $\nabla g(\boldsymbol{x}_{m+1} \mid \boldsymbol{x}_m) = \boldsymbol{0}$ and $\nabla g(x_m \mid x_m) = \nabla f(\boldsymbol{x}_m)$ and the strict convexity of $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$. By the same token, stationary points and only stationary points give equality in the descent inequality $f[M(\boldsymbol{x})] \leq f(\boldsymbol{x})$.

The next technical proposition prepares the ground for a proof of global convergence. We remind the reader that a point $\boldsymbol{y}$ is a cluster point of a sequence $\boldsymbol{x}_m$ provided there is a subsequence $\boldsymbol{x}_{m_k}$ that tends to $\boldsymbol{y}$. One can easily verify that any limit of a sequence of cluster points is also a cluster point and that a bounded sequence has a limit if and only if it has at most one cluster point. See Problem 21.

Proposition 12.4.1 *If a bounded sequence $\boldsymbol{x}_m$ in R^n satisfies*

$$\lim_{m\to\infty} \|\boldsymbol{x}_{m+1} - \boldsymbol{x}_m\| = 0, \tag{12.9}$$

then its set T of cluster points is connected. If T is finite, then T reduces to a single point, and $\lim_{m\to\infty} \boldsymbol{x}_m = \boldsymbol{y}$ exists.

Proof: It is straightforward to prove that T is a compact set. If it is disconnected, then there is a continuous disconnecting function $\phi(\boldsymbol{x})$ having exactly the two values 0 and 1. The inverse images of the closed sets 0 and 1 under $\phi(\boldsymbol{x})$ can be represented as the intersections $T_0 = T \cap C_0$ and $T_1 = T \cap C_1$ of T with two closed sets C_0 and C_1. Because T is compact, T_0 and T_1 are closed, nonempty, and disjoint. Furthermore, the distance

$$\operatorname{dist}(T_0, T_1) = \inf_{\boldsymbol{u}\in T_0} \operatorname{dist}(\boldsymbol{u}, T_1) = \inf_{\boldsymbol{u}\in T_0, \boldsymbol{v}\in T_1} \|\boldsymbol{u} - \boldsymbol{v}\|$$

separating T_0 and T_1 is positive. Indeed, the continuous function $\operatorname{dist}(\boldsymbol{u}, T_1)$ attains its minimum at some point $\boldsymbol{u}$ of the compact set T_0, and the distance $\operatorname{dist}(\boldsymbol{u}, T_1)$ separating that $\boldsymbol{u}$ from T_1 must be positive because T_1 is closed.

Now consider the sequence $\boldsymbol{x}_m$ in the statement of the proposition. For large enough m, we have $\|\boldsymbol{x}_{m+1} - \boldsymbol{x}_m\| < \operatorname{dist}(T_0, T_1)/4$. As the sequence $\boldsymbol{x}_m$ bounces back and forth between cluster points in T_0 and T_1, it must enter the closed set $W = \{\boldsymbol{u} : \operatorname{dist}(\boldsymbol{u}, T) \geq \operatorname{dist}(T_0, T_1)/4\}$ infinitely often. But this means that W contains a cluster point of $\boldsymbol{x}_m$. Because W is disjoint from T_0 and T_1, and these two sets are postulated to contain all of the cluster points of $\boldsymbol{x}_m$, this contradiction implies that T is connected.

Because a finite set with more than one point is necessarily disconnected, T can be a finite set only if it consists of a single point. Finally, a bounded sequence with only a single cluster point has that point as its limit. ■

With these facts in mind, we now state and prove a version of Liapunov's theorem for discrete dynamical systems [183].

Proposition 12.4.2 (Liapunov) *Let Γ be the set of cluster points generated by the MM sequence $\boldsymbol{x}_{m+1} = M(\boldsymbol{x}_m)$ starting from some initial $\boldsymbol{x}_0$. Then Γ is contained in the set S of stationary points of $f(\boldsymbol{x})$.*

Proof: The sequence $\boldsymbol{x}_m$ stays within the compact set

$$\{\boldsymbol{x} \in U : f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}.$$

Consider a cluster point $\boldsymbol{z} = \lim_{k\to\infty} \boldsymbol{x}_{m_k}$. Since the sequence $f(\boldsymbol{x}_m)$ is monotonically decreasing and bounded below, $\lim_{m\to\infty} f(\boldsymbol{x}_m)$ exists. Hence, taking limits in the inequality $f[M(\boldsymbol{x}_{m_k})] \leq f(\boldsymbol{x}_{m_k})$ and invoking the continuity of $M(\boldsymbol{x})$ and $f(\boldsymbol{x})$ imply $f[M(\boldsymbol{z})] = f(\boldsymbol{z})$. Thus, $\boldsymbol{z}$ is a fixed point of $M(\boldsymbol{x})$ and consequently also a stationary point of $f(\boldsymbol{x})$. ■

The next two propositions are adapted from the reference [195]. In the second of these, recall that a point $\boldsymbol{x}$ in a set S is isolated if and only if there exists a radius $r > 0$ such that $S \cap B(\boldsymbol{x}, r) = \{\boldsymbol{x}\}$.

Proposition 12.4.3 *The set of cluster points Γ of $\boldsymbol{x}_{m+1} = M(\boldsymbol{x}_m)$ is compact and connected.*

Proof: Γ is a closed subset of the compact set $\{\boldsymbol{x} \in U : f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$ and is therefore itself compact. According to Proposition 12.4.1, Γ is connected provided $\lim_{m\to\infty} \|\boldsymbol{x}_{m+1} - \boldsymbol{x}_m\| = 0$. If this sufficient condition fails, then the compactness of $\{\boldsymbol{x} \in U : f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$ makes it possible to extract a subsequence $\boldsymbol{x}_{m_k}$ such that $\lim_{k\to\infty} \boldsymbol{x}_{m_k} = \boldsymbol{u}$ and $\lim_{k\to\infty} \boldsymbol{x}_{m_k+1} = \boldsymbol{v}$ both exist, but $\boldsymbol{v} \neq \boldsymbol{u}$. However, the continuity of $M(\boldsymbol{x})$ requires $\boldsymbol{v} = M(\boldsymbol{u})$ while the descent condition implies

$$f(\boldsymbol{v}) \quad = \quad f(\boldsymbol{u}) \quad = \quad \lim_{m\to\infty} f(\boldsymbol{x}_m).$$

The equality $f(\boldsymbol{v}) = f(\boldsymbol{u})$ forces the contradictory conclusion that $\boldsymbol{u}$ is a fixed point of $M(\boldsymbol{x})$. Hence, the sufficient condition (12.9) for connectivity holds. ■

Proposition 12.4.4 *Suppose that all stationary points of $f(\boldsymbol{x})$ are isolated and that the stated differentiability, coerciveness, and convexity assumptions are true. Then any sequence of iterates $\boldsymbol{x}_{m+1} = M(\boldsymbol{x}_m)$ generated by the iteration map $M(\boldsymbol{x})$ of the MM algorithm possesses a limit, and that limit is a stationary point of $f(\boldsymbol{x})$. If $f(\boldsymbol{x})$ is strictly convex, then $\lim_{m\to\infty} \boldsymbol{x}_m$ is the minimum point.*

Proof: In the compact set $\{\boldsymbol{x} \in U : f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$ there can only be a finite number of stationary points. An infinite number of stationary points would admit a convergent sequence whose limit would not be isolated. Since the set of cluster points Γ is a connected subset of this finite set of stationary points, Γ reduces to a single point. ■

Two remarks on Proposition 12.4.4 are in order. First, except when strict convexity prevails for $f(\boldsymbol{x})$, the proposition offers no guarantee that the limit $\boldsymbol{y}$ of the sequence $\boldsymbol{x}_m$ furnishes a global minimum. Problem 11 contains a counterexample of Wu [275] exhibiting convergence to a saddle point in the EM algorithm. Fortunately, in practice, descent algorithms almost always converge to at least a local minimum of the objective function. Second, suppose that the set S of stationary points possesses a sequence $\boldsymbol{z}_m \in S$ converging to $\boldsymbol{z} \in S$ with $\boldsymbol{z}_m \neq \boldsymbol{z}$ for all m. Because the unit sphere in R^n is compact, we can extract a subsequence such that

$$\lim_{k\to\infty} \frac{\boldsymbol{z}_{m_k} - \boldsymbol{z}}{\|\boldsymbol{z}_{m_k} - \boldsymbol{z}\|} = \boldsymbol{v}$$

exists and is nontrivial. Now let $s_{\nabla f}(\boldsymbol{y}, \boldsymbol{x})$ be a slope function for $\nabla f(\boldsymbol{x})$. Taking limits in

$$\begin{aligned} \mathbf{0} &= \frac{1}{\|\boldsymbol{z}_{m_k} - \boldsymbol{z}\|}[\nabla f(\boldsymbol{z}_{m_k}) - \nabla f(\boldsymbol{z})] \\ &= \frac{1}{\|\boldsymbol{z}_{m_k} - \boldsymbol{z}\|} s_{\nabla f}(\boldsymbol{z}_{m_k}, \boldsymbol{z})(\boldsymbol{z}_{m_k} - \boldsymbol{z}) \end{aligned}$$

then produces $\mathbf{0} = d^2 f(\boldsymbol{z})\boldsymbol{v}$. In other words, the second differential at $\boldsymbol{z}$ is singular. If one can rule out such degeneracies, then all stationary points are isolated. Interested readers can consult the literature on Morse functions for further commentary on this subject [115].

12.5 Global Convergence of Block Relaxation

Verification of global convergence of block relaxation parallels the MM algorithm case. Careful scrutiny of the proof of Proposition 12.4.4 shows that it relies on five properties of the objective function $f(\boldsymbol{x})$ and the iteration map $M(\boldsymbol{x})$:

(a) $f(\boldsymbol{x})$ is coercive on its convex open domain U,

(b) $f(\boldsymbol{x})$ has only isolated stationary points,

(c) $M(\boldsymbol{x})$ is continuous,

(d) $\boldsymbol{y}$ is a fixed point of $M(\boldsymbol{x})$ if and only if it is a stationary point of $f(\boldsymbol{x})$,

(e) $f[M(\boldsymbol{y})] \leq f(\boldsymbol{y})$, with equality if and only if $\boldsymbol{y}$ is a fixed point of $M(\boldsymbol{x})$.

Let us suppose for notational simplicity that the argument $\boldsymbol{x} = (\boldsymbol{v}, \boldsymbol{w})$ breaks into just two blocks. Criteria (a) and (b) can be demonstrated for many objective functions and are independent of the algorithm chosen to minimize $f(\boldsymbol{x})$. In block relaxation we ordinarily take U to be the Cartesian product $V \times W$ of two convex open sets. If we assume that $f(\boldsymbol{v}, \boldsymbol{w})$ is strictly convex in $\boldsymbol{v}$ for fixed $\boldsymbol{w}$ and vice versa, then the block relaxation updates are well defined. If $f(\boldsymbol{v}, \boldsymbol{w})$ is twice continuously differentiable, and $d^2_{\boldsymbol{v}} f(\boldsymbol{v}, \boldsymbol{w})$ and $d^2_{\boldsymbol{w}} f(\boldsymbol{v}, \boldsymbol{w})$ are invertible matrices, then application of the implicit function theorem demonstrates that the iteration map $M(\boldsymbol{x})$ is a composition of two differentiable maps. Criterion (c) is therefore valid. A fixed point $\boldsymbol{x} = (\boldsymbol{v}, \boldsymbol{w})$ satisfies the two equation $\nabla_{\boldsymbol{v}} f(\boldsymbol{v}, \boldsymbol{w}) = \boldsymbol{0}$ and $\nabla_{\boldsymbol{w}} f(\boldsymbol{v}, \boldsymbol{w}) = \boldsymbol{0}$, and criterion (d) follows. Finally, both block updates decrease $f(\boldsymbol{x})$. They give a strict decrease if and only if they actually change either argument $\boldsymbol{v}$ or $\boldsymbol{w}$. Hence, criterion (e) is true. We emphasize that collectively these are sufficient but not necessary conditions. Observe that we have not assumed that $f(\boldsymbol{v}, \boldsymbol{w})$ is convex in both variables simultaneously.

12.6 Global Convergence of Gradient Algorithms

We now turn to the question of global convergence for gradient algorithms of the sort

$$\boldsymbol{x}_{m+1} \quad = \quad \boldsymbol{x}_m - A(\boldsymbol{x}_m)^{-1} \nabla f(\boldsymbol{x}_m).$$

The assumptions concerning $f(\boldsymbol{x})$ made in the previous section remain in force. A major impediment to establishing the global convergence of any minimization algorithm is the possible failure of the descent property

$$f(\boldsymbol{x}_{m+1}) \quad \leq \quad f(\boldsymbol{x}_m)$$

enjoyed by the MM algorithm. Provided the matrix $A(\boldsymbol{x}_m)$ is positive definite, the direction $\boldsymbol{v}_m = -A(\boldsymbol{x}_m)^{-1} \nabla f(\boldsymbol{x}_m)$ is guaranteed to point locally downhill. Hence, if we elect the natural strategy of instituting a limited line search along the direction $\boldsymbol{v}_m$ emanating from $\boldsymbol{x}_m$, then we can certainly find an $\boldsymbol{x}_{m+1}$ that decreases $f(\boldsymbol{x})$.

Although an exact line search is tempting, we may pay too great a price for precision when we merely need progress. The step-halving tactic mentioned in Chap. 10 is better than a full line search but not quite adequate for theoretical purposes. Instead, we require a sufficient decrease along a descent direction $\boldsymbol{v}$. This is summarized by the Armijo rule of considering only steps $t\boldsymbol{v}$ satisfying the inequality

$$f(\boldsymbol{x} + t\boldsymbol{v}) \quad \leq \quad f(\boldsymbol{x}) + \alpha t df(\boldsymbol{x})\boldsymbol{v} \tag{12.10}$$

for $t > 0$ and some fixed α in $(0, 1)$. To avoid too stringent a test, we take a low value of α such as 0.01. In combining Armijo's rule with regular step decrementing, we first test the step $\boldsymbol{v}$. If it satisfies Armijo's rule we are done. If it fails, we choose $\sigma \in (0, 1)$ and test $\sigma\boldsymbol{v}$. It this fails, we test $\sigma^2\boldsymbol{v}$, and so forth until we encounter and take the first partial step $\sigma^k\boldsymbol{v}$ that works. In step halving, obviously $\sigma = 1/2$.

Step halving can be combined with a partial line search. For instance, suppose the line search has been confined to the interval $t \in [0, s]$. If the point $\boldsymbol{x} + s\boldsymbol{v}$ passes Armijo's test, then we accept it. Otherwise, we fit a cubic to the function $t \mapsto f(\boldsymbol{x} + t\boldsymbol{v})$ on the interval $[0, s]$ as described in Sect. 11.4. If the minimum point t of the cubic approximation satisfies $t \geq \sigma s$ and passes Armijo's test, then we accept $\boldsymbol{x} + t\boldsymbol{v}$. Otherwise, we replace the interval $[0, s]$ by the interval $[0, \sigma s]$ and proceed inductively. For the sake of simplicity in the sequel, we will ignore this elaboration of step halving and concentrate on the unadorned version.

We would like some guarantee that the exponent k of the step decrementing power σ^k does not grow too large. Mindful of this criterion, we suppose that the positive definite matrix $A(\boldsymbol{x})$ depends continuously on $\boldsymbol{x}$. This is not much of a restriction for Newton's method, the Gauss-Newton algorithm, the MM gradient algorithm, or scoring. If we combine continuity with coerciveness, then we can conclude that there exist positive constants β, γ, δ, and ϵ with

$$\begin{aligned} \|A(\boldsymbol{x})\| &\leq \beta, \qquad \|A(\boldsymbol{x})^{-1}\| \leq \gamma \\ \|\nabla f(\boldsymbol{x})\| &\leq \epsilon, \qquad \|s_f^2(\boldsymbol{y}, \boldsymbol{x})\| \leq \delta \end{aligned}$$

for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in the compact set $D = \{\boldsymbol{x} \in U : f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$ where any descent algorithm acts. Here $s_f^2(\boldsymbol{y}, \boldsymbol{x})$ is the second slope of $f(\boldsymbol{x})$.

Before we tackle Armijo's rule, let us consider the more pressing question of whether the proposed points $\boldsymbol{x} + \boldsymbol{v}$ lie in the domain U of $f(\boldsymbol{x})$. This is too much to hope for, but it is worth considering whether $x + \sigma^d\boldsymbol{v}$ always lies in U for some fixed power σ^d. Fortunately, $v(\boldsymbol{x}) = -A(\boldsymbol{x})^{-1}\nabla f(\boldsymbol{x})$ satisfies the bound

$$\|v(\boldsymbol{x})\| \leq \gamma\epsilon$$

on D. Now suppose no single power σ^k is adequate for all $\boldsymbol{x} \in D$. Then there exists a sequence of points $\boldsymbol{x}_k \in D$ with $\boldsymbol{y}_k = \boldsymbol{x}_k + \sigma^k v(\boldsymbol{x}_k) \notin U$. Passing to a subsequence if necessary, we can assume that $\boldsymbol{x}_k$ converges to $\boldsymbol{x} \in D$. Because σ^k is tending to 0, and $v(\boldsymbol{x})$ is bounded on D, the sequence $\boldsymbol{y}_k$ likewise converges to $\boldsymbol{x}$. Since the complement of U is closed, $\boldsymbol{x}$ must lie in the complement of U as well as in D. This contradiction proves our contention.

To use these bounds, let $\boldsymbol{v} = -A(\boldsymbol{x})^{-1}\nabla f(\boldsymbol{x})$ and consider the inequality

$$f(\boldsymbol{x} + t\boldsymbol{v}) = f(\boldsymbol{x}) + tdf(\boldsymbol{x})\boldsymbol{v} + \frac{1}{2}t^2\boldsymbol{v}^* s_f^2(\boldsymbol{x} + t\boldsymbol{v}, \boldsymbol{x})\boldsymbol{v}$$

$$\leq \quad f(\boldsymbol{x}) + tdf(\boldsymbol{x})\boldsymbol{v} + \frac{1}{2}t^2\delta\|\boldsymbol{v}\|^2 \tag{12.11}$$

for $\boldsymbol{x}$ and $\boldsymbol{x} + t\boldsymbol{v}$ in D. Taking into account the bound on $\|A(\boldsymbol{x})\|$ and the identity

$$\|A(\boldsymbol{x})^{1/2}\| \quad = \quad \|A(\boldsymbol{x})\|^{1/2}$$

entailed by Proposition 2.2.1, we also have

$$\begin{aligned} \|\nabla f(\boldsymbol{x})\|^2 &= \|A(\boldsymbol{x})^{1/2}A(\boldsymbol{x})^{-1/2}\nabla f(\boldsymbol{x})\|^2 \\ &\leq \|A(\boldsymbol{x})^{1/2}\|^2\|A(\boldsymbol{x})^{-1/2}\nabla f(\boldsymbol{x})\|^2 \\ &\leq \beta df(\boldsymbol{x})A(\boldsymbol{x})^{-1}\nabla f(\boldsymbol{x}). \end{aligned} \tag{12.12}$$

It follows that

$$\begin{aligned} \|\boldsymbol{v}\|^2 &= \|A(\boldsymbol{x})^{-1}\nabla f(\boldsymbol{x})\|^2 \\ &\leq \gamma^2\|\nabla f(\boldsymbol{x})\|^2 \\ &\leq -\beta\gamma^2 df(\boldsymbol{x})\boldsymbol{v}. \end{aligned}$$

Combining this last inequality with inequality (12.11) yields

$$f(\boldsymbol{x} + t\boldsymbol{v}) \quad \leq \quad f(\boldsymbol{x}) + t\left(1 - \frac{\beta\gamma^2\delta}{2}t\right)df(\boldsymbol{x})\boldsymbol{v}.$$

Hence, as soon as σ^k satisfies

$$1 - \frac{\beta\gamma^2\delta}{2}\sigma^k \quad \geq \quad \alpha,$$

Armijo's rule (12.10) holds. In terms of k, backtracking is guaranteed to succeed in at most

$$k_{\max} \quad = \quad \max\left\{\left\lceil \frac{1}{\ln\sigma}\ln\frac{2(1-\alpha)}{\beta\gamma^2\delta}\right\rceil, d\right\}$$

decrements. Of course, a lower value of k may suffice.

Proposition 12.6.1 *Suppose that all stationary points of $f(\boldsymbol{x})$ are isolated and that the stated continuity, differentiability, positive definiteness, and coerciveness assumptions are true. Then any sequence of iterates $\boldsymbol{x}_m$ generated by the iteration map $M(\boldsymbol{x}) = \boldsymbol{x} - tA(\boldsymbol{x})^{-1}\nabla f(\boldsymbol{x})$ with t chosen by step decrementing possesses a limit, and that limit is a stationary point of $f(\boldsymbol{x})$. If $f(\boldsymbol{x})$ is strictly convex, then $\lim_{m\to\infty}\boldsymbol{x}_m$ is the minimum point.*

Proof: Let $\boldsymbol{v}_m = -A(\boldsymbol{x}_m)^{-1}\nabla f(\boldsymbol{x}_m)$ and $\boldsymbol{x}_{m+1} = \boldsymbol{x}_m + \sigma^{k_m}\boldsymbol{v}_m$. The sequence $f(\boldsymbol{x}_m)$ is decreasing by construction. Because the function $f(\boldsymbol{x})$ is bounded below on the compact set $D = \{\boldsymbol{x} \in U : f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$,

$f(\boldsymbol{x}_m)$ is bounded below as well and possesses a limit. Based on Armijo's rule (12.10) and inequality (12.12), we calculate

$$\begin{aligned} f(\boldsymbol{x}_m) - f(\boldsymbol{x}_{m+1}) &\geq -\alpha\sigma^{k_m} df(\boldsymbol{x}_m)\boldsymbol{v}_m \\ &= \alpha\sigma^{k_m} df(\boldsymbol{x}_m)A(\boldsymbol{x}_m)^{-1}\nabla f(\boldsymbol{x}_m) \\ &\geq \frac{\alpha\sigma^{k_m}}{\beta}\|\nabla f(\boldsymbol{x}_m)\|^2. \end{aligned}$$

Since $\sigma^{k_m} \geq \sigma^{k_{\max}}$, and the difference $f(\boldsymbol{x}_m) - f(\boldsymbol{x}_{m+1})$ tends to 0, we deduce that $\|\nabla f(\boldsymbol{x}_m)\|$ tends to 0. This conclusion and the inequality

$$\begin{aligned} \|\boldsymbol{x}_{m+1} - \boldsymbol{x}_m\| &= \sigma^{k_m}\|A(\boldsymbol{x}_m)^{-1}\nabla f(\boldsymbol{x}_m)\| \\ &\leq \sigma^{k_m}\gamma\|\nabla f(\boldsymbol{x}_m)\|, \end{aligned}$$

demonstrate that $\|\boldsymbol{x}_{m+1} - \boldsymbol{x}_m\|$ tends to 0 as well. Given these results, Propositions 12.4.2 and 12.4.3 are true. All claims of the current proposition now follow as in the proof of Proposition 12.4.4. ■

12.7 Problems

1. Consider the functions $f(x) = x - x^3$ and $g(x) = x + x^3$ on R. Show that the iterates $x_{m+1} = f(x_m)$ are locally attracted to 0 and that the iterates $x_{m+1} = g(x_m)$ are locally repelled by 0. In both cases $f'(0) = g'(0) = 1$.

2. Consider the iteration map $h(x) = \sqrt{a + x}$ on $(0, \infty)$ for $a > 0$. Find the fixed point of $h(x)$ and show that it is locally attractive. Is it also globally attractive?

3. In Example 10.2.1 suppose $x_0 = 1$ and $a \in (0, 2)$. Demonstrate that

$$\begin{aligned} x_m &= \frac{1 - (1-a)^{2^m}}{a} \\ \left|x_{m+1} - \frac{1}{a}\right| &= a\left|x_m - \frac{1}{a}\right|^2. \end{aligned}$$

   This shows very explicitly that x_m converges to $1/a$ at a quadratic rate.

4. In Example 10.2.2 prove that

$$\begin{aligned} x_m &= \sqrt{a} + \frac{2\sqrt{a}}{\left[\left(1 + \frac{2\sqrt{a}}{x_0 - \sqrt{a}}\right)^{2^m} - 1\right]} \\ \left|x_{m+1} - \sqrt{a}\right| &\leq \frac{1}{2\sqrt{a}}\left|x_m - \sqrt{a}\right|^2 \end{aligned}$$

when $n = 2$ and $x_0 > 0$. Thus, Newton's method converges at a quadratic rate. Use the first of these formulas or the iteration equation directly to show that $\lim_{m\to\infty} x_m = -\sqrt{a}$ for $x_0 < 0$.

5. Suppose the real-valued function $f(x)$ is twice continuously differentiable on the interval (a, b) with a root y where $f(y) = 0$ and $f'(y) \neq 0$. Show that the iteration scheme

$$x_{n+1} = x_n - \frac{f(x_n)^2}{f[x_n + f(x_n)] - f(x_n)}$$

converges at a quadratic rate to y if x_0 is sufficiently close to y. In particular, demonstrate that

$$\lim_{n\to\infty} \frac{x_{n+1} - y}{(x_n - y)^2} = \frac{f''(y)[f'(y) - 1]}{2f'(y)}.$$

6. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $n \times n$ matrices. If $\boldsymbol{A}$ and $\boldsymbol{A} - \boldsymbol{B}^*\boldsymbol{A}\boldsymbol{B}$ are both positive definite, then show that $\boldsymbol{B}$ has spectral radius $\rho(\boldsymbol{B}) < 1$. Note that $\boldsymbol{A}$ is symmetric, but $\boldsymbol{B}$ need not be symmetric. (Hint: Consider the quadratic form $\boldsymbol{v}^*(\boldsymbol{A} - \boldsymbol{B}^*\boldsymbol{A}\boldsymbol{B})\boldsymbol{v}$ for an eigenvector $\boldsymbol{v}$ of $\boldsymbol{B}$.)

7. In block relaxation with b blocks, let $B_i(\boldsymbol{x})$ be the map that updates block i and leaves the other blocks fixed. Show that the overall iteration map $M(\boldsymbol{x}) = B_b \circ \cdots \circ B_1(\boldsymbol{x})$ has differential $dB_b(\boldsymbol{y}) \cdots dB_1(\boldsymbol{y})$ at a fixed point $\boldsymbol{y}$. Write $dB_i(\boldsymbol{y})$ as a block matrix and identify the blocks by applying the implicit function theorem as needed. Do not confuse $B_i(\boldsymbol{x})$ with the update $M_i(\boldsymbol{x})$ of the text. In fact, $M_i(\boldsymbol{x})$ only summarizes the update of block i, and its argument is the value of $\boldsymbol{x}$ at the start of the current round of updates.

8. Consider a Poisson-distributed random variable Y with mean $a\theta + b$, where a and b are known positive constants and $\theta \geq 0$ is a parameter to be estimated. An EM algorithm for estimating θ can be concocted that takes as complete data independent Poisson random variables U and V with means $a\theta$ and b and sum $U + V = Y$. If $Y = y$ is observed, then show that the EM iterates are defined by

$$\theta_{m+1} = \frac{y\theta_m}{a\theta_m + b}.$$

Show that these iterates converge monotonically to the maximum likelihood estimate $\max\{0, (y - b)/a\}$. When $y = b$, verify that convergence to the boundary value 0 occurs at a rate slower than linear [90]. (Hint: When $y = b$, check that $\theta_{m+1} = b\theta_0/(ma\theta_0 + b)$.)

9. The sublinear convergence of the EM algorithm exhibited in the previous problem occurs in other problems. Here is a conceptually harder

example by Robert Jennrich. Suppose that $W_1, \ldots, W_n$ and B are independent normally distributed random variables with 0 means. Let σ_w^2 be the common variance of the W_i and σ_b^2 be the variance of B. If the values y_i of the linear combinations $Y_i = B + W_i$ are observed, then show that the EM algorithm amounts to

$$\begin{aligned}
\sigma_{m+1,b}^2 &= \left(\frac{n\sigma_{mb}^2\bar{y}}{n\sigma_{mb}^2+\sigma_{mw}^2}\right)^2 + \frac{\sigma_{mb}^2\sigma_{nw}^2}{n\sigma_{mb}^2+\sigma_{mw}^2} \\
\sigma_{m+1,w}^2 &= \frac{n-1}{n}s_y^2 + \left(\frac{\sigma_{mw}^2\bar{y}}{n\sigma_{mb}^2+\sigma_{mw}^2}\right)^2 + \frac{\sigma_{mb}^2\sigma_{mw}^2}{n\sigma_{mb}^2+\sigma_{mw}^2},
\end{aligned}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $s_y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$ are the sample mean and variance. Although one can formally calculate the maximum likelihood estimates $\hat{\sigma}_w^2 = s_y^2$ and $\hat{\sigma}_b^2 = \bar{y}^2 - s_y^2/n$, these are only valid provided $\hat{\sigma}_b^2 \geq 0$. If for instance $\bar{y} = 0$, then the EM iterates will converge to $\sigma_w^2 = (n-1)s_y^2/n$ and $\sigma_b^2 = 0$. Show that convergence is sublinear when $\bar{y} = 0$.

10. Consider independent observations $y_1, \ldots, y_n$ from the univariate t-distribution. These data have loglikelihood

$$\begin{aligned}
L &= -\frac{n}{2}\ln\sigma^2 - \frac{\nu+1}{2}\sum_{i=1}^n \ln(\nu+\delta_i^2) \\
\delta_i^2 &= \frac{(y_i-\mu)^2}{\sigma^2}.
\end{aligned}$$

To illustrate the occasionally bizarre behavior of the MM algorithm, we take $\nu = 0.05$ and the data vector $y = (-20, 1, 2, 3)^*$ with $n = 4$ observations. Devise an MM maximum likelihood algorithm for estimating μ with σ^2 fixed at 1. Show that the iteration map is

$$\begin{aligned}
\mu_{m+1} &= \frac{\sum_{i=1}^n w_{mi}y_i}{\sum_{i=1}^n w_{mi}} \\
w_{mi} &= \frac{\nu+1}{\nu+(y_i-\mu_m)^2}.
\end{aligned}$$

Plot the likelihood curve and show that it has the four local maxima -19.993, 1.086, 1.997, and 2.906 and the three local minima -14.516, 1.373, and 2.647. Demonstrate numerically convergence to a local maximum that is not the global maximum. Show that the algorithm converges to a local minimum in one step starting from -1.874 or -0.330 [191].

11. Suppose the data displayed in Table 12.1 constitute a random sample from a bivariate normal distribution with both means 0, variances σ_1^2

TABLE 12.1. Bivariate normal data for the EM algorithm

Obs	Obs	Obs	Obs	Obs	Obs
(1,1)	$(1,-1)$	$(-1,1)$	$(-1,-1)$	(2,∗)	(2,∗)
$(-2,*)$	$(-2,*)$	(∗,2)	(∗,2)	(∗,−2)	(∗,−2)

and σ_2^2, and correlation coefficient ρ. The asterisks indicate missing values. Specify the EM algorithm for estimating σ_1^2, σ_2^2, and ρ. Show that the observed loglikelihood has a saddle point at the point where $\rho = 0$ and $\sigma_1^2 = \sigma_2^2 = \frac{5}{2}$. If the EM algorithm starts with $\rho = 0$, prove that convergence to the saddle point occurs [275]. (Hints: At the given point, show that the observed loglikelihood achieves a global maximum in σ_1^2 and σ_2^2 with ρ fixed at 0 and a local minimum in ρ with σ_1^2 and σ_2^2 fixed at $\frac{5}{2}$. Also show that the EM algorithm iterates satisfy

$$\begin{aligned} \sigma_{m+1,1}^2 - \frac{5}{2} &= \frac{1}{3}\left(\sigma_{m1}^2 - \frac{5}{2}\right) \\ \sigma_{m+1,2}^2 - \frac{5}{2} &= \frac{1}{3}\left(\sigma_{m2}^2 - \frac{5}{2}\right) \\ \rho_{m+1} &= 0 \end{aligned}$$

along the slice $\rho = 0$.)

12. Under the hypotheses of Proposition 12.4.4, if the MM gradient algorithm is started close enough to a local minimum $\boldsymbol{y}$ of $f(\boldsymbol{x})$, then the iterates $\boldsymbol{x}_m$ converge to $\boldsymbol{y}$ without step decrementing. Prove that for all sufficiently large m, either $\boldsymbol{x}_m = \boldsymbol{y}$ or $f(\boldsymbol{x}_{m+1}) < f(\boldsymbol{x}_m)$ [163]. (Hints: Let $\boldsymbol{v}_m = \boldsymbol{x}_{m+1} - \boldsymbol{x}_m$, $\boldsymbol{C}_m = s_f^2(\boldsymbol{x}_{m+1}, \boldsymbol{x}_m)$, and $\boldsymbol{D}_m = d^2 g(\boldsymbol{x}_m \mid \boldsymbol{x}_m)$. Show that

$$f(\boldsymbol{x}_{m+1}) = f(\boldsymbol{x}_m) + \frac{1}{2}\boldsymbol{v}_m^*(\boldsymbol{C}_m - 2\boldsymbol{D}_m)\boldsymbol{v}_m.$$

Then use a continuity argument, noting that $d^2 g(\boldsymbol{y} \mid \boldsymbol{y}) - d^2 f(\boldsymbol{y})$ is positive semidefinite and $d^2 g(\boldsymbol{y} \mid \boldsymbol{y})$ is positive definite.)

13. Let $M(\boldsymbol{x})$ be the MM algorithm or MM gradient algorithm map. Consider the modified algorithm $M_t(\boldsymbol{x}) = \boldsymbol{x} + t[M(\boldsymbol{x}) - \boldsymbol{x}]$ for $t > 0$. At a local optimum $\boldsymbol{y}$, show that the spectral radius ρ_t of the differential $dM_t(\boldsymbol{y}) = (1-t)\boldsymbol{I} + t dM(\boldsymbol{y})$ satisfies $\rho_t < 1$ when $0 < t < 2$. Hence, Ostrowski's theorem implies local attraction of $M_t(\boldsymbol{x})$ to $\boldsymbol{y}$. If the largest and smallest eigenvalues of $dM(\boldsymbol{y})$ are $\omega_{\max}$ and $\omega_{\min}$, then prove that ρ_t is minimized by taking $t = [1-(\omega_{\min}+\omega_{\max})/2]^{-1}$. In practice, the eigenvalues of $dM(\boldsymbol{y})$ are impossible to predict without advance knowledge of $\boldsymbol{y}$, but for many problems the value $t = 2$

works well [163]. (Hints: All eigenvalues of $dM(\boldsymbol{y})$ occur on $[0,1)$. To every eigenvalue ω of $dM(\boldsymbol{y})$, there corresponds an eigenvalue $\omega_t = 1 - t + t\omega$ of $dM_t(\boldsymbol{y})$ and vice versa.)

14. In the notation of Chap. 9, prove the EM algorithm formula

$$d^2L(\boldsymbol{\theta}) \;=\; d^{20}Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}) + \mathrm{Var}[\nabla \ln f(X \mid \boldsymbol{\theta}) \mid Y, \boldsymbol{\theta}]$$

of Louis [180].

15. Which of the following functions is coercive on its domain?

(a) $f(x) = x + 1/x$ on $(0,\infty)$,
(b) $f(x) = x - \ln x$ on $(0,\infty)$,
(c) $f(\boldsymbol{x}) = x_1^2 + x_2^2 - 2x_1x_2$ on R^2,
(d) $f(\boldsymbol{x}) = x_1^4 + x_2^4 - 3x_1x_2$ on R^2,
(e) $f(\boldsymbol{x}) = x_1^2 + x_2^2 + x_3^2 - \sin(x_1x_2x_3)$ on R^3.

Give convincing reasons in each case.

16. Consider a polynomial $p(\boldsymbol{x})$ in n variables $x_1, \ldots, x_n$. Suppose that $p(\boldsymbol{x}) = \sum_{i=1}^n c_i x_i^{2m} +$ lower-order terms, where all $c_i > 0$ and where a lower-order term is a product $bx_1^{m_1} \cdots x_n^{m_n}$ with $\sum_{i=1}^n m_i < 2m$. Prove rigorously that $p(\boldsymbol{x})$ is coercive on R^n.

17. Demonstrate that $h(\boldsymbol{x}) + k(\boldsymbol{x})$ is coercive on R^n if $k(\boldsymbol{x})$ is convex and $h(\boldsymbol{x})$ satisfies $\lim_{\|\boldsymbol{x}\|\to\infty} \|\boldsymbol{x}\|^{-1} h(\boldsymbol{x}) = \infty$. Problem 31 of Chap. 6 is a special case. (Hint: Apply definition (6.4) of Chap. 6.)

18. In some problems it is helpful to broaden the notion of coerciveness. Consider a continuous function $f : \mathsf{R}^n \mapsto \mathsf{R}$ such that the limit

$$c \;=\; \lim_{r\to\infty} \inf_{\|\boldsymbol{x}\|\geq r} f(\boldsymbol{x})$$

exists. The value of c can be finite or ∞ but not $-\infty$. Now let $\boldsymbol{y}$ be any point with $f(\boldsymbol{y}) < c$. Show that the set $S_{\boldsymbol{y}} = \{\boldsymbol{x} : f(\boldsymbol{x}) \leq f(\boldsymbol{y})\}$ is compact and that $f(\boldsymbol{x})$ attains its global minimum on $S_{\boldsymbol{y}}$. The particular function

$$g(\boldsymbol{x}) \;=\; \frac{x_1 + 2x_2}{1 + x_1^2 + x_2^2}$$

furnishes an example when $n = 2$. Demonstrate that the limit c equals 0. What is the minimum value and minimum point of $g(\boldsymbol{x})$? (Hint: What is the minimum value of $g(\boldsymbol{x})$ on the circle $\{\boldsymbol{x} : \|\boldsymbol{x}\| = r\}$?)

19. Let $f(\boldsymbol{x})$ be a convex function on R^n. Prove that all sublevel sets $\{\boldsymbol{x} \in \mathsf{R}^n : f(\boldsymbol{x}) \leq b\}$ are bounded if and only if

$$\lim_{r\to\infty} \inf_{\|\boldsymbol{x}\|\geq r} f(\boldsymbol{x}) \quad > \quad 0.$$

20. Assume that the function $f(\boldsymbol{x})$ is (a) continuously differentiable, (b) maps R^n into itself, (c) has Jacobian $df(\boldsymbol{x})$ of full rank at each $\boldsymbol{x}$, and (d) has coercive norm $\|f(\boldsymbol{x})\|$. Show that the image of $f(\boldsymbol{x})$ is all of R^n. (Hints: The image is connected. Show that it is open via the inverse function theorem and closed because of coerciveness.)

21. Consider a sequence $\boldsymbol{x}_m$ in R^n. Verify that the set of cluster points of $\boldsymbol{x}_m$ is closed. If $\boldsymbol{x}_m$ is bounded, then show that it has a limit if and only if it has at most one cluster point.

22. In our exposition of least absolute deviation regression, we considered in Problem 14 of Chap. 8 a modified iteration scheme that minimizes the criterion

$$h_\epsilon(\boldsymbol{\theta}) \quad = \quad \sum_{i=1}^p \left\{[y_i - \mu_i(\boldsymbol{\theta})]^2 + \epsilon\right\}^{1/2}. \tag{12.13}$$

For a sequence of constants ϵ_m tending to 0, let $\boldsymbol{\theta}_m$ be a corresponding sequence minimizing (12.13). If $\boldsymbol{\phi}$ is a cluster point of this sequence and the regression functions $\mu_i(\boldsymbol{\theta})$ are continuous, then show that $\boldsymbol{\phi}$ minimizes $h_0(\boldsymbol{\theta}) = \sum_{i=1}^p |y_i - \mu_i(\boldsymbol{\theta})|$. If, in addition, the minimum point $\boldsymbol{\phi}$ of $h_0(\boldsymbol{\theta})$ is unique and $\lim_{\|\boldsymbol{\theta}\|\to\infty} \sum_{i=1}^p |\mu_i(\boldsymbol{\theta})| = \infty$, then prove that $\lim_{m\to\infty} \boldsymbol{\theta}_m = \boldsymbol{\phi}$. (Hints: For the first assertion, take limits in

$$h_\epsilon(\boldsymbol{\theta}_m) \quad \leq \quad h_\epsilon(\boldsymbol{\theta}).$$

For the second assertion, it suffices to prove that the sequence $\boldsymbol{\theta}_m$ is confined to a bounded set. To prove this fact, demonstrate the inequalities $|\mu| + |y| + \sqrt{\epsilon} \geq \sqrt{r^2 + \epsilon} \geq |\mu| - |y|$ for $r = y - \mu$.)

23. Example 10.2.5 and Problem 9 of Chap. 10 suggest a method of accelerating the MM gradient algorithm. Suppose we are maximizing the loglikelihood $L(\boldsymbol{\theta})$ using the surrogate function is $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n)$. To accelerate the MM gradient algorithm, we can replace the positive definite matrix $B(\boldsymbol{\theta})^{-1} = -d^{20}g(\boldsymbol{\theta} \mid \boldsymbol{\theta})$ by a matrix that better approximates the observed information $A(\boldsymbol{\theta}) = -d^2L(\boldsymbol{\theta})$. Note that often $d^{20}g(\boldsymbol{\theta} \mid \boldsymbol{\theta})$ is diagonal and therefore trivial to invert. Now consider the formal expansion

$$\boldsymbol{A}^{-1} \quad = \quad (\boldsymbol{B}^{-1} + \boldsymbol{A} - \boldsymbol{B}^{-1})^{-1}$$

$$\begin{aligned} &= \{\boldsymbol{B}^{-\frac{1}{2}}[\boldsymbol{I} - \boldsymbol{B}^{\frac{1}{2}}(\boldsymbol{B}^{-1} - \boldsymbol{A})\boldsymbol{B}^{\frac{1}{2}}]\boldsymbol{B}^{-\frac{1}{2}}\}^{-1} \\ &= \boldsymbol{B}^{\frac{1}{2}} \sum_{i=0}^{\infty} [\boldsymbol{B}^{\frac{1}{2}}(\boldsymbol{B}^{-1} - \boldsymbol{A})\boldsymbol{B}^{\frac{1}{2}}]^i \boldsymbol{B}^{\frac{1}{2}}. \end{aligned}$$

If we truncate this series after a finite number of terms, then we recover the first iterate of equation (10.18) in the disguised form

$$\boldsymbol{S}_j = \boldsymbol{B}^{\frac{1}{2}} \sum_{i=0}^{j} [\boldsymbol{B}^{\frac{1}{2}}(\boldsymbol{B}^{-1} - \boldsymbol{A})\boldsymbol{B}^{\frac{1}{2}}]^i \boldsymbol{B}^{\frac{1}{2}}.$$

The accelerated algorithm

$$\boldsymbol{\theta}_{m+1} = \boldsymbol{\theta}_m + S_j(\boldsymbol{\theta}_m)\nabla L(\boldsymbol{\theta}_m) \tag{12.14}$$

has several desirable properties.

(a) Show that $\boldsymbol{S}_j$ is positive definite and hence that the update (12.14) is an ascent algorithm. (Hint: Use the fact that $\boldsymbol{B}^{-1} - \boldsymbol{A}$ is positive semidefinite.)

(b) Algorithm (12.14) has differential

$$\boldsymbol{I} + S_j(\boldsymbol{\theta}_\infty)d^2L(\boldsymbol{\theta}_\infty) = \boldsymbol{I} - S_j(\boldsymbol{\theta}_\infty)A(\boldsymbol{\theta}_\infty)$$

at a local maximum $\boldsymbol{\theta}_\infty$. If $d^2L(\boldsymbol{\theta}_\infty)$ is negative definite, then prove that all eigenvalues of this differential lie on $[0, 1)$. (Hint: The eigenvalues are determined by the stationary points of the Rayleigh quotient $\boldsymbol{v}^*[A^{-1}(\boldsymbol{\theta}_\infty) - S_j(\boldsymbol{\theta}_\infty)]\boldsymbol{v}/\boldsymbol{v}^*A^{-1}(\boldsymbol{\theta}_\infty)\boldsymbol{v}$.)

(c) If ρ_j is the spectral radius of the differential, then demonstrate that $\rho_j \leq \rho_{j-1}$, with strict inequality when $B^{-1}(\boldsymbol{\theta}_\infty) - A(\boldsymbol{\theta}_\infty)$ is positive definite.

In other words, the accelerated algorithm (12.14) is guaranteed to converge faster than the MM gradient algorithm. It will be particularly useful for maximum likelihood problems with many parameters because it entails no matrix inversion or multiplication, just matrix times vector multiplication. When $j = 1$, it takes the simple form

$$\boldsymbol{\theta}_{m+1} = \boldsymbol{\theta}_m + [2B(\boldsymbol{\theta}_m) - B(\boldsymbol{\theta}_m)A(\boldsymbol{\theta}_m)B(\boldsymbol{\theta}_m)]\nabla L(\boldsymbol{\theta}_m).$$

13
Penalty and Barrier Methods

13.1 Introduction

Penalties and barriers feature prominently in two areas of modern optimization theory. First, both devices are employed to solve constrained optimization problems [96, 183, 226]. The general idea is to replace hard constraints by penalties or barriers and then exploit the well-oiled machinery for solving unconstrained problems. Penalty methods operate on the exterior of the feasible region and barrier methods on the interior. The strength of a penalty or barrier is determined by a tuning constant. In classical penalty methods, a single global tuning constant is gradually sent to ∞; in barrier methods, it is gradually sent to 0. Nothing prevents one from assigning different tuning constants to different penalties or barriers in the same problem. Either strategy generates a sequence of solutions that converges in practice to the solution of the original constrained optimization problem.

One of the lessons of the current chapter is that it is profitable to view penalties and barriers from the perspective of the MM algorithm. For example, this mental exercise suggests a way of engineering barrier tuning constants in a constrained minimization problem so that the objective function is forced steadily downhill [41, 162, 253]. Over time the tuning constant for each inequality constraint adapts to the need to avoid the constraint or converge to it.

A detailed study of penalty methods in constrained optimization requires considerable knowledge of the convex calculus. For this reason we defer

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_13,

to Chap. 16 our presentation of exact penalty methods. These methods substitute absolute value and hinge penalties for square penalties. Numerical analysts have shied away from exact penalty methods because of the difficulties in working with non-differentiable functions. Such pessimism is unwarranted, however. In convex programs one can easily follow the solution path as the penalty constant increases. Certainly, path following has been highly successful in interior point programming. Readers who want to pursue this traditional application of path following should consult one or more of the superb references [19, 30, 96, 203, 205, 274].

Imposition of penalties has other beneficial effects. The addition of convex penalties can regularize a problem by reducing the effect of noise and eliminating spurious local minima. Penalties also steer solutions to unconstrained optimization problems in productive directions. Recall, for instance, the beneficial effects of smoothing penalties in transmission tomography. Bayesian statistics is predicated on the philosophy that prior knowledge should never be ignored. Beyond smoothing and exploitation of prior knowledge, penalties play a role in model selection. In lasso penalized estimation, statisticians impose an ℓ_1 penalty that shrinks parameter estimates toward zero and performs a kind of continuous model selection [75, 256]. The predictors whose estimated regression coefficients are exactly zero are candidates for elimination from the model. With the enormous data sets now confronting statisticians, considerations of model parsimony have taken on greater urgency. In addition to this philosophical justification, imposition of lasso penalties also has an huge impact on computational speed. Standard methods of regression require matrix diagonalization, matrix inversion, or, at the very least, the solution of large systems of linear equations. Because the number of arithmetic operations for these processes scales as the cube of the number of predictors, problems with tens of thousands of predictors appear intractable. Recent research has shown this assessment to be too pessimistic [36, 83, 119, 143, 210, 267]. Coordinate descent methods mesh well with the lasso and are simple, fast, and stable. We will see how their potential to transform data mining plays out in both ℓ_1 and ℓ_2 regression.

13.2 Rudiments of Barrier and Penalty Methods

In general, unconstrained optimization problems are easier to solve than constrained optimization problems, and equality constrained problems are easier to solve than inequality constrained problems. To simplify analysis, mathematical scientists rely on several devices. For instance, one can replace the inequality constraint $g(\boldsymbol{x}) \leq 0$ by the equality constraint $g_+(\boldsymbol{x}) = 0$, where $g_+(\boldsymbol{x}) = \max\{g(\boldsymbol{x}), 0\}$. This tactic is not entirely satisfactory because $g_+(\boldsymbol{x})$ has kinks along the boundary $g(\boldsymbol{x}) = 0$. The smoother substitute $g_+(\boldsymbol{x})^2$ avoids the kinks in first derivatives. Alternatively, one

can introduce an extra parameter y and require $g(\boldsymbol{x})+y = 0$ and $y \geq 0$. This tactic substitutes a simple inequality constraint for a complex inequality constraint.

The addition of barrier and penalty terms to the objective function $f(\boldsymbol{x})$ is a more systematic approach. Later in the chapter we will discuss the role of penalties in producing sparse solutions. In the current section, penalties are introduced to steer the optimization process toward the feasible region. In the penalty method we construct a continuous nonnegative penalty $p(\boldsymbol{x})$ that is 0 on the feasible region and positive outside it. We then optimize the functions $f(\boldsymbol{x}) + \lambda_m p(\boldsymbol{x})$ for an increasing sequence of tuning constants λ_m that tend to ∞. The penalty method works from the outside of the feasible region inward. Under the right hypotheses, the sequence of unconstrained solutions $\boldsymbol{x}_m$ tends to a solution of the constrained optimization problem.

Example 13.2.1 *Linear Regression with Linear Constraints*

Consider the problem of minimizing $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ subject to the linear constraints $\boldsymbol{V}\boldsymbol{\beta} = \boldsymbol{d}$. If we take the penalty function $p(\boldsymbol{\beta}) = \|\boldsymbol{V}\boldsymbol{\beta} - \boldsymbol{d}\|^2$, then we must minimize at each stage the function

$$h_m(\boldsymbol{\beta}) \;=\; \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_m \|\boldsymbol{V}\boldsymbol{\beta} - \boldsymbol{d}\|^2.$$

Setting the gradient

$$\nabla h_m(\boldsymbol{\beta}) \;=\; -2\boldsymbol{X}^*(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + 2\lambda_m \boldsymbol{V}^*(\boldsymbol{V}\boldsymbol{\beta} - \boldsymbol{d})$$

equal to $\mathbf{0}$ yields the sequence of solutions

$$\boldsymbol{\beta}_m \;=\; (\boldsymbol{X}^*\boldsymbol{X} + \lambda_m \boldsymbol{V}^*\boldsymbol{V})^{-1}(\boldsymbol{X}^*\boldsymbol{y} + \lambda_m \boldsymbol{V}^*\boldsymbol{d}). \tag{13.1}$$

In a moment we will demonstrate that the $\boldsymbol{\beta}_m$ tend to the constrained solution as λ_m tends to ∞. ■

In contrast, the barrier method works from the inside of the feasible region outward by introducing a continuous barrier function $b(\boldsymbol{x})$ that is finite on the interior of the feasible region and infinite on its boundary. We then optimize the sequence of functions $f(\boldsymbol{x}) + \mu_m b(\boldsymbol{x})$ as the decreasing sequence of tuning constants μ_m tends to 0. Again under the right hypotheses, the sequence of unconstrained solutions $\boldsymbol{x}_m$ tends to the solution of the constrained optimization problem.

Example 13.2.2 *Estimation of Multinomial Proportions*

In estimating multinomial proportions, we minimize the negative loglikelihood $-\sum_{i=1}^d n_i \ln p_i$ subject to the constraints $\sum_{i=1}^d p_i = 1$ and $p_i \geq 0$ for d categories. An appropriate barrier function is $-\sum_{i=1}^d \ln p_i$. The minimum of the function

$$h_m(\boldsymbol{p}) \;=\; -\sum_{i=1}^d n_i \ln p_i - \mu_m \sum_{i=1}^d \ln p_i$$

subject to the constraint $\sum_{i=1}^d p_i = 1$ occurs at the point with coordinates

$$p_{mi} = \frac{n_i + \mu_m}{n + d\mu_m},$$

where $n = \sum_{i=1}^d n_i$. In this example, it is clear that the solution vector $\boldsymbol{p}_m$ occurs on the interior of the parameter space and tends to the maximum likelihood estimate as μ_m tends to 0. ■

The next proposition highlights the ascent and descent properties of the penalty and barrier methods.

Proposition 13.2.1 *Consider two real-valued functions $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ on a common domain and two positive constants $\alpha < \beta$. Suppose the linear combination $f(\boldsymbol{x}) + \alpha g(\boldsymbol{x})$ attains its minimum value at $\boldsymbol{y}$ and the linear combination $f(\boldsymbol{x}) + \beta g(\boldsymbol{x})$ attains its minimum value at $\boldsymbol{z}$. Then we have $f(\boldsymbol{y}) \leq f(\boldsymbol{z})$ and $g(\boldsymbol{y}) \geq g(\boldsymbol{z})$.*

Proof: Adding the two inequalities

$$\begin{aligned} f(\boldsymbol{z}) + \beta g(\boldsymbol{z}) &\leq f(\boldsymbol{y}) + \beta g(\boldsymbol{y}) \\ -f(\boldsymbol{z}) - \alpha g(\boldsymbol{z}) &\leq -f(\boldsymbol{y}) - \alpha g(\boldsymbol{y}) \end{aligned}$$

and dividing by the constant $\beta - \alpha$ validates the claim $g(\boldsymbol{y}) \geq g(\boldsymbol{z})$. The claim $f(\boldsymbol{y}) \leq f(\boldsymbol{z})$ is proved by interchanging the roles of $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ and considering the functions $g(\boldsymbol{x}) + \alpha^{-1} f(\boldsymbol{x})$ and $g(\boldsymbol{x}) + \beta^{-1} f(\boldsymbol{x})$. ■

It is now fairly easy to prove a version of global convergence for the penalty method.

Proposition 13.2.2 *Suppose that both the objective function $f(\boldsymbol{x})$ and the penalty function $p(\boldsymbol{x})$ are continuous on R^n and that the penalized functions $h_m(\boldsymbol{x}) = f(\boldsymbol{x}) + \lambda_m p(\boldsymbol{x})$ are coercive on R^n. Then one can extract a corresponding sequence of minimum points $\boldsymbol{x}_m$ such that $f(\boldsymbol{x}_m) \leq f(\boldsymbol{x}_{m+1})$. Furthermore, any cluster point of this sequence resides in the feasible region $C = \{\boldsymbol{x} : p(\boldsymbol{x}) = 0\}$ and attains the minimum value of $f(\boldsymbol{x})$ there. Finally, if $f(\boldsymbol{x})$ is coercive and possesses a unique minimum point in C, then the sequence $\boldsymbol{x}_m$ converges to that point.*

Proof: In view of the coerciveness assumption, the minimum points $\boldsymbol{x}_m$ exist. Proposition 13.2.1 confirms the ascent property. Now suppose that $\boldsymbol{z}$ is a cluster point of the sequence $\boldsymbol{x}_m$ and $\boldsymbol{y}$ is any point in C. If we take limits in the inequalities

$$f(\boldsymbol{y}) = h_m(\boldsymbol{y}) \geq h_m(\boldsymbol{x}_m) \geq f(\boldsymbol{x}_m)$$

along the subsequence $\boldsymbol{x}_{m_l}$ tending to $\boldsymbol{z}$, then the inequality $f(\boldsymbol{y}) \geq f(\boldsymbol{z})$ follows. Furthermore, because the λ_m tend to infinity, the bound

$$\limsup_{l \to \infty} \lambda_{m_l} p(\boldsymbol{x}_{m_l}) \leq f(\boldsymbol{y}) - \lim_{l \to \infty} f(\boldsymbol{x}_{m_l}) = f(\boldsymbol{y}) - f(\boldsymbol{z})$$

can only hold if $p(\boldsymbol{z}) = \lim_{l\to\infty} p(\boldsymbol{x}_{m_l}) = 0$.

If $f(\boldsymbol{x})$ possesses a unique minimum point $\boldsymbol{y}$ in C, then to prove that $\boldsymbol{x}_m$ converges to $\boldsymbol{y}$, it suffices to prove that $\boldsymbol{x}_m$ is bounded. Given that $f(\boldsymbol{x})$ is coercive, it is possible to choose r so that $f(\boldsymbol{x}) > f(\boldsymbol{y})$ for all $\boldsymbol{x}$ with $\|\boldsymbol{x}\| \geq r$. The assumption $\|\boldsymbol{x}_m\| \geq r$ consequently implies

$$h_m(\boldsymbol{x}_m) \quad \geq \quad f(\boldsymbol{x}_m) \quad > \quad f(\boldsymbol{y}) \quad = \quad h_m(\boldsymbol{y}),$$

which contradicts the assumption that $\boldsymbol{x}_m$ minimizes $h_m(\boldsymbol{x})$. Hence, all $\boldsymbol{x}_m$ satisfy $\|\boldsymbol{x}_m\| < r$. ■

Here is the corresponding result for the barrier method.

Proposition 13.2.3 *Suppose the real-valued function $f(\boldsymbol{x})$ is continuous on the bounded open set $U \subset \mathsf{R}^n$ and its closure V. Also suppose the barrier function $b(\boldsymbol{x})$ is continuous and coercive on U. If the tuning constants μ_m decrease to 0, then the linear combinations $h_m(\boldsymbol{x}) = f(\boldsymbol{x}) + \mu_m b(\boldsymbol{x})$ attain their minima at a sequence of points $\boldsymbol{x}_m$ in U satisfying the descent property $f(\boldsymbol{x}_{m+1}) \leq f(\boldsymbol{x}_m)$. Furthermore, any cluster point of the sequence furnishes the minimum value of $f(\boldsymbol{x})$ on V. If the minimum point of $f(\boldsymbol{x})$ in V is unique, then the sequence $\boldsymbol{x}_m$ converges to this point.*

Proof: Each of the continuous functions $h_m(\boldsymbol{x})$ is coercive on U, being the sum of a coercive function and a function bounded below. Therefore, the sequence $\boldsymbol{x}_m$ exists. An appeal to Proposition 13.2.1 establishes the descent property. If $\boldsymbol{z}$ is a cluster point of $\boldsymbol{x}_m$ and $\boldsymbol{x}$ is any point of U, then taking limits in the inequality

$$f(\boldsymbol{x}_m) + \mu_m b(\boldsymbol{x}_m) \quad \leq \quad f(\boldsymbol{x}) + \mu_m b(\boldsymbol{x})$$

along the relevant subsequence $\boldsymbol{x}_{m_l}$ produces

$$f(\boldsymbol{z}) \quad \leq \quad \lim_{l\to\infty} f(\boldsymbol{x}_{m_l}) + \limsup_{l\to\infty} \mu_{m_l} b(\boldsymbol{x}_{m_l}) \quad \leq \quad f(\boldsymbol{x}).$$

It follows that $f(\boldsymbol{z}) \leq f(\boldsymbol{x})$ for every $\boldsymbol{x}$ in V as well. If the minimum point of $f(\boldsymbol{x})$ on V is unique, then every cluster point of the bounded sequence $\boldsymbol{x}_m$ coincides with this point. Hence, the sequence itself converges to the point. ■

Despite the elegance of the penalty and barrier methods, they suffer from three possible defects. First, they are predicated on finding the minimum point of the surrogate function for each value of the tuning constant. This entails iterations within iterations. Second, there is no obvious prescription for deciding how fast to send the tuning constants to their limits. Third, too large a value of λ_m in the penalty method or too small a value μ_m in the barrier method can lead to numerical instability in finding $\boldsymbol{x}_m$.

13.3 An Adaptive Barrier Method

The standard convex programming problem involves minimizing a convex function $f(\boldsymbol{x})$ subject to affine equality constraints $\boldsymbol{a}_i^*\boldsymbol{x} - b_i = 0$ for $1 \leq i \leq p$ and convex inequality constraints $h_j(\boldsymbol{x}) \leq 0$ for $1 \leq j \leq q$. This formulation renders the feasible region convex. To avoid distracting negative signs in this section, we will replace the constraint $h_j(\boldsymbol{x}) \leq 0$ by the constraint $v_j(\boldsymbol{x}) \geq 0$ for $v_j(\boldsymbol{x}) = -h_j(\boldsymbol{x})$. In the logarithmic barrier method, we define the barrier function

$$b(\boldsymbol{x}) \quad = \quad \sum_{j=1}^{q} \ln v_j(\boldsymbol{x}) \tag{13.2}$$

and optimize $g_m(\boldsymbol{x}) = f(\boldsymbol{x}) + \mu_m b(\boldsymbol{x})$ subject to the equality constraints. The presence of the barrier term $\ln v_j(\boldsymbol{x})$ keeps an initially inactive constraint $v_j(\boldsymbol{x})$ inactive throughout the search. Proposition 13.2.3 demonstrates convergence under specific hypotheses.

One way of improving the barrier method is to change the barrier constant as the iterations proceed [41, 162, 253]. This sounds vague, but matters simplify enormously if we view the construction of an adaptive barrier method from the perspective of the MM algorithm. Consider the following inequalities

$$\begin{aligned} & -v_j(\boldsymbol{x}_m) \ln v_j(\boldsymbol{x}) + v_j(\boldsymbol{x}_m) \ln v_j(\boldsymbol{x}_m) + dv_j(\boldsymbol{x}_m)(\boldsymbol{x} - \boldsymbol{x}_m) \\ \geq & -\frac{v_j(\boldsymbol{x}_m)}{v_j(\boldsymbol{x}_m)}[v_j(\boldsymbol{x}) - v_j(\boldsymbol{x}_m)] + dv_j(\boldsymbol{x}_m)(\boldsymbol{x} - \boldsymbol{x}_m) \\ = & -v_j(\boldsymbol{x}) + v_j(\boldsymbol{x}_m) + dv_j(\boldsymbol{x}_m)(\boldsymbol{x} - \boldsymbol{x}_m) \\ \geq & 0 \end{aligned} \tag{13.3}$$

based on the concavity of the functions $\ln y$ and $v_j(\boldsymbol{x})$. Because equality holds throughout when $\boldsymbol{x} = \boldsymbol{x}_m$, we have identified a novel function majorizing 0 and incorporating a barrier for $v_j(\boldsymbol{x})$. (Such functions are known as Bregman distances in the literature [24].) The significance of this discovery is that the surrogate function

$$\begin{aligned} g(\boldsymbol{x} \mid \boldsymbol{x}_m) \quad = \quad & f(\boldsymbol{x}) - \gamma \sum_{j=1}^{q} v_j(\boldsymbol{x}_m) \ln v_j(\boldsymbol{x}) \\ & + \gamma \sum_{j=1}^{q} dv_j(\boldsymbol{x}_m)(\boldsymbol{x} - \boldsymbol{x}_m) \end{aligned} \tag{13.4}$$

majorizes $f(\boldsymbol{x})$ up to an irrelevant additive constant. Here γ is a fixed positive constant. Minimization of the surrogate function drives $f(\boldsymbol{x})$ downhill while keeping the inequality constraints inactive. In the limit, one or more of the inequality constraints may become active.

Because minimization of the surrogate function $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ cannot be accomplished in closed form, we must revert to the MM gradient algorithm. In performing one step of Newton's method, we need the first and second differentials

$$\begin{aligned} dg(\boldsymbol{x}_m \mid \boldsymbol{x}_m) &= df(\boldsymbol{x}_m) \\ d^2g(\boldsymbol{x}_m \mid \boldsymbol{x}_m) &= d^2f(\boldsymbol{x}_m) - \gamma \sum_{j=1}^{q} d^2v_j(\boldsymbol{x}_m) \\ &\quad + \gamma \sum_{j=1}^{q} \frac{1}{v_j(\boldsymbol{x}_m)} \nabla v_j(\boldsymbol{x}_m) dv_j(\boldsymbol{x}_m). \end{aligned}$$

In view of the convexity of $f(\boldsymbol{x})$ and the concavity of the $v_j(\boldsymbol{x})$, it is obvious that $d^2g(\boldsymbol{x}_m \mid \boldsymbol{x}_m)$ is positive semidefinite. It can be positive definite even if $d^2f(\boldsymbol{x}_m)$ is not.

As a safeguard in Newton's method, it is always a good idea to contract any proposed step so that simultaneously $f(\boldsymbol{x}_{m+1}) < f(\boldsymbol{x}_m)$ and $v_j(\boldsymbol{x}_{m+1}) \geq \delta v_j(\boldsymbol{x}_m)$ for all j and a small δ such as 0.1. It is also prudent to guard against ill conditioning of the matrix $d^2g(\boldsymbol{x}_m \mid \boldsymbol{x}_m)$ as a boundary $v_j(\boldsymbol{x}) = 0$ is approached and the multiplier $v_j(\boldsymbol{x}_m)^{-1}$ tends to ∞. If one inverts $d^2g(\boldsymbol{x}_m \mid \boldsymbol{x}_m)$ or a bordered version of it by sweeping [170], then ill conditioning can be monitored as successive diagonal entries are swept. When the jth diagonal entry is dangerously close to 0, a small positive ϵ can be added to it just prior to sweeping. This apparently ad hoc remedy corresponds to adding the penalty $\frac{\epsilon}{2}(x_j - x_{mj})^2$ to $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$. Although this action does not compromise the descent property, it does attenuate the parameter increment along the jth coordinate.

The surrogate function (13.4) does not exhaust the possibilities for majorizing the objective function. If we replace the concave function $\ln y$ by the concave function $-y^{-\alpha}$ in our derivation (13.3), then we can construct for each $\alpha > 0$ and β the alternative surrogate

$$\begin{aligned} g(\boldsymbol{x} \mid \boldsymbol{x}_m) &= f(\boldsymbol{x}) + \gamma \sum_{j=1}^{q} v_j(\boldsymbol{x}_m)^{\alpha+\beta} v_j(\boldsymbol{x})^{-\alpha} \\ &\quad + \gamma\alpha \sum_{j=1}^{q} v_j(\boldsymbol{x}_m)^{\beta-1} dv_j(\boldsymbol{x}_m)(\boldsymbol{x} - \boldsymbol{x}_m) \end{aligned} \tag{13.5}$$

majorizing $f(\boldsymbol{x})$ up to an irrelevant additive constant. This surrogate also exhibits an adaptive barrier that prevents the constraint $v_j(\boldsymbol{x})$ from becoming prematurely active. Imposing the condition $\alpha + \beta > 0$ is desirable because we want a barrier to relax as its boundary is approached. For this particular surrogate, straightforward differentiation yields

$$dg(\boldsymbol{x}_m \mid \boldsymbol{x}_m) = df(\boldsymbol{x}_m) \tag{13.6}$$

$$\begin{aligned} d^2 g(\boldsymbol{x}_m \mid \boldsymbol{x}_m) &= d^2 f(\boldsymbol{x}_m) - \gamma\alpha \sum_{j=1}^{q} v_j(\boldsymbol{x}_m)^{\beta-1} d^2 v_j(\boldsymbol{x}_m) \\ &\quad + \gamma\alpha(\alpha+1) \sum_{j=1}^{q} v_j(\boldsymbol{x}_m)^{\beta-2} \nabla v_j(\boldsymbol{x}_m) dv_j(\boldsymbol{x}_m). \end{aligned} \tag{13.7}$$

Example 13.3.1 *A Geometric Programming Example*

Consider the typical geometric programming problem of minimizing

$$f(\boldsymbol{x}) = \frac{1}{x_1 x_2 x_3} + x_2 x_3$$

subject to

$$v(\boldsymbol{x}) = 4 - 2x_1 x_3 - x_1 x_2 \geq 0$$

and positive values for the x_i. Making the change of variables $x_i = e^{y_i}$ transforms the problem into a convex program. With the choice $\gamma = 1$, the MM gradient algorithm with the exponential parameterization and the log surrogate (13.4) produces the iterates displayed in the top half of Table 13.1. In this case Newton's method performs well, and none of the safeguards is needed. The MM gradient algorithm with the power surrogate (13.5) does somewhat better. The results shown in the bottom half of Table 13.1 reflect the choices $\gamma = 1$, $\alpha = 1/2$, and $\beta = 1$. ■

In the presence of linear constraints, both updates for the adaptive barrier method rely on the quadratic approximation of the surrogate function $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ using the calculated first and second differentials. This quadratic approximation is then minimized subject to the equality constraints as prescribed in Example 5.2.6.

Example 13.3.2 *Linear Programming*

Consider the standard linear programming problem of minimizing $\boldsymbol{c}^*\boldsymbol{x}$ subject to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and $\boldsymbol{x} \geq \boldsymbol{0}$ [87]. At iteration $m+1$ of the adaptive barrier method with the power surrogate (13.5), we minimize the quadratic approximation

$$\boldsymbol{c}^*\boldsymbol{x}_m + \boldsymbol{c}^*(\boldsymbol{x} - \boldsymbol{x}_m) + \tfrac{1}{2}\gamma\alpha(\alpha+1) \textstyle\sum_{j=1}^{n} x_{mj}^{\beta-2}(x_j - x_{mj})^2$$

to the surrogate subject to $\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_m) = \boldsymbol{0}$. Note here the application of the two identities (13.6) and (13.7). According to Example 5.2.6 and equation (10.5), this minimization problem has solution

$$\boldsymbol{x}_{m+1} = \boldsymbol{x}_m - [\boldsymbol{D}_m^{-1} - \boldsymbol{D}_m^{-1}\boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{D}_m^{-1}\boldsymbol{A}^*)^{-1}\boldsymbol{A}\boldsymbol{D}_m^{-1}]\boldsymbol{c},$$

where $\boldsymbol{D}_m$ is a diagonal matrix with jth diagonal entry $\gamma\alpha(\alpha+1)x_{mj}^{\beta-2}$. It is convenient here to take $\gamma\alpha(\alpha+1) = 1$ and to step halve along the search

TABLE 13.1. Solution of a geometric programming problem

Iterates for the log surrogate

Iteration m	$f(\boldsymbol{x}_m)$	x_{m1}	x_{m2}	x_{m3}
1	2.0000	1.0000	1.0000	1.0000
2	1.7299	1.4386	0.9131	0.6951
3	1.6455	1.6562	0.9149	0.6038
4	1.5993	1.7591	0.9380	0.5685
5	1.5700	1.8256	0.9554	0.5478
10	1.5147	1.9614	0.9903	0.5098
15	1.5034	1.9910	0.9977	0.5023
20	1.5008	1.9979	0.9995	0.5005
25	1.5002	1.9995	0.9999	0.5001
30	1.5000	1.9999	1.0000	0.5000
35	1.5000	2.0000	1.0000	0.5000

Iterates for the power surrogate

1	2.0000	1.0000	1.0000	1.0000
2	1.6478	1.5732	1.0157	0.6065
3	1.5817	1.7916	0.9952	0.5340
4	1.5506	1.8713	1.0011	0.5164
5	1.5324	1.9163	1.0035	0.5090
10	1.5040	1.9894	1.0011	0.5008
15	1.5005	1.9986	1.0002	0.5001
20	1.5001	1.9998	1.0000	0.5000
25	1.5000	2.0000	1.0000	0.5000

direction $\boldsymbol{x}_{m+1} - \boldsymbol{x}_m$ whenever necessary. The case $\beta = 0$ bears a strong resemblance to Karmarkar's celebrated method of linear programming. ■

We now show that the MM algorithms based on the surrogates (13.4) and (13.5) converge under fairly natural conditions. In the interests of generality, we will not require the objective function $f(\boldsymbol{x})$ to be convex. However, we will retain the assumptions of linear equality constraints $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and concave inequality constraints $v_j(\boldsymbol{x}) \geq 0$. To carry out our agenda, we assume that (a) $f(\boldsymbol{x})$ and the constraint functions $v_j(\boldsymbol{x})$ are continuously differentiable, (b) $f(\boldsymbol{x})$ is coercive, and (c) the second differential of $-\sum_{j=1}^{q} v_j(\boldsymbol{x})$ is positive definite on the affine subspace $\{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$. For simplicity, the objective function and the constraint functions are defined throughout R^n. Either algorithm starts with a feasible point with all inequality constraints inactive.

For a subset $S \subset \{1, \ldots, q\}$, let M_S be the active manifold defined by the equalities $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and $v_j(\boldsymbol{x}) = 0$ for $j \in S$ and the inequalities $v_j(\boldsymbol{x}) > 0$ for $j \notin S$. If M_S is empty, then we can safely ignore it in the sequel. Let $P_S(\boldsymbol{x})$ denote the projection matrix satisfying $dv_j(\boldsymbol{x})P_S(\boldsymbol{x}) = \boldsymbol{0}^*$ for every

$j \in S$ and defined by

$$P_S(\boldsymbol{x}) \;=\; \boldsymbol{I} - dV_S(\boldsymbol{x})^*[dV_S(\boldsymbol{x})dV_S(\boldsymbol{x})^*]^{-1}dV_S(\boldsymbol{x}), \tag{13.8}$$

where $dV_S(\boldsymbol{x})$ consists of the row vectors $dv_j(\boldsymbol{x})$ with $j \in S$ stacked one atop another. For the matrix inverse appearing in equation (13.8) to make sense, the matrix $dV_S(\boldsymbol{x})$ should have full row rank. The matrix $P_S(\boldsymbol{x})$ projects a row vector onto the subspace perpendicular to the differentials $dv_j(\boldsymbol{x})$ of the active constraints. For reasons that will become clear later, we insist that $\boldsymbol{A}P_S(\boldsymbol{x})$ have full row rank for each nonempty manifold M_S. When S is the empty set, we interpret $P_S(\boldsymbol{x})$ as the identity matrix $\boldsymbol{I}$.

We will call a point $\boldsymbol{x} \in M_S$ a stationary point if it satisfies the multiplier rule

$$df(\boldsymbol{x}) + \boldsymbol{\lambda}^*\boldsymbol{A} - \sum_{j\in S}\mu_j dv_j(\boldsymbol{x}) \;=\; \boldsymbol{0}^* \tag{13.9}$$

for some vector $\boldsymbol{\lambda}$ and collection of nonnegative coefficients μ_j. According to Proposition 6.5.3, a stationary point furnishes a global minimum of $f(\boldsymbol{x})$ when $f(\boldsymbol{x})$ is convex. We will assume that each manifold M_S possesses at most a finite number of stationary points. This is certainly the case when $f(\boldsymbol{x})$ is strictly convex, but it can also hold for linear or even non-convex objective functions [87].

Proposition 13.3.1 *Under the conditions just sketched, the adaptive barrier algorithm based on either the surrogate function (13.4) or the surrogate function (13.5) with $\beta = 1$ converges to a stationary point of $f(\boldsymbol{x})$. If $f(\boldsymbol{x})$ is convex, then the algorithms converge to the unique global minimum $\boldsymbol{y}$ of $f(\boldsymbol{x})$ subject to the constraints.*

Proof: For the sake of brevity, we consider only the surrogate function (13.4). The coerciveness assumption guarantees that $f(\boldsymbol{x})$ possesses a minimum and that all iterates of a descent algorithm remain within a compact set. Because $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ majorizes $f(\boldsymbol{x})$, it is coercive and attains its minimum value as well. Unless $f(\boldsymbol{x})$ is convex, the minimum point $\boldsymbol{x}_{m+1}$ of $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ may fail to be unique. When $f(\boldsymbol{x})$ is convex, assumption (c) implies that the quadratic form

$$\begin{aligned}\boldsymbol{u}^*d^2g(\boldsymbol{x} \mid \boldsymbol{x}_m)\boldsymbol{u} \;&=\; \boldsymbol{u}^*d^2f(\boldsymbol{x})\boldsymbol{u} - \gamma\sum_{j=1}^{q}\frac{v_j(\boldsymbol{x}_m)}{v(\boldsymbol{x})}\boldsymbol{u}^*d^2v_j(\boldsymbol{x})\boldsymbol{u} \\ &\quad + \gamma\sum_{j=1}^{q}\frac{v_j(\boldsymbol{x}_m)}{v_j(\boldsymbol{x})^2}[dv_j(\boldsymbol{x})\boldsymbol{u}]^2\end{aligned}$$

is positive whenever $\boldsymbol{u} \neq \boldsymbol{0}$. Hence, $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ is strictly convex on the affine subspace $\{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$.

Given these preliminaries, our attack is based on taking limits in the stationarity equation

$$\begin{aligned}\mathbf{0}^* &= df(\boldsymbol{x}_{m+1}) + \boldsymbol{\lambda}_m^* A \\ &\quad - \gamma \sum_{j=1}^{q} \left[\frac{v_j(\boldsymbol{x}_m)}{v_j(\boldsymbol{x}_{m+1})} dv_j(\boldsymbol{x}_{m+1}) - dv_j(\boldsymbol{x}_m) \right]\end{aligned} \tag{13.10}$$

satisfied by $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ and recovering the Lagrange multiplier rule satisfied by $f(\boldsymbol{x})$. If we are to be successful in this regard, then we must show that

$$\lim_{m\to\infty} \|\boldsymbol{x}_{m+1} - \boldsymbol{x}_m\| = 0. \tag{13.11}$$

Suppose the contrary is true. Then there exists a subsequence $\boldsymbol{x}_{m_k}$ such that

$$\liminf_{k\to\infty} \|\boldsymbol{x}_{m_k+1} - \boldsymbol{x}_{m_k}\| > 0.$$

Invoking compactness and passing to a subsubsequence if necessary, we can also assume that $\lim_{k\to\infty} \boldsymbol{x}_{m_k} = \boldsymbol{u}$ and $\lim_{k\to\infty} \boldsymbol{x}_{m_k+1} = \boldsymbol{w}$ with $\boldsymbol{u} \neq \boldsymbol{w}$. In view of the inequalities (13.3) and $g(\boldsymbol{x}_{m+1} \mid \boldsymbol{x}_m) \leq g(\boldsymbol{x}_m \mid \boldsymbol{x}_m)$ and the concavity of $v_j(\boldsymbol{x})$ and $\ln t$, we deduce the further inequalities

$$\begin{aligned}0 &\leq \gamma \sum_{j=1}^{q} [v_j(\boldsymbol{x}_{m_k}) - v_j(\boldsymbol{x}_{m_k+1}) + dv_j(\boldsymbol{x}_{m_k})(\boldsymbol{x}_{m_k+1} - \boldsymbol{x}_{m_k})] \\ &\leq \gamma \sum_{j=1}^{q} \left[v_j(\boldsymbol{x}_{m_k}) \ln \frac{v_j(\boldsymbol{x}_{m_k})}{v_j(\boldsymbol{x}_{m_k+1})} + dv_j(\boldsymbol{x}_{m_k})(\boldsymbol{x}_{m_k+1} - \boldsymbol{x}_{m_k}) \right] \\ &= g(\boldsymbol{x}_{m_k+1} \mid \boldsymbol{x}_{m_k}) - f(\boldsymbol{x}_{m_k+1}) - g(\boldsymbol{x}_{m_k} \mid \boldsymbol{x}_{m_k}) + f(\boldsymbol{x}_{m_k}) \\ &\leq f(\boldsymbol{x}_{m_k}) - f(\boldsymbol{x}_{m_k+1}).\end{aligned}$$

Given that $f(\boldsymbol{x}_m)$ is bounded and decreasing, in the limit the difference $f(\boldsymbol{x}_{m_k}) - f(\boldsymbol{x}_{m_k+1})$ tends to 0. It follows that

$$\gamma \sum_{j=1}^{q} [v_j(\boldsymbol{u}) - v_j(\boldsymbol{w}) + dv_j(\boldsymbol{u})(\boldsymbol{w} - \boldsymbol{u})] = 0,$$

contradicting the strict concavity of the sum $\sum_{j=1}^{q} v_j(\boldsymbol{x})$ on the affine subspace $\{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ and the hypothesis $\boldsymbol{u} \neq \boldsymbol{w}$.

Because the iterates $\boldsymbol{x}_m$ all belong to the same compact set, the proposition can be proved by demonstrating that every convergent subsequence $\boldsymbol{x}_{m_k}$ converges to the same stationary point $\boldsymbol{y}$. Consider such a subsequence with limit $\boldsymbol{z}$. Let us divide the constraint functions $v_j(\boldsymbol{x})$ into those that are active at $\boldsymbol{z}$ and those that are inactive at $\boldsymbol{z}$. In the former case, we take $j \in S$, and in the latter case, we take $j \in S^c$. In a moment we will

demonstrate that $\boldsymbol{z}$ is a stationary point of M_S. Because by hypothesis there are only a finite number of manifolds M_S and only a finite number of stationary points per manifold, Proposition 12.4.1 implies that all cluster points coincide and that the full sequence $\boldsymbol{x}_m$ tends to a limit $\boldsymbol{y}$. To finish the proof, we must demonstrate that $\boldsymbol{y}$ satisfies the multiplier rule for a constrained minimum. This last step is accomplished by taking limits in equality (13.10), assuming that $\boldsymbol{\lambda}_m$ and the ratios $v_j(\boldsymbol{x}_m)/v_j(\boldsymbol{x}_{m+1})$ all have limits. To avoid breaking the flow of our argument, we defer proof of these assertions. If $v_j(\boldsymbol{y}) > 0$, then it is obvious that $v_j(\boldsymbol{x}_m)/v_j(\boldsymbol{x}_{m+1})$ tends to 1, corresponding to a multiplier $\mu_j = 0$ for the inactive constraint j. If $v_j(\boldsymbol{y}) = 0$, then we must show that the limit of $v_j(\boldsymbol{x}_m)/v_j(\boldsymbol{x}_{m+1})$ exceeds 1. Otherwise, the multiplier μ_j is negative. But this limit relationship is valid because $v_j(\boldsymbol{x}_{m+1}) \leq v_j(\boldsymbol{x}_m)$ must hold for infinitely many m in order for $v_j(\boldsymbol{x}_m)$ to tend to 0.

We now return to the question of whether the limit $\boldsymbol{z}$ of the convergent subsequence $\boldsymbol{x}_{m_k}$ is a stationary point. To demonstrate that the subsequence $\boldsymbol{\lambda}_{m_k}$ converges, we multiply equation (13.10) on the right by the matrix $P_S(\boldsymbol{x}_{m_k+1})\boldsymbol{A}^*$ and solve for $\boldsymbol{\lambda}_{m_k}$. This is possible because the matrix

$$B(\boldsymbol{x}_{m_k}) \;=\; \boldsymbol{A}P_S(\boldsymbol{x}_{m_k+1})\boldsymbol{A}^* \;=\; \boldsymbol{A}P_S(\boldsymbol{x}_{m_k+1})P_S(\boldsymbol{x}_{m_k+1})^*\boldsymbol{A}^*$$

has full rank by assumption. Since $P_S(\boldsymbol{x})$ annihilates $dv_j(\boldsymbol{x})$ for $j \in S$, and since $\boldsymbol{x}_{m_k+1}$ converges to $\boldsymbol{z}$, a brief calculation shows that

$$\boldsymbol{\lambda}^* \;=\; \lim_{k\to\infty} \boldsymbol{\lambda}_{m_k} \;=\; -df(\boldsymbol{z})P_S(\boldsymbol{z})\boldsymbol{A}^*B(\boldsymbol{z})^{-1}.$$

To prove that the ratio $v_j(\boldsymbol{x}_{m_k})/v_j(\boldsymbol{x}_{m_k+1})$ has a limit for $j \in S$, we multiply equation (13.10) on the right by the matrix-vector product

$$P_{S_{-j}}(\boldsymbol{x}_{m_k+1})\nabla v_j(\boldsymbol{x}_{m_k+1}),$$

where $S_{-j} = S \setminus \{j\}$. This action annihilates all $dv_i(\boldsymbol{x}_{m_k+1})$ with $i \in S_{-j}$ and makes it possible to express

$$\lim_{k\to\infty} \frac{v_j(\boldsymbol{x}_{m_k})}{v_j(\boldsymbol{x}_{m_k+1})} \;=\; \frac{[df(\boldsymbol{z}) + \lambda^* A + \gamma dv_j(\boldsymbol{z})]P_{S_{-j}}(\boldsymbol{z})\nabla v_j(\boldsymbol{z})}{\gamma dv_j(\boldsymbol{z})P_{S_{-j}}(\boldsymbol{z})\nabla v_j(\boldsymbol{z})}.$$

Note that the denominator $\gamma dv_j(\boldsymbol{z})P_{S_{-j}}(\boldsymbol{z})\nabla v_j(\boldsymbol{z}) > 0$ because $\gamma > 0$ and $dV_{S_{-j}}(\boldsymbol{z})$ has full row rank. Given these results, we can legitimately take limits in equation (13.10) along the given subsequence and recover the multiplier rule (13.9) at $\boldsymbol{z}$.

Now that we have demonstrated that $\boldsymbol{x}_m$ tends to a unique limit $\boldsymbol{y}$, we can show that $\boldsymbol{\lambda}_m$ and the ratios $v_j(\boldsymbol{x}_m)/v_j(\boldsymbol{x}_{m+1})$ tend to well-defined limits by the logic employed with the subsequence $\boldsymbol{x}_{m_k}$. As noted earlier, this permits us to take limits in equation (13.10) and recover the multiplier

rule at $\boldsymbol{y}$. When $f(\boldsymbol{x})$ is convex, $\boldsymbol{y}$ furnishes the global minimum. If there is another global minimum $\boldsymbol{w}$, then the entire line segment between $\boldsymbol{w}$ and $\boldsymbol{y}$ consists of minimum points. This contradicts the assumption that there are at most a finite number of stationary points throughout the feasible region. Hence, the minimum point $\boldsymbol{y}$ is unique. ■

The next example illustrates that the local rate of convergence can be linear even when one of the constraints $v_i(\boldsymbol{x}) \geq 0$ is active at the minimum.

Example 13.3.3 *Convergence for the Multinomial Distribution*

As pointed out in Example 1.4.2, the loglikelihood for a multinomial distribution with d categories reduces to $\sum_{i=1}^d n_i \ln p_i$, where n_i is the observed number of counts in category i and p_i is the probability attached to category i. Maximizing the loglikelihood subject to the constraints $p_i \geq 0$ and $\sum_{i=1}^d p_i = 1$ gives the explicit maximum likelihood estimates $p_i = n_i/n$ for n trials. To compute the maximum likelihood estimates iteratively using the surrogate function (13.4), we find a stationary point of the Lagrangian

$$-\sum_{i=1}^d n_i \ln p_i - \gamma \sum_{i=1}^d p_{mi} \ln p_i + \gamma \sum_{i=1}^d (p_i - p_{mi}) + \lambda\Big(\sum_{i=1}^d p_i - 1\Big).$$

Setting the ith partial derivative of the Lagrangian equal to 0 gives

$$-\frac{n_i}{p_i} - \frac{\gamma p_{mi}}{p_i} + \gamma + \lambda \;=\; 0. \tag{13.12}$$

Multiplying equation (13.12) by p_i, summing on i, and solving for λ yield $\lambda = n$. Substituting this value back in equation (13.12) produces

$$p_{m+1,i} \;=\; \frac{n_i + \gamma p_{mi}}{n + \gamma}.$$

At first glance it is not obvious that p_{mi} tends to n_i/n, but the algebraic rearrangement

$$\begin{aligned} p_{m+1,i} - \frac{n_i}{n} &= \frac{n_i + \gamma p_{mi}}{n + \gamma} - \frac{n_i}{n} \\ &= \frac{\gamma}{n + \gamma}\Big(p_{mi} - \frac{n_i}{n}\Big) \end{aligned}$$

shows that p_{mi} approaches n_i/n at the linear rate $\gamma/(n + \gamma)$. This is true regardless of whether $n_i/n = 0$ or $n_i/n > 0$. ■

13.4 Imposition of a Prior in EM Clustering

Priors imposed in Bayesian models play out as penalties in maximum a posteriori estimation. As an example, consider the EM clustering model studied in Sect. 9.5. There we postulated that each normally distributed

cluster had the same variance matrix. Relaxing this assumption sometimes causes the likelihood to become unbounded. Imposing a prior improves inference and stabilizes numerical estimation of parameters [44]. Let us review the derivation of the EM algorithm with these benefits in mind. The form of the EM updates

$$\pi_{n+1,j} = \frac{1}{m}\sum_{i=1}^{m} w_{ij}$$

for the admixture proportions π_j depend only on Bayes' rule and is valid regardless of the particular cluster densities. Here w_{ij} is the posterior probability that observation i comes from cluster j. To estimate the cluster means and common variance, we formed the surrogate function

$$\begin{aligned} & Q(\{\boldsymbol{\mu}_j, \boldsymbol{\Omega}_j\}_{j=1}^k \mid \{\boldsymbol{\mu}_{nj}, \boldsymbol{\Omega}_{nj}\}_{j=1}^k) \\ = & -\frac{1}{2}\sum_{j=1}^{k}\Big(\sum_{i=1}^{m} w_{ij}\Big)\ln\det\boldsymbol{\Omega}_j \\ & -\frac{1}{2}\sum_{j=1}^{k}\operatorname{tr}\Big[\boldsymbol{\Omega}_j^{-1}\sum_{i=1}^{m} w_{ij}(\boldsymbol{y}_i-\boldsymbol{\mu}_j)(\boldsymbol{y}_i-\boldsymbol{\mu}_j)^*\Big] \end{aligned}$$

with all $\boldsymbol{\Omega}_j = \boldsymbol{\Omega}$.

It is mathematically convenient to relax the common variance assumption and impose independent inverse Wishart priors on the different variance matrices $\boldsymbol{\Omega}_j$. In view of Problem 37 of Chap. 4, this amounts to adding the logprior

$$-\sum_{j=1}^{k}\left[\frac{a}{2}\ln\det\boldsymbol{\Omega}_j + \frac{b}{2}\operatorname{tr}(\boldsymbol{\Omega}_j^{-1}\boldsymbol{S}_j)\right]$$

to the surrogate function. Here the positive constants a and b and the positive definite matrices $\boldsymbol{S}_j$ must be determined. Regardless of how these choices are made, we derive the usual EM updates

$$\boldsymbol{\mu}_{n+1,j} = \frac{1}{\sum_{i=1}^{m} w_{ij}}\sum_{i=1}^{m} w_{ij}\boldsymbol{y}_i$$

of the cluster means. Note that the constants a and b and the matrices $\boldsymbol{S}_j$ have no influence on the weights w_{ij} computed via Bayes' rule.

The most natural choice is to take all $\boldsymbol{S}_j$ equal to the sample variance matrix

$$\boldsymbol{S} = \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{y}_i-\bar{\boldsymbol{y}})(\boldsymbol{y}_i-\bar{\boldsymbol{y}})^*.$$

This choice is probably too diffuse, but it is better to err on the side of vagueness and avoid getting trapped at a local mode of the likelihood surface. In the absence of a prior, Example 4.7.6 implies that the EM update of $\boldsymbol{\Omega}_j$ is

$$\tilde{\boldsymbol{\Omega}}_{n+1,j} \;=\; \frac{1}{\sum_{i=1}^m w_{ij}} \sum_{i=1}^m w_{ij}(\boldsymbol{y}_i - \boldsymbol{\mu}_{n+1,j})(\boldsymbol{y}_i - \boldsymbol{\mu}_{n+1,j})^*$$

By the same reasoning, the maximum of the penalized surrogate function with respect to $\boldsymbol{\Omega}_j$ is

$$\boldsymbol{\Omega}_{n+1,j} \;=\; \frac{a}{a + \sum_{i=1}^m w_{ij}} \left(\frac{b}{a}\boldsymbol{S}\right) + \frac{\sum_{i=1}^m w_{ij}}{a + \sum_{i=1}^m w_{ij}} \tilde{\boldsymbol{\Omega}}_{n+1,j}.$$

In other words, the penalized EM update is a convex combination of the standard EM update and the mode $\frac{b}{a}\boldsymbol{S}$ of the prior. Chen and Tan [44] tentatively recommend the choice $a = b = 2/\sqrt{m}$. As m tends to ∞, the influence of the prior diminishes.

13.5 Model Selection and the Lasso

We now turn to penalized regression and continuous model selection. Our focus will be on the lasso penalty and its application in regression problems where the number of predictors p exceeds the number of cases n [45, 49, 229, 252, 256]. The lasso also finds applications in generalized linear models. In each of these contexts, let y_i be the response for case i, x_{ij} be the value of predictor j for case i, and β_j be the regression coefficient corresponding to predictor j. In practice one should standardize each predictor to have mean 0 and variance 1. Standardization puts all regression coefficients on a common scale as implicitly demanded by the lasso penalty.

The intercept α is ignored in the lasso penalty, whose strength is determined by the positive tuning constant λ. If $\boldsymbol{\theta} = (\alpha, \beta_1, \ldots, \beta_p)^*$ is the parameter vector and $g(\boldsymbol{\theta})$ is the loss function ignoring the penalty, then the lasso minimizes the criterion

$$f(\boldsymbol{\theta}) \;=\; g(\boldsymbol{\theta}) + \lambda \sum_{j=1}^p |\beta_j|,$$

where $g(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^n (y_i - \boldsymbol{x}_i^*\boldsymbol{\theta})^2$ in ℓ_2 regression and $g(\boldsymbol{\theta}) = \sum_{i=1}^n |y_i - \boldsymbol{x}_i^*\boldsymbol{\theta}|$ in ℓ_1 regression. The penalty $\lambda \sum_j |\beta_j|$ shrinks each β_j toward the origin and tends to discourage models with large numbers of irrelevant predictors. The lasso penalty is more effective in this regard than the ridge penalty $\lambda \sum_j \beta_j^2$ because $|b|$ is much bigger than b^2 for small b.

Lasso penalized estimation raises two issues. First, what is the most effective method of minimizing the objective function $f(\boldsymbol{\theta})$? In the current

section we highlight the method of coordinate descent [55, 98, 100, 276]. Second, how does one choose the tuning constant λ? The standard answer is cross-validation. Although this is a good reply, it does not resolve the problem of how to minimize average cross-validation error as measured by the loss function. Recall that in k-fold cross-validation, one divides the data into k equal batches (subsamples) and estimates parameters k times, leaving one batch out per time. The testing error (total loss) for each omitted batch is computed using the estimates derived from the remaining batches, and the cross-validation error $c(\lambda)$ is computed by averaging testing error across the k batches.

Unless carefully planned, evaluation of $c(\lambda)$ on a grid of points may be computationally costly, particularly if grid points occur near $\lambda = 0$. Because coordinate descent is fastest when λ is large and the vast majority of β_j are estimated as 0, it makes sense to start with a very large value and work downward. One advantage of this tactic is that parameter estimates for a given λ can be used as parameter starting values for the next lower λ. For the initial value of λ, the starting value $\boldsymbol{\theta} = \mathbf{0}$ is recommended. It is also helpful to set an upper bound on the number of active parameters allowed and abort downward sampling of λ when this bound is exceeded. Once a fine enough grid is available, visual inspection usually suggests a small interval flanking the minimum. Application of golden section search over the flanking interval will then quickly lead to the minimum.

Coordinate descent comes in several varieties. The standard version cycles through the parameters and updates each in turn. An alternative version is greedy and updates the parameter giving the largest decrease in the objective function. Because it is impossible to tell in advance the extent of each decrease, the greedy version uses the surrogate criterion of steepest descent. In other words, for each parameter we compute forward and backward directional derivatives and update the parameter with the most negative directional derivative, either forward or backward. The overhead of keeping track of the directional derivative works to the detriment of the greedy method. For ℓ_1 regression, the overhead is relatively light, and greedy coordinate descent converges faster than cyclic coordinate descent.

Although the lasso penalty is nondifferentiable, it does possess directional derivatives along each forward or backward coordinate direction. For instance, if $\boldsymbol{e}_j$ is the coordinate direction along which β_j varies, then

$$d_{\boldsymbol{e}_j} f(\boldsymbol{\theta}) = \lim_{t \downarrow 0} \frac{f(\boldsymbol{\theta} + t\boldsymbol{e}_j) - f(\boldsymbol{\theta})}{t} = d_{\boldsymbol{e}_j} g(\boldsymbol{\theta}) + \begin{cases} \lambda & \beta_j \geq 0 \\ -\lambda & \beta_j < 0, \end{cases}$$

and

$$d_{-\boldsymbol{e}_j} f(\boldsymbol{\theta}) = \lim_{t \downarrow 0} \frac{f(\boldsymbol{\theta} - t\boldsymbol{e}_j) - f(\boldsymbol{\theta})}{t} = d_{-\boldsymbol{e}_j} g(\boldsymbol{\theta}) + \begin{cases} -\lambda & \beta_j > 0 \\ \lambda & \beta_j \leq 0. \end{cases}$$

In ℓ_1 regression, the loss function is also nondifferentiable, and a brief calculation shows that the coordinate directional derivatives are

$$d_{\boldsymbol{e}_j} \sum_{i=1}^n |y_i - \boldsymbol{x}_i^* \boldsymbol{\theta}| = \sum_{i=1}^n \begin{cases} -x_{ij} & y_i - \boldsymbol{x}_i^* \boldsymbol{\theta} > 0 \\ x_{ij} & y_i - \boldsymbol{x}_i^* \boldsymbol{\theta} < 0 \\ |x_{ij}| & y_i - \boldsymbol{x}_i^* \boldsymbol{\theta} = 0 \end{cases}$$

and

$$d_{-\boldsymbol{e}_j} \sum_{i=1}^n |y_i - \boldsymbol{x}_i^* \boldsymbol{\theta}| = \sum_{i=1}^n \begin{cases} x_{ij} & y_i - \boldsymbol{x}_i^* \boldsymbol{\theta} > 0 \\ -x_{ij} & y_i - \boldsymbol{x}_i^* \boldsymbol{\theta} < 0 \\ |x_{ij}| & y_i - \boldsymbol{x}_i^* \boldsymbol{\theta} = 0 \end{cases}$$

with predictor vector $\boldsymbol{x}_i^* = (1, \boldsymbol{z}_i^*)$ for case i. Fortunately, when a function is differentiable, its directional derivative along $\boldsymbol{e}_j$ coincides with its ordinary partial derivative, and its directional derivative along $-\boldsymbol{e}_j$ coincides with the negative of its ordinary partial derivative.

When we visit parameter β_j in cyclic coordinate descent, we evaluate $d_{\boldsymbol{e}_j} f(\boldsymbol{\theta})$ and $d_{-\boldsymbol{e}_j} f(\boldsymbol{\theta})$. If both are nonnegative, then we skip the update for β_j. This decision is defensible when $g(\boldsymbol{\theta})$ is convex because the sign of a directional derivative fully determines whether improvement can be made in that direction. If either directional derivative is negative, then we must solve for the minimum in that direction. When the current slope parameter β_j is parked at 0 and the partial derivative $\frac{\partial}{\partial \beta_j} g(\boldsymbol{\theta})$ exists,

$$d_{\boldsymbol{e}_j} f(\boldsymbol{\theta}) = \frac{\partial}{\partial \beta_j} g(\boldsymbol{\theta}) + \lambda, \quad d_{-\boldsymbol{e}_j} f(\boldsymbol{\theta}) = -\frac{\partial}{\partial \beta_j} g(\boldsymbol{\theta}) + \lambda.$$

Hence, β_j moves to the right if $\frac{\partial}{\partial \beta_j} g(\boldsymbol{\theta}) < -\lambda$, to the left if $\frac{\partial}{\partial \beta_j} g(\boldsymbol{\theta}) > \lambda$, and stays fixed otherwise. In underdetermined problems with just a few relevant predictors, most updates are skipped, and the parameters never budge from their starting values of 0. This simple fact plus the complete absence of matrix operations explains the speed of coordinate descent. It inherits its numerical stability from the descent property of each update.

13.6 Lasso Penalized ℓ_1 Regression

In lasso constrained ℓ_1 regression, greedy coordinate descent is quick because directional derivatives are trivial to update. Indeed, if updating β_j does not alter the sign of the residual $y_i - \boldsymbol{x}_i^* \boldsymbol{\theta}$ for case i, then the contributions of case i to the various directional derivatives do not change. When the residual $y_i - \boldsymbol{x}_i^* \boldsymbol{\theta}$ changes sign, these contributions change by $\pm 2x_{ij}$. When a residual changes from 0 to nonzero or vice versa, the increment depends on the sign of the nonzero residual and the sign of x_{ij}.

Updating the value of the chosen parameter can be achieved by the nearly forgotten algorithm of Edgeworth [80, 81], which for a long time was

considered a competitor of least squares. Portnoy and Koenker [214] trace the history of the algorithm from Boscovich to Laplace to Edgeworth. It is fair to say that the algorithm has managed to cling to life despite decades of obscurity both before and after its rediscovery by Edgeworth.

To illustrate Edgeworth's algorithm in operation, consider minimizing the two-parameter model

$$g(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{n} |y_i - \alpha - z_i\beta|$$

with a single slope β. To update α, we recall the well-known connection between ℓ_1 regression and medians and replace α for fixed β by the sample median of the numbers $v_i = y_i - z_i\beta$. This action drives $g(\boldsymbol{\theta})$ downhill. Updating β for α fixed depends on writing

$$g(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{n} |z_i| \left| \frac{y_i - \alpha}{z_i} - \beta \right|,$$

sorting the numbers $v_i = (y_i - \alpha)/z_i$, and finding the weighted median with weight $w_i = |z_i|$ assigned to v_i. We replace β by the order statistic $v_{[i]}$ with weight $w_{[i]}$ whose index i satisfies

$$\sum_{j=1}^{i-1} w_{[j]} \;<\; \frac{1}{2}\sum_{j=1}^{n} w_{[j]}, \quad \sum_{j=1}^{i} w_{[j]} \;\geq\; \frac{1}{2}\sum_{j=1}^{n} w_{[j]}.$$

Problem 6 demonstrates that this choice is valid. Edgeworth's algorithm easily generalizes to multiple linear regression. Implementing the algorithm with a lasso penalty requires viewing the penalty terms as the absolute values of pseudo-residuals. Thus, we write

$$\lambda|\beta_j| \;=\; |y - \boldsymbol{x}^*\boldsymbol{\theta}|$$

by taking $y = 0$ and $x_k = \lambda 1_{\{k=j\}}$.

Two criticisms have been leveled at Edgeworth's algorithm. First, although it drives the objective function steadily downhill, it sometimes stalls at an inferior point. See Problem 8 for an example. The second criticism is that convergence often occurs in a slow seesaw pattern. These defects are not completely fatal. As late as 1978, Armstrong and Kung published a computer implementation of Edgeworth's algorithm in the journal *Applied Statistics* [4].

13.7 Lasso Penalized ℓ_2 Regression

In ℓ_2 regression with a lasso penalty, we minimize the objective function

$$f(\boldsymbol{\theta}) \;=\; \frac{1}{2}\sum_{i=1}^{n}(y_i - \alpha - \boldsymbol{z}_i^*\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}|\beta_j| \;=\; g(\boldsymbol{\theta}) + \lambda\sum_{j=1}^{p}|\beta_j|.$$

The update of the intercept parameter can be written as

$$\hat{\alpha} \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{z}_i^*\boldsymbol{\beta}) \;=\; \alpha - \frac{1}{n}\frac{\partial}{\partial\alpha}g(\boldsymbol{\theta}).$$

For the parameter β_k, there are separate solutions to the left and right of 0. These boil down to

$$\begin{aligned}\hat{\beta}_{k,-} &= \min\left\{0, \beta_k - \frac{\frac{\partial}{\partial\beta_k}g(\boldsymbol{\theta}) - \lambda}{\sum_{i=1}^{n} z_{ik}^2}\right\} \\ \hat{\beta}_{k,+} &= \max\left\{0, \beta_k - \frac{\frac{\partial}{\partial\beta_k}g(\boldsymbol{\theta}) + \lambda}{\sum_{i=1}^{n} z_{ik}^2}\right\}.\end{aligned}$$

The reader can check that only one of these two solutions can be nonzero. The partial derivatives

$$\frac{\partial}{\partial\alpha}g(\boldsymbol{\theta}) \;=\; -\sum_{i=1}^{n} r_i, \qquad \frac{\partial}{\partial\beta_k}g(\boldsymbol{\theta}) \;=\; -\sum_{i=1}^{n} r_i z_{ik}$$

of $g(\boldsymbol{\theta})$ are easy to compute provided we keep track of all of the residuals $r_i = y_i - \alpha - \boldsymbol{z}_i^*\boldsymbol{\beta}$. The residual r_i starts with the value y_i and is reset to $r_i + \alpha - \hat{\alpha}$ when α is updated and to $r_i + z_{ij}(\beta_j - \hat{\beta}_j)$ when β_j is updated. Organizing all updates around residuals promotes fast evaluation of $g(\boldsymbol{\theta})$. At the expense of somewhat more complex code [99], a better tactic is to exploit the identity

$$\sum_{i=1}^{n} r_i z_{ik} \;=\; \sum_{i=1}^{n} y_i z_{ik} - \alpha\sum_{i=1}^{n} z_{ik} - \sum_{j:|\beta_j|>0}\left(\sum_{i=1}^{n} z_{ij}z_{ik}\right)\beta_j.$$

This representation suggests storing and reusing the inner products

$$\sum_{i=1}^{n} y_i z_{ik}, \qquad \sum_{i=1}^{n} z_{ik}, \qquad \sum_{i=1}^{n} z_{ij}z_{ik}$$

for the active predictors.

Example 13.7.1 *Obesity and Gene Expression in Mice*

Consider a genetics example involving gene expression levels and obesity in mice. Wang et al. [268] measured abdominal fat mass on $n = 311$ F2 mice (155 males and 156 females). The F2 mice were created by mating two inbred strains and then mating brother-sister pairs from the resulting offspring. Wang et al. [268] also recorded the expression levels in liver of $p = 23{,}388$ genes in each mouse. A reasonable model postulates

$$y_i \;=\; 1_{\{i\,\text{male}\}}\alpha_1 + 1_{\{i\,\text{female}\}}\alpha_2 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i,$$

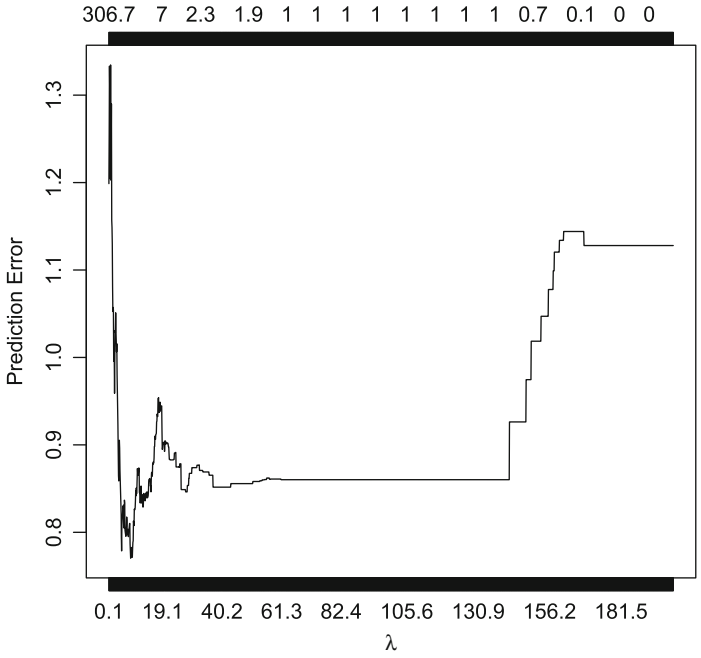

FIGURE 13.1. The cross-validation curve $c(\lambda)$ for obesity in mice

where y_i measures fat mass on mouse i, x_{ij} is the expression level of gene j in mouse i, and ϵ_i is random error. Since male and female mice exhibit across the board differences in size and physiology, it is prudent to estimate a different intercept for each sex. Figure 13.1 plots average prediction error as a function of λ (lower horizontal axis) and the average number of nonzero predictors (upper horizontal axis). Here we use ℓ_2 penalized regression and 10-fold cross-validation. Examination of the cross-validation curve $c(\lambda)$ over a fairly dense grid shows an optimal λ of 7.8 with 41 nonzero predictors. For ℓ_1 penalized regression, the optimal λ is around 3.5 with 77 nonzero predictors. The preferred ℓ_1 and ℓ_2 models share 27 predictors in common. Several of the genes identified are known or suspected to be involved in lipid metabolism, adipose deposition, and impaired insulin sensitivity in mice. More details can be found in the paper [276].

The tactics described for ℓ_2 regression carry over to generalized linear models. In this setting, the loss function $g(\boldsymbol{\theta})$ is the negative loglikelihood. In many cases, $g(\boldsymbol{\theta})$ is convex, and it is possible to determine whether progress can be made along a forward or backward coordinate direction without actually minimizing the objective function. It is clearly computationally beneficial to organize parameter updates by tracking the linear predictor $\alpha+\boldsymbol{z}_i^*\boldsymbol{\beta}$ of each case. Although we no longer have explicit solutions

to fall back on, the scoring algorithm serves as a substitute. Since it usually converges in a few iterations, the computational overhead of cyclic coordinate descent remains manageable.

13.8 Penalized Discriminant Analysis

Discriminant analysis is another attractive candidate for penalized estimation. In discriminant analysis with two categories, each case i is characterized by a feature vector $\boldsymbol{z}_i$ and a category membership indicator y_i taking the values -1 or 1. In the machine learning approach to discriminant analysis [231, 264], the hinge loss function $[1-y_i(\alpha+\boldsymbol{z}_i^*\boldsymbol{\beta})]_+$ plays a prominent role. Here u_+ is shorthand for the convex function $\max\{u, 0\}$. Just as in ordinary regression, we can penalize the overall loss

$$g(\boldsymbol{\theta}) \quad = \quad \sum_{i=1}^{n}[1 - y_i(\alpha + \boldsymbol{z}_i^*\boldsymbol{\beta})]_+$$

by imposing a lasso or ridge penalty. Note that the linear regression function $h_i(\boldsymbol{\theta}) = \alpha + \boldsymbol{z}_i^*\boldsymbol{\beta}$ predicts either -1 or 1. If $y_i = 1$ and $h_i(\boldsymbol{\theta})$ over-predicts in the sense that $h_i(\boldsymbol{\theta}) > 1$, then there is no loss. Similarly, if $y_i = -1$ and $h_i(\boldsymbol{\theta})$ under-predicts in the sense that $h_i(\boldsymbol{\theta}) < -1$, then there is no loss.

Most strategies for estimating $\boldsymbol{\theta}$ pass to the dual of the original minimization problem. A simpler strategy is to majorize each contribution to the loss by a quadratic and minimize the surrogate loss plus penalty [114]. A little calculus shows that $(u)_+$ is majorized at $u_m \neq 0$ by the quadratic

$$q(u \mid u_m) \quad = \quad \frac{1}{4|u_m|}\left(u + |u_m|\right)^2. \tag{13.13}$$

(See Problem 13.) In fact, this is the best quadratic majorizer of u_+ [62]. Both of the majorizations (8.12) and (13.13) have singularities at the point $u_m = 0$. One simple fix is to replace $|u_m|$ by $|u_m| + \epsilon$ wherever $|u_m|$ appears in a denominator in either formula. We recommend double precision arithmetic with $0 < \epsilon \leq 10^{-5}$. Problem 14 explores a more sophisticated remedy that replaces the functions $|u|$ and u_+ by differentiable approximations.

In any case, if we impose a ridge penalty, then the hinge majorization leads to a pure MM algorithm exploiting weighted least squares. Coordinate descent algorithms with a lasso or ridge penalty are also enabled by majorization, but each coordinate update merely decreases the objective function along the given coordinate direction. Fortunately, this drawback is outweighed by the gain in numerical simplicity in majorizing hinge loss. The decisions to use a lasso or ridge penalty and apply pure MM or coordinate descent with majorization will be dictated in practical problems by the number of potential predictors. If a lasso penalty is imposed to

eliminate irrelevant predictors, then cyclic coordinate descent is preferable, with the surrogate function substituting for the objective function in each parameter update.

In discriminant analysis with more than two categories, it is convenient to pass to ϵ-insensitive loss and multiple linear regression. The story is too long to tell here, but it is worth mentioning that the conjunction of a parsimonious loss function and efficient MM or coordinate descent algorithms produce some of the most effective discriminant analysis methods tested [169, 277].

13.9 Problems

1. In Example 13.2.1 prove directly that the solution displayed in equation (13.1) converges to the minimum point of $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ subject to the linear constraints $\boldsymbol{V}\boldsymbol{\beta} = \boldsymbol{d}$. (Hints: Assume that the matrix $\boldsymbol{V}$ has full column rank and consult Example 5.2.6, Proposition 5.2.2, and Problem 10 of Chap. 11.)

2. Prove that the surrogate function (13.5) majorizes $f(\boldsymbol{x})$ up to an irrelevant additive constant.

3. The power plant production problem [226] involves minimizing

$$f(\boldsymbol{x}) \;=\; \sum_{i=1}^{n} f_i(x_i), \quad f_i(x_i) \;=\; a_i x_i + \frac{1}{2} b_i x_i^2$$

subject to the constraints $0 \leq x_i \leq u_i$ for each i and $\sum_{i=1}^{n} x_i \geq d$. For plant i, x_i is the power output, u_i is the capacity, and $f_i(x_i)$ is the cost. The total demand is d, and the cost constants a_i and b_i are positive. This problem can be solved by the adaptive barrier algorithm. Program this algorithm and test it on a simple example with at least two power plants. Argue that the minimum is unique. Example 15.6.1 sketches another approach.

4. In Problem 3 investigate the performance of cyclic coordinate descent. Explain why it fails.

5. Implement and test the EM clustering algorithm with a Bayesian prior. Apply the algorithm to Fisher's classic iris data set. Fisher's data can be downloaded from the web. See the book [191] for commentary and further references.

6. Show that $\hat{\mu}$ minimizes $f(\mu) = \sum_{i=1}^{n} w_i |x_i - \mu|$ if and only if

$$\sum_{x_i < \hat{\mu}} w_i \;\leq\; \frac{1}{2} \sum_{i=1}^{n} w_i \quad \text{and} \quad \sum_{x_i \leq \hat{\mu}} w_i \;\geq\; \frac{1}{2} \sum_{i=1}^{n} w_i.$$

Assume that the weights w_i are positive. (Hint: Apply Proposition 6.5.2.)

7. Consider the piecewise linear function

$$f(\mu) \;=\; c\mu + \sum_{i=1}^{n} w_i |x_i - \mu|,$$

where the positive weights satisfy $\sum_{i=1}^{n} w_i = 1$ and the points satisfy $x_1 < x_2 < \cdots < x_n$. Show that $f(\mu)$ has no minimum when $|c| > 1$. What happens when $c = 1$ or $c = -1$? This leaves the case $|c| < 1$. Show that a minimum occurs when

$$\sum_{x_i > \mu} w_i - \sum_{x_i \le \mu} w_i \;\le\; c \text{ and } \sum_{x_i \ge \mu} w_i - \sum_{x_i < \mu} w_i \;\ge\; c.$$

(Hints: A crude plot of $f(\mu)$ might help. What conditions on the right-hand and left-hand derivatives of $f(\mu)$ characterize a minimum?)

8. Show that Edgeworth's algorithm [178] for ℓ_1 regression converges to an inferior point for the data values $(0.3, -1.0), (-0.4, -0.1), (-2.0, -2.9), (-0.9, -2.4)$, and $(-1.1, 2.2)$ for the pairs (x_i, y_i) and parameter starting values $(\alpha, \beta) = (3.5, -1.0)$.

9. Implement and test greedy coordinate descent for lasso penalized ℓ_1 regression or cyclic coordinate descent for lasso penalized ℓ_2 regression.

10. In lasso penalized regression, suppose the convex loss function $g(\boldsymbol{\theta})$ is differentiable. A stationary point $\boldsymbol{\theta}$ of coordinate descent satisfies the conditions $d_{\boldsymbol{e}_j} f(\boldsymbol{\theta}) \ge 0$ and $d_{-\boldsymbol{e}_j} f(\boldsymbol{\theta}) \ge 0$ for all j. Here the intercept α varies along the coordinate direction $\boldsymbol{e}_0$. Calculate the general directional derivative

$$d_{\boldsymbol{v}} f(\boldsymbol{\theta}) \;=\; \sum_j \frac{\partial}{\partial \theta_j} g(\boldsymbol{\theta}) v_j + \lambda \sum_{j>0} \begin{cases} v_j & \theta_j > 0 \\ -v_j & \theta_j < 0 \\ |v_j| & \theta_j = 0 \end{cases}$$

and show that

$$d_{\boldsymbol{v}} f(\boldsymbol{\theta}) \;=\; \sum_{v_j > 0} d_{\boldsymbol{e}_j} f(\boldsymbol{\theta}) v_j + \sum_{v_j < 0} d_{-\boldsymbol{e}_j} f(\boldsymbol{\theta}) |v_j|.$$

Conclude that every directional derivative is nonnegative at a stationary point. In view of Proposition 6.5.2, stationary points therefore coincide with minimum points. This result does not hold for lasso penalized ℓ_1 regression.

11. Show that the function $\|\boldsymbol{x}\|_0 = \sum_{i=1}^n 1_{\{x_i \neq 0\}}$ satisfies the properties:

    (a) $\|\boldsymbol{x}\|_0$ is nonnegative and equal to 0 if and only if $\boldsymbol{x} = \boldsymbol{0}$,

    (b) $\|\boldsymbol{x}\|_0 = \|-\boldsymbol{x}\|_0$,

    (c) $\|\boldsymbol{x}+\boldsymbol{y}\|_0 \leq \|\boldsymbol{x}\|_0 + \|\boldsymbol{y}\|_0$,

    (d) The function $\boldsymbol{x} \mapsto \|\boldsymbol{x}\|_0$ is lower semicontinuous.

    What norm property fails?

12. For the ℓ_0 "norm" $\|\boldsymbol{x}\|_0$ defined in the previous problem, demonstrate that

$$\|\boldsymbol{x}\|_0 \quad = \quad \lim_{\epsilon \downarrow 0} \sum_{i=1}^n \frac{\ln\left(1 + \frac{|x_i|}{\epsilon}\right)}{\ln\left(1 + \frac{1}{\epsilon}\right)}.$$

    Note that the same limit applies if one substitutes x_i^2 for $|x_i|$. Now prove the majorization

$$\ln(\epsilon + y) \quad \leq \quad \ln(\epsilon + y_m) + \frac{1}{\epsilon + y_m}(y - y_m)$$

    for nonnegative scalars y and y_m, and show how it can be employed to majorize approximations to $\|\boldsymbol{x}\|_0$ based on the choices $|x_i|$ and x_i^2. See the references [39, 88, 272] for applications to sparse estimation and machine learning.

13. Show that the function $u_+ = \max\{u, 0\}$ is majorized by the quadratic function (13.13) at a point $u_m \neq 0$. Why does it suffice to prove that u_+ and $q(u \mid u_m)$ have the same value and same derivative at u_m and $-u_m$? Also check that u_+^2 is majorized by u^2 for $u_m \geq 0$ and by $(u - u_m)^2$ for $u_m < 0$. (Hint: Draw rough graphs of u_+ and $q(u \mid u_m)$.)

14. For a small $\epsilon > 0$, the functions $\sqrt{u^2 + \epsilon} - \sqrt{\epsilon}$ and $\sqrt{u_+^2 + \epsilon} - \sqrt{\epsilon}$ are excellent differentiable approximations to $|u|$ and u_+, respectively. Derive the majorizations

$$\begin{aligned}
\sqrt{u^2 + \epsilon} - \sqrt{\epsilon} \quad &\leq \quad \sqrt{u_m^2 + \epsilon} - \sqrt{\epsilon} + \frac{1}{2\sqrt{u_m^2 + \epsilon}}(u^2 - u_m^2) \\
\sqrt{u_+^2 + \epsilon} - \sqrt{\epsilon} \quad &\leq \quad \begin{cases} \frac{1}{2\sqrt{\epsilon}} u^2 & u_m = 0 \\ \frac{\sqrt{u_m^2+\epsilon}+\sqrt{\epsilon}}{4(u_m^2+\epsilon)} \left[u + \frac{u_m^2}{(\sqrt{u_m^2+\epsilon}+\sqrt{\epsilon})^2}\right]^2 & u_m < 0 \\ \frac{\sqrt{r_m^2+\epsilon}-\sqrt{\epsilon}}{(r_m-u_m)^2}(u - u_m)^2 & u_m > 0, \end{cases}
\end{aligned}$$

where in the last case r_m is the largest real root of the cubic equation $u^3+2u_mu^2+u_m^2u+4\epsilon u_m = 0$. (Hints: In majorizing the approximation to u_+, in each case assume that $q(u \mid u_m) = c(u-d)^2$. Choose c and d to give one or two tangency points.)

15. Implement and test one of the discriminant analysis algorithms that depend on quadratic majorization of hinge loss.

16. Nonnegative matrix factorization was introduced by Lee and Seung [174, 175] as an analog of principal components and vector quantization with applications in data compression and clustering. In mathematical terms, one approximates a matrix $\boldsymbol{U}$ with nonnegative entries u_{ij} by a product $\boldsymbol{VW}$ of two low-rank matrices with nonnegative entries v_{ij} and w_{ij}. If the entries u_{ij} are integers, then they can be viewed as realizations of independent Poisson random variables with means $\sum_k v_{ik}w_{kj}$. In this setting the loglikelihood is

$$L(\boldsymbol{V},\boldsymbol{W}) \;=\; \sum_i \sum_j \Big[u_{ij} \ln\Big(\sum_k v_{ik}w_{kj}\Big) - \sum_k v_{ik}w_{kj}\Big].$$

Maximization with respect to $\boldsymbol{V}$ and $\boldsymbol{W}$ should lead to a good factorization. Lee and Seung construct a block ascent algorithm that hinges on the minorization

$$\ln\Big(\sum_k v_{ik}w_{kj}\Big) \;\geq\; \sum_k \frac{a_{ikj}^n}{b_{ij}^n} \ln\Big(\frac{b_{ij}^n}{a_{ikj}^n} v_{ik}w_{kj}\Big),$$

where

$$a_{ikj}^n \;=\; v_{ik}^n w_{kj}^n, \qquad b_{ij}^n \;=\; \sum_k v_{ik}^n w_{kj}^n,$$

and n indicates the current iteration. Prove this minorization and derive the Lee-Seung algorithm with alternating multiplicative updates

$$v_{ik}^{n+1} \;=\; v_{ik}^n \frac{\sum_j u_{ij} \frac{w_{kj}^n}{b_{ij}^n}}{\sum_j w_{kj}^n}$$

and

$$w_{kj}^{n+1} \;=\; w_{kj}^n \frac{\sum_i u_{ij} \frac{v_{ik}^n}{b_{ij}^n}}{\sum_i v_{ik}^n}.$$

17. Continuing Problem 16, consider minimizing the squared Frobenius norm

$$\|\boldsymbol{U}-\boldsymbol{VW}\|_F^2 \;=\; \sum_i \sum_j \Big(u_{ij} - \sum_k v_{ik}w_{kj}\Big)^2.$$

Demonstrate the majorization

$$\Big(u_{ij} - \sum_k v_{ik}w_{kj}\Big)^2 \quad \leq \quad \sum_k \frac{a_{ikj}^n}{b_{ij}^n}\Big(u_{ij} - \frac{b_{ij}^n}{a_{ikj}^n}v_{ik}w_{kj}\Big)^2$$

based on the notation of Problem 16. Now derive the block descent algorithm with multiplicative updates

$$v_{ik}^{n+1} \quad = \quad v_{ik}^n \frac{\sum_j u_{ij}w_{kj}^n}{\sum_j b_{ij}^n w_{kj}^n}$$

and

$$w_{kj}^{n+1} \quad = \quad w_{kj}^n \frac{\sum_i u_{ij}v_{ik}^n}{\sum_i b_{ij}^n v_{ik}^n}.$$

18. In Problem 16 calculate the partial derivative

$$\frac{\partial}{\partial v_{il}} L(\boldsymbol{V}, \boldsymbol{W}) \quad = \quad \sum_j w_{lj}\Big(\frac{u_{ij}}{\sum_k v_{ik}w_{kj}} - 1\Big).$$

Show that the conditions $\min\{v_{il}, -\frac{\partial}{\partial v_{il}} L(\boldsymbol{V}, \boldsymbol{W})\} = 0$ for all pairs (i, l) are both necessary and sufficient for $\boldsymbol{V}$ to maximize $L(\boldsymbol{V}, \boldsymbol{W})$ when $\boldsymbol{W}$ is fixed. The same conditions apply in minimizing the criterion $\|\boldsymbol{U} - \boldsymbol{V}\boldsymbol{W}\|_F^2$ of Problem 17 with different partial derivatives.

19. In the matrix factorizations described in Problems 16 and 17, it may be worthwhile shrinking the estimates of the entries of $\boldsymbol{V}$ and $\boldsymbol{W}$ toward 0 [211]. Let λ and μ be positive constants, and consider the penalized objective functions

$$\begin{aligned} l(\boldsymbol{V}, \boldsymbol{W}) \quad &= \quad L(\boldsymbol{V}, \boldsymbol{W}) - \lambda \sum_i \sum_k v_{ik} - \mu \sum_k \sum_j w_{kj} \\ r(\boldsymbol{V}, \boldsymbol{W}) \quad &= \quad \|\boldsymbol{U} - \boldsymbol{V}\boldsymbol{W}\|_F^2 + \lambda \sum_i \sum_k v_{ik}^2 + \mu \sum_k \sum_j w_{kj}^2 \end{aligned}$$

with lasso and ridge penalties, respectively. Derive the block ascent updates

$$v_{ik}^{n+1} \quad = \quad v_{ik}^n \frac{\sum_j u_{ij} \frac{w_{kj}^n}{b_{ij}^n}}{\sum_j w_{kj}^n + \lambda}, \quad w_{kj}^{n+1} \quad = \quad w_{kj}^n \frac{\sum_i u_{ij} \frac{v_{ik}^n}{b_{ij}^n}}{\sum_i v_{ik}^n + \mu}$$

for $l(\boldsymbol{V}, \boldsymbol{W})$ and the block descent updates

$$v_{ik}^{n+1} \quad = \quad v_{ik}^n \frac{\sum_j u_{ij}w_{kj}^n}{\sum_j b_{ij}^n w_{kj}^n + \lambda v_{ik}^n}, \quad w_{kj}^{n+1} \quad = \quad w_{kj}^n \frac{\sum_i u_{ij}v_{ik}^n}{\sum_i b_{ij}^n v_{ik}^n + \mu w_{kj}^n}$$

for $r(\boldsymbol{V}, \boldsymbol{W})$. These updates maintain positivity. Shrinkage is obvious, with stronger shrinkage for the lasso penalty with small parameters.

20. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ be a random sample from a multivariate normal distribution on R^p. Example 6.5.7 demonstrates that the sample mean $\bar{\boldsymbol{y}}$ and sample variance matrix $\boldsymbol{S}$ are the maximum likelihood estimates of the theoretical mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Omega}$. The implicit assumption here is that $m \geq p$ and $\boldsymbol{S}$ is invertible. Unfortunately, $\boldsymbol{S}$ is singular whenever $m < p$. Furthermore, the entries of $\boldsymbol{S}$ typically have high variance when $m \geq p$. To avoid these problems, Levina et al. [177] pursue lasso penalized estimation of $\boldsymbol{\Omega}^{-1}$. If we assume that $\boldsymbol{\Omega}$ is invertible and let $\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{L}^*$ be its Cholesky decomposition, then $\boldsymbol{\Omega}^{-1} = (\boldsymbol{L}^*)^{-1}\boldsymbol{L}^{-1} = \boldsymbol{R}\boldsymbol{R}^*$ for the upper triangular matrix $\boldsymbol{R} = (r_{ij}) = (\boldsymbol{L}^*)^{-1}$. With the understanding $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$, show that the loglikelihood of the sample is

$$m \ln \det \boldsymbol{R} - \frac{m}{2} \operatorname{tr}(\boldsymbol{R}^* \boldsymbol{S} \boldsymbol{R}) \quad = \quad m \sum_i \ln r_{ii} - \frac{m}{2} \sum_j \boldsymbol{r}_j^* \boldsymbol{S} \boldsymbol{r}_j,$$

where $\boldsymbol{r}_j$ is column j of $\boldsymbol{R}$. In lasso penalized estimation of $\boldsymbol{R}$, we minimize the objective function

$$f(\boldsymbol{R}) \quad = \quad -m \sum_i \ln r_{ii} + \frac{m}{2} \sum_j \boldsymbol{r}_j^* \boldsymbol{S} \boldsymbol{r}_j + \lambda \sum_{j>i} |r_{ij}|.$$

The diagonal entries of $\boldsymbol{R}$ are not penalized because we want $\boldsymbol{R}$ to be invertible. Why is $f(\boldsymbol{R})$ a convex function? For $r_{ij} \neq 0$, show that

$$\begin{aligned} \frac{\partial}{\partial r_{ij}} f(\boldsymbol{R}) \quad &= \quad -1_{\{j=i\}} \frac{m}{r_{ii}} + m s_{ii} r_{ij} + m \sum_{k \neq i} s_{ik} r_{kj} \\ &\quad +1_{\{j \neq i\}} \begin{cases} \lambda & r_{ij} > 0 \\ -\lambda & r_{ij} < 0. \end{cases} \end{aligned}$$

Demonstrate that this leads to the cyclic coordinate descent update

$$\hat{r}_{ii} \quad = \quad \frac{-\sum_{k \neq i} s_{ik} r_{ki} + \sqrt{(\sum_{k \neq i} s_{ik} r_{ki})^2 + 4 s_{ii}}}{2 s_{ii}}.$$

Finally for $j \neq i$, demonstrate that the cyclic coordinate descent update chooses

$$\hat{r}_{ij} \quad = \quad -\frac{m \sum_{k \neq i} s_{ik} r_{kj} + \lambda}{m s_{ii}}$$

when this quantity is positive, it chooses

$$\hat{r}_{ij} \quad = \quad -\frac{m \sum_{k \neq i} s_{ik} r_{kj} - \lambda}{m s_{ii}}$$

when this second quantity is negative, and it chooses 0 otherwise. In organizing cyclic coordinate descent, it is helpful to retain and periodically update the sums $\sum_{k \neq i} s_{ik} r_{kj}$. The matrix $\boldsymbol{R}$ can be traversed column by column.

14
Convex Calculus

14.1 Introduction

Two generations of mathematicians have labored to extend the machinery of differential calculus to convex functions. For many purposes it is convenient to generalize the definition of a convex function $f(\boldsymbol{x})$ to include the possibility that $f(\boldsymbol{x}) = \infty$. This maneuver has the advantage of allowing one to enlarge the domain of a convex function $f(\boldsymbol{x})$ defined on a convex set $C \subset \mathsf{R}^n$ to all of R^n by the simple device of setting $f(\boldsymbol{x}) = \infty$ for $\boldsymbol{x} \notin C$. Many of the results for finite-valued convex functions generalize successfully in this setting. For instance, convex functions can still be characterized by their epigraphs and their satisfaction of Jensen's inequality.

The notion of the subdifferential $\partial f(\boldsymbol{x})$ of a convex function $f(\boldsymbol{x})$ has been a particularly fertile idea. This set consists of all vectors $\boldsymbol{g}$ satisfying the supporting hyperplane inequality $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{g}^*(\boldsymbol{y} - \boldsymbol{x})$ for all $\boldsymbol{y}$. These vectors $\boldsymbol{g}$ are called subgradients. If $f(\boldsymbol{x})$ is differentiable at $\boldsymbol{x}$, then $\partial f(\boldsymbol{x})$ reduces to the single vector $\nabla f(\boldsymbol{x})$. The subdifferential enjoys such familiar properties as

$$\begin{aligned} \partial[\alpha f(\boldsymbol{x})] &= \alpha\, \partial f(\boldsymbol{x}), \quad \alpha > 0 \\ \partial[f(\boldsymbol{x}) + g(\boldsymbol{x})] &= \partial f(\boldsymbol{x}) + \partial g(\boldsymbol{x}) \\ \partial[f \circ g(\boldsymbol{x})] &= dg(\boldsymbol{x})^* \partial f(\boldsymbol{y})|_{\boldsymbol{y}=g(\boldsymbol{x})} \end{aligned}$$

under the right hypotheses. Fermat's principle generalizes in the sense that $\boldsymbol{y}$ furnishes a minimum of $f(\boldsymbol{x})$ if and only if $\mathbf{0} \in \partial f(\boldsymbol{y})$. A version

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_14,

of the mean value theorem is true, and, properly interpreted, the Lagrange multiplier rule for a minimum remains valid. Perhaps more remarkable are the Fenchel conjugate and the formula for the subdifferential of the maximum of a finite collection of functions. The price of these successes is a theory more complicated than that encountered in classical calculus.

This chapter takes up the expository challenge of explaining these new concepts in the simplest possible terms. Fortunately, sacrificing generality for clarity does not mean losing sight of interesting applications. Convex calculus is an incredibly rich amalgam of ideas from analysis, linear algebra, and geometry. It has been instrumental in the construction of new algorithms for the solution of convex programs and their duals. Many readers will want to follow our brief account by pursuing the deeper treatises [13, 17, 131, 221, 226].

14.2 Notation

Although we allow ∞ for the value of a convex function, a function that is everywhere infinite is too boring to be of much interest. We will rule out such improper convex functions and the value $-\infty$ for a convex function. The convex set $\{\boldsymbol{x} \in \mathsf{R}^n : f(\boldsymbol{x}) < \infty\}$ is called the essential domain of $f(\boldsymbol{x})$ and abbreviated $\operatorname{dom}(f)$. We denote the closure of a set C by $\operatorname{cl} C$ and the convex hull of C by $\operatorname{conv} C$.

The notion of lower semicontinuity introduced in Sect. 2.6 turns out to be crucial in many convexity arguments. Lower semicontinuity is equivalent to the epigraph $\operatorname{epi}(f)$ being closed, and for this reason mathematicians call a lower semicontinuous function closed. Again, it makes sense to extend a finite closed function $f(\boldsymbol{x})$ with closed domain to all of R^n by setting $f(\boldsymbol{x}) = \infty$ outside $\operatorname{dom}(f)$. The fact that a closed convex function $f(\boldsymbol{x})$ has a closed convex epigraph permits application of the geometric separation property of convex sets described in Proposition 6.2.3. As an example of a closed convex function, consider

$$f(x) = \begin{cases} x \ln x - x & x > 0 \\ 0 & x = 0 \\ \infty & x < 0. \end{cases}$$

If on the one hand we redefine $f(0) < 0$, then $f(x)$ fails to be convex. If on the other hand we redefine $f(0) > 0$, then $f(x)$ fails to be closed.

14.3 Fenchel Conjugates

The Fenchel conjugate was defined in Example 1.2.6 for real-valued functions of a real argument. This transform, which is in some ways the

optimization analogue of the Fourier transform, has profound consequences in many branches of mathematics. The Fenchel conjugate generalizes to

$$f^\star(\boldsymbol{y}) \quad = \quad \sup_{\boldsymbol{x}} [\boldsymbol{y}^* \boldsymbol{x} - f(\boldsymbol{x})] \tag{14.1}$$

for functions $f(\boldsymbol{x})$ mapping R^n into $(-\infty, \infty]$. The conjugate $f^\star(\boldsymbol{y})$ is always closed and convex even when $f(\boldsymbol{x})$ is neither. Condition (j) in the next proposition rules out improper conjugate functions.

On first contact definition (14.1) is frankly a little mysterious. So too is the definition of the Fourier transform. Readers are advised to exercise patience and suspend their initial skepticism for several reasons. The Fenchel conjugate encodes the solutions to a family of convex optimization problems. It is one of the keys to understanding convex duality and serves as a device for calculating subdifferentials. Finally, the Fenchel biconjugate $f^{\star\star}(\boldsymbol{x})$ provides a practical way of convexifying $f(\boldsymbol{x})$. Indeed, the biconjugate has the geometric interpretation of falling below $f(\boldsymbol{x})$ and above any supporting hyperplane minorizing $f(\boldsymbol{x})$. This claim follows from our subsequent proof of the Fenchel-Moreau theorem and a double application of item (f) in the next proposition.

Proposition 14.3.1 *The Fenchel conjugate enjoys the following properties:*

(a) *If* $g(\boldsymbol{x}) = f(\boldsymbol{x} - \boldsymbol{v})$, *then* $g^\star(\boldsymbol{y}) = f^\star(\boldsymbol{y}) + \boldsymbol{v}^* \boldsymbol{y}$.

(b) *If* $g(\boldsymbol{x}) = f(\boldsymbol{x}) - \boldsymbol{v}^* \boldsymbol{x} - c$, *then* $g^\star(\boldsymbol{y}) = f^\star(\boldsymbol{y} + \boldsymbol{v}) + c$.

(c) *If* $g(\boldsymbol{x}) = \alpha f(\boldsymbol{x})$ *for* $\alpha > 0$, *then* $g^\star(\boldsymbol{y}) = \alpha f^\star(\alpha^{-1} \boldsymbol{y})$.

(d) *If* $g(\boldsymbol{x}) = f(\boldsymbol{M}\boldsymbol{x})$, *then* $g^\star(\boldsymbol{y}) = f^\star[(\boldsymbol{M}^{-1})^* \boldsymbol{y}]$ *for* $\boldsymbol{M}$ *invertible.*

(e) *If* $g(\boldsymbol{x}) = \sum_{i=1}^n f_i(x_i)$ *for* $\boldsymbol{x} \in \mathsf{R}^n$, *then* $g^\star(\boldsymbol{y}) = \sum_{i=1}^n f_i^\star(y_i)$.

(f) *If* $g(\boldsymbol{x}) \leq h(\boldsymbol{x})$ *for all* $\boldsymbol{x}$, *then* $g^\star(\boldsymbol{y}) \geq h^\star(\boldsymbol{y})$ *for all* $\boldsymbol{y}$.

(g) *For all* $\boldsymbol{x}$ *and* $\boldsymbol{y}$, $f^\star(\boldsymbol{y}) + f(\boldsymbol{x}) \geq \boldsymbol{y}^* \boldsymbol{x}$.

(h) *The conjugate* $f^\star(\boldsymbol{y})$ *is convex.*

(i) *The conjugate* $f^\star(\boldsymbol{y})$ *is closed.*

(j) *If* $f(\boldsymbol{x})$ *satisfies* $f(\boldsymbol{x}) \geq \boldsymbol{z}^* \boldsymbol{x} + c$ *for all* $\boldsymbol{x}$, *then* $f^\star(\boldsymbol{z}) \leq -c$.

(k) $f^\star(\boldsymbol{0}) = -\inf_{\boldsymbol{x}} f(\boldsymbol{x})$.

Proof: All of these claims are direct consequences of definition (14.1). For instance, claim (d) follows from

$$g^\star(\boldsymbol{y}) \quad = \quad \sup_{\boldsymbol{x}} [\boldsymbol{y}^* \boldsymbol{x} - f(\boldsymbol{M}\boldsymbol{x})] \quad = \quad \sup_{\boldsymbol{z}} [\boldsymbol{y}^* \boldsymbol{M}^{-1} \boldsymbol{z} - f(\boldsymbol{z})].$$

Claims (h) and (i) stem from the convexity and continuity of the affine functions $\boldsymbol{y} \mapsto \boldsymbol{y}^*\boldsymbol{x} - f(\boldsymbol{x})$ and the closure properties of convex and lower semicontinuous functions under suprema. Finally, claim (j) follows from the inequality $f^\star(\boldsymbol{z}) \leq \sup_{\boldsymbol{x}}[(\boldsymbol{z}-\boldsymbol{z})^*\boldsymbol{x} - \boldsymbol{c}] = -c$. ■

Example 14.3.1 *Conjugate of a Strictly Convex Quadratic*

For a well-behaved function $f(\boldsymbol{x})$, one can find $f^\star(\boldsymbol{y})$ by setting the gradient $\boldsymbol{y} - \nabla f(\boldsymbol{x})$ equal to $\boldsymbol{0}$ and solving for $\boldsymbol{x}$. For example, consider the quadratic $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}$ defined by the positive definite matrix $\boldsymbol{A}$. The gradient condition becomes $\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ with solution $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}$. This result gives the conjugate $f^\star(\boldsymbol{y}) = \frac{1}{2}\boldsymbol{y}^*\boldsymbol{A}^{-1}\boldsymbol{y}$. For the general convex quadratic

$$f(\boldsymbol{x}) \;=\; \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x} + c,$$

rule (b) of Proposition 14.3.1 implies

$$f^\star(\boldsymbol{y}) \;=\; \frac{1}{2}(\boldsymbol{y}-\boldsymbol{b})^*\boldsymbol{A}^{-1}(\boldsymbol{y}-\boldsymbol{b}) - c.$$

For instance, the univariate function $f(x) = \frac{1}{2}x^2$ is self-conjugate. According to rule (e) of Proposition 14.3.1, the multivariate function $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$ is also self-conjugate. Rule (g) is called the Fenchel-Young inequality. In the current example it amounts to

$$\frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \frac{1}{2}\boldsymbol{y}^*\boldsymbol{A}^{-1}\boldsymbol{y} \;\geq\; \boldsymbol{y}^*\boldsymbol{x},$$

a surprising result in its own right. ■

Example 14.3.2 *Entropy as a Fenchel Conjugate*

Consider the convex function $f(\boldsymbol{x}) = \ln(\sum_{j=1}^n e^{x_j})$. If $f^\star(\boldsymbol{y})$ is finite, then setting the gradient of $\boldsymbol{y}^*\boldsymbol{x} - f(\boldsymbol{x})$ with respect to $\boldsymbol{x}$ equal to $\boldsymbol{0}$ entails

$$y_i \;=\; \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}.$$

It follows that the y_i are positive and sum to 1. Furthermore,

$$f^\star(\boldsymbol{y}) \;=\; \sum_{i=1}^n y_i\Big[\ln y_i + \ln\Big(\sum_{j=1}^n e^{x_i}\Big)\Big] - \ln\Big(\sum_{j=1}^n e^{x_i}\Big) \;=\; \sum_{i=1}^n y_i \ln y_i.$$

With the understanding that $0 \ln 0 = 0$, the entropy formula

$$f^\star(\boldsymbol{y}) \;=\; \sum_{i=1}^n y_i \ln y_i$$

remains valid when $\boldsymbol{y}$ occurs on the boundary of the unit simplex. Indeed, if $\boldsymbol{y}$ belongs to the unit simplex but $y_i = 0$, then we send x_i to $-\infty$ and reduce calculation of the conjugate to the setting $\boldsymbol{y} \in \mathsf{R}^{n-1}$. If $y_i < 0$, then the sequence $\boldsymbol{x}_k = -k\boldsymbol{e}_i$ gives

$$\lim_{k\to\infty} \boldsymbol{y}^*\boldsymbol{x}_k - f(\boldsymbol{x}_k) \quad = \quad \lim_{k\to\infty} \ln\left(\frac{e^{-ky_i}}{n-1+e^{-k}}\right) \quad = \quad \infty.$$

Similarly, if all $y_i \geq 0$ but $\sum_{j=1}^n y_j \neq 1$, then one of the two sequences $\boldsymbol{x}_k = \pm k\mathbf{1}$ compels the same conclusion $f^\star(\boldsymbol{y}) = \infty$. ■

Examples 1.2.6 and 14.3.1 obey the curious rule $f^{\star\star}(\boldsymbol{x}) = f(\boldsymbol{x})$. This duality relation is more widely true. An affine function $f(\boldsymbol{x}) = \boldsymbol{z}^*\boldsymbol{x} + c$ provides another example. This fact follows readily from the form

$$f^\star(\boldsymbol{y}) \quad = \quad \begin{cases} -c & \boldsymbol{y} = \boldsymbol{z} \\ \infty & \boldsymbol{y} \neq \boldsymbol{z} \end{cases}$$

of the conjugate. One cannot expect duality to hold for all functions because Proposition 14.3.1 requires $f^{\star\star}(\boldsymbol{x})$ to be convex and closed. Remarkably, the combination of these two conditions is both necessary and sufficient for proper functions.

Proposition 14.3.2 (Fenchel-Moreau) *A proper function $f(\boldsymbol{x})$ from R^n to $(-\infty, \infty]$ satisfies the duality relation $f^{\star\star}(\boldsymbol{x}) = f(\boldsymbol{x})$ for all $\boldsymbol{x}$ if and only if it is closed and convex.*

Proof: Suppose the duality relation holds. Being the conjugate of a conjugate, $f(\boldsymbol{x}) = f^{\star\star}(\boldsymbol{x})$ is closed and convex. This proves that the stated conditions are necessary for duality.

To demonstrate the converse, first note that the Fenchel-Young inequality $f(\boldsymbol{x}) \geq \boldsymbol{y}^*\boldsymbol{x} - f^\star(\boldsymbol{y})$ implies

$$f(\boldsymbol{x}) \quad \geq \quad \sup_{\boldsymbol{y}} [\boldsymbol{y}^*\boldsymbol{x} - f^\star(\boldsymbol{y})] \quad = \quad f^{\star\star}(\boldsymbol{x}). \tag{14.2}$$

In other words, the epigraph $\operatorname{epi}(f^{\star\star})$ of $f^{\star\star}(\boldsymbol{x})$ contains the epigraph $\operatorname{epi}(f)$ of $f(\boldsymbol{x})$. Verifying the reverse containment $\operatorname{epi}(f^{\star\star}) \subset \operatorname{epi}(f)$ proves the converse of the proposition. Our general strategy for establishing containment is to exploit the separation properties of closed convex sets. As already mentioned, the convexity of $f(\boldsymbol{x})$ entails the convexity of $\operatorname{epi}(f)$, and the lower semicontinuity of $f(\boldsymbol{x})$ entails the closedness of $\operatorname{epi}(f)$.

We first show that $f(\boldsymbol{x})$ dominates some affine function and hence that the conjugate function $f^\star(\boldsymbol{y})$ is proper. Suppose $\boldsymbol{x}_0$ satisfies $f(\boldsymbol{x}_0) < \infty$. Given that $\operatorname{epi}(f)$ is closed and convex, we can separate it from the exterior point $[\boldsymbol{x}_0, f(\boldsymbol{x}_0) - 1]$ by a hyperplane. Thus, there exists a vector $\boldsymbol{v}$ and scalars η and ν such that

$$\boldsymbol{v}^*\boldsymbol{x} + \eta r \quad \geq \quad \nu \quad > \quad \boldsymbol{v}^*\boldsymbol{x}_0 + \eta[f(\boldsymbol{x}_0) - 1] \tag{14.3}$$

for all $\boldsymbol{x}$ and $r \geq f(\boldsymbol{x})$. Sending r to ∞ demonstrates that $\eta \geq 0$. Setting $\boldsymbol{x} = \boldsymbol{x}_0$ rules out the possibility $\eta = 0$. Finally, dividing inequality (14.3) by $\eta > 0$, replacing r by $f(\boldsymbol{x})$, and rearranging the result yield an affine function $\boldsymbol{z}^*\boldsymbol{x} + c$ positioned below $f(\boldsymbol{x})$.

Now suppose $(\boldsymbol{y}, \beta)$ is in $\text{epi}(f^{\star\star})$ but not in $\text{epi}(f)$. Proposition 6.2.3 guarantees the existence of a vector-scalar pair $(\boldsymbol{v}, \gamma)$ and a constant $\epsilon > 0$ such that

$$\boldsymbol{v}^*\boldsymbol{y} + \gamma\beta \quad \leq \quad \boldsymbol{v}^*\boldsymbol{x} + \gamma\alpha - \epsilon$$

for all $(\boldsymbol{x}, \alpha) \in \text{epi}(f)$. Sending α to ∞ shows that $\gamma \geq 0$. If $\gamma > 0$, then

$$g(\boldsymbol{x}) \quad = \quad \gamma^{-1}\boldsymbol{v}^*(\boldsymbol{y} - \boldsymbol{x}) + \beta + \gamma^{-1}\epsilon \quad \leq \quad \alpha$$

for all $\alpha \geq f(\boldsymbol{x})$. Hence, $g(\boldsymbol{x})$ is an affine function positioned below $f(\boldsymbol{x})$, and a double application of part (f) of Proposition 14.3.1 implies

$$f^{\star\star}(\boldsymbol{y}) \quad \geq \quad g^{\star\star}(\boldsymbol{y}) \quad = \quad g(\boldsymbol{y}) \quad > \quad \beta,$$

contradicting the choice of $(\boldsymbol{y}, \beta) \in \text{epi}(f^{\star\star})$.

Completing the proof now requires eliminating the possibility $\gamma = 0$. If we multiply the inequality $\boldsymbol{v}^*\boldsymbol{y} - \boldsymbol{v}^*\boldsymbol{x} + \epsilon \leq 0$ by $\delta > 0$ and add it to the previous inequality $\boldsymbol{z}^*\boldsymbol{x} + c \leq f(\boldsymbol{x})$, then we arrive at

$$h(\boldsymbol{x}) \quad = \quad \boldsymbol{z}^*\boldsymbol{x} + c + \delta\boldsymbol{v}^*(\boldsymbol{y} - \boldsymbol{x}) + \delta\epsilon \quad \leq \quad f(\boldsymbol{x}).$$

The conclusion

$$f^{\star\star}(\boldsymbol{y}) \quad \geq \quad h^{\star\star}(\boldsymbol{y}) \quad = \quad h(\boldsymbol{y}) \quad = \quad \boldsymbol{z}^*\boldsymbol{y} + c + \delta\epsilon$$

for all $\delta > 0$ can only be true if $f^{\star\star}(\boldsymbol{y}) = \infty$, which is also incompatible with $(\boldsymbol{y}, \beta) \in \text{epi}(f^{\star\star})$. ■

The left panel of Fig. 14.1 illustrates the relationship between a function $f(x)$ on the real line and its Fenchel conjugate $f^\star(y)$. According to the Fenchel-Young inequality, the line with slope y and intercept $-f^\star(y)$ falls below $f(x)$. The curve and the line intersect when $y = f'(x)$. The right panel of Fig. 14.1 shows that the biconjugate $f^{\star\star}(y)$ is the greatest convex function lying below $f(x)$. The biconjugate is formed by taking the pointwise supremum of the supporting lines.

Example 14.3.3 *Perspective of a Function*

Let $f(\boldsymbol{x})$ be a closed convex function. The perspective of $f(\boldsymbol{x})$ is the function $g(\boldsymbol{x}, t) = tf(t^{-1}\boldsymbol{x})$ defined for $t > 0$. On this domain $g(\boldsymbol{x}, t)$ is closed and convex owing to the representation

$$g(\boldsymbol{x}, t) \quad = \quad t \sup_{\boldsymbol{y}} [t^{-1}\boldsymbol{x}^*\boldsymbol{y} - f^\star(\boldsymbol{y})] \quad = \quad \sup_{\boldsymbol{y}} [\boldsymbol{x}^*\boldsymbol{y} - tf^\star(\boldsymbol{y})]$$

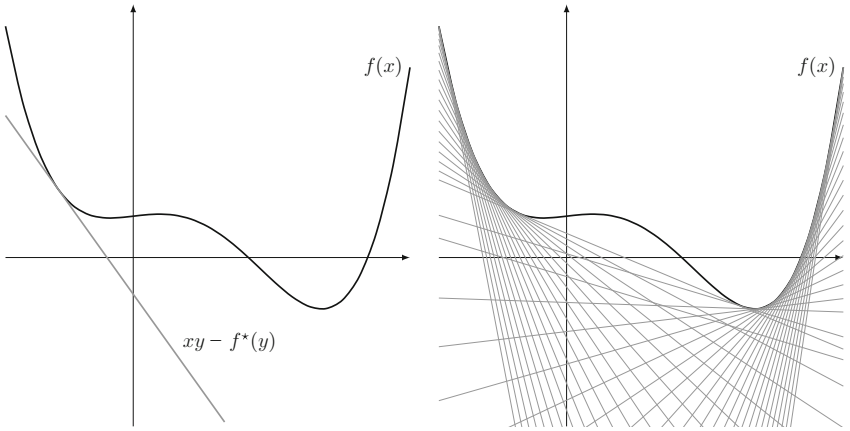

FIGURE 14.1. *Left panel*: The Fenchel-Young inequality for $f(x)$ and its conjugate $f^\star(y)$. *Right panel*: The envelope of supporting lines defined by $f^\star(y)$

and the linearity of the map $(\boldsymbol{x}, t) \mapsto \boldsymbol{x}^*\boldsymbol{y} - tf^\star(\boldsymbol{y})$. For instance, the choice $f(\boldsymbol{x}) = \boldsymbol{x}^*\boldsymbol{M}\boldsymbol{x}$ for $\boldsymbol{M}$ positive semidefinite yields the convexity of $t^{-1}\boldsymbol{x}^*\boldsymbol{M}\boldsymbol{x}$. The choice $f(x) = -\ln x$ shows that the relative entropy $g(x,t) = t\ln t - t\ln x$ is convex. Finally, the function

$$g(\boldsymbol{x}) \quad = \quad (\boldsymbol{c}^*\boldsymbol{x} + d) f\left(\frac{\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}}{\boldsymbol{c}^*\boldsymbol{x} + d}\right)$$

is closed and convex on the domain $\boldsymbol{c}^*\boldsymbol{x} + d > 0$ whenever the function $f(\boldsymbol{x})$ is closed and convex. ■

Example 14.3.4 *Indicator and Support Functions*

Every set C can be represented by its indicator function

$$\delta_C(\boldsymbol{x}) \quad = \quad \begin{cases} 0 & \boldsymbol{x} \in C \\ \infty & \boldsymbol{x} \notin C. \end{cases}$$

If C is closed and convex, then it is easy to check that $\delta_C(\boldsymbol{x})$ is a closed convex function. One reason for making the substitution of ∞ for 0 and 0 for 1 in defining an indicator is that it simplifies the Fenchel conjugate

$$\delta_C^\star(\boldsymbol{y}) \quad = \quad \sup_{\boldsymbol{x}} [\boldsymbol{y}^*\boldsymbol{x} - \delta_C(\boldsymbol{x})] \quad = \quad \sup_{\boldsymbol{x} \in C} \boldsymbol{y}^*\boldsymbol{x}.$$

The function $\delta_C^\star(\boldsymbol{y})$ is called the support function of C. Proposition 14.3.2 implies that the Fenchel biconjugate $\delta_C^{\star\star}(\boldsymbol{z})$ equals $\delta_C(\boldsymbol{z})$.

It turns out that support functions with full essential domains are the same as sublinear functions with full essential domains. A function $h(\boldsymbol{u})$ is said to be sublinear whenever

$$h(\alpha\boldsymbol{u} + \beta\boldsymbol{v}) \quad \leq \quad \alpha h(\boldsymbol{u}) + \beta h(\boldsymbol{v})$$

holds for all points $\boldsymbol{u}$ and $\boldsymbol{v}$ and nonnegative scalars α and β. Sublinearity is an amalgam of homogeneity, $h(\lambda \boldsymbol{v}) = \lambda h(\boldsymbol{v})$ for $\lambda > 0$, and convexity, $h[\lambda \boldsymbol{u} + (1-\lambda)\boldsymbol{v}] \leq \lambda h(\boldsymbol{u}) + (1-\lambda)h(\boldsymbol{v})$ for $\lambda \in [0,1]$. One can easily check that a support function is sublinear. To prove the converse, we first demonstrate that the conjugate $h^\star(\boldsymbol{y})$ of a sublinear function is an indicator function. Indeed, the identity

$$\begin{aligned} \lambda h^\star(\boldsymbol{y}) &= \lambda \sup_{\boldsymbol{x}} [\boldsymbol{y}^* \boldsymbol{x} - h(\boldsymbol{x})] \\ &= \sup_{\boldsymbol{x}} [\boldsymbol{y}^* (\lambda \boldsymbol{x}) - h(\lambda \boldsymbol{x})] \\ &= \sup_{\boldsymbol{x}} [\boldsymbol{y}^* \boldsymbol{x} - h(\boldsymbol{x})] \\ &= h^\star(\boldsymbol{y}) \end{aligned}$$

compels $h^\star(\boldsymbol{y})$ to equal 0 or ∞. When $h(\boldsymbol{x})$ is finite-valued, Proposition 6.4.1 requires it to be continuous and consequently closed. Thus, Fenchel duality implies that $h(\boldsymbol{x})$ equals the support function of the closed convex set $C = \{\boldsymbol{y} \in \mathsf{R}^n : h^\star(\boldsymbol{y}) = 0\}$.

The support function $h(\boldsymbol{u})$ of a closed convex set C is finite valued if and only if C is bounded. Boundedness is clearly sufficient to guarantee that $h(\boldsymbol{u})$ is finite valued. To show that boundedness is necessary as well, suppose C is unbounded. Then part (c) of Problem 5 of Chap. 6 says that C contains a ray $\{\boldsymbol{u} + t\boldsymbol{v} : t \in [0, \infty)\}$. This forces $h(\boldsymbol{v}) = \sup_{\boldsymbol{x} \in C} \boldsymbol{v}^* \boldsymbol{x}$ to be infinite. ■

Example 14.3.5 *Vector Dual Norms*

If C equals the closed unit ball $B = \{\boldsymbol{x} : \|\boldsymbol{x}\|_\dagger \leq 1\}$ associated with a norm $\|\boldsymbol{x}\|_\dagger$ on R^n, then the support function $\|\boldsymbol{y}\|_\star = \delta_B^\star(\boldsymbol{y}) = \sup_{\boldsymbol{x} \in B} \boldsymbol{y}^* \boldsymbol{x}$ also qualifies as a norm. Verification of the norm properties for the dual norm $\|\boldsymbol{y}\|_\star$ is left to the reader as Problem 13. For a sublinear function such as $\|\boldsymbol{y}\|_\star$, the only things to check are that whether it is nonnegative and vanishes if and only if $\boldsymbol{x} = \boldsymbol{0}$. In defining $\|\boldsymbol{y}\|_\star$ one can clearly confine $\boldsymbol{x}$ to the boundary of B. The generalized Cauchy-Schwarz inequality

$$\boldsymbol{y}^* \boldsymbol{x} \leq \|\boldsymbol{y}\|_\star \|\boldsymbol{x}\|_\dagger \tag{14.4}$$

follows directly from this observation. Equality in the generalized Cauchy-Schwarz inequality is attained for some $\boldsymbol{x}$ on the boundary of B because the maximum of a continuous function over a compact set is attained.

There are many concrete examples of norms and their duals. The dual norm of $\|\boldsymbol{x}\|_1$ is $\|\boldsymbol{x}\|_\infty$ and vice versa. For $p^{-1} + q^{-1} = 1$, Example 6.6.3 and Problem 21 of Chap. 5 show that the norms $\|\boldsymbol{x}\|_p$ and $\|\boldsymbol{y}\|_q$ constitute another dual pair, with the generalized Cauchy-Schwarz inequality reducing to Hölder's inequality. Of course, the Euclidean norm $\|\boldsymbol{x}\| = \|\boldsymbol{x}\|_2$ is its own dual.

These examples are not accidents. The primary reason for calling $\|\boldsymbol{y}\|_\star$ a dual norm is that taking the dual of the dual gives us back the original norm $\|\boldsymbol{x}\|_\dagger$. One way of deducing this fact is to construct the Fenchel conjugate of the convex function $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_\dagger^2$. In view of the generalized Cauchy-Schwarz inequality and the inequality $(\|\boldsymbol{x}\|_\dagger - \|\boldsymbol{y}\|_\star)^2 \geq 0$, we have

$$\boldsymbol{y}^*\boldsymbol{x} - \frac{1}{2}\|\boldsymbol{x}\|_\dagger^2 \quad \leq \quad \|\boldsymbol{y}\|_\star\|\boldsymbol{x}\|_\dagger - \frac{1}{2}\|\boldsymbol{x}\|_\dagger^2 \quad \leq \quad \frac{1}{2}\|\boldsymbol{y}\|_\star^2.$$

On the other hand, suppose we choose a vector $\boldsymbol{z}$ with $\|\boldsymbol{z}\|_\dagger = 1$ such that equality is attained in the generalized Cauchy-Schwarz inequality. Then for any scalar $s > 0$, we find

$$\boldsymbol{y}^*(s\boldsymbol{z}) - \frac{1}{2}\|s\boldsymbol{z}\|_\dagger^2 \quad = \quad s\|\boldsymbol{y}\|_\star\|\boldsymbol{z}\|_\dagger - \frac{s^2}{2}\|\boldsymbol{z}\|_\dagger^2.$$

If we take $s = \|\boldsymbol{y}\|_\star$, then this equality gives

$$\boldsymbol{y}^*(s\boldsymbol{z}) - \frac{1}{2}\|s\boldsymbol{z}\|_\dagger^2 \quad = \quad \frac{1}{2}\|\boldsymbol{y}\|_\star^2.$$

In other words, $f^\star(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}\|_\star^2$. Taking the conjugate of $f^\star(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}\|_\star^2$ yields $f^{\star\star}(\boldsymbol{x}) = f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_\dagger^2$. Thus, the original norm $\|\boldsymbol{x}\|_\dagger$ is dual to the dual norm $\|\boldsymbol{y}\|_\star$. ■

Example 14.3.6 *Matrix Dual Norms*

One can also define dual norms of matrix norms using the Frobenius inner product $\langle \boldsymbol{Y}, \boldsymbol{X}\rangle = \operatorname{tr}(\boldsymbol{Y}^*\boldsymbol{X})$. Under this matrix inner product, the Frobenius matrix norm $\|\boldsymbol{X}\|_F$ is self-dual. The easiest way to deduce this fact is to observe that the Frobenius norm can be calculated by stacking the columns of $\boldsymbol{X}$ to form a vector and then taking the Euclidean norm of the vector. Column stacking is clearly compatible with the exchange of the Frobenius inner product for the Euclidean inner product.

Under the Frobenius inner product, calculation of the dual norm of the matrix spectral norm is more subtle. The most illuminating approach takes a detour through Fan's inequality and the singular value decomposition (svd) covered in Appendices A.4 and A.5. Any matrix $\boldsymbol{X}$ has an svd representation $\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{Q}^*$, where $\boldsymbol{P}$ and $\boldsymbol{Q}$ are orthogonal matrices and $\boldsymbol{\Sigma}$ is a diagonal matrix with nonnegative entries σ_i arranged in decreasing order along its diagonal. The columns of $\boldsymbol{P}$ and $\boldsymbol{Q}$ are referred to as singular vectors and the diagonal entries of $\boldsymbol{\Sigma}$ as singular values. The svd immediately yields the spectral decompositions $\boldsymbol{X}\boldsymbol{X}^* = \boldsymbol{P}\boldsymbol{\Sigma}^2\boldsymbol{P}^*$ and $\boldsymbol{X}^*\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{\Sigma}^2\boldsymbol{Q}^*$ and consequently the spectral norm $\|\boldsymbol{X}\| = \sigma_1$. If $\boldsymbol{Y}$ has svd $\boldsymbol{R}\boldsymbol{\Omega}\boldsymbol{S}^*$ with $\boldsymbol{\Omega} = \operatorname{diag}(\omega_i)$, then equality is attained in Fan's inequality

$$\operatorname{tr}(\boldsymbol{Y}^*\boldsymbol{X}) \quad \leq \quad \sum_i \omega_i\sigma_i$$

when $\boldsymbol{R} = \boldsymbol{P}$ and $\boldsymbol{S} = \boldsymbol{Q}$. The matrix $\boldsymbol{X}$ with $\|\boldsymbol{X}\| = \sigma_1 \leq 1$ giving the maximum value of $\operatorname{tr}(\boldsymbol{Y}^*\boldsymbol{X})$ has the same singular vectors as $\boldsymbol{Y}$ and the singular values $\sigma_i = 1$ for $\omega_i > 0$. It follows that the dual norm of the spectral norm equals

$$\|\boldsymbol{Y}\|_\star \;=\; \sum_i \omega_i.$$

This dual norm is also called the nuclear norm or the trace norm. ■

Example 14.3.7 *Cones and Polar Cones*

If C is a cone, then the Fenchel conjugate of its indicator function $\delta_C(\boldsymbol{x})$ turns out to be the indicator function of the polar cone

$$C^\circ \;=\; \{\boldsymbol{y} : \boldsymbol{y}^*\boldsymbol{x} \leq 0, \;\; \forall \boldsymbol{x} \in C\}.$$

This assertion follows from the limits

$$\begin{aligned}
\lim_{c \downarrow 0} \boldsymbol{y}^*(c\boldsymbol{x}) &= 0 \\
\lim_{c \uparrow \infty} \boldsymbol{y}^*(c\boldsymbol{x}) &= \begin{cases} \infty & \boldsymbol{y}^*\boldsymbol{x} > 0 \\ 0 & \boldsymbol{y}^*\boldsymbol{x} = 0 \\ -\infty & \boldsymbol{y}^*\boldsymbol{x} < 0 \end{cases}
\end{aligned}$$

for any $\boldsymbol{x} \in C$. Although C may be neither convex nor closed, its polar C° is always both. The duality relation $\delta_C^{\star\star}(\boldsymbol{x}) = \delta_C(\boldsymbol{x})$ for a closed convex cone C is equivalent to the set relation $C^{\circ\circ} = C$. Notice the analogy here to the duality relation $S^{\perp\perp} = S$ for subspaces under the orthogonal complement operator $\perp$.

As a concrete example, let us calculate the polar cone of the set S_+^n of $n \times n$ positive semidefinite matrices under the Frobenius inner product $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \operatorname{tr}(\boldsymbol{A}\boldsymbol{B})$. If $\boldsymbol{A} \in S_+^n$ has eigenvalues $\lambda_1, \ldots, \lambda_n$ with corresponding unit eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$, then

$$\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) \;=\; \operatorname{tr}\Big(\sum_{i=1}^n \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^* \boldsymbol{B}\Big) \;=\; \sum_{i=1}^n \lambda_i \boldsymbol{u}_i^* \boldsymbol{B} \boldsymbol{u}_i.$$

If $\boldsymbol{B} \in -S_+^n$, then it is clear that $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) \leq 0$. Thus, the polar cone contains $-S_+^n$. Conversely, suppose $\boldsymbol{B}$ is in the polar cone of S_+^n, and choose $\boldsymbol{A} = \boldsymbol{v}\boldsymbol{v}^*$ for some nontrivial vector $\boldsymbol{v}$. The inequality $\operatorname{tr}(\boldsymbol{v}\boldsymbol{v}^*\boldsymbol{B}) = \boldsymbol{v}^*\boldsymbol{B}\boldsymbol{v} \leq 0$ for all such $\boldsymbol{v}$ implies that $\boldsymbol{B} \in -S_+^n$. Thus, the polar cone of S_+^n equals $-S_+^n$. ■

Example 14.3.8 *Log Determinant*

Minimization of the objective function (4.18) featured in Example 4.7.6 can be rephrased as calculating the Fenchel conjugate

$$g^\star(\boldsymbol{N}) \;=\; \sup_{\boldsymbol{P}} \left[\operatorname{tr}(\boldsymbol{N}\boldsymbol{P}) + c \ln \det \boldsymbol{P}\right]$$

for c positive and $\boldsymbol{P}$ an $n \times n$ positive definite matrix. The essential domain of $g^\star(\boldsymbol{N})$ is the set of negative definite matrices. Indeed, if $\boldsymbol{N}$ falls outside this domain, then it possesses a unit eigenvector $\boldsymbol{v}$ with nonnegative eigenvalue λ. The choice $\boldsymbol{P} = \boldsymbol{I} + s\boldsymbol{v}\boldsymbol{v}^*$ has eigenvalues 1 (multiplicity $n-1$) and $1+s$ (multiplicity 1). For $s > 0$ we calculate

$$\begin{aligned} \operatorname{tr}(\boldsymbol{N}\boldsymbol{P}) + c \ln \det \boldsymbol{P} &= \operatorname{tr} \boldsymbol{N} + s\lambda + c \ln \det(\boldsymbol{I} + s\boldsymbol{v}\boldsymbol{v}^*) \\ &= \operatorname{tr} \boldsymbol{N} + s\lambda + c \ln(1+s), \end{aligned}$$

which tends to ∞ as s tends to ∞. When $\boldsymbol{N}$ is negative definite, our previous calculations gave the gradient $\boldsymbol{N} + c\boldsymbol{P}^{-1}$ of $\operatorname{tr}(\boldsymbol{N}\boldsymbol{P}) + c \ln \det \boldsymbol{P}$. Setting the gradient to $\boldsymbol{0}$ yields $\boldsymbol{P} = -c\boldsymbol{N}^{-1}$ and $g^\star(\boldsymbol{N}) = -cn - c \ln \det[-c^{-1}\boldsymbol{N}]$. Because a Fenchel conjugate is convex, this line of argument establishes the log-concavity of $\det \boldsymbol{P}$ for $\boldsymbol{P}$ positive definite. Example 6.3.12 presents a different proof of this fact. ■

14.4 Subdifferentials

Convex calculus revolves around the ideas of forward directional derivatives and supporting hyperplanes. At this juncture the reader may want to review Sect. 6.4 on the former topic. Appendix A.6 develops the idea of a semidifferential, the single most fruitful generalization of forward directional derivatives to date. For the sake of brevity henceforth, we will drop the adjective forward and refer to forward directional derivatives simply as directional derivatives.

Consider the absolute value function $f(x) = |x|$. At the point $x = 0$, the derivative $f'(x)$ does not exist. However, the directional derivatives $d_v f(0) = |v|$ are all well defined. Furthermore, the supporting hyperplane inequality $f(x) \geq f(0) + gx$ is valid for all x and all g with $|g| \leq 1$. The set $\partial f(0) = \{g : |g| \leq 1\}$ is called the subdifferential of $f(x)$ at $x = 0$. The notion of subdifferential is hardly limited to functions defined on the real line. Consider a convex function $f(\boldsymbol{x}) : \mathsf{R}^n \mapsto (-\infty, \infty]$. By convention the subdifferential $\partial f(\boldsymbol{x})$ is empty for $\boldsymbol{x} \notin \operatorname{dom}(f)$. For $\boldsymbol{x} \in \operatorname{dom}(f)$, the subdifferential $\partial f(\boldsymbol{x})$ is the set of vectors $\boldsymbol{g}$ in R^n such that the supporting hyperplane inequality

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{g}^*(\boldsymbol{y} - \boldsymbol{x})$$

is valid for all $\boldsymbol{y}$. An element $\boldsymbol{g}$ of the subdifferential is called a subgradient. If $f(\boldsymbol{x})$ is differentiable at $\boldsymbol{x}$, then as we prove later, its subdifferential collapses to its gradient.

For a nontrivial example, consider $f(\boldsymbol{x}) = \max\{x_1, x_2\}$ with domain R^2. Off the diagonal $D = \{\boldsymbol{x} \in \mathsf{R}^2 : x_1 = x_2\}$, the convex function $f(\boldsymbol{x})$ is differentiable. On D the subdifferential $\partial f(\boldsymbol{x})$ equals the unit simplex

$S = \{\boldsymbol{w} \in \mathsf{R}^2 : w_1 + w_2 = 1,\ w_1 \geq 0,\ w_2 \geq 0\}$. Indeed for points $\boldsymbol{x} \in D$ and $\boldsymbol{w} \in S$, one can easily demonstrate that

$$\max\{y_1, y_2\} \quad \geq \quad \max\{x_1, x_2\} + w_1(y_1 - x_1) + w_2(y_2 - x_2)$$

is valid for all $\boldsymbol{y}$ by closely examining the two extreme cases $\boldsymbol{w} = (1,0)^*$ and $\boldsymbol{w} = (0,1)^*$. Conversely, we will prove later that any point $\boldsymbol{w}$ outside S fails the supporting hyperplane test for some $\boldsymbol{y}$. The directional derivative $d_{\boldsymbol{v}} f(\boldsymbol{x})$ equals $(1,0)\boldsymbol{v} = v_1$ for $x_1 > x_2$ and $(0,1)\boldsymbol{v} = v_2$ for $x_2 > x_1$. Example 4.4.4 shows that the directional derivative is $d_{\boldsymbol{v}} f(\boldsymbol{x}) = \max\{v_1, v_2\}$ on D. It is no accident that

$$d_{\boldsymbol{v}} f(\boldsymbol{x}) \quad = \quad \max\{\boldsymbol{w}^* \boldsymbol{v} : \boldsymbol{w} \in \partial f(\boldsymbol{x}) = S\} \tag{14.5}$$

on D. Indeed, the relationship (14.5) is generally true.

As a prelude to proving this fact, let us study the directional derivative $d_{\boldsymbol{v}} f(\boldsymbol{x})$ more thoroughly. For $f(\boldsymbol{x})$ convex and $\boldsymbol{x}$ an interior point of $\mathrm{dom}(f)$, one can argue that $d_{\boldsymbol{v}} f(\boldsymbol{x})$ exists and is finite for all $\boldsymbol{v}$ because any line segment starting at $\boldsymbol{x}$ can be extended backwards and remain in $\mathrm{dom}(f)$. Given that $\boldsymbol{x}$ is internal to the segment, the results of Sect. 6.4 apply. In addition, one can show that $d_{\boldsymbol{v}} f(\boldsymbol{x})$ is sublinear in its argument $\boldsymbol{v}$. See Example 14.3.4 for the definition of sublinearity. A directional derivative $d_{\boldsymbol{v}} f(\boldsymbol{x})$ is obviously homogeneous in $\boldsymbol{v}$. Because

$$d_{\boldsymbol{v}} f(\boldsymbol{x}) \quad = \quad \lim_{t \downarrow 0} \frac{f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x})}{t}$$

is a pointwise limit of its difference quotients, and these are convex functions of $\boldsymbol{v}$, $d_{\boldsymbol{v}} f(\boldsymbol{x})$ is also convex. The monotonicity of the difference quotient implies that

$$f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x}) \quad \geq \quad t d_{\boldsymbol{v}} f(\boldsymbol{x}).$$

Any vector $\boldsymbol{g}$ satisfying $\boldsymbol{g}^* \boldsymbol{v} \leq d_{\boldsymbol{v}} f(\boldsymbol{x})$ for all $\boldsymbol{v}$ therefore acts as a subgradient. Conversely, any subgradient $\boldsymbol{g}$ must satisfy $\boldsymbol{g}^* \boldsymbol{v} \leq d_{\boldsymbol{v}} f(\boldsymbol{x})$ for all $\boldsymbol{v}$. Hence, we have the relation

$$\partial f(\boldsymbol{x}) \quad = \quad \{\boldsymbol{g} : \boldsymbol{g}^* \boldsymbol{v} \leq d_{\boldsymbol{v}} f(\boldsymbol{x}) \text{ for all } \boldsymbol{v}\}. \tag{14.6}$$

Despite this interesting logical equivalence, the question remains of whether any subgradient $\boldsymbol{g}$ satisfies $\boldsymbol{g}^* \boldsymbol{v} = d_{\boldsymbol{v}} f(\boldsymbol{x})$ for a particular direction $\boldsymbol{v}$.

In constructing a subgradient $\boldsymbol{g}$ with $\boldsymbol{g}^* \boldsymbol{v} = d_{\boldsymbol{v}} f(\boldsymbol{x})$, we view the map $\boldsymbol{u} \mapsto \boldsymbol{g}^* \boldsymbol{u}$ as a linear functional $\ell(\boldsymbol{u})$. Initially $\ell(\boldsymbol{u})$ is defined for the vector $\boldsymbol{v}$ of interest by the requirement $\ell(\boldsymbol{v}) = d_{\boldsymbol{v}} f(\boldsymbol{x})$. We now show how to extend $\ell(\boldsymbol{u})$ to all of R^n. By homogeneity,

$$\ell(\lambda \boldsymbol{v}) \quad = \quad \lambda \ell(\boldsymbol{v}) \quad = \quad d_{\lambda \boldsymbol{v}} f(\boldsymbol{x})$$

for $\lambda \geq 0$. This defines $\ell(\boldsymbol{u})$ on the ray $\{\lambda \boldsymbol{v} : \lambda \geq 0\}$. For $\lambda < 0$ we continue to define $\ell(\lambda \boldsymbol{v})$ by $\lambda \ell(\boldsymbol{v})$. Jensen's inequality

$$0 \quad = \quad d_{\boldsymbol{0}} f(\boldsymbol{x}) \leq \frac{1}{2} d_{-\boldsymbol{v}} f(\boldsymbol{x}) + \frac{1}{2} d_{\boldsymbol{v}} f(\boldsymbol{x})$$

then implies that

$$\ell(-\boldsymbol{v}) \quad = \quad -d_{\boldsymbol{v}} f(\boldsymbol{x}) \quad \leq \quad d_{-\boldsymbol{v}} f(\boldsymbol{x})$$

and assures us that $\ell(\boldsymbol{u})$ is dominated by $d_{\boldsymbol{u}} f(\boldsymbol{x})$ on the 1-dimensional subspace $\{\lambda \boldsymbol{v} : \lambda \in \mathsf{R}\}$. The remainder of the proof is supplied by the finite-dimensional version of the Hahn-Banach theorem.

Proposition 14.4.1 (Hahn-Banach) *Suppose the linear function $\ell(\boldsymbol{v})$ is defined on a subspace S of R^n and dominated there by the sublinear function $h(\boldsymbol{v})$ defined on all of R^n. Then $\ell(\boldsymbol{v})$ can be extended to a linear function that is dominated throughout R^n by $h(\boldsymbol{v})$.*

Proof: The proof proceeds by induction on the dimension of S. Let $\boldsymbol{u}$ be any point in R^n outside S. It suffices to show that $\ell(\boldsymbol{v})$ can be consistently defined on the subspace T spanned by S and $\boldsymbol{u}$. For $\boldsymbol{v} \in S$ linearity requires

$$\ell(\boldsymbol{v} + \lambda \boldsymbol{u}) \quad = \quad \ell(\boldsymbol{v}) + \lambda \ell(\boldsymbol{u}),$$

and the crux of the matter is properly defining $\ell(\boldsymbol{u})$. For two points $\boldsymbol{v}$ and $\boldsymbol{w}$ in S, we have

$$\ell(\boldsymbol{v}) + \ell(\boldsymbol{w}) \quad = \quad \ell(\boldsymbol{v} + \boldsymbol{w}) \quad \leq \quad h(\boldsymbol{v} + \boldsymbol{w}) \quad \leq \quad h(\boldsymbol{v} - \boldsymbol{u}) + h(\boldsymbol{w} + \boldsymbol{u}).$$

It follows that

$$\ell(\boldsymbol{v}) - h(\boldsymbol{v} - \boldsymbol{u}) \quad \leq \quad h(\boldsymbol{w} + \boldsymbol{u}) - \ell(\boldsymbol{w}).$$

Because $\boldsymbol{v}$ and $\boldsymbol{w}$ are arbitrary, the left-hand side of this inequality is bounded above for $\boldsymbol{u}$ fixed and the right-hand side is bounded below for $\boldsymbol{u}$ fixed. The idea now is to define $\ell(\boldsymbol{u})$ to be any number α satisfying

$$\sup_{\boldsymbol{v} \in S} [\ell(\boldsymbol{v}) - h(\boldsymbol{v} - \boldsymbol{u})] \quad \leq \quad \alpha \quad \leq \quad \inf_{\boldsymbol{w} \in S} [h(\boldsymbol{w} + \boldsymbol{u}) - \ell(\boldsymbol{w})].$$

If $\lambda > 0$, then our choice of α entails

$$\begin{aligned} \ell(\boldsymbol{v} + \lambda \boldsymbol{u}) &= \lambda \left[\alpha + \ell(\lambda^{-1} \boldsymbol{v})\right] \\ &\leq \lambda \left[h(\lambda^{-1} \boldsymbol{v} + \boldsymbol{u}) - \ell(\lambda^{-1} \boldsymbol{v}) + \ell(\lambda^{-1} \boldsymbol{v})\right] \\ &= h(\boldsymbol{v} + \lambda \boldsymbol{u}). \end{aligned}$$

Similarly if $\lambda < 0$, then our choice of α entails

$$\begin{aligned} \ell(\boldsymbol{v} + \lambda \boldsymbol{u}) &= -\lambda \left[-\alpha + \ell(-\lambda^{-1} \boldsymbol{v})\right] \\ &\leq -\lambda \left[h(-\lambda^{-1} \boldsymbol{v} - \boldsymbol{u}) - \ell(-\lambda^{-1} \boldsymbol{v}) + \ell(-\lambda^{-1} \boldsymbol{v})\right] \\ &= h(\boldsymbol{v} + \lambda \boldsymbol{u}). \end{aligned}$$

Thus, $\ell(\boldsymbol{x})$ is dominated by $h(\boldsymbol{x})$ on T. ■

The next proposition summarizes the previous discussion and collects some further pertinent facts about subdifferentials. Recall that the notion of a support function mentioned in the proposition was defined in Example 14.3.4.

Proposition 14.4.2 *The subdifferential $\partial f(\boldsymbol{x})$ of a convex function $f(\boldsymbol{x})$ is a closed convex set for all $\boldsymbol{x}$. If $\boldsymbol{x}$ is an interior point of* $\operatorname{dom}(f)$*, then $\partial f(\boldsymbol{x})$ is nonempty and compact, the directional derivative $d_{\boldsymbol{v}} f(\boldsymbol{x})$ exists and is finite for all $\boldsymbol{v}$, and $d_{\boldsymbol{v}} f(\boldsymbol{x})$ is the support function of $\partial f(\boldsymbol{x})$. If $f(\boldsymbol{x})$ is differentiable at $\boldsymbol{x}$, then $\partial f(\boldsymbol{x})$ coincides with the singleton set $\{\nabla f(\boldsymbol{x})\}$.*

Proof: The supporting hyperplane inequality is preserved under limits and convex combinations of subgradients. Hence, $\partial f(\boldsymbol{x})$ is closed and convex. For an interior point $\boldsymbol{x}$, the Hahn-Banach theorem shows that $\partial f(\boldsymbol{x})$ is nonempty and that

$$d_{\boldsymbol{v}} f(\boldsymbol{x}) \quad = \quad \max_{\boldsymbol{g} \in \partial f(\boldsymbol{x})} \boldsymbol{g}^* \boldsymbol{v},$$

so by definition $d_{\boldsymbol{v}} f(\boldsymbol{x})$ is the support function of $\partial f(\boldsymbol{x})$. To demonstrate that $\partial f(\boldsymbol{x})$ is compact, it suffices to show that it is bounded. To reach a contradiction, assume that $\boldsymbol{g}_m \in \partial f(\boldsymbol{x})$ satisfies $\lim_{m\to\infty} \|\boldsymbol{g}_m\| = \infty$. By passing to a subsequence if necessary, one can further assume that the sequence of unit vectors $\boldsymbol{v}_m = \|\boldsymbol{g}_m\|^{-1}\boldsymbol{g}_m$ converges to a unit vector $\boldsymbol{v}$. If the ball of radius ϵ centered at $\boldsymbol{x}$ lies wholly within $\operatorname{dom}(f)$, then the inequality

$$f(\boldsymbol{x} + \epsilon \boldsymbol{v}_m) \quad \geq \quad f(\boldsymbol{x}) + \epsilon \boldsymbol{g}_m^* \boldsymbol{v}_m \quad = \quad f(\boldsymbol{x}) + \epsilon \|\boldsymbol{g}_m\|$$

contradicts the boundedness of $f(\boldsymbol{y})$ within the ball guaranteed by Proposition 6.4.1. Finally, suppose $f(\boldsymbol{x})$ is differentiable at $\boldsymbol{x}$. For $\boldsymbol{g} \in \partial f(\boldsymbol{x})$ we have

$$f(\boldsymbol{x}) + \alpha \boldsymbol{g}^* \boldsymbol{v} \quad \leq \quad f(\boldsymbol{x} + \alpha \boldsymbol{v}) \quad = \quad f(\boldsymbol{x}) + \alpha df(\boldsymbol{x})\boldsymbol{v} + o(|\alpha|).$$

If we let $\boldsymbol{v} = \nabla f(\boldsymbol{x}) - \boldsymbol{g}$, then this inequality implies

$$0 \quad \leq \quad \alpha[\nabla f(\boldsymbol{x}) - \boldsymbol{g}]^*[\nabla f(\boldsymbol{x}) - \boldsymbol{g}] + o(|\alpha|) \quad = \quad \alpha \|\nabla f(\boldsymbol{x}) - \boldsymbol{g}\|^2 + o(|\alpha|).$$

Taking $\alpha < 0$ small now yields a contradiction unless $\nabla f(\boldsymbol{x}) - \boldsymbol{g} = \boldsymbol{0}$. ■

For a convex function $f(x)$ with domain the real line, the subdifferential $\partial f(x)$ equals the interval $[-d_{-1}f(x), d_1 f(x)]$. For instance, the function $f(x) = |y - x|$ has

$$\partial f(x) \quad = \quad \begin{cases} -1 & x < y \\ [-1, 1] & x = y \\ 1 & x > y. \end{cases}$$

The behavior of $\partial f(x)$ at boundary points of $\text{dom}(f)$ is more erratic than the behavior at interior points. For example, the choice

$$f(x) = \begin{cases} -\sqrt{x} & x \in [0,1] \\ \infty & \text{otherwise} \end{cases}$$

has subdifferential

$$\partial f(x) = \begin{cases} -\frac{1}{2\sqrt{x}} & x \in (0,1) \\ [-1/2, \infty) & x = 1 \\ \emptyset & x \leq 0 \text{ or } x > 1. \end{cases}$$

Thus, at one boundary point of $\text{dom}(f)$ the subdifferential $\partial f(x)$ is empty, and at the other it is unbounded.

Here is the convex generalization of Fermat's stationarity condition.

Proposition 14.4.3 *A convex function $f(\boldsymbol{y})$ possesses a minimum at the point $\boldsymbol{x}$ if and only if $\mathbf{0} \in \partial f(\boldsymbol{x})$.*

Proof: The inequality $f(\boldsymbol{y}) \geq f(\boldsymbol{x})$ for all $\boldsymbol{y}$ is trivially equivalent to the inequality $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \mathbf{0}^*(\boldsymbol{y} - \boldsymbol{x})$ for all $\boldsymbol{y}$. ■

The next proposition highlights the importance of the Fenchel conjugate in calculating subdifferentials.

Proposition 14.4.4 *For a convex function $f(\boldsymbol{x})$ with conjugate $f^\star(\boldsymbol{y})$, assertions (a) and (b) from the following list*

(a) $f(\boldsymbol{x}) + f^\star(\boldsymbol{y}) = \boldsymbol{y}^*\boldsymbol{x}$

(b) $\boldsymbol{y} \in \partial f(\boldsymbol{x})$

(c) $\boldsymbol{x} \in \partial f^\star(\boldsymbol{y})$

are logically equivalent for a vector pair $(\boldsymbol{x}, \boldsymbol{y})$. If $f(\boldsymbol{x})$ is closed, then both assertions are logically equivalent to assertion (c). Furthermore, the set of minima of $f(\boldsymbol{x})$ coincides with $\partial f^\star(\mathbf{0})$.

Proof: A pair $(\boldsymbol{x}, \boldsymbol{y})$ satisfies condition (a) if and only if $\boldsymbol{y}^*\boldsymbol{x} - f(\boldsymbol{x})$ attains its maximum at $\boldsymbol{x}$ for $\boldsymbol{y}$ fixed. The latter condition is equivalent to the convex function $h(\boldsymbol{x}) = f(\boldsymbol{x}) - \boldsymbol{y}^*\boldsymbol{x}$ attaining is minimum. A brief calculation shows that $\partial h(\boldsymbol{x}) = \partial f(\boldsymbol{x}) - \boldsymbol{y}$. Proposition 14.4.3 therefore implies that condition (a) is equivalent to $\mathbf{0} \in \partial f(\boldsymbol{x}) - \boldsymbol{y}$, which is just a restatement of condition (b). When $f(\boldsymbol{x})$ is closed, $f^{\star\star}(\boldsymbol{x}) = f(\boldsymbol{x})$, and we can reverse the roles of $f(\boldsymbol{x})$ and $f^\star(\boldsymbol{y})$ and deduce the equivalence of condition (c). The final assertion of the proposition follows from the observation that $\mathbf{0} \in \partial f(\boldsymbol{x})$ if and only if $\boldsymbol{x} \in \partial f^\star(\mathbf{0})$. ■

Example 14.4.1 *Subdifferential of the Indicator of a Closed Convex Set*

Let $\delta_C(\boldsymbol{x})$ be the indicator of a closed convex set C. Outside C the subdifferential $\partial\delta_C(\boldsymbol{x})$ is empty. For $\boldsymbol{x} \in C$ the calculation

$$\begin{aligned}\partial\delta_C(\boldsymbol{x}) &= \{\boldsymbol{y} : \delta_C(\boldsymbol{x}) + \delta_C^\star(\boldsymbol{y}) = \boldsymbol{y}^*\boldsymbol{x}\} \\ &= \{\boldsymbol{y} : \sup_{\boldsymbol{z}\in C} \boldsymbol{y}^*\boldsymbol{z} = \boldsymbol{y}^*\boldsymbol{x}\} \\ &= \{\boldsymbol{y} : \boldsymbol{y}^*(\boldsymbol{z}-\boldsymbol{x}) \le 0 \text{ for all } \boldsymbol{z} \in C\}\end{aligned}$$

identifies the subdifferential as the polar cone to the translated set $C - \boldsymbol{x}$. This polar cone is denoted $N_C(\boldsymbol{x})$ and called the normal cone to C at $\boldsymbol{x}$. The double polar cone $N_C(\boldsymbol{x})^\circ$ is termed the tangent cone to C at $\boldsymbol{x}$; it is the smallest closed convex cone containing $C - \boldsymbol{x}$. Various special cases of $N_C(\boldsymbol{x})$ readily come to mind. For example, if $\boldsymbol{x}$ belongs to the interior of C, then $N_C(\boldsymbol{x}) = \{\boldsymbol{0}\}$. Alternatively, if C is an affine subspace $S + \boldsymbol{x}$, then $N_C(\boldsymbol{x}) = S^\perp$, the orthogonal complement of the subspace S. When C is the affine subspace $\{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ corresponding to solutions of the linear equation $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, then S equals the kernel of $\boldsymbol{A}$, and the fundamental theorem of linear algebra implies that $S^\perp$ equals the range of the transpose $\boldsymbol{A}^*$ [248]. ■

Example 14.4.2 *Subdifferential of a Norm*

Let B be the closed unit ball associated with a norm $\|\boldsymbol{x}\|_\dagger$ on R^n. The dual norm $\|\boldsymbol{y}\|_\star$ defined in Example 14.3.5 coincides with the support function $\delta_B^\star(\boldsymbol{y})$. These considerations imply

$$\begin{aligned}\partial\|\boldsymbol{y}\|_\star &= \{\boldsymbol{x} : \delta_B(\boldsymbol{x}) + \delta_B^\star(\boldsymbol{y}) = \boldsymbol{y}^*\boldsymbol{x}\} \\ &= \{\boldsymbol{x} \in B : \|\boldsymbol{y}\|_\star = \boldsymbol{y}^*\boldsymbol{x}\}.\end{aligned}$$

The dual relation

$$\partial\|\boldsymbol{x}\|_\dagger = \{\boldsymbol{y} \in U : \|\boldsymbol{x}\|_\dagger = \boldsymbol{y}^*\boldsymbol{x}\}$$

holds for U the closed unit ball associated with $\|\boldsymbol{y}\|_\star$. In the case of the Euclidean norm, which is its own dual, the subdifferential coincides with the gradient $\|\boldsymbol{x}\|^{-1}\boldsymbol{x}$ when $\boldsymbol{x} \neq \boldsymbol{0}$. At the origin $\partial\|\boldsymbol{0}\| = B$. In general for any norm, $\boldsymbol{0} \in \partial\|\boldsymbol{x}\|_\dagger$ if and only if $\boldsymbol{x} = \boldsymbol{0}$. This is just a manifestation of the fact that $\boldsymbol{x} = \boldsymbol{0}$ is the unique minimum point of the norm. ■

Example 14.4.3 *Subdifferential of the Distance to a Closed Convex Set*

Let C be a closed convex set in R^n. The distance $f(\boldsymbol{x}) = \min_{\boldsymbol{z}\in C} \|\boldsymbol{z} - \boldsymbol{x}\|_\dagger$ from $\boldsymbol{x}$ to C under any norm $\|\boldsymbol{z}\|_\dagger$ is attained and finite. Furthermore, as observed in Problem 17 of Chap. 6, $f(\boldsymbol{x})$ is a convex function. In contrast to Euclidean distance, the closest point $\boldsymbol{z}$ in C to $\boldsymbol{x}$ may not be unique. Despite

this complication, one can calculate the conjugate $f^\star(\boldsymbol{y})$ and subdifferential $\partial f(\boldsymbol{x})$. If U is the closed unit ball of the dual norm $\|\boldsymbol{y}\|_\star$, then the conjugate amounts to

$$\begin{aligned} f^\star(\boldsymbol{y}) &= \sup_{\boldsymbol{x}} (\boldsymbol{y}^*\boldsymbol{x} - \min_{\boldsymbol{z}\in C} \|\boldsymbol{z}-\boldsymbol{x}\|_\dagger) \\ &= \sup_{\boldsymbol{x}} \sup_{\boldsymbol{z}\in C} (\boldsymbol{y}^*\boldsymbol{x} - \|\boldsymbol{z}-\boldsymbol{x}\|_\dagger) \\ &= \sup_{\boldsymbol{z}\in C} \{\boldsymbol{y}^*\boldsymbol{z} + \sup_{\boldsymbol{x}} [\boldsymbol{y}^*(\boldsymbol{x}-\boldsymbol{z}) - \|\boldsymbol{z}-\boldsymbol{x}\|_\dagger]\} \\ &= \sup_{\boldsymbol{z}\in C} [\boldsymbol{y}^*\boldsymbol{z} + \sup_{\boldsymbol{x}} (\boldsymbol{y}^*\boldsymbol{x} - \|\boldsymbol{x}\|_\dagger)] \\ &= \sup_{\boldsymbol{z}\in C} [\boldsymbol{y}^*\boldsymbol{z} + \delta_U(\boldsymbol{y})] \\ &= \delta_C^\star(\boldsymbol{y}) + \delta_U(\boldsymbol{y}). \end{aligned} \tag{14.7}$$

A subgradient $\boldsymbol{y} \in \partial f(\boldsymbol{x})$ is characterized by the equation

$$\boldsymbol{y}^*\boldsymbol{x} = f^\star(\boldsymbol{y}) + f(\boldsymbol{x}) = \delta_C^\star(\boldsymbol{y}) + \delta_U(\boldsymbol{y}) + f(\boldsymbol{x}),$$

which clearly forces $\boldsymbol{y}$ to reside in U. To make further progress, select a point $\boldsymbol{z} \in C$ with $f(\boldsymbol{x}) = \|\boldsymbol{z}-\boldsymbol{x}\|_\dagger$. Because $\delta_C^\star(\boldsymbol{y})$ is the support function of C, this choice yields the inequality

$$\|\boldsymbol{z}-\boldsymbol{x}\|_\dagger = f(\boldsymbol{x}) = \boldsymbol{y}^*\boldsymbol{x} - \delta_C^\star(\boldsymbol{y}) \leq \boldsymbol{y}^*(\boldsymbol{x}-\boldsymbol{z}). \tag{14.8}$$

Fortunately, the restriction $\boldsymbol{y} \in U$ and the generalized Cauchy-Schwarz inequality (14.4) imply the opposite inequality, and we conclude that

$$\boldsymbol{y}^*(\boldsymbol{x}-\boldsymbol{z}) = \|\boldsymbol{x}-\boldsymbol{z}\|_\dagger. \tag{14.9}$$

The previous example now implies that $\boldsymbol{y} \in \partial\|\boldsymbol{x}-\boldsymbol{z}\|_\dagger$. Duplicating the reasoning that led to inequality (14.8) proves that any other point $\boldsymbol{w} \in C$ satisfies $\boldsymbol{y}^*(\boldsymbol{x}-\boldsymbol{w}) \geq \|\boldsymbol{x}-\boldsymbol{z}\|_\dagger$. Subtracting this inequality from equality (14.9) gives $\boldsymbol{y}^*(\boldsymbol{w}-\boldsymbol{z}) \leq 0$. Hence, $\boldsymbol{y}$ belongs to the normal cone $N_C(\boldsymbol{z})$ to C at $\boldsymbol{z}$, and we have proved the containment

$$\partial f(\boldsymbol{x}) \subset U \cap N_C(\boldsymbol{z}) \cap \partial\|\boldsymbol{x}-\boldsymbol{z}\|_\dagger.$$

The reverse containment is also true. Suppose $\boldsymbol{y}$ belongs to the intersection $U \cap N_C(\boldsymbol{z}) \cap \partial\|\boldsymbol{x}-\boldsymbol{z}\|_\dagger$. Then the normal cone condition $\boldsymbol{y}^*\boldsymbol{w} \leq \boldsymbol{y}^*\boldsymbol{z}$ for all $\boldsymbol{w} \in C$ implies the equality $\boldsymbol{y}^*\boldsymbol{z} = \sup_{\boldsymbol{w}\in C} \boldsymbol{y}^*\boldsymbol{w} = \delta_C^\star(\boldsymbol{y})$. Together with the subgradient condition $\boldsymbol{y}^*(\boldsymbol{x}-\boldsymbol{z}) = \|\boldsymbol{x}-\boldsymbol{z}\|_\dagger$ of Example 14.4.2, this gives

$$f(\boldsymbol{x}) = \|\boldsymbol{z}-\boldsymbol{x}\|_\dagger = \boldsymbol{y}^*(\boldsymbol{x}-\boldsymbol{z}) = \boldsymbol{y}^*\boldsymbol{x} - \delta_C^\star(\boldsymbol{y}).$$

In view of the identity $f^\star(\boldsymbol{y}) = \delta_C^\star(\boldsymbol{y}) + \delta_U(\boldsymbol{y})$ of equation (14.7) and part (b) of Proposition 14.4.4, we are justified in concluding that $\boldsymbol{y} \in \partial f(\boldsymbol{x})$.

In the case of the Euclidean norm and a point $\boldsymbol{x} \notin C$, the point $\boldsymbol{z}$ is the projection onto C, and the subdifferential $\partial\|\boldsymbol{x}-\boldsymbol{z}\|_\dagger = \|\boldsymbol{x}-\boldsymbol{z}\|^{-1}(\boldsymbol{x}-\boldsymbol{z})$ reduces to a unit vector consistent with the normal cone $N_C(\boldsymbol{z})$. In other words, $f(\boldsymbol{x})$ is differentiable with gradient $\|\boldsymbol{x}-\boldsymbol{z}\|^{-1}(\boldsymbol{x}-\boldsymbol{z})$. ■

14.5 The Rules of Convex Differentiation

Some, but not all, of the usual rules of differentiation carry over to convex functions. The simplest rule that successfully generalizes is

$$\partial[\alpha f(\boldsymbol{x})] \quad = \quad \alpha \partial f(\boldsymbol{x})$$

for α a positive scalar. This result is just another way of saying that the two supporting plane inequalities

$$\begin{aligned} f(\boldsymbol{y}) &\geq f(\boldsymbol{x}) + \boldsymbol{g}^*(\boldsymbol{y}-\boldsymbol{x}) \\ \alpha f(\boldsymbol{y}) &\geq \alpha f(\boldsymbol{x}) + (\alpha \boldsymbol{g})^*(\boldsymbol{y}-\boldsymbol{x}) \end{aligned}$$

are logically equivalent. For functions that depend on a single coordinate, another basic rule that the reader can readily verify is

$$\partial f(x_i) \quad = \quad \partial_i f(x_i)\boldsymbol{e}_i,$$

where ∂ takes the subdifferential with respect to $\boldsymbol{x} \in \mathsf{R}^n$, ∂_i takes the subdifferential with respect to $x_i \in \mathsf{R}$, and $\boldsymbol{e}_i$ denotes the ith unit coordinate vector.

Our strategy for deriving other rules is indirect and relies on the characterization (14.5) of the directional derivative as the support function of the subdifferential. Consider a convex function $f(\boldsymbol{x})$ all of whose directional derivatives $d_{\boldsymbol{v}} f(\boldsymbol{z})$ exist and are finite at some point $\boldsymbol{z} \in \operatorname{dom}(f)$. In this circumstance, equation (14.6) defines the subdifferential $\partial f(\boldsymbol{z})$ in terms of the directional derivatives $d_{\boldsymbol{v}} f(\boldsymbol{z})$. Although definition (14.6) is still operative even when $d_{\boldsymbol{v}} f(\boldsymbol{z}) = \infty$ for some $\boldsymbol{v}$, we will ignore this possibility. Example 14.3.4 documents the fact that every nonempty closed convex set C is uniquely characterized by its support function

$$\delta_C^\star(\boldsymbol{y}) \quad = \quad \sup_{\boldsymbol{x} \in C} \boldsymbol{y}^* \boldsymbol{x}.$$

We will also need the fact that any nonempty set K with closed convex hull C generates the same support function as C. This is true because

$$\delta_K^\star(\boldsymbol{y}) \quad = \quad \sup_{\boldsymbol{x} \in K} \boldsymbol{y}^* \boldsymbol{x} \quad = \quad \sup_{\boldsymbol{x} \in C} \boldsymbol{y}^* \boldsymbol{x} \quad = \quad \delta_C^\star(\boldsymbol{y}),$$

where the middle equality follows from the identity

$$\max\{a_1, \ldots, a_m\} \quad = \quad \max\left\{ \sum_{i=1}^m \lambda_i a_i : \text{all } \lambda_i \geq 0, \sum_{i=1}^m \lambda_i = 1 \right\}$$

and the continuity and linearity of the map $\boldsymbol{x} \mapsto \boldsymbol{y}^*\boldsymbol{x}$.

To generalize the chain rule, consider a convex function $f(\boldsymbol{y})$ and a differentiable function $g(\boldsymbol{x})$ whose composition $f \circ g(\boldsymbol{x})$ with $f(\boldsymbol{y})$ is convex. If $\boldsymbol{x} \in \text{dom}(g)$ and $g(\boldsymbol{x}) \in \text{dom}(f)$, then the difference quotient

$$\begin{aligned}\frac{f\circ g(\boldsymbol{x}+t\boldsymbol{v}) - f \circ g(\boldsymbol{x})}{t} &= \frac{f\circ g(\boldsymbol{x}+t\boldsymbol{v}) - f[g(\boldsymbol{x}) + tdg(\boldsymbol{x})\boldsymbol{v}]}{t} \\ &\quad + \frac{f[g(\boldsymbol{x}) + tdg(\boldsymbol{x})\boldsymbol{v}] - f\circ g(\boldsymbol{x})}{t}\end{aligned}$$

yields in the limit the directional derivative. The first term on the right of this equation vanishes when $g(\boldsymbol{x})$ is a linear function. The second term tends to $d_{dg(\boldsymbol{x})\boldsymbol{v}} f[g(\boldsymbol{x})]$, the directional derivative of $f(\boldsymbol{y})$ at the point $\boldsymbol{y} = g(\boldsymbol{x})$ in the direction $dg(\boldsymbol{x})\boldsymbol{v}$. Even when $g(\boldsymbol{x})$ is nonlinear, there is still hope that the first term vanishes in the limit. Suppose $g(\boldsymbol{x})$ belongs to the interior of $\text{dom}(f)$. If we let L be the Lipschitz constant for $f(\boldsymbol{y})$ near the point $g(\boldsymbol{x})$ guaranteed by Proposition 6.4.1, then the inequality

$$\begin{aligned}\left|\frac{f\circ g(\boldsymbol{x}+t\boldsymbol{v}) - f[g(\boldsymbol{x}) + tdg(\boldsymbol{x})\boldsymbol{v}]}{t}\right| &\leq \frac{L\|g(\boldsymbol{x}+t\boldsymbol{v}) - g(\boldsymbol{x}) - tdg(\boldsymbol{x})\boldsymbol{v}\|}{t} \\ &= \frac{o(t)}{t}\end{aligned}$$

shows once again that $d_{\boldsymbol{v}}(f \circ g)(\boldsymbol{x}) = d_{dg(\boldsymbol{x})\boldsymbol{v}} f(\boldsymbol{y})$ for $\boldsymbol{y} = g(\boldsymbol{x})$. This directional derivative identity translates into the further identity

$$\sup_{\boldsymbol{z}\in\partial(f\circ g)(\boldsymbol{x})} \boldsymbol{v}^*\boldsymbol{z} = \sup_{\boldsymbol{z}\in\partial f(\boldsymbol{y})} \boldsymbol{v}^* dg(\boldsymbol{x})^*\boldsymbol{z} = \sup_{\boldsymbol{w}\in dg(\boldsymbol{x})^*\partial f(\boldsymbol{y})} \boldsymbol{v}^*\boldsymbol{w}$$

connecting the corresponding support functions. In view of our earlier remarks regarding support functions and closed convex sets, the convex set $\partial(f \circ g)(\boldsymbol{x})$ is the closure of the convex set $dg(\boldsymbol{x})^*\partial f(\boldsymbol{y})$.

To show that the two sets coincide, it suffices to show that $dg(\boldsymbol{x})^*\partial f(\boldsymbol{y})$ is closed. If $\partial f(\boldsymbol{y})$ is compact, say when $\boldsymbol{y}$ is an interior point of $\text{dom}(f)$, then $dg(\boldsymbol{x})^*\partial f(\boldsymbol{y})$ is compact as well, and the set equality

$$\partial f \circ g(\boldsymbol{x}) = dg(\boldsymbol{x})^*\partial f(\boldsymbol{y})$$

holds. Equality is also true in some circumstances when $\partial f(\boldsymbol{y})$ is not compact. For example, when $\partial f(\boldsymbol{y})$ is polyhedral, the image set $\boldsymbol{A}^*\partial f(\boldsymbol{y})$ generated by the composition $f(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})$ is closed for every compatible matrix $\boldsymbol{A}$. A polyhedral set is the nonempty intersection of a finite number of closed halfspaces. Appendix A.3 takes up the subject of polyhedral sets. Proposition A.3.4 is particularly pertinent to the current discussion. The next proposition summaries our conclusions.

Proposition 14.5.1 (Convex Chain Rule) *Let $f(\boldsymbol{y})$ be a convex function and $g(\boldsymbol{x})$ be a differentiable function. If the composition $f \circ g(\boldsymbol{x})$ is well defined and convex, then the chain rule $\partial f \circ g(\boldsymbol{x}) = dg(\boldsymbol{x})^*\partial f(\boldsymbol{y})$ is*

valid whenever (a) $\boldsymbol{y} = g(\boldsymbol{x})$*, (b) all possible directional derivatives* $d_{\boldsymbol{v}} f(\boldsymbol{y})$ *exist and are finite, and (c)* $dg(\boldsymbol{x})^* \partial f(\boldsymbol{y})$ *is a closed set.*

Proof: See the previous discussion. ■

The composition $f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ of $f(\boldsymbol{y})$ with an affine function $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ is not the only case of interest. Recall that $f \circ g(\boldsymbol{x})$ is convex when $f(y)$ is convex and increasing and $g(\boldsymbol{x})$ is convex. In particular, the composite function $g(\boldsymbol{x})_+ = \max\{0, g(\boldsymbol{x})\}$ is convex whenever $g(\boldsymbol{x})$ is differentiable and convex. Its subdifferential amounts to

$$\partial g(\boldsymbol{x})_+ \quad = \quad dg(\boldsymbol{x})^* \begin{cases} 0 & g(\boldsymbol{x}) < 0 \\ [0,1] & g(\boldsymbol{x}) = 0 \\ 1 & g(\boldsymbol{x}) > 0. \end{cases}$$

In less favorable circumstances, $f \circ g(\boldsymbol{x})$ is not convex. However, if the directional derivative $d_{\boldsymbol{v}}(f \circ g)(\boldsymbol{x})$ is sublinear in $\boldsymbol{v}$, then the subdifferential $\partial(f \circ g)(\boldsymbol{x})$ still makes sense [220].

The sum rule of differentiation is equally easy to prove. Consider two convex functions $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ possessing all possible directional derivatives at the point $\boldsymbol{x}$. The identity

$$d_{\boldsymbol{v}}[f(\boldsymbol{x}) + g(\boldsymbol{x})] \quad = \quad d_{\boldsymbol{v}} f(\boldsymbol{x}) + d_{\boldsymbol{v}} g(\boldsymbol{x})$$

entails the support function identity

$$\sup_{\boldsymbol{z} \in \partial[f(\boldsymbol{x})+g(\boldsymbol{x})]} \boldsymbol{v}^* \boldsymbol{z} \quad = \quad \sup_{\boldsymbol{u} \in \partial f(\boldsymbol{x})} \boldsymbol{v}^* \boldsymbol{u} + \sup_{\boldsymbol{w} \in \partial g(\boldsymbol{x})} \boldsymbol{v}^* \boldsymbol{w} \quad = \quad \sup_{\boldsymbol{z} \in \partial f(\boldsymbol{x}) + \partial g(\boldsymbol{x})} \boldsymbol{v}^* \boldsymbol{z}.$$

It follows that $\partial[f(\boldsymbol{x}) + g(\boldsymbol{x})]$ is the closure of the convex set $\partial f(\boldsymbol{x}) + \partial g(\boldsymbol{x})$. If either of the two subdifferentials $\partial f(\boldsymbol{x})$ and $\partial g(\boldsymbol{x})$ is compact, then the identity $\partial[f(\boldsymbol{x}) + g(\boldsymbol{x})] = \partial f(\boldsymbol{x}) + \partial g(\boldsymbol{x})$ holds. Indeed, the sum of a closed set and a compact set is always a closed set. Again compactness is not necessary. For example, Proposition A.3.4 demonstrates that the sum of two polyhedral sets is closed. The sum rule is called the Moreau-Rockafellar theorem in the convex calculus literature. Let us again summarize our conclusions by a formal proposition.

Proposition 14.5.2 (Moreau-Rockafellar) *Let* $f(\boldsymbol{x})$ *and* $g(\boldsymbol{x})$ *be convex functions defined on* R^n*. If all possible directional derivatives* $d_{\boldsymbol{v}} f(\boldsymbol{x})$ *and* $d_{\boldsymbol{v}} g(\boldsymbol{x})$ *exist and are finite at a point* $\boldsymbol{x}$*, and the set* $\partial f(\boldsymbol{x}) + \partial g(\boldsymbol{x})$ *is closed, then the sum rule* $\partial[f(\boldsymbol{x}) + g(\boldsymbol{x})] = \partial f(\boldsymbol{x}) + \partial g(\boldsymbol{x})$ *is valid.*

Proof: See the foregoing discussion. ■

Example 14.5.1 *Mean Value Theorem for Convex Functions.*

Let $f(\boldsymbol{x})$ be a convex function. If two points $\boldsymbol{y}$ and $\boldsymbol{z}$ belong to the interior of $\mathrm{dom}(f)$, then the line segment connecting them does as well. Based on

the function $g(t) = f[t\boldsymbol{y} + (1-t)\boldsymbol{z}]$, define the continuous convex function

$$h(t) \quad = \quad g(t) - g(0) - t[g(1) - g(0)].$$

The sum rule and the chain rule imply that $h(t)$ has subdifferential

$$\partial g(t) - g(1) + g(0) \quad = \quad (\boldsymbol{y} - \boldsymbol{z})^* \partial f[t\boldsymbol{y} + (1-t)\boldsymbol{z}] - g(1) + g(0).$$

Now $h(0) = h(1) = 0$, so $h(t)$ attains its minimum on the open interval $(0, 1)$. At a minimum point t we have $0 \in \partial h(t)$. It follows that

$$f(\boldsymbol{y}) - f(\boldsymbol{z}) \quad = \quad g(1) - g(0) \quad = \quad \boldsymbol{v}^*(\boldsymbol{y} - \boldsymbol{z})$$

for some $\boldsymbol{v} \in \partial f[t\boldsymbol{y} + (1-t)\boldsymbol{z}]$. ■

Example 14.5.2 *Quantiles and Subdifferentials*

A median μ of n numbers $x_1 \leq x_2 \leq \cdots \leq x_n$ satisfies the two inequalities

$$\frac{1}{n} \sum_{x_i \leq \mu} 1 \quad \geq \quad \frac{1}{2} \quad \text{and} \quad \frac{1}{n} \sum_{x_i \geq \mu} 1 \quad \geq \quad \frac{1}{2}.$$

One can relate this to the minimum of the function

$$f(x) \quad = \quad \sum_{i=1}^{n} |x_i - x|.$$

According to the sum rule, the differential of $f(x)$ equals the set

$$\partial f(x) \quad = \quad -\sum_{x_i > x} 1 + \sum_{x_i = x} [-1, 1] + \sum_{x_i < x} 1.$$

A minimum point μ of $f(x)$ is determined by the condition $0 \in \partial f(\mu)$, which is a disguised form of the median definition just given.

As a generalization take $q \in (0, 1)$ and define

$$\rho_q(r) \quad = \quad \begin{cases} qr & r \geq 0 \\ -(1-q)r & r < 0. \end{cases}$$

The subdifferential of the function $f_q(x) = \sum_{i=1}^{n} \rho_q(x_i - x)$ is the set

$$\partial f_q(x) \quad = \quad -\sum_{x_i > x} q + \sum_{x_i = x} [-q, 1-q] + \sum_{x_i < x} (1-q).$$

Again the minimum points μ_q of $f_q(x)$ satisfy $0 \in \partial f_q(\mu_q)$, which is equivalent to the q-quantile conditions

$$\frac{1}{n} \sum_{x_i \leq \mu_q} 1 \quad \geq \quad q \quad \text{and} \quad \frac{1}{n} \sum_{x_i \geq \mu_q} 1 \quad \geq \quad 1 - q.$$

Problem 13 of Chap. 8 suggests an MM algorithm for finding μ_q that involves no sorting of the list $(x_1, \ldots, x_n)$. More importantly, one can devise similar MM algorithms for the wider class of quantile regression problems [141]. ■

Example 14.5.3 *Lasso Penalized Estimation*

Lasso penalized estimation minimizes the criterion

$$g(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) + \lambda \sum_{i=1}^{p} |\theta_i|,$$

where $\lambda \geq 0$ and $f(\boldsymbol{\theta})$ is a convex differentiable loss function. The choice $f(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2$ corresponds to ℓ_2 linear regression. In this case, $\boldsymbol{\theta}$ has $p+1$ components. Penalties are only imposed on the slope parameters $\theta_1, \ldots, \theta_p$ and not on the intercept parameter θ_0. The sum rule implies that $g(\boldsymbol{\theta})$ has subdifferential

$$\partial g(\boldsymbol{\theta}) = \nabla f(\boldsymbol{\theta}) + \lambda \sum_{i=1}^{p} \partial |\theta_i|.$$

It follows that the stationarity condition $\mathbf{0} \in \partial g(\boldsymbol{\theta})$ amounts to

$$\begin{aligned} \frac{\partial}{\partial \theta_0} f(\boldsymbol{\theta}) &= 0 \\ \frac{\partial}{\partial \theta_i} f(\boldsymbol{\theta}) &\in \lambda \begin{cases} -1 & \theta_i > 0 \\ [-1, 1] & \theta_i = 0 \\ 1 & \theta_i < 0 \end{cases} \end{aligned}$$

for $1 \leq i \leq p$. Hence, a minimum point $\hat{\boldsymbol{\theta}}$ satisfies $|\frac{\partial}{\partial \theta_i} f(\hat{\boldsymbol{\theta}})| \leq \lambda$ for all i. This rather painless deduction extends well beyond ℓ_2 regression. ■

Example 14.5.4 *Euclidean Shrinkage*

Minimization of the strictly convex function

$$f(\boldsymbol{y}) = \|\boldsymbol{y}\| + \boldsymbol{w}^* \boldsymbol{y} + \frac{\lambda}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2$$

for $\lambda > 0$ illustrates the notion of shrinkage. In the absence of the term $\|\boldsymbol{y}\|$, the minimum of $f(\boldsymbol{y})$ occurs at $\boldsymbol{y} = \lambda^{-1}(\lambda \boldsymbol{x} - \boldsymbol{w})$. The presence of the term $\|\boldsymbol{y}\|$ shrinks the solution to

$$\boldsymbol{y} = \left[\frac{\|\lambda \boldsymbol{x} - \boldsymbol{w}\| - 1}{\|\lambda \boldsymbol{x} - \boldsymbol{w}\|} \right]_+ \frac{\lambda \boldsymbol{x} - \boldsymbol{w}}{\lambda}.$$

To prove this claim, observe that the subdifferential of $f(\boldsymbol{y})$ for $\boldsymbol{y} \neq \mathbf{0}$ collapses to the gradient

$$\nabla f(\boldsymbol{y}) = \|\boldsymbol{y}\|^{-1} \boldsymbol{y} + \boldsymbol{w} + \lambda(\boldsymbol{y} - \boldsymbol{x}).$$

The stationarity condition $\mathbf{0} \in \partial f(\boldsymbol{y})$ defining a minimum implies

$$\left(\lambda + \|\boldsymbol{y}\|^{-1}\right)\boldsymbol{y} \quad = \quad \lambda \boldsymbol{x} - \boldsymbol{w},$$

which in turn implies

$$\|\boldsymbol{y}\| \quad = \quad \frac{\|\lambda \boldsymbol{x} - \boldsymbol{w}\| - 1}{\lambda}.$$

Provided $\|\lambda \boldsymbol{x} - \boldsymbol{w}\| > 1$, this is consistent with $\|\boldsymbol{y}\| > 0$, and we can assert that

$$\boldsymbol{y} \quad = \quad \|\boldsymbol{y}\| \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} \quad = \quad \frac{\|\lambda \boldsymbol{x} - \boldsymbol{w}\| - 1}{\lambda} \cdot \frac{\lambda \boldsymbol{x} - \boldsymbol{w}}{\|\lambda \boldsymbol{x} - \boldsymbol{w}\|}$$

furnishes the minimum. If $\|\lambda \boldsymbol{x} - \boldsymbol{w}\| \leq 1$, then we must look elsewhere for the minimum. The only other possibility is $\boldsymbol{y} = \mathbf{0}$. In view of Example 14.4.2, $\partial f(\mathbf{0}) = B + \boldsymbol{w} - \lambda \boldsymbol{x}$. The required inclusion $\mathbf{0} \in B + \boldsymbol{w} - \lambda \boldsymbol{x}$ now reduces to a tautology. ■

Finally, let us consider a rule with no classical analogue. Generalizing Example 4.4.4, one can easily demonstrate that the maximum function $f(\boldsymbol{x}) = \max_{1 \leq i \leq p} g_i(\boldsymbol{x})$ has directional derivative

$$d_{\boldsymbol{v}} f(\boldsymbol{x}) \quad = \quad \max_{i \in I(\boldsymbol{x})} d_{\boldsymbol{v}} g_i(\boldsymbol{x}), \tag{14.10}$$

where $I(\boldsymbol{x})$ is the set of indices i such that $f(\boldsymbol{x}) = g_i(\boldsymbol{x})$. All that is required for the validity of formula (14.10) is the upper semicontinuity of the functions $g_i(\boldsymbol{x})$ at $\boldsymbol{x}$ and the existence of the directional derivatives $d_{\boldsymbol{v}} g_i(\boldsymbol{x})$ for $i \in I(\boldsymbol{x})$. If we further assume that the $g_i(\boldsymbol{x})$ are closed and convex, then $f(\boldsymbol{x})$ is closed and convex, and

$$\sup_{z \in \partial f(\boldsymbol{x})} \boldsymbol{v}^* \boldsymbol{z} \quad = \quad \max_{i \in I(\boldsymbol{x})} \sup_{\boldsymbol{z}_i \in \partial g_i(\boldsymbol{x})} \boldsymbol{v}^* \boldsymbol{z}_i.$$

In view of our earlier remarks about support functions and convex hulls, we also have

$$\max_{i \in I(\boldsymbol{x})} \sup_{\boldsymbol{z}_i \in \partial g_i(\boldsymbol{x})} \boldsymbol{v}^* \boldsymbol{z}_i \quad = \quad \sup_{\boldsymbol{z} \in \operatorname{conv}[\cup_{i \in I(\boldsymbol{x})} \partial g_i(\boldsymbol{x})]} \boldsymbol{v}^* \boldsymbol{z}.$$

If each subdifferential $\partial g_i(\boldsymbol{x})$ is compact, then the finite union $\cup_{i \in I(\boldsymbol{x})} \partial g_i(\boldsymbol{x})$ is also compact. Because the convex hull of a compact set is compact (Proposition 6.2.4), the conclusion

$$\partial f(\boldsymbol{x}) \quad = \quad \operatorname{conv}[\cup_{i \in I(\boldsymbol{x})} \partial g_i(\boldsymbol{x})] \tag{14.11}$$

emerges. This formula is valid in more general circumstances. For example, if i is a continuous variable, then it suffices for the $g_i(\boldsymbol{x})$ to be convex, the set of indices i to be compact, and the function $i \mapsto g_i(\boldsymbol{x})$ to be upper semicontinuous for each $\boldsymbol{x}$ [131, 226]. Proposition A.6.6 in the Appendix also addresses this topic.

Example 14.5.5 *Minima of Max Functions*

When the functions $g_i(\boldsymbol{x})$ are convex, a point $\boldsymbol{y}$ minimizes

$$f(\boldsymbol{x}) \quad = \quad \max_{1 \le i \le n} g_i(\boldsymbol{x})$$

if and only if $\mathbf{0} \in \partial f(\boldsymbol{y})$. Equivalently, $\mathbf{0}$ belongs to the convex hull of the vectors $\{\nabla g_i(\boldsymbol{y})\}_{i \in I(\boldsymbol{y})}$. According to Proposition 5.3.2, a necessary and sufficient condition for the latter event to occur is that no vector $\boldsymbol{u}$ exists with $dg_i(\boldsymbol{y})\boldsymbol{u} < 0$ for all $i \in I(\boldsymbol{y})$. Of course, such a vector $\boldsymbol{u}$ would constitute a descent direction along which $f(\boldsymbol{x})$ could be locally reduced. One can reformulate the problem of minimization of $f(\boldsymbol{x})$ as a linear program when the $g_i(\boldsymbol{x}) = \boldsymbol{w}_i^* \boldsymbol{x} + a_i$ are affine. In this case one just minimizes the scalar t subject to the inequality constraints $\boldsymbol{w}_i^* \boldsymbol{x} + a_i \le t$ for all i. ■

Example 14.5.6 *Sum of the m Largest Order Statistics*

Consider the order statistics $x_{(1)} \le \cdots \le x_{(n)}$ corresponding to a point $\boldsymbol{x} \in \mathsf{R}^n$. The sum

$$s_m(\boldsymbol{x}) \quad = \quad \sum_{i=n-m+1}^{n} x_{(i)} \quad = \quad \max_{|T|=m} \sum_{i \in T} x_i$$

of the m largest order statistics is a convex function of $\boldsymbol{x}$. Here T is any subset of $\{1, \ldots, n\}$ of size $|T| = m$. The gradient of $\sum_{i \in T} x_i$ is the vector 1_T whose ith entry is 1 if $i \in T$ and 0 otherwise. The subdifferential $\partial s_m(\boldsymbol{x})$ is therefore

$$\operatorname{conv}\left\{1_T : \textstyle\sum_{i \in T} x_i = s_m(\boldsymbol{x}),\ |T| = m\right\}$$

When all of the components of $\boldsymbol{x}$ are unique, $s_m(\boldsymbol{x})$ is differentiable. ■

Example 14.5.7 *Maximum Eigenvalue of a Symmetric Matrix*

As mentioned in Example 6.3.8, the maximum eigenvalue $\lambda_{\max}(\boldsymbol{M})$ of a symmetric matrix is a convex function whose value is determined by the maximum Rayleigh quotient

$$\lambda_{\max}(\boldsymbol{M}) \quad = \quad \sup_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^* \boldsymbol{M} \boldsymbol{x} \quad = \quad \sup_{\|\boldsymbol{x}\|=1} \operatorname{tr}(\boldsymbol{M} \boldsymbol{x} \boldsymbol{x}^*).$$

The Frobenius inner product map $\boldsymbol{M} \mapsto \operatorname{tr}(\boldsymbol{M} \boldsymbol{x} \boldsymbol{x}^*)$ is linear in $\boldsymbol{M}$ and has differential $\boldsymbol{x} \boldsymbol{x}^*$. Hence, the subdifferential $\partial \lambda_{\max}(\boldsymbol{M})$ is the convex hull of the set $\{\boldsymbol{x} \boldsymbol{x}^* : \|\boldsymbol{x}\| = 1,\ \boldsymbol{M} \boldsymbol{x} = \lambda_{\max}(\boldsymbol{M}) \boldsymbol{x}\}$. When there is a unique unit eigenvector $\boldsymbol{x}$ up to sign, $\lambda_{\max}(\boldsymbol{M})$ is differentiable with gradient $\boldsymbol{x} \boldsymbol{x}^*$. ■

14.6 Spectral Functions

We are already acquainted with some of the connections between eigenvalues and convexity. The lovely theory of Lewis [17] extends these previously scattered results. Once again the Fenchel conjugate is the tool of choice. Lewis' theory covers the composition of a symmetric function $f : \mathsf{R}^n \mapsto \mathsf{R}$ with the eigenvalue map $\lambda(\boldsymbol{X})$ defined on $n \times n$ symmetric matrices $\boldsymbol{X}$. Recall that $f(\boldsymbol{x})$ is symmetric if $f(\boldsymbol{w}) = f(\boldsymbol{x})$ whenever the vector arguments $\boldsymbol{w}$ and $\boldsymbol{x}$ agree up to a permutation of their entries. The vector $\lambda(\boldsymbol{X})$ presents the spectrum of $\boldsymbol{X}$ ordered from the largest to the smallest eigenvalue. The next proposition shows how to calculate the Fenchel conjugate of the composite function $f \circ \lambda(\boldsymbol{X})$. In the proposition the operator $\operatorname{diag}(\boldsymbol{x})$ promotes a vector $\boldsymbol{x}$ to a diagonal matrix with x_i as its ith diagonal entry.

Proposition 14.6.1 *Suppose $f(\boldsymbol{x})$ has Fenchel conjugate $f^\star(\boldsymbol{y})$. Then the spectral function $g(\boldsymbol{X}) = f \circ \lambda(\boldsymbol{X})$ has Fenchel conjugate $f^\star \circ \lambda(\boldsymbol{Y})$. Hence, if $f(\boldsymbol{x})$ is closed and convex, then $f \circ \lambda(\boldsymbol{X})$ is closed and convex and equals its Fenchel biconjugate.*

Proof: Verification relies heavily on Fan's inequality (A.4.2) proved in Appendix A.4. If we denote the set of $n \times n$ symmetric matrices by S^n, then Fan's inequality implies

$$\begin{aligned}
(f \circ \lambda)^\star(\boldsymbol{Y}) &= \sup_{\boldsymbol{X} \in S^n} [\operatorname{tr}(\boldsymbol{Y}\boldsymbol{X}) - f \circ \lambda(\boldsymbol{X})] \\
&\leq \sup_{\boldsymbol{X} \in S^n} [\lambda(\boldsymbol{Y})^* \lambda(\boldsymbol{X}) - f \circ \lambda(\boldsymbol{X})] \\
&\leq \sup_{\boldsymbol{x} \in \mathsf{R}^n} [\lambda(\boldsymbol{Y})^* \boldsymbol{x} - f(\boldsymbol{x})] \\
&= f^\star[\lambda(\boldsymbol{Y})].
\end{aligned}$$

The reverse inequality follows from the spectral decomposition

$$\boldsymbol{Y} = \boldsymbol{U}^* \operatorname{diag}(\boldsymbol{y}) \boldsymbol{U}$$

of $\boldsymbol{Y}$ in terms of an orthogonal matrix $\boldsymbol{U}$. This leads to

$$\begin{aligned}
f^\star[\lambda(\boldsymbol{Y})] &= \sup_{\boldsymbol{x} \in \mathsf{R}^n} [\lambda(\boldsymbol{Y})^* \boldsymbol{x} - f(\boldsymbol{x})] \\
&= \sup_{\boldsymbol{x} \in \mathsf{R}^n} \{\operatorname{tr}[\boldsymbol{U}\boldsymbol{Y}\boldsymbol{U}^* \operatorname{diag}(\boldsymbol{x})] - f(\boldsymbol{x})\} \\
&= \sup_{\boldsymbol{x} \in \mathsf{R}^n} \{\operatorname{tr}[\boldsymbol{Y}\boldsymbol{U}^* \operatorname{diag}(\boldsymbol{x})\boldsymbol{U}] - f[\lambda(\boldsymbol{U}^* \operatorname{diag}(\boldsymbol{x})\boldsymbol{U})]\} \\
&\leq \sup_{\boldsymbol{X} \in S^n} \{\operatorname{tr}(\boldsymbol{Y}^*\boldsymbol{X}) - f[\lambda(\boldsymbol{X})]\} \\
&= (f \circ \lambda)^\star(\boldsymbol{Y}).
\end{aligned}$$

Thus, the two extremes of each inequality are equal. The second claim of the proposition follows from a double application of Proposition 14.3.2. ■

The maximum eigenvalue $\lambda(\boldsymbol{M})$ of a symmetric matrix is a spectral function. We have already calculated its subdifferential. The next proposition allows us to calculate the subdifferentials of more complicated spectral functions.

Proposition 14.6.2 *Suppose $f(\boldsymbol{x})$ is closed and convex. A symmetric matrix $\boldsymbol{Y}$ belongs to the subdifferential $\partial(f \circ \lambda)(\boldsymbol{X})$ if and only if the ordered vector $\lambda(\boldsymbol{Y})$ belongs to the subdifferential $\partial f[\lambda(\boldsymbol{X})]$ and $\boldsymbol{X}$ and $\boldsymbol{Y}$ have simultaneous ordered spectral decompositions $\boldsymbol{X} = \boldsymbol{U}^* \operatorname{diag} \circ \lambda(\boldsymbol{X})\boldsymbol{U}$ and $\boldsymbol{Y} = \boldsymbol{U}^* \operatorname{diag} \circ \lambda(\boldsymbol{Y})\boldsymbol{U}$.*

Proof: The combination of Proposition 14.6.1, the Fenchel-Young inequality, and Fan's inequality yields

$$\begin{aligned}(f \circ \lambda)(\boldsymbol{X}) + (f \circ \lambda)^\star(\boldsymbol{Y}) &= f[\lambda(\boldsymbol{X})] + f^\star[\lambda(\boldsymbol{Y})] \\ &\geq \lambda(\boldsymbol{X})^* \lambda(\boldsymbol{Y}) \\ &\geq \operatorname{tr}(\boldsymbol{Y}\boldsymbol{X}).\end{aligned}$$

The matrix $\boldsymbol{Y}$ belongs to $\partial(f \circ \lambda)(\boldsymbol{X})$ if and only if the extreme members of these two inequalities agree. The first inequality is an equality if and only if $\lambda(\boldsymbol{Y})$ belongs to the subdifferential $\partial f[\lambda(\boldsymbol{X})]$. As observed in Proposition A.4.2, the second inequality is an equality if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ have simultaneous ordered spectral decompositions. ■

Example 14.6.1 *Sample Calculations with Spectral Functions*

Various matrix norms can be identified as spectral functions. For instance, the nuclear and Euclidean matrix norms originate from the symmetric convex functions

$$f_1(\boldsymbol{x}) = \sum_{i=1}^n |x_i| \quad \text{and} \quad f_2(\boldsymbol{x}) = \max_{1 \leq i \leq n} |x_i|.$$

Example 14.5.7 and Proposition 14.6.2 clearly produce the same subdifferential for the Euclidean matrix norm. The spectral function

$$f_3(\boldsymbol{x}) = \sum_{i=1}^n \max\{-x_i, 0\},$$

arises when negative eigenvalues are downplayed.

The forgoing theory can be applied to penalized approximation of symmetric matrices. Indeed, consider minimization of

$$\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\|_F^2 + \rho f_i \circ \lambda(\boldsymbol{Y}) \tag{14.12}$$

as a function of the symmetric matrix $\boldsymbol{Y}$ for one of the three functions just defined. Fermat's rule requires

$$\boldsymbol{0} \in \boldsymbol{Y} - \boldsymbol{X} + \rho\, \partial(f_i \circ \lambda)(\boldsymbol{Y}).$$

In view of Proposition 14.6.2,

$$\boldsymbol{X} \quad \in \quad \boldsymbol{U}^* \operatorname{diag}(\boldsymbol{y})\boldsymbol{U} + \rho \boldsymbol{U}^* \operatorname{diag}[\partial f_i(\boldsymbol{y})]\boldsymbol{U}$$

for $\boldsymbol{y} = \lambda(\boldsymbol{Y})$ and some orthogonal matrix $\boldsymbol{U}$. It follows that $\boldsymbol{X}$ and $\boldsymbol{Y}$ have simultaneous spectral decompositions under $\boldsymbol{U}$. If we let $\boldsymbol{x} = \lambda(\boldsymbol{X})$ and multiply the relation

$$\boldsymbol{0} \quad \in \quad \boldsymbol{U}^* \operatorname{diag}(\boldsymbol{y})\boldsymbol{U} - \boldsymbol{U}^* \operatorname{diag}(\boldsymbol{x})\boldsymbol{U} + \rho \boldsymbol{U}^* \operatorname{diag}[\partial f_i(\boldsymbol{y})]\boldsymbol{U},$$

by $\boldsymbol{U}$ on the left and $\boldsymbol{U}^*$ on the right, then the relation

$$\boldsymbol{0} \quad \in \quad \boldsymbol{y} - \boldsymbol{x} + \rho \partial f_i(\boldsymbol{y}),$$

emerges. But this is just Fermat's rule for minimizing the spectral function $g(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 + \rho f_i(\boldsymbol{y})$.

The problem of minimizing $g(\boldsymbol{y})$ is separable for the penalty $f_1(\boldsymbol{y})$. Its solution has components

$$y_i \quad = \quad \begin{cases} x_i - \rho & x_i > \rho \\ 0 & |x_i| \le \rho \\ x_i + \rho & x_i < -\rho \end{cases} \tag{14.13}$$

exhibiting shrinkage. For the penalty $f_3(\boldsymbol{y})$, the minimum point of $g(\boldsymbol{y})$ has a solution with components

$$y_i \quad = \quad \begin{cases} x_i & x_i > 0 \\ 0 & -\rho \le x_i \le 0 \\ x_i + \rho & x_i < -\rho \end{cases}$$

exhibiting one-sided shrinkage.

Finding the minimum of $g(\boldsymbol{y})$ for the $f_2(\boldsymbol{y})$ penalty is harder because the objective function is no longer separable. Inspection of $g(\boldsymbol{y})$ shows that the solution must satisfy $\operatorname{sgn}(y_i) = \operatorname{sgn}(x_i)$ and $|y_i| \le |x_i|$ for all i. Thus, there is no loss in generality in assuming that $0 \le y_i \le x_i$ for all i. Instead of exploiting subdifferentials directly, let us focus on forward directional derivatives. At the point $\boldsymbol{y} = \boldsymbol{0}$ an easy calculation shows that

$$d_{\boldsymbol{v}} g(\boldsymbol{0}) \quad = \quad \sum_{i=1}^n (0 - x_i)v_i + \rho \max_{1 \le i \le n} |v_i|.$$

According to Proposition 6.5.2, $\boldsymbol{0}$ furnishes the minimum of $g(\boldsymbol{y})$ if and only if all of these directional derivatives are nonnegative. In view of the generalized Cauchy-Schwarz inequality

$$\left| \sum_{i=1}^n (0 - x_i)v_i \right| \quad \le \quad \|\boldsymbol{x}\|_1 \|\boldsymbol{v}\|_\infty,$$

the vector $\boldsymbol{0}$ qualifies provided $\|\boldsymbol{x}\|_1 \le \rho$.

When $\|\boldsymbol{x}\|_1 > \rho$, there is a simple recipe for constructing the solution $\boldsymbol{y}$. Gradually decrease $r = \|\boldsymbol{y}\|_\infty$ from $\|\boldsymbol{x}\|_\infty$ to 0 until the condition

$$\sum_{x_i \geq r} (x_i - r) \quad = \quad \rho$$

is met. We claim that the vector $\boldsymbol{y}$ with components

$$y_i \quad = \quad \begin{cases} r & x_i \geq r \\ x_i & 0 \leq x_i < r \end{cases}$$

provides the minimum. It suffices to prove that the directional derivative

$$d_{\boldsymbol{v}} g(\boldsymbol{y}) \quad = \quad \sum_{x_i \geq r} (y_i - x_i) v_i + \rho \max_{x_i \geq r} v_i$$

is nonnegative regardless of how we choose $\boldsymbol{v}$. But this fact follows from the inequality

$$\begin{aligned} \sum_{x_i \geq r} (y_i - x_i) v_i &= \sum_{x_i \geq r} (x_i - y_i)(-v_i) \\ &\geq \min_{x_i \geq r} (-v_i) \sum_{x_i \geq r} (x_i - y_i) \\ &= \rho \min_{x_i \geq r} (-v_i) \\ &= -\rho \max_{x_i \geq r} v_i. \end{aligned}$$

Given this solution for $\boldsymbol{x}$ with nonnegative components, it is straightforward to recover the general solution. ■

Example 14.6.2 *Matrix Completion Problem*

Similar calculations are pertinent to matrix completion [37]. Consider a matrix $\boldsymbol{Y} = (y_{ij})$, some of whose entries are unobserved. If Δ denotes the set of index pairs (i, j) such that y_{ij} is observed, then it is convenient to define the projected matrix $P_\Delta(\boldsymbol{Y})$ with entries

$$P_\Delta(\boldsymbol{Y}) \quad = \quad \begin{cases} y_{ij} & (i,j) \in \Delta \\ 0 & (i,j) \notin \Delta. \end{cases}$$

Its orthogonal complement $P_\Delta^\perp(\boldsymbol{Y})$ satisfies $P_\Delta^\perp(\boldsymbol{Y}) + P_\Delta(\boldsymbol{Y}) = \boldsymbol{Y}$. In the matrix completion problem one seeks to minimize the criterion

$$\frac{1}{2} \sum_{(i,j) \in \Delta} (y_{ij} - x_{ij})^2 + \rho \|\boldsymbol{X}\|_*$$

involving the nuclear norm of $\boldsymbol{X} = (x_{ij})$. One way of attacking the problem is to majorize the objective function at the current iterate $\boldsymbol{X}_m$ by

$$g(\boldsymbol{X} \mid \boldsymbol{X}_m) \quad = \quad \frac{1}{2} \|P_\Delta(\boldsymbol{Y}) + P_\Delta^\perp(\boldsymbol{X}_m) - \boldsymbol{X}\|_F^2 + \rho \|\boldsymbol{X}\|_*.$$

In light of the MM principle, minimizing $g(\boldsymbol{X} \mid \boldsymbol{X}_m)$ drives the matrix completion criterion downhill [189].

Now set $\boldsymbol{Z}_m = P_\Delta(\boldsymbol{Y}) + P_\Delta^\perp(\boldsymbol{X}_m)$ and contemplate the expansion

$$g(\boldsymbol{X} \mid \boldsymbol{X}_m) = \frac{1}{2}\|\boldsymbol{Z}_m\|_F^2 - \operatorname{tr}(\boldsymbol{Z}_m^*\boldsymbol{X}) + \frac{1}{2}\|\boldsymbol{X}\|_F^2 + \rho\|\boldsymbol{X}\|_*.$$

Suppose $\boldsymbol{X}$ has svd $\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{Q}^*$ and $\boldsymbol{Z}_m$ has svd $\boldsymbol{U}\boldsymbol{\Omega}\boldsymbol{V}^*$. According to Fan's inequality as discussed in Appendix A.5, $-\operatorname{tr}(\boldsymbol{Z}_m^*\boldsymbol{X}) \geq -\sum_i \omega_i\sigma_i$, with equality when $\boldsymbol{U} = \boldsymbol{P}$ and $\boldsymbol{V} = \boldsymbol{Q}$. Here, the ω_i and σ_i are the ordered singular values of $\boldsymbol{X}$ and $\boldsymbol{Z}_m$, respectively. Furthermore, neither the Frobenius norm $\|\boldsymbol{X}\|_F = (\sum_i \sigma_i^2)^{1/2}$ nor the nuclear norm $\|\boldsymbol{X}\|_* = \sum_i \sigma_i$ depends on the orthogonal matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ of singular vectors. Hence, it is optimal to take $\boldsymbol{P} = \boldsymbol{U}$ and $\boldsymbol{Q} = \boldsymbol{V}$. The problem therefore becomes one of minimizing

$$\frac{1}{2}\sum_i (\omega_i - \sigma_i)^2 + \rho\sum_i \sigma_i$$

subject to the nonnegativity constraints $\sigma_i \geq 0$. The solution is given by equation (14.13) with $x_i = \omega_i \geq 0$ and $y_i = \sigma_i$. In practice only the largest singular values of Ω need be extracted. The Lanczos procedure [107] efficiently computes these and their corresponding singular vectors.

A related problem is to minimize $\|\boldsymbol{X}\|_*$ subject to the quadratic constraint $\frac{1}{2}\sum_{(i,j)\in\Delta}(y_{ij} - x_{ij})^2 \leq \epsilon$ [35]. This problem can be recast as minimizing the penalized convex function $\|\boldsymbol{X}\|_* + \delta_C(\boldsymbol{X})$, where C is the closed convex set

$$C = \Big\{\boldsymbol{X} = (x_{ij}) : \frac{1}{2}\sum_{(i,j)\in\Delta}(y_{ij} - x_{ij})^2 \leq \epsilon\Big\}.$$

To derive an MM algorithm, suppose that the current iterate is $\boldsymbol{X}_m$ and that $\boldsymbol{Z}_m = P_\Delta(\boldsymbol{Y}) + P_\Delta^\perp(\boldsymbol{X}_m)$. If we define the closed convex set

$$C_m = \Big\{\boldsymbol{X} = (x_{ij}) : \frac{1}{2}\sum_i\sum_j (z_{mij} - x_{ij})^2 \leq \epsilon\Big\},$$

then it is obvious that $\|\boldsymbol{X}\|_* + \delta_{C_m}(\boldsymbol{X})$ majorizes $\|\boldsymbol{X}\|_* + \delta_C(\boldsymbol{X})$ around the point $\boldsymbol{X}_m$. In this majorization we allow infinite function values.

Minimization of the surrogate function $\|\boldsymbol{X}\|_* + \delta_{C_m}(\boldsymbol{X})$ again relies on the fact that the nuclear and Frobenius norms are invariant under left and right multiplication of their arguments by orthogonal matrices. As we have just argued, we can assume that $\boldsymbol{X}$ and $\boldsymbol{Z}_m$ have svds $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ and $\boldsymbol{U}\boldsymbol{\Omega}\boldsymbol{V}^*$ involving shared orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$. Thus, the current problem reduces to minimizing $\sum_i \sigma_i$ subject to $\frac{1}{2}\sum_i(\omega_i - \sigma_i)^2 \leq \epsilon$ and $\sigma_i \geq 0$ for all i. Of course, the singular values ω_i are nonnegative as well. If $\frac{1}{2}\sum_i \omega_i^2 \leq \epsilon$, then the trivial solution $\sigma_i = 0$ for all i holds.

Hence, suppose that $\frac{1}{2}\sum_i \omega_i^2 > \epsilon$ and form the Lagrangian

$$\mathcal{L}(\boldsymbol{\sigma}, \lambda, \boldsymbol{\mu}) \;=\; \sum_i \sigma_i + \lambda\Big[\frac{1}{2}\sum_i (\omega_i - \sigma_i)^2 - \epsilon\Big] - \sum_i \mu_i \sigma_i.$$

If we assume $\frac{1}{2}\sum_i(\omega_i - \sigma_i)^2 = \epsilon$ and $\lambda > 0$, then the stationarity condition

$$0 \;=\; 1 + \lambda(\sigma_i - \omega_i) - \mu_i$$

and complementary slackness imply

$$\sigma_i \;=\; \begin{cases} \omega_i - \lambda^{-1} & \omega_i - \lambda^{-1} > 0 \\ 0 & \omega_i - \lambda^{-1} \leq 0 \end{cases}.$$

The condition $\frac{1}{2}\sum_i(\omega_i - \sigma_i)^2 = \epsilon$ then amounts to the identity

$$\frac{1}{2}\sum_i \min\{\omega_i, \lambda^{-1}\}^2 \;=\; \epsilon.$$

The continuous function $f(\beta) = \frac{1}{2}\sum_i \min\{\omega_i, \beta\}^2$ is strictly increasing on the interval $[0, \max_i \omega_i]$. Since $f(0) = 0$ and $f(\max_i \omega_i) > \epsilon$ by assumption, there is a unique positive solution. For n singular values, the solution belongs to the interval $[\omega_k, \omega_{k+1}]$ if and only if

$$\sum_{i=1}^k \omega_i^2 + (n-k)\omega_k^2 \;\leq\; 2\epsilon \;\leq\; \sum_{i=1}^k \omega_i^2 + (n-k)\omega_{k+1}^2.$$

Because this is equivalent to

$$(n-k)\omega_k^2 \;\leq\; 2\epsilon - \sum_i \omega_i^2 + \sum_{i>k} \omega_i^2 \;\leq\; (n-k)\omega_{k+1}^2$$

and $\sum_i \omega_i^2 = \|\boldsymbol{Z}_m\|_F^2$, the solution again depends on only the largest singular values. ■

Example 14.6.3 *Stable Estimation of a Covariance Matrix*

Example 6.5.7 demonstrates that the sample mean and sample covariance matrix

$$\begin{aligned} \bar{\boldsymbol{y}} &= \frac{1}{k}\sum_{j=1}^k \boldsymbol{y}_j \\ \boldsymbol{S} &= \frac{1}{k}\sum_{j=1}^k (\boldsymbol{y}_j - \bar{\boldsymbol{y}})(\boldsymbol{y}_j - \bar{\boldsymbol{y}})^* \end{aligned}$$

are the maximum likelihood estimates of the theoretical mean $\boldsymbol{\mu}$ and theoretical covariance $\boldsymbol{\Omega}$ of a random sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ from a multivariate

normal distribution. When the number of components n of $\boldsymbol{y}$ exceeds the sample size k, this analysis breaks down because it assumes that $\boldsymbol{S}$ is invertible. To deal with this dilemma and to stabilize estimation generally, we add a penalty. The estimate of $\boldsymbol{\mu}$ remains $\bar{\boldsymbol{y}}$.

As in Sect. 13.4, we impose a prior $p(\boldsymbol{\Omega})$ on $\boldsymbol{\Omega}$. Now, however, the prior is designed to steer the eigenvalues of $\boldsymbol{\Omega}$ away from the extremes of 0 and ∞. The reasonable choice

$$p(\boldsymbol{\Omega}) \quad \propto \quad e^{-\frac{\lambda}{2}\left[\alpha\|\boldsymbol{\Omega}\|_*+(1-\alpha)\|\boldsymbol{\Omega}^{-1}\|_*\right]},$$

relies on the nuclear norms of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}^{-1}$, a positive strength constant λ, and an admixture constant $\alpha \in (0,1)$. This is a proper prior on the set of invertible matrices because

$$e^{-\frac{\lambda}{2}\left[\alpha\|\boldsymbol{\Omega}\|_*+(1-\alpha)\|\boldsymbol{\Omega}^{-1}\|_*\right]} \quad \leq \quad e^{-c\lambda\|\boldsymbol{\Omega}\|_F}$$

for some positive constant c by virtue of the equivalence of any two vector norms on R^{n^2}. The normalizing constant of $p(\boldsymbol{\Omega})$ is irrelevant in the ensuing discussion. Consider therefore minimization of the function

$$f(\boldsymbol{\Omega}) \quad = \quad \frac{k}{2}\ln\det\boldsymbol{\Omega}+\frac{k}{2}\operatorname{tr}(\boldsymbol{S}\boldsymbol{\Omega}^{-1})+\frac{\lambda}{2}\left[\alpha\|\boldsymbol{\Omega}\|_*+(1-\alpha)\|\boldsymbol{\Omega}^{-1}\|_*\right].$$

The maximum of $-f(\boldsymbol{\Omega})$ occurs at the posterior mode. In the limit as λ tends to 0, $-f(\boldsymbol{\Omega})$ reduces to the loglikelihood.

Fortunately, three of the four terms of $f(\boldsymbol{\Omega})$ can be expressed as functions of the eigenvalues e_i of $\boldsymbol{\Omega}$. The trace contribution presents a greater challenge. Let $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^*$ denote the spectral decomposition of $\boldsymbol{S}$ with nonnegative diagonal entries d_i ordered from largest to smallest. Likewise, let $\boldsymbol{\Omega} = \boldsymbol{V}\boldsymbol{E}\boldsymbol{V}^*$ denote the spectral decomposition of $\boldsymbol{\Omega}$ with positive diagonal entries e_i ordered from largest to smallest. In view of Fan's inequality (A.6), we can assert that

$$-\operatorname{tr}(\boldsymbol{S}\boldsymbol{\Omega}^{-1}) \quad \leq \quad -\sum_{i=1}^{n}\frac{d_i}{e_i},$$

with equality if and only if $\boldsymbol{V} = \boldsymbol{U}$. Consequently, we make the latter assumption and replace $f(\boldsymbol{\Omega})$ by

$$g(\boldsymbol{E}) \quad = \quad \frac{k}{2}\sum_{i=1}^{n}\ln e_i+\frac{k}{2}\sum_{i=1}^{n}\frac{d_i}{e_i}+\frac{\lambda}{2}\left[\alpha\sum_{i=1}^{n}e_i+(1-\alpha)\sum_{i=1}^{n}\frac{1}{e_i}\right].$$

At a stationary point of $g(\boldsymbol{E})$, we have

$$0 \quad = \quad \frac{k}{e_i}-\frac{kd_i+\lambda(1-\alpha)}{e_i^2}+\lambda\alpha.$$

The solution to this essentially quadratic equation is

$$e_i = \frac{-k + \sqrt{k^2 + 4\lambda\alpha[kd_i + \lambda(1-\alpha)]}}{2\lambda\alpha}. \tag{14.14}$$

We reject the negative root as inconsistent with $\boldsymbol{\Omega}$ being positive definite. For the special case $k = 0$ of no data, all $e_i = \sqrt{(1-\alpha)/\alpha}$, and the prior mode occurs at a multiple of the identity matrix.

Holding all but one variable fixed in formula (14.14), one can demonstrate after a fair amount of algebra that

$$\begin{aligned} e_i &= d_i + \frac{\lambda(1-\alpha-\alpha d_i^2)}{k} + O\left(\frac{1}{k^2}\right), \quad k \to \infty \\ e_i &= \sqrt{\frac{1-\alpha}{\alpha}} + \left[\sqrt{\frac{1-\alpha}{\alpha}}\frac{kd_i}{2(1-\alpha)} - \frac{k}{2\alpha}\right]\frac{1}{\lambda} + O\left(\frac{1}{\lambda^2}\right), \quad \lambda \to \infty. \end{aligned}$$

These asymptotic expansions accord with common sense. Namely, the data eventually overwhelms a fixed prior, and increasing the penalty strength for a fixed amount of data pulls the estimate of $\boldsymbol{\Omega}$ toward the prior mode. Choice of the constants λ and α is an issue. To match the prior to the scale of the data, we recommend determining α as the solution to the equation

$$n\sqrt{\frac{1-\alpha}{\alpha}} = \operatorname{tr}\left(\sqrt{\frac{1-\alpha}{\alpha}}\boldsymbol{I}\right) = \operatorname{tr}(\boldsymbol{S}).$$

Cross validation leads to a reasonable choice of λ. For the sake of brevity, we omit further details. For a summary of other approaches to this subject, consult the reference [173]. ■

14.7 A Convex Lagrange Multiplier Rule

The Lagrange multiplier rule is one of the dominant themes of optimization theory. In convex programming it represents both a necessary and sufficient condition for a minimum point. Our previous proofs of the rule invoked differentiability. In fact, differentiability assumptions can be dismissed in deriving the multiplier rule for convex programs. Recall that we posed convex programming as the problem of minimizing a convex function $f(\boldsymbol{y})$ subject to the constraints

$$g_i(\boldsymbol{y}) = 0, \quad 1 \le i \le p, \qquad h_j(\boldsymbol{y}) \le 0, \quad 1 \le j \le q,$$

where the $g_i(\boldsymbol{y})$ are affine functions and the $h_j(\boldsymbol{y})$ are convex functions. One can simplify the statement of many convex programs by intersecting the essential domains of the objective and constraint functions with a

closed convex set C. For example, although the set of positive semidefinite matrices is convex, it is awkward to represent it as an intersection of convex sets determined by affine equality constraints and simple convex inequality constraints.

In proving the multiplier rule anew, we will call on Slater's constraint qualification. In the current context, this entails postulating the existence of a point $\boldsymbol{z} \in C$ such that $g_i(\boldsymbol{z}) = 0$ for all i and $h_j(\boldsymbol{z}) < 0$ for all j. In addition we assume that the constraint gradient vectors $\nabla g_i(\boldsymbol{y})$ are linearly independent. These preliminaries put us into position to restate and prove the Lagrange multiplier rule for convex programs.

Proposition 14.7.1 *Suppose that $f(\boldsymbol{y})$ achieves its minimum subject to the constraints at the point $\boldsymbol{x} \in C$. Then there exists a Lagrangian function*

$$\mathcal{L}(\boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad = \quad \lambda_0 f(\boldsymbol{y}) + \sum_{i=1}^{p} \lambda_i g_i(\boldsymbol{y}) + \sum_{j=1}^{q} \mu_j h_j(\boldsymbol{y})$$

characterized by the following three properties: (a) $\mathcal{L}(\boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ achieves its minimum over C at $\boldsymbol{x}$, (b) the multipliers λ_0 and μ_j are nonnegative, and (c) the complementary slackness conditions $\mu_j h_j(\boldsymbol{x}) = 0$ hold. If Slater's constraint qualification is true and $\boldsymbol{x}$ is an interior point of C, then one can take $\lambda_0 = 1$. Conversely, if properties (a) through (c) hold with $\lambda_0 = 1$, then $\boldsymbol{x}$ minimizes $f(\boldsymbol{y})$ subject to the constraints.

Proof: To prove the necessity of the three properties, we separate a convex set and a point by a hyperplane. Accordingly, define the set S to consist of all points $\boldsymbol{u} \in \mathsf{R}^{p+q+1}$ such that for some $\boldsymbol{y} \in C$, we have $u_0 \geq f(\boldsymbol{y})$, $u_i = g_i(\boldsymbol{y})$ for $1 \leq i \leq p$, and $u_{p+j} \geq h_j(\boldsymbol{y})$ for $1 \leq j \leq q$. To show that the set S is convex, suppose $\boldsymbol{y} \in C$ corresponds to $\boldsymbol{u} \in S$, $\boldsymbol{z} \in C$ corresponds to $\boldsymbol{v} \in S$, and α and β are nonnegative numbers summing to 1. The relations

$$\begin{aligned}
f(\alpha\boldsymbol{y} + \beta\boldsymbol{z}) &\leq \alpha f(\boldsymbol{y}) + \beta f(\boldsymbol{z}) \leq \alpha u_0 + \beta v_0 \\
g_i(\alpha\boldsymbol{y} + \beta\boldsymbol{z}) &= \alpha g_i(\boldsymbol{y}) + \beta g_i(\boldsymbol{z}) = \alpha u_i + \beta v_i, \ 1 \leq i \leq p \\
h_j(\alpha\boldsymbol{y} + \beta\boldsymbol{z}) &\leq \alpha h_j(\boldsymbol{y}) + \beta h_j(\boldsymbol{z}) \leq \alpha u_{p+j} + \beta v_{p+j}, \ 1 \leq j \leq q
\end{aligned}$$

and the convexity of C prove the convexity claim. The point to be separated from S is $[f(\boldsymbol{x}), \mathbf{0}^*]^*$. It belongs to S because $\boldsymbol{x}$ is feasible. It lies on the boundary of S because the point $[f(\boldsymbol{x}) - \epsilon, \mathbf{0}^*]^*$ does not belong to S for any $\epsilon > 0$.

Application of Proposition 6.2.3 shows that there exists a nontrivial vector $\boldsymbol{\omega}$ such that $\omega_0 f(\boldsymbol{x}) \leq \boldsymbol{\omega}^* \boldsymbol{u}$ for all $\boldsymbol{u} \in S$. Identify the entries of $\boldsymbol{\omega}$ with $\lambda_0, \lambda_1, \ldots, \lambda_p, \mu_1, \ldots, \mu_q$, in that order. Sending u_0 to ∞ implies $\lambda_0 \geq 0$; similarly, sending u_{p+j} to ∞ implies $\mu_j \geq 0$. If $h_j(\boldsymbol{x}) < 0$, then the vector $\boldsymbol{u} \in S$ with $f(\boldsymbol{x})$ as entry 0, $h_j(\boldsymbol{x})$ as entry $p+j$, and 0's elsewhere demonstrates that $\mu_j = 0$. This proves properties (b) and (c). To verify property

(a), take $\boldsymbol{y} \in C$ and put $u_0 = f(\boldsymbol{y})$, $u_i = g_i(\boldsymbol{y})$, and $u_{p+j} = h_j(\boldsymbol{y})$. Then $\boldsymbol{u} \in S$, and the separating hyperplane condition reads

$$\mathcal{L}(\boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad \geq \quad \lambda_0 f(\boldsymbol{x}) \quad = \quad \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}),$$

proving property (a).

Next suppose $\lambda_0 = 0$ and $\boldsymbol{z}$ is a Slater point. The inequality

$$0 \quad = \quad \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad \leq \quad \mathcal{L}(\boldsymbol{z}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad = \quad \sum_{j=1}^{q} \mu_j h_j(\boldsymbol{z})$$

is inconsistent with at least one μ_j being positive and all $h_j(\boldsymbol{z})$ being negative. Hence, it suffices to assume all $\mu_j = 0$. For any $\boldsymbol{y} \in C$ we now find that

$$0 \quad = \quad \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad \leq \quad \mathcal{L}(\boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad = \quad \sum_{i=1}^{p} \lambda_i g_i(\boldsymbol{y}).$$

Let $a(\boldsymbol{y})$ denote the affine function $\sum_{i=1}^{p} \lambda_i g_i(\boldsymbol{y})$. We have just shown that $a(\boldsymbol{y}) \geq 0$ for all $\boldsymbol{y}$ in a neighborhood of $\boldsymbol{x}$. Because $a(\boldsymbol{x}) = 0$, Fermat's rule requires the vector $\nabla a(\boldsymbol{x})$ to vanish. Finally, the fact that some of the λ_i are nonzero contradicts the assumed linear independence of the gradient vectors $\nabla g_i(\boldsymbol{x})$. The only possibility left is $\lambda_0 > 0$. Divide $\mathcal{L}(\boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ by λ_0 to achieve the canonical form of the multiplier rule.

Finally for the converse, let $\boldsymbol{y} \in C$ be any feasible point. The inequalities

$$f(\boldsymbol{x}) \quad = \quad \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad \leq \quad \mathcal{L}(\boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad \leq \quad f(\boldsymbol{y})$$

follow directly from properties (a) through (c) and demonstrate that $\boldsymbol{x}$ furnishes the constrained minimum of $f(\boldsymbol{y})$. ■

Proposition 14.7.1 falls short of our expectations in the sense that the usual multiplier rule involves a stationarity condition. For convex programs, the required gradients do not necessarily exist. Fortunately, subgradients provide a suitable substitute. One can better understand the situation by exploiting the fact that $\boldsymbol{x}$ minimizes the Lagrangian over the convex set C. This suggests that we replace the Lagrangian by the related function

$$\mathcal{K}(\boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad = \quad \mathcal{L}(\boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\mu}) + \delta_C(\boldsymbol{y}).$$

Because $\mathcal{K}(\boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ achieves its unconstrained minimum over R^n at $\boldsymbol{x}$, we have $\mathbf{0} \in \partial \mathcal{K}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$. If the sum rule for subdifferentials is justified, then in view of Example 14.4.1 we recover the generalization

$$\mathbf{0} \quad \in \quad \partial f(\boldsymbol{x}) + \sum_{i=1}^{p} \lambda_i \partial g_i(\boldsymbol{x}) + \sum_{j=1}^{q} \mu_j \partial h_j(\boldsymbol{x}) + N_C(\boldsymbol{x})$$

of the usual multiplier rule. As we have previously stressed, simple sufficient conditions ensure the validity of the sum rule.

As an example, consider minimizing the function

$$f(\boldsymbol{x}) \quad = \quad \sum_i a_i|x_i| + \sum_i b_i x_i$$

over the closed unit ball, where each constant $a_i \geq 0$. The multiplier conditions are

$$0 \quad \in \quad b_i + \mu x_i + a_i \begin{cases} 1 & x_i > 0 \\ [-1,1] & x_i = 0 \\ -1 & x_i < 0. \end{cases}$$

If all $|b_i| \leq a_i$, then these are satisfied at the origin $\mathbf{0}$ with the choice $\mu = 0$ dictated by complementary slackness. If any $|b_i| > a_i$, then we take μ to be the positive square root of

$$\mu^2 \quad = \quad \sum_{|b_i|>a_i} [b_i - a_i \operatorname{sgn}(b_i)]^2.$$

Those components x_i with $|b_i| \leq a_i$ we assign the value 0. The remaining components we assign the value

$$x_i \quad = \quad \frac{1}{\mu}[-b_i + a_i \operatorname{sgn}(b_i)].$$

Because the function $f(\boldsymbol{x})$ is homogeneous and its minimum is negative, the optimal point $\boldsymbol{x}$ occurs on the boundary of the unit ball.

14.8 Problems

1. Derive the Fenchel conjugates displayed in Table 14.1 for functions on the real line.

2. Show that the function $f_1(x) = (x^2 - 1)^2$ has Fenchel biconjugate

$$f_1^{\star\star}(x) \quad = \quad \begin{cases} 0 & |x| \leq 1 \\ (x^2-1)^2 & |x| > 1 \end{cases}$$

and that the function

$$f_2(x) \quad = \quad \begin{cases} |x| & |x| \leq 1 \\ 2 - |x| & 1 < |x| \leq 3/2 \\ |x| - 1 & |x| > 3/2 \end{cases}$$

has Fenchel biconjugate

$$f_2^{\star\star}(x) \quad = \quad \begin{cases} \frac{|x|}{3} & |x| \leq 3/2 \\ f_2(x) & |x| > 3/2\,. \end{cases}$$

TABLE 14.1. Some specific Fenchel conjugates

$f(x)$	$f^\star(y)$
x	$\begin{cases} 0 & y = 1 \\ \infty & y \neq 1 \end{cases}$
$\lvert x \rvert$	$\begin{cases} 0 & \lvert y \rvert \leq 1 \\ \infty & \lvert y \rvert > 1 \end{cases}$
$\begin{cases} \frac{1}{x} & x > 0 \\ \infty & x \leq 0 \end{cases}$	$\begin{cases} \infty & y > 0 \\ -2\sqrt{-y} & y \leq 0 \end{cases}$
$\begin{cases} x \ln x & x > 0 \\ \infty & x \leq 0 \end{cases}$	e^{y-1}
e^x	$\begin{cases} y \ln y - y & y > 0 \\ 0 & y = 0 \\ \infty & y < 0 \end{cases}$
$\begin{cases} x \ln x + (1-x)\ln(1-x) & x \in (0,1) \\ 0 & x \in \{0,1\} \\ \infty & \text{otherwise} \end{cases}$	$\ln(1 + e^y)$

(Hints: There is no need to calculate the conjugate $f_i^\star(y)$. According to inequality (14.2), $f_i^{\star\star}(x)$ falls below $f_i(x)$. It also falls above any line supporting $f_i(x)$.)

3. Prove the Fenchel-Young inequality $xy \leq e^x - y + y \ln y$ for x and y nonnegative.

4. Find the support function of the convex set $C = \{\boldsymbol{x} : x_2 + e^{x_1} \leq 0\}$ in R^2. Explain why this is pertinent to Problem 23 of Chap. 6.

5. Show that the Fenchel conjugate of $f(\boldsymbol{x}) = \max_{1 \leq i \leq n} x_i$ is

$$f^\star(\boldsymbol{y}) = \begin{cases} 0 & \text{all } y_i \geq 0 \text{ and } \sum_{i=1}^n y_i = 1 \\ \infty & \text{otherwise .} \end{cases}$$

6. Assume $f(x)$ is a continuous, strictly increasing function with functional inverse $g(y)$ and value $f(0) = 0$. Show that the even functions defined by

$$F(x) = \int_0^x f(u)du \quad \text{and} \quad G(y) = \int_0^y g(v)dv$$

for $x \geq 0$ and $y \geq 0$ constitute a conjugate pair. Why is Example 1.2.6 a special case? Give a graphical interpretation of Young's inequality $xy \leq F(x) + G(y)$. (Hint: Equality holds in Young's inequality when $x = g(y)$. Interpret graphically and in terms of $F^\star(y)$.)

7. Suppose $f(\boldsymbol{x})$ is a convex function and $\boldsymbol{u} \in \partial f(\boldsymbol{x})$ and $\boldsymbol{v} \in \partial f(\boldsymbol{y})$. Prove that $(\boldsymbol{u}-\boldsymbol{v})^*(\boldsymbol{x}-\boldsymbol{y}) \geq 0$.

8. For some orthogonal matrix $\boldsymbol{O}$, suppose $f(\boldsymbol{x}) = f(\boldsymbol{O}\boldsymbol{x})$ for all $\boldsymbol{x}$. Prove that $f^\star(\boldsymbol{y}) = f^\star(\boldsymbol{O}\boldsymbol{y})$ for all $\boldsymbol{y}$ as well.

9. Suppose the continuous function $f(\boldsymbol{x})$ defined on R^n satisfies the condition $\lim_{\|\boldsymbol{x}\|\to\infty} \|\boldsymbol{x}\|^{-1} f(\boldsymbol{x}) = \infty$. Show that $f^\star(\boldsymbol{y}) < \infty$ for all $\boldsymbol{y}$.

10. Demonstrate that $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$ is the only function satisfying the identity $f(\boldsymbol{x}) = f^\star(\boldsymbol{x})$ for all $\boldsymbol{x}$. (Hint: Use the trivial inequality $f^\star(\boldsymbol{y}) + f(\boldsymbol{x}) \geq \boldsymbol{y}^*\boldsymbol{x}$ to prove that $f(\boldsymbol{x}) \geq \frac{1}{2}\|\boldsymbol{x}\|^2$ when $f(\boldsymbol{x}) = f^\star(\boldsymbol{x})$. For the reverse inequality, substitute this inequality in the definition of $f^\star(\boldsymbol{y})$.)

11. Let $f(\boldsymbol{y})$ be a differentiable function from R^n to R. Prove that $\boldsymbol{x}$ is a global minimum of $f(\boldsymbol{y})$ if and only if $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$ and $f^{\star\star}(\boldsymbol{x}) = f(\boldsymbol{x})$ [130]. (Hints: If $\boldsymbol{x}$ is a global minimum, then $\boldsymbol{0}$ is a subgradient of $f(\boldsymbol{y})$ at $\boldsymbol{x}$. Conversely, if the two conditions hold, then show that every directional derivative $d_{\boldsymbol{v}} f(\boldsymbol{x})$ satisfies $d_{\boldsymbol{v}} f^{\star\star}(\boldsymbol{x}) \leq 0$. Because $-d_{-\boldsymbol{v}} f^{\star\star}(\boldsymbol{x}) \leq d_{\boldsymbol{v}} f^{\star\star}(\boldsymbol{x})$, we have in fact $d_{\boldsymbol{v}} f^{\star\star}(\boldsymbol{x}) = 0$ for every direction $\boldsymbol{v}$. Now use Problem 21 of Chap. 4 to establish that $\nabla f^{\star\star}(\boldsymbol{x}) = \boldsymbol{0}$. Because $f^{\star\star}(\boldsymbol{y})$ is convex, $\boldsymbol{x}$ minimizes $f^{\star\star}(\boldsymbol{y})$.)

12. Let the convex function $f(\boldsymbol{x}, \boldsymbol{y})$ have Fenchel conjugate $f^\star(\boldsymbol{u}, \boldsymbol{v})$. Demonstrate that the function $g(\boldsymbol{x}) = \inf_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$ has Fenchel conjugate $f^\star(\boldsymbol{u}, \boldsymbol{0})$.

13. Let $B = \{\boldsymbol{x} : \|\boldsymbol{x}\|_\dagger \leq 1\}$ be the closed unit ball associated with a norm $\|\boldsymbol{x}\|_\dagger$ on R^n. Show that $\|\boldsymbol{y}\|_\star = \delta_B^\star(\boldsymbol{y}) = \sup_{\boldsymbol{x}\in B} \boldsymbol{y}^*\boldsymbol{x}$ also qualifies as a norm.

14. Assume $f(t)$ is a proper even function from R to $(-\infty, \infty]$. Let $\|\boldsymbol{x}\|_\dagger$ be a norm on R^n with dual norm $\|\boldsymbol{y}\|_\star$. Prove that the composite function $g(\boldsymbol{x}) = f(\|\boldsymbol{x}\|_\dagger)$ has Fenchel conjugate $g^\star(\boldsymbol{y}) = f^\star(\|\boldsymbol{y}\|_\star)$. We have already considered the special case of the self-conjugate function $f(t) = \frac{1}{2}t^2$. (Hint: $g^\star(\boldsymbol{y}) = \sup_{t\geq o} \sup_{\|\boldsymbol{x}\|=t} [\boldsymbol{y}^*\boldsymbol{x} - f(t)]$.)

15. The infimal convolution of two convex functions $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ is defined by $(f \odot g)(\boldsymbol{x}) = \inf_{\boldsymbol{w}} [f(\boldsymbol{w}) + g(\boldsymbol{x}-\boldsymbol{w})]$. Prove that $(f \odot g)(\boldsymbol{x})$ is convex and has Fenchel conjugate $f^\star(\boldsymbol{y}) + g^\star(\boldsymbol{y})$. Calculate $(f \odot g)(\boldsymbol{x})$ when $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ and $g(\boldsymbol{x}) = \delta_U(\boldsymbol{y})$ for a nonempty set U. What is the Fenchel conjugate of this particular infimal convolution?

16. Suppose that the convex function $f(\boldsymbol{x})$ is coercive and twice continuously differentiable with $d^2 f(\boldsymbol{x})$ positive definite for all $\boldsymbol{x}$. Argue via the implicit function theorem that the stationarity equation $\boldsymbol{0} = \boldsymbol{y} - \nabla f(\boldsymbol{x})$ can be solved for $\boldsymbol{x}$ in terms of $\boldsymbol{y}$. Furthermore, show that

$\boldsymbol{x}(\boldsymbol{y}) = \nabla f^\star(\boldsymbol{y})$ by differentiating the equation $f^\star(\boldsymbol{y}) = \boldsymbol{y}^*\boldsymbol{x}(\boldsymbol{y}) - f[\boldsymbol{x}(\boldsymbol{y})]$. Finally, show that $d^2 f^\star(\boldsymbol{y})$ exists and equals the matrix inverse of $d^2 f[\boldsymbol{x}(\boldsymbol{y})]$. (Hints: For the last assertion, consider the equation $\nabla f[\nabla f^\star(\boldsymbol{y})] = \boldsymbol{y}$, and apply Problem 20 of Chap. 4. Section 12.3 defines coercive functions.)

17. Let $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x}$ for $\boldsymbol{A}$ positive semidefinite but not positive definite. Prove that $f^\star(\boldsymbol{A}\boldsymbol{z}+\boldsymbol{b}) = \frac{1}{2}\boldsymbol{z}^*\boldsymbol{A}\boldsymbol{z}$ and $f^\star(\boldsymbol{y}) = \infty$ if $\boldsymbol{y}$ cannot be represented as $\boldsymbol{A}\boldsymbol{z} + \boldsymbol{b}$ for some $\boldsymbol{z}$. (Hint: Invoke Problem 28 of Chap. 5.)

18. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be positive semidefinite matrices of the same dimension. Show that the matrix $a\boldsymbol{A} + b\boldsymbol{B}$ is positive semidefinite for every pair of nonnegative scalars a and b. Thus, the set of positive semidefinite matrices is a convex cone. Why is it a closed set as well?

19. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be positive definite matrices of the same dimension. We write $\boldsymbol{A} \succeq \boldsymbol{B}$ provided $\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{x}^*\boldsymbol{B}\boldsymbol{x}$ for all vectors $\boldsymbol{x}$. Demonstrate via the Fenchel conjugate that $\boldsymbol{A} \succeq \boldsymbol{B}$ implies $\boldsymbol{B}^{-1} \succeq \boldsymbol{A}^{-1}$.

20. Show that the two closed convex cones

$$\begin{aligned} C_1 &= \{\boldsymbol{x} \in \mathsf{R}^n : x_i \geq 0, \ \forall i\} \\ C_2 &= \{\boldsymbol{x} \in \mathsf{R}^n : x_i \leq x_{i+1}, \ \forall i < n\} \end{aligned}$$

have the polar cones

$$\begin{aligned} C_1^\circ &= \{\boldsymbol{y} \in \mathsf{R}^n : y_i \leq 0, \ \forall i\} \\ C_2^\circ &= \{\boldsymbol{y} \in \mathsf{R}^n : \sum_{i=1}^{j} y_i \geq 0, \ \forall j < n, \text{ and } \sum_{i=1}^{n} y_i = 0\}. \end{aligned}$$

21. For a closed convex set C and $\boldsymbol{x} \in C$, prove that the normal cone

$$N_C(\boldsymbol{x}) = \{\boldsymbol{y} : P_C(\boldsymbol{x} + \boldsymbol{y}) = \boldsymbol{x}\},$$

where $P_C(\boldsymbol{z})$ projects a point $\boldsymbol{z}$ onto the closest point in C.

22. Let C be a closed convex cone in R^n. Find the normal cone $N_C(\boldsymbol{x})$ when $\boldsymbol{x} \in C$.

23. Let C be a closed convex cone, For any point $\boldsymbol{x}$, verify the representation $\boldsymbol{x} = P_C(\boldsymbol{x}) + P_{C^\circ}(\boldsymbol{x})$ with $P_C(\boldsymbol{x})^* P_{C^\circ}(\boldsymbol{x}) = 0$. Here P_K denotes projection onto the closed convex set K. (Hints: Let $\boldsymbol{a} = P_C(\boldsymbol{x})$ and $\boldsymbol{b} = P_{C^\circ}(\boldsymbol{x})$. Prove that $(r\boldsymbol{a}-\boldsymbol{a})^*(\boldsymbol{x}-\boldsymbol{a}) \leq 0$ and $(r\boldsymbol{b}-\boldsymbol{b})^*(\boldsymbol{x}-\boldsymbol{b}) \leq 0$ for $r \in \{0, 2\}$. Next show that $\boldsymbol{x}-\boldsymbol{a} \in C^\circ$ and that $\boldsymbol{x}-\boldsymbol{b} \in C$. Finally, show that $\boldsymbol{x} - \boldsymbol{a} = \boldsymbol{b}$. Throughout apply the obtuse angle criterion of Example 6.5.3 and the definition of a polar cone.)

24. Suppose the $n \times n$ matrix $\boldsymbol{A}$ is positive definite. Show that the function $\|\boldsymbol{x}\|_\dagger = \sqrt{\boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x}}$ is a norm. Find its dual norm. (Hint: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^* \boldsymbol{A} \boldsymbol{y}$ defines an inner product.)

25. A norm $\|\boldsymbol{x}\|_\dagger$ is said to be strictly convex if the conditions

$$\|\boldsymbol{x}\|_\dagger \;=\; \|\boldsymbol{y}\|_\dagger \;=\; \left\|\frac{1}{2}(\boldsymbol{x}+\boldsymbol{y})\right\|_\dagger \;=\; 1$$

imply $\boldsymbol{x} = \boldsymbol{y}$. Which of the norms $\|\boldsymbol{x}\|_1$, $\|\boldsymbol{x}\|$, and $\|\boldsymbol{x}\|_\infty$ is strictly convex? When $\|\boldsymbol{x}\|_\dagger$ is strictly convex, prove that the closest point in a closed convex set C to an external point $\boldsymbol{z}$ is unique. (Hint: Reduce the problem to the case $\boldsymbol{z} = \boldsymbol{0}$.)

26. Show that a closed convex function $f(\boldsymbol{x})$ satisfies

$$f(\boldsymbol{x}) \;=\; \sup_{r>0} \inf_{0<\|\boldsymbol{y}-\boldsymbol{x}\|\le r} f(\boldsymbol{y})$$

provided $\operatorname{dom} f$ contains at least two points. (Hint: Convexity gives a lower bound and lower semicontinuity an upper bound.)

27. Find the subdifferentials of the functions

$$f_1(x) \;=\; \begin{cases} 0 & x \in [-1,1] \\ |x|-1 & x \in [-2,-1) \cup (1,2] \\ \infty & \text{otherwise} \end{cases}$$

$$f_2(x) \;=\; \begin{cases} 1-\sqrt{1-x^2} & x \in [-1,1] \\ \infty & \text{otherwise.} \end{cases}$$

28. Use the representation $|x| = \max\{-x, x\}$ to find the subdifferential of $|x|$.

29. Calculate the subdifferentials $\partial\|\boldsymbol{x}\|_1$ and $\partial\|\boldsymbol{x}\|_\infty$ for $\boldsymbol{x} \in \mathsf{R}^2$ at the points $\boldsymbol{x} = (0,0)$, $\boldsymbol{x} = (1,0)$, and $\boldsymbol{x} = (1,1)$.

30. Let $f(x)$ and $g(x)$ be positive increasing convex functions with common essential domain equal to an interval J. Prove that the product $f(x)g(x)$ is convex with subdifferential $f(x)\partial g(x) + g(x)\partial f(x)$ on the interior of J. (Hints: First prove convexity, and then take forward directional derivatives.)

31. Consider a positive concave function $f(x)$ with essential domain an interval J. Show that the function $g(x) = f(x)^{-1}$ is convex with subdifferential $f(x)^{-2}\partial[-f(x)]$ for x interior to J. (Hints: First prove convexity, and then take forward directional derivatives.)

32. Consider the indicator function $\delta_C(\boldsymbol{x})$ of the set

$$C = \{\boldsymbol{x} \in \mathsf{R}^2 : x_1^2 + (x_2 - 1)^2 \leq 1\}.$$

Prove that

$$\partial\delta_C(\mathbf{0}) = \{\boldsymbol{g} \in \mathsf{R}^2 : g_1 = 0,\ g_2 \leq 0\}$$

and that

$$d_{\boldsymbol{v}}\delta_C(\mathbf{0}) = \infty \neq 0 = \sup_{\boldsymbol{g} \in \partial\delta_C(\mathbf{0})} \boldsymbol{g}^* \boldsymbol{v}$$

for $\boldsymbol{v} = (1, 0)^*$. This result is inconsistent with the support function of $\partial\delta_C(\mathbf{0})$ attaining the upper bound $d_{\boldsymbol{v}}\delta_C(\mathbf{0})$. In this case $\mathbf{0}$ does not belong to the interior of $\operatorname{dom}(\delta_C)$.

33. Demonstrate that the sum $C + D$ of a compact set C and a closed set D is closed.

34. Let $f(x)$ equal $-\sqrt{x}$ for $x \geq 0$ and ∞ for $x < 0$. If $g(x) = f(-x)$, then show that $\partial f(0) = \emptyset$ and $\partial g(0) = \emptyset$ but $\partial[f(0) + g(0)] = \mathsf{R}$. This result appears to contradict the sum rule. What assumption in our derivation of the sum rule fails?

35. A counterexample to the chain rule can be constructed by considering the closed convex set $C = \{\boldsymbol{x} \in \mathsf{R}^2 : x_1 x_2 \geq 1, x_1 > 0, x_2 > 0\}$. Show that the Fenchel conjugate of the indicator $\delta_C(\boldsymbol{y})$ equals the support function

$$\delta^\star(\boldsymbol{y}) = \begin{cases} -2\sqrt{y_1 y_2} & y_1 \leq 0 \text{ and } y_2 \leq 0 \\ \infty & \text{otherwise.} \end{cases}$$

Given the symmetric matrix

$$\boldsymbol{P} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

that projects a point $\boldsymbol{y}$ onto the y_2 axis, prove that

$$\begin{aligned} \partial(\delta_C^\star \circ \boldsymbol{P})(\mathbf{0}) &= \{\boldsymbol{x} : x_1 = 0, x_2 \geq 0\} \\ \boldsymbol{P}^* \partial\delta_C^\star(\boldsymbol{P}\mathbf{0}) &= \{\boldsymbol{x} : x_1 = 0, x_2 > 0\}. \end{aligned}$$

Thus, the two sets differ by the presence and absence of $\mathbf{0}$.

36. Let $f(\boldsymbol{y})$ be a convex function and C be a closed convex set in R^m. For $\boldsymbol{x} \in \operatorname{dom}(f) \cap C$, prove the equivalence of the following three statements:

(a) $\boldsymbol{x}$ minimizes $f(\boldsymbol{y})$ on C.

(b) $d_{\boldsymbol{v}} f(\boldsymbol{x}) \geq 0$ for all directions $\boldsymbol{v} = \boldsymbol{y} - \boldsymbol{x}$ defined by $\boldsymbol{y} \in C$.

(c) $\mathbf{0} \in \partial[f(\boldsymbol{x}) + \delta_C(\boldsymbol{x})]$.

Furthermore, deduce that

$$\partial[f(\boldsymbol{x}) + \delta_C(\boldsymbol{x})] \quad = \quad \partial f(\boldsymbol{x}) + N_C(\boldsymbol{x})$$

when $\boldsymbol{x}$ belongs to the interior of either $\text{dom}(f)$ or C.

37. Let $f(\boldsymbol{y}) : \mathsf{R}^n \mapsto \mathsf{R}$ be a convex differentiable function, $\boldsymbol{A}$ an $n \times m$ matrix with jth column $\boldsymbol{a}_j$, and $\lambda > 0$ a constant. Prove the following assertions about the function $g(\boldsymbol{z}) = f(\boldsymbol{A}\boldsymbol{z}) + \lambda\|\boldsymbol{z}\|_1$: (a) $g(\boldsymbol{z})$ achieves its global minimum at some point $\boldsymbol{x}$ whenever $f(\boldsymbol{y})$ is bounded below or λ is sufficiently large, (b) at the point $\boldsymbol{x}$

$$\boldsymbol{a}_j^* \nabla f(\boldsymbol{A}\boldsymbol{x}) \quad = \quad \begin{cases} -\lambda & x_j > 0 \\ u \in [-\lambda, \lambda] & x_j = 0 \\ \lambda & x_j < 0 \end{cases}$$

for all j, and (c) for the choice $f(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{y}\|^2$ the components of the minimum point satisfy

$$x_j \quad = \quad \frac{S[\boldsymbol{a}_j^*(\boldsymbol{u} - \sum_{i \neq j} x_i \boldsymbol{a}_i), \lambda]}{\|\boldsymbol{a}_j\|^2},$$

where $S(v, \lambda) = \text{sgn}(v) \max\{|v| - \lambda, 0\}$ is the soft threshold operator.

38. Suppose the matrix $\boldsymbol{X}$ has singular value decomposition $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ with all diagonal entries of $\boldsymbol{\Sigma}$ positive. Calculate the subdifferential

$$\partial\|\boldsymbol{X}\|_\star \quad = \quad \{\boldsymbol{U}\boldsymbol{V}^* + \boldsymbol{W} : \boldsymbol{U}^*\boldsymbol{W} = \boldsymbol{0}, \boldsymbol{W}\boldsymbol{V} = \boldsymbol{0}, \|\boldsymbol{W}\| \leq 1\}$$

of the nuclear norm [269]. (Hints: See Examples 14.3.6 and 14.4.2. The equality $\text{tr}(\boldsymbol{Y}^*\boldsymbol{X}) = \|\boldsymbol{X}\|_\star \|\boldsymbol{Y}\|_2$ entails equality in Fan's inequality.)

39. Suppose the convex function $g(\boldsymbol{y} \mid \boldsymbol{x})$ majorizes the convex function $f(\boldsymbol{y})$ around the point $\boldsymbol{x}$. Demonstrate that $\partial f(\boldsymbol{x}) \subset \partial g(\boldsymbol{x} \mid \boldsymbol{x})$. Give an example where $\partial g(\boldsymbol{x} \mid \boldsymbol{x}) \neq \partial f(\boldsymbol{x})$. If $g(\boldsymbol{y} \mid \boldsymbol{x})$ is differentiable at $\boldsymbol{x}$, then equality holds.

40. The $\ell_{p,q}$ norm on R^n is useful in group penalties [230]. Suppose the sets σ_g partition $\{1, \ldots, n\}$. For $\boldsymbol{x} \in \mathsf{R}^n$ let $\boldsymbol{x}_{\sigma_g}$ denote the vector formed by taking the components of $\boldsymbol{x}$ derived from σ_g. For p and q between 1 and ∞, the $\ell_{p,q}$ norm equals

$$\|\boldsymbol{x}\|_{p,q} \quad = \quad \left(\sum_g \|\boldsymbol{x}_{\sigma_g}\|_q^p \right)^{1/p},$$

Demonstrate that the $\ell_{r,s}$ norm is dual to the $\ell_{p,q}$ norm, where r and s satisfy $p^{-1} + r^{-1} = 1$ and $q^{-1} + s^{-1} = 1$.

15
Feasibility and Duality

15.1 Introduction

This chapter provides a concrete introduction to several advanced topics in optimization theory. Specifying an interior feasible point is the first issue that must be faced in applying a barrier method. Given an exterior point, Dykstra's algorithm [21, 70, 79] finds the closest point in the intersection $\cap_{i=0}^{r-1} C_i$ of a finite number of closed convex sets. If C_i is defined by the convex constraint $h_i(\boldsymbol{x}) \leq 0$, then one obvious tactic for finding an interior point is to replace C_i by the set $C_i(\epsilon) = \{\boldsymbol{x} : h_j(\boldsymbol{x}) \leq -\epsilon\}$ for some small $\epsilon > 0$. Projecting onto the intersection of the $C_i(\epsilon)$ then produces an interior point.

The method of alternating projections is faster in practice than Dykstra's algorithm, but it is only guaranteed to find some feasible point, not the closest feasible point to a given exterior point. Projection operators are specific examples of paracontractions. We study these briefly and their classical counterparts, contractions and strict contractions. Under the right hypotheses, a contraction T possesses a unique fixed point, and the sequence $\boldsymbol{x}_{m+1} = T(\boldsymbol{x}_m)$ converges to it regardless of the initial point $\boldsymbol{x}_0$.

Duality is one of the deepest and most pervasive themes of modern optimization theory. It takes considerable mathematical maturity to appreciate this subtle topic, and it is impossible to do it justice in a short essay. Every convex program generates a corresponding dual program, which can be simpler to solve than the original or primal program. We show how to

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_15,

construct dual programs and relate the absence of a duality gap to Slater's constraint qualification. We also point out important connections between duality and the Fenchel conjugate.

15.2 Dykstra's Algorithm

Example 2.5.2 demonstrates that the distance $\operatorname{dist}(\boldsymbol{x}, C)$ from a point $\boldsymbol{x}$ in R^n to a set C is a uniformly continuous function of $\boldsymbol{x}$. If the set C is closed and convex, then $\operatorname{dist}(\boldsymbol{x}, C)$ is convex in $\boldsymbol{x}$ and $\operatorname{dist}(\boldsymbol{x}, C) = \|P_C(\boldsymbol{x}) - \boldsymbol{x}\|$ for exactly one projected point $P_C(\boldsymbol{x}) \in C$. (See Examples 2.5.5 and 6.3.3 and Proposition 6.2.2.) Although it is generally impossible to calculate $P_C(\boldsymbol{x})$ explicitly, some specific projection operators are well known.

Example 15.2.1 *Examples of Projection Operators*

Closed Euclidean Ball: If $C = \{\boldsymbol{y} \in \mathsf{R}^n : \|\boldsymbol{y} - \boldsymbol{z}\| \leq r\}$, then

$$P_C(\boldsymbol{x}) = \begin{cases} \boldsymbol{z} + \frac{r}{\|\boldsymbol{x}-\boldsymbol{z}\|}(\boldsymbol{x} - \boldsymbol{z}) & \boldsymbol{x} \notin C \\ \boldsymbol{x} & \boldsymbol{x} \in C\,. \end{cases}$$

Closed Rectangle: If $C = [a, b]$ is a closed rectangle in R^n, then

$$P_C(\boldsymbol{x})_i = \begin{cases} a_i & x_i < a_i \\ x_i & x_i \in [a_i, b_i] \\ b_i & x_i > b_i\,. \end{cases}$$

Hyperplane: If $C = \{\boldsymbol{y} \in \mathsf{R}^n : \boldsymbol{a}^*\boldsymbol{y} = b\}$ for $\boldsymbol{a} \neq \boldsymbol{0}$, then

$$P_C(\boldsymbol{x}) = \boldsymbol{x} - \frac{\boldsymbol{a}^*\boldsymbol{x} - b}{\|\boldsymbol{a}\|^2}\boldsymbol{a}.$$

Closed Halfspace: If $C = \{\boldsymbol{y} \in \mathsf{R}^n : \boldsymbol{a}^*\boldsymbol{y} \leq b\}$ for $\boldsymbol{a} \neq \boldsymbol{0}$, then

$$P_C(\boldsymbol{x}) = \begin{cases} \boldsymbol{x} - \frac{\boldsymbol{a}^*\boldsymbol{x}-b}{\|\boldsymbol{a}\|^2}\boldsymbol{a} & \boldsymbol{a}^*\boldsymbol{x} > b \\ \boldsymbol{x} & \boldsymbol{a}^*\boldsymbol{x} \leq b\,. \end{cases}$$

Subspace: If C is the range of a matrix $\boldsymbol{A}$ with full column rank, then

$$P_C(\boldsymbol{x}) = \boldsymbol{A}(\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*\boldsymbol{x}.$$

Positive Semidefinite Matrices: Let $\boldsymbol{M}$ be an $n \times n$ symmetric matrix with spectral decomposition $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^*$, where $\boldsymbol{U}$ is an orthogonal matrix and $\boldsymbol{D}$ is a diagonal matrix with ith diagonal entry d_i. The projection of $\boldsymbol{M}$ onto the set S of positive semidefinite matrices is given by $P_S(\boldsymbol{M}) = \boldsymbol{U}\boldsymbol{D}_+\boldsymbol{U}^*$, where $\boldsymbol{D}_+$ is diagonal with ith diagonal entry $\max\{d_i, 0\}$. Problem 7 asks the reader to check this fact.

Unit Simplex: The problem of projecting a point $\boldsymbol{x}$ onto the unit simplex

$$S = \Big\{\boldsymbol{y} : \sum_{i=1}^{n} y_i = 1,\ y_i \geq 0 \text{ for } 1 \leq i \leq n\Big\}.$$

can be solved by a simple algorithm [77]. If we apply Gibbs lemma as sketched in Problem 18 of Chap. 5 to $f(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2$, then the coordinates of the minimum point $\boldsymbol{y}$ satisfy

$$y_i = \begin{cases} x_i - \lambda & y_i > 0 \\ x_i - \lambda + \mu_i & y_i = 0 \end{cases}$$

for Lagrange multipliers λ and $\mu_i \geq 0$. Setting $I_+ = \{i : y_i > 0\}$ and invoking the equality constraint

$$1 = \sum_{i \in I_+} y_i = \sum_{i \in I_+} x_i - |I_+|\lambda$$

then imply

$$\lambda = \frac{1}{|I_+|}\Big(\sum_{i \in I_+} x_i - 1\Big).$$

The catch, of course, is that we do not know I_+. The key to avoiding all 2^n possible subsets is the simple observation that the y_i and x_i are consistently ordered. Suppose on the contrary that $x_i < x_j$ and $y_j < y_i$. For small $s > 0$ substitute $y_j + s$ for y_j and $y_i - s$ for y_i. The objective function $f(\boldsymbol{y})$ then changes by the amount

$$\begin{aligned} & \frac{1}{2}\Big[(y_i - s - x_i)^2 + (y_j + s - x_j)^2 - (y_i - x_i)^2 - (y_j - x_j)^2\Big] \\ = \ & s(x_i - x_j + y_j - y_i) + s^2, \end{aligned}$$

which is negative for small s. Thus, let $z_1, \ldots, z_n$ denote the x_i ordered from largest to smallest. For each integer j between 1 and n, it suffices to set $\lambda = \frac{1}{j}(\sum_{i=1}^{j} z_i - 1)$ and check whether $z_j > \lambda$ and $z_{j+1} \leq \lambda$. When these two conditions are met, we put $y_i = (x_i - \lambda)_+$ for all i. Michelot's [196] algorithm as described in Problem 13 also solves this projection problem.

Closed ℓ_1 Ball: Projecting a point $\boldsymbol{x}$ onto the ℓ_1 ball

$$C = \{\boldsymbol{y} : \|\boldsymbol{y} - \boldsymbol{z}\|_1 \leq r\}.$$

yields to a variation of the previous projection algorithm as sketched in Problem 14.

Isotone Convex Cone: Projection onto the convex cone

$$C = \{\boldsymbol{y} : y_i \leq y_{i+1} \text{ for } 1 \leq i < n\}.$$

is accomplished by the pool adjacent violators algorithm as discussed later in Example 15.2.3. ■

Dykstra's algorithm [21, 70, 79] is designed to find the projection of a point $\boldsymbol{x}$ onto a finite intersection $C = \cap_{i=0}^{r-1} C_i$ of r closed convex sets. Here are some possible situations where Dykstra's algorithm applies.

Example 15.2.2 *Applications of Dykstra's Algorithm*

Linear Equalities: Any solution of the system of linear equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ belongs to the intersection of the hyperplanes $\boldsymbol{a}_i^* \boldsymbol{x} = b_i$, where $\boldsymbol{a}_i^*$ is the ith row of $\boldsymbol{A}$.

Linear Inequalities: Any solution of the system of linear inequalities $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$ belongs to the intersection of the halfspaces $\boldsymbol{a}_i^* \boldsymbol{x} \leq b_i$, where $\boldsymbol{a}_i^*$ is the ith row of $\boldsymbol{A}$.

Isotone Regression: The least squares problem of minimizing the sum $\sum_{i=1}^n (x_i - w_i)^2$ subject to the constraints $w_i \leq w_{i+1}$ corresponds to projection of $\boldsymbol{x}$ onto the intersection of the halfspaces

$$C_i = \{\boldsymbol{w} \in \mathsf{R}^n : w_i - w_{i+1} \leq 0\}, \quad 1 \leq i \leq n-1.$$

Convex Regression: The least squares problem of minimizing the sum $\sum_{i=1}^n (x_i - w_i)^2$ subject to the constraints $w_i \leq \frac{1}{2}(w_{i-1} + w_{i+1})$ corresponds to projection of $\boldsymbol{x}$ onto the intersection of the halfspaces

$$C_i = \left\{\boldsymbol{w} \in \mathsf{R}^n : w_i - \frac{1}{2}(w_{i-1} + w_{i+1}) \leq 0\right\}, \quad 2 \leq i \leq n-1.$$

Quadratic Programming: To minimize the strictly convex quadratic form $\frac{1}{2}\boldsymbol{x}^* \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^* \boldsymbol{x} + c$ subject to $\boldsymbol{D}\boldsymbol{x} = \boldsymbol{e}$ and $\boldsymbol{F}\boldsymbol{x} \leq \boldsymbol{g}$, we make the change of variables $\boldsymbol{y} = \boldsymbol{A}^{1/2}\boldsymbol{x}$. This transforms the problem to one of minimizing

$$\begin{aligned} \frac{1}{2}\boldsymbol{x}^* \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^* \boldsymbol{x} + c &= \frac{1}{2}\|\boldsymbol{y}\|^2 + \boldsymbol{b}^* \boldsymbol{A}^{-1/2}\boldsymbol{y} + c \\ &= \frac{1}{2}\|\boldsymbol{y} + \boldsymbol{A}^{-1/2}\boldsymbol{b}\|^2 - \frac{1}{2}\boldsymbol{b}^* \boldsymbol{A}^{-1}\boldsymbol{b} + c \end{aligned}$$

subject to $\boldsymbol{D}\boldsymbol{A}^{-1/2}\boldsymbol{y} = \boldsymbol{e}$ and $\boldsymbol{F}\boldsymbol{A}^{-1/2}\boldsymbol{y} \leq \boldsymbol{g}$. The solution in the $\boldsymbol{y}$ coordinates is determined by projecting $-\boldsymbol{A}^{-1/2}\boldsymbol{b}$ onto the convex feasible region determined by the revised constraints. Instead of the symmetric square root transformation $\boldsymbol{y} = \boldsymbol{A}^{1/2}\boldsymbol{x}$, one can employ the asymmetric square root transformation $\boldsymbol{y} = \boldsymbol{U}\boldsymbol{x}$ furnished by the Cholesky decomposition $\boldsymbol{L} = \boldsymbol{U}^*$ of $\boldsymbol{A}$. ■

To state Dykstra's algorithm, it is helpful to label the closed convex sets $C_0, \ldots, C_{r-1}$ and denote their intersection by $C = \cap_{i=0}^{r-1} C_i$. The algorithm keeps track of a primary sequence $\boldsymbol{x}_m$ and a companion sequence $\boldsymbol{\omega}_m$. In the limit, $\boldsymbol{x}_m$ tends to $P_C(\boldsymbol{x})$. To initiate the process, we set $\boldsymbol{x}_0 = \boldsymbol{x}$ and $\boldsymbol{\omega}_{-r+1} = \cdots = \boldsymbol{\omega}_0 = \mathbf{0}$. For $m \geq 0$ we then iterate via

$$\begin{aligned} \boldsymbol{x}_{m+1} &= P_{C_{m \bmod r}}(\boldsymbol{x}_m + \boldsymbol{\omega}_{m-r+1}) \\ \boldsymbol{\omega}_{m+1} &= \boldsymbol{x}_m + \boldsymbol{\omega}_{m-r+1} - \boldsymbol{x}_{m+1}. \end{aligned}$$

Here $m \bmod r$ is the nonnegative remainder after dividing m by r. In essence, the algorithm cycles through the convex sets and projects the sum of the current vector and the relevant previous companion vector onto the current convex set.

TABLE 15.1. Iterates of Dykstra's algorithm

Iteration m	x_{m1}	x_{m2}
0	−1.00000	2.00000
1	−0.44721	0.89443
2	0.00000	0.89443
3	−0.26640	0.96386
4	0.00000	0.96386
5	−0.14175	0.98990
10	0.00000	0.99934
15	−0.00454	0.99999
25	−0.00014	1.00000
30	0.00000	1.00000
35	0.00000	1.00000

As an example, suppose $r = 2$, C_0 is the closed unit ball in R^2, and C_1 is the closed halfspace with $x_1 \geq 0$. The intersection C is the right half ball centered at the origin. Table 15.1 records the iterates of Dykstra's algorithm starting from the point $\boldsymbol{x}_0 = (-1, 2)^*$ and their eventual convergence to the geometrically obvious solution $(0, 1)^*$.

When C_i is a subspace, Dykstra's algorithm can dispense with the corresponding companion subsequence $\boldsymbol{\omega}_m$. In this case, $\boldsymbol{\omega}_{m+1}$ is perpendicular to C_i whenever $m \bmod r = i$. Indeed, since $P_{C_i}(\boldsymbol{y})$ is a linear projection, we have

$$\begin{aligned} \boldsymbol{x}_{m+1} &= P_{C_i}(\boldsymbol{x}_m + \boldsymbol{\omega}_{m-r+1}) \\ &= P_{C_i}(\boldsymbol{x}_m) + P_{C_i}(\boldsymbol{\omega}_{m-r+1}) \\ &= P_{C_i}(\boldsymbol{x}_m) \end{aligned}$$

under the perpendicularity assumption. The initial condition $\boldsymbol{\omega}_{i-r} = \mathbf{0}$, the identity

$$\begin{aligned}\boldsymbol{\omega}_{m+1} &= \boldsymbol{x}_m - \boldsymbol{x}_{m+1} + \boldsymbol{\omega}_{m-r+1} \\ &= \boldsymbol{x}_m - P_{C_i}(\boldsymbol{x}_m) + \boldsymbol{\omega}_{m-r+1} \\ &= P_{C_i^\perp}(\boldsymbol{x}_m) + \boldsymbol{\omega}_{m-r+1},\end{aligned}$$

and induction show that $\boldsymbol{\omega}_{m+1}$ belongs to the perpendicular complement $C_i^\perp$ if $m \bmod r = i$. When all of the C_i are subspaces, Dykstra's algorithm reduces to the method of alternating projections first studied by von Neumann. Example 15.5.6 will justify Dykstra's algorithm theoretically.

Example 15.2.3 *Pool Adjacent Violators Algorithm*

Dykstra's algorithm can be beat in specific problems. Consider weighted isotone regression with objective function $f(\boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^n w_i(y_i - x_i)^2$ and arbitrary positive weights w_i. The Lagrangian for this problem reads

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\mu}) = \frac{1}{2}\sum_{i=1}^n w_i(y_i - x_i)^2 + \sum_{i=1}^{n-1} \mu_i(x_i - x_{i+1}),$$

for nonnegative multipliers μ_i. The optimal point is determined by the stationarity conditions

$$0 = w_i(x_i - y_i) + \mu_i - \mu_{i-1}$$

and the complementary slackness conditions $\mu_i(x_i - x_{i+1}) = 0$. Here we define $\mu_0 = \mu_n = 0$ for convenience. Because of telescoping, one can solve for the multipliers in either of the equivalent forms

$$\mu_j = \sum_{i=1}^j w_i(y_i - x_i) = -\sum_{i=j+1}^n w_i(y_i - x_i). \tag{15.1}$$

since $\mu_n = \sum_{i=1}^n w_i(y_i - x_i) = 0$.

The pool adjacent violators algorithm [6, 157] exploits the fact that whole blocks of the solution vector $\boldsymbol{x}$ are constant. From one block to the next, this constant increases. The algorithm starts with $x_i = y_i$ and $\mu_i = 0$ for all i and all blocks reducing to singletons. It then marches through the blocks and considers whether to consolidate or pool adjacent blocks. Let

$$B = \{l, l+1, \ldots, r-1, r\}$$

denote a generic block. In view of complementary slackness, the right multiplier μ_r is assumed to be 0. According to the above calculations, the constant assigned to block B is the weighted average

$$x_B = \frac{1}{\sum_{i=l}^r w_i}\sum_{i=l}^r w_i y_i.$$

Now the only thing that prevents a block decomposition from supplying the solution vector is a reversal of two adjacent block constants. Suppose $B_1 = \{l_1, \ldots, r_1\}$ and $B_2 = \{l_2, \ldots, r_2\}$ are adjacent violating blocks. We pool the two blocks and assign the constant

$$x_{B_1 \cup B_2} = \frac{1}{\sum_{i=l_1}^{r_2} w_i} \sum_{i=l_1}^{r_2} w_i y_i$$

to $B_1 \cup B_2$. Note here that $r_1 + 1 = l_2$. We must also recalculate the multipliers associated with the combined block and check that they are nonnegative. Equation (15.1) is instrumental in achieving this goal. By induction we assume that it holds when 1 is replaced by l_i and n replaced by r_i for $i \in \{1, 2\}$. In recalculating the μ_j for j between l_1 and r_2, we can take the last multiplier μ_{r_2} to be 0 by virtue of the definition of $x_{B_1 \cup B_2}$. For j between l_1 and r_1, the left definition in equation (15.1) and the inequality $x_{B_1} > x_{B_1 \cup B_2}$ imply that the new $\mu_j \geq 0$. For j between l_2 and $r_2 - 1$, the right definition in equation (15.1) and the inequality $x_{B_1 \cup B_2} > x_{B_2}$ again imply that the new $\mu_j \geq 0$. Thus, the pooled block satisfies the multiplier conditions. Pooling continues until all blocks coalesce or no violations occur. At that point the multiplier conditions hold, and the minimum has been reached. ■

15.3 Contractive Maps

We met locally contractive maps in our study of convergence in Chap. 12. Here we discuss a generalization that is helpful in finding initial feasible points in nonlinear programming. First recall that a map $T : D \subset \mathsf{R}^n \mapsto \mathsf{R}^n$ is contractive relative to a norm $\|\boldsymbol{x}\|_\dagger$ if $\|T(\boldsymbol{y}) - T(\boldsymbol{z})\|_\dagger < \|\boldsymbol{y} - \boldsymbol{z}\|_\dagger$ for all $\boldsymbol{y} \neq \boldsymbol{z}$ in D. It is strictly contractive if there exists a constant $c \in [0, 1)$ with $\|T(\boldsymbol{y}) - T(\boldsymbol{z})\|_\dagger \leq c\|\boldsymbol{y} - \boldsymbol{z}\|_\dagger$ for all such pairs. Finally, T is said to be paracontractive provided for every fixed point $\boldsymbol{y}$ of $T(\boldsymbol{x})$ the inequality $\|T(\boldsymbol{x}) - \boldsymbol{y}\|_\dagger < \|\boldsymbol{x} - \boldsymbol{y}\|_\dagger$ holds unless $\boldsymbol{x}$ is itself a fixed point. A strictly contractive map is contractive, and a contractive map is paracontractive.

For instance, the affine map $\boldsymbol{M}\boldsymbol{x} + \boldsymbol{v}$ is contractive under some induced matrix norm whenever the spectral radius $\rho(\boldsymbol{M}) < 1$. (See Proposition 6.3.2 of the reference [166].) Projection onto a closed convex set C containing more than a single point is paracontractive but not contractive under the standard Euclidean norm. To validate paracontraction, note that the obtuse angle criterion stated in Example 6.5.3 implies

$$\begin{aligned} [\boldsymbol{x} - P_C(\boldsymbol{x})]^*[P_C(\boldsymbol{y}) - P_C(\boldsymbol{x})] &\leq 0 \\ [\boldsymbol{y} - P_C(\boldsymbol{y})]^*[P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})] &\leq 0. \end{aligned}$$

Adding these inequalities, rearranging, and applying the Cauchy-Schwarz inequality give

$$\begin{aligned} \|P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})\|^2 &\leq (\boldsymbol{x}-\boldsymbol{y})^*[P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})] \\ &\leq \|\boldsymbol{x}-\boldsymbol{y}\| \cdot \|P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})\|. \end{aligned} \tag{15.2}$$

Dividing the extremes of inequality (15.2) by $\|P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})\|$ now demonstrates that $\|P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})\| \leq \|\boldsymbol{x}-\boldsymbol{y}\|$. Equality holds in the Cauchy-Schwarz half of inequality (15.2) if and only if $P_C(\boldsymbol{x}) - P_C(\boldsymbol{y}) = c(\boldsymbol{x}-\boldsymbol{y})$ for some constant c. If $\boldsymbol{y}$ is a fixed point, then overall equality in inequality (15.2) entails

$$\|P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})\|^2 = (\boldsymbol{x}-\boldsymbol{y})^*[P_C(\boldsymbol{x}) - P_C(\boldsymbol{y})] = c\|\boldsymbol{x}-\boldsymbol{y}\|^2.$$

This precipitates a cascade of deductions. First that $c = 1$, second that $P_C(\boldsymbol{x}) - \boldsymbol{x} = P_C(\boldsymbol{y}) - \boldsymbol{y} = \boldsymbol{0}$, third that $\boldsymbol{x}$ is a fixed point, and fourth that the projection map $P_C(\boldsymbol{x})$ is paracontractive under the Euclidean norm.

With this background in mind, we now state and prove an important result due to Elsner, Koltracht, and Neumann [85].

Proposition 15.3.1 *Suppose the continuous maps $T_0, \ldots, T_{r-1}$ of a set into itself are paracontractive under the norm $\|\boldsymbol{x}\|_\dagger$. Let F_i denote the set of fixed points of T_i. If the intersection $F = \cap_{i=0}^{r-1} F_i$ is nonempty, then the sequence*

$$\boldsymbol{x}_{m+1} = T_{m \bmod r}(\boldsymbol{x}_m)$$

converges to a limit in F. In particular, if $r = 1$ and $T = T_0$ has a nonempty set of fixed points F, then $\boldsymbol{x}_{m+1} = T(\boldsymbol{x}_m)$ converges to a point in F.

Proof: Let $\boldsymbol{y}$ be any point in F. The scalar sequence $\|\boldsymbol{x}_m - \boldsymbol{y}\|_\dagger$ satisfies

$$\|\boldsymbol{x}_{m+1} - \boldsymbol{y}\|_\dagger = \|T_{m \bmod r}(\boldsymbol{x}_m) - \boldsymbol{y}\|_\dagger \leq \|\boldsymbol{x}_m - \boldsymbol{y}\|_\dagger$$

and therefore possesses a limit $d \geq 0$. Because the sequence $\boldsymbol{x}_m$ is bounded, it possesses a cluster point $\boldsymbol{x}_\infty$. Furthermore, $\|\boldsymbol{x}_\infty - \boldsymbol{y}\|_\dagger$ attains the lower bound d. But this implies that $\|T_i(\boldsymbol{x}_\infty) - \boldsymbol{y}\|_\dagger = \|\boldsymbol{x}_\infty - \boldsymbol{y}\|_\dagger$ for all i, and therefore $\boldsymbol{x}_\infty \in F$. For the choice $\boldsymbol{y} = \boldsymbol{x}_\infty$, the corresponding constant d equals 0. Finally, the monotone convergence of $\|\boldsymbol{x}_m - \boldsymbol{x}_\infty\|_\dagger$ to 0 implies $\lim_{m\to\infty} \boldsymbol{x}_m = \boldsymbol{x}_\infty$. ■

There are two corollaries to Proposition 15.3.1. First, the set of fixed points of the composite map $S = T_{r-1} \circ \cdots \circ T_0$ equals $F = \cap_{i=0}^{r-1} F_i$. Second, S itself is paracontractive. Problem 21 asks the reader to prove these facts. As a trivial application of the proposition, consider the toy example of projection onto the half ball appearing in Table 15.1. For the chosen initial point $\boldsymbol{x}_0$, a single round $P_{C_1} \circ P_{C_0}(\boldsymbol{x}_0)$ of projection lands in the half ball. In contrast to Dykstra's algorithm, the limit is not the closest point in $C = C_0 \cap C_1$ to $\boldsymbol{x}_0$.

Alternatively, one can iterate via a convex combination

$$R(\boldsymbol{x}) \quad = \quad \sum_{i=0}^{r-1} \lambda_i T_i(\boldsymbol{x})$$

with positive coefficients. Clearly, any point $\boldsymbol{x} \in F$ is also a fixed point of $R(\boldsymbol{x})$. Conversely, if $\boldsymbol{x}$ is a fixed point and $\boldsymbol{y}$ is any point in F, then

$$\begin{aligned} \|\boldsymbol{x}-\boldsymbol{y}\|_\dagger &= \left\| \sum_{i=0}^{r-1} \lambda_i T_i(\boldsymbol{x}) - \sum_{i=0}^{r-1} \lambda_i T_i(\boldsymbol{y}) \right\|_\dagger \\ &\leq \sum_{i=0}^{r-1} \lambda_i \|T_i(\boldsymbol{x}) - T_i(\boldsymbol{y})\|_\dagger \\ &\leq \sum_{i=0}^{r-1} \lambda_i \|\boldsymbol{x}-\boldsymbol{y}\|_\dagger \\ &= \|\boldsymbol{x}-\boldsymbol{y}\|_\dagger. \end{aligned}$$

Because equality must occur throughout, it follows from the paracontractiveness of T_i that $\boldsymbol{x} \in F_i$ for each i. Hence, the fixed points of $R(\boldsymbol{x})$ coincide with F. If we take $\boldsymbol{x} \notin F$ and $\boldsymbol{y} \in F$, then basically the same argument demonstrates the paracontractiveness requirement $\|R(\boldsymbol{x})-\boldsymbol{y}\|_\dagger < \|\boldsymbol{x}-\boldsymbol{y}\|_\dagger$. Convergence of the iterates $\boldsymbol{x}_{n+1} = \sum_{i=0}^{r-1} \lambda_i T_i(\boldsymbol{x}_n)$ is now a consequence of the paracontractiveness of the map $R(\boldsymbol{x})$. The evidence suggests that simultaneous projection converges more slowly than alternating projection [42, 109]. However, simultaneous projection enjoys the advantage of being parallelizable.

Proposition 15.3.1 postulates the existence of fixed points. The next proposition introduces simple sufficient conditions guaranteeing existence.

Proposition 15.3.2 *Suppose T is contractive under a norm $\|\boldsymbol{x}\|_\dagger$ and maps the nonempty compact set $D \subset \mathsf{R}^n$ into itself. Then T has a unique fixed point in D. If T is a strict contraction with contraction constant c, then the assumption that D is compact can be relaxed to the assumption that D is closed.*

Proof: We first demonstrate that there is at most one fixed point $\boldsymbol{y}$. If there is a second fixed point $\boldsymbol{z} \neq \boldsymbol{y}$, then

$$\|\boldsymbol{y}-\boldsymbol{z}\|_\dagger \quad = \quad \|T(\boldsymbol{y}) - T(\boldsymbol{z})\|_\dagger \quad < \quad \|\boldsymbol{y}-\boldsymbol{z}\|_\dagger,$$

which is a contradiction.

Now define $d = \inf_{\boldsymbol{x} \in D} f(\boldsymbol{x})$ for the function $f(\boldsymbol{x}) = \|T(\boldsymbol{x}) - \boldsymbol{x}\|_\dagger$. Since $f(\boldsymbol{x})$ is continuous, its infimum is attained at some point $\boldsymbol{y}$ in the compact set D. If $\boldsymbol{y}$ is not a fixed point, then

$$\|T \circ T(\boldsymbol{y}) - T(\boldsymbol{y})\|_\dagger \quad < \quad \|T(\boldsymbol{y}) - \boldsymbol{y}\|_\dagger,$$

contradicting the definition of $\boldsymbol{y}$. On the other hand, if D is not compact, but T is a strict contraction, then choose any point $\boldsymbol{y} \in D$, and define the set $C = \{\boldsymbol{x} \in D : f(\boldsymbol{x}) \leq f(\boldsymbol{y})\}$. For $\boldsymbol{x} \in C$ we then have

$$\begin{aligned} \|\boldsymbol{x} - \boldsymbol{y}\|_\dagger &\leq \|\boldsymbol{x} - T(\boldsymbol{x})\|_\dagger + \|T(\boldsymbol{x}) - T(\boldsymbol{y})\|_\dagger + \|T(\boldsymbol{y}) - \boldsymbol{y}\|_\dagger \\ &\leq 2f(\boldsymbol{y}) + c\|\boldsymbol{x} - \boldsymbol{y}\|_\dagger. \end{aligned}$$

It follows that

$$\|\boldsymbol{x} - \boldsymbol{y}\|_\dagger \leq \frac{2f(\boldsymbol{y})}{1-c}.$$

Thus, the closed set C is bounded and hence compact. Furthermore, the inequality $f[T(\boldsymbol{x})] \leq cf(\boldsymbol{x})$ indicates that T maps C into itself. The rest of the argument proceeds as before. ■

Example 15.3.1 *Stationary Distribution of a Markov Chain*

Example 6.2.1 demonstrates that every finite state Markov chain possesses a stationary distribution. Under an appropriate ergodic hypothesis, this distribution is unique, and the chain converges to it. In understanding these phenomena, it simplifies notation to pass to column vectors and replace $\boldsymbol{P} = (p_{ij})$ by its transpose $\boldsymbol{Q} = (q_{ij})$. It is easy to check that $\boldsymbol{Q}$ maps the standard simplex

$$S = \left\{\boldsymbol{x} : x_i \geq 0,\ i = 1, \ldots, n,\ \sum_{i=1}^n x_i = 1\right\}$$

into itself and that candidate vectors belong to S. The natural norm on S is $\|\boldsymbol{x}\|_1 = \sum_{i=1}^n |x_i|$. According to Problem 9 of Chap. 2, the corresponding induced matrix norm of $\boldsymbol{Q}$ is

$$\|\boldsymbol{Q}\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |q_{ij}| = \max_{1 \leq j \leq n} \sum_{i=1}^n p_{ji} = 1.$$

Thus, $\boldsymbol{Q}$ is almost a contraction.

The standard ergodic hypothesis says that some power $\boldsymbol{P}^k$ of $\boldsymbol{P}$ has all entries positive. If we define $\boldsymbol{R} = \boldsymbol{Q}^k - c\mathbf{1}\mathbf{1}^*$ for c the minimum entry of $\boldsymbol{P}^k$, then the matrix $\boldsymbol{R}$ has all entries nonnegative and norm $\|\boldsymbol{R}\|_1 < 1$. Since $\mathbf{1}^*(\boldsymbol{y} - \boldsymbol{x}) = 0$ for all pairs $\boldsymbol{x}$ and $\boldsymbol{y}$ in S, we have

$$||\boldsymbol{Q}^k\boldsymbol{x} - \boldsymbol{Q}^k\boldsymbol{y}||_1 = \|\boldsymbol{R}(\boldsymbol{x} - \boldsymbol{y})\|_1 \leq \|\boldsymbol{R}\|_1\|\boldsymbol{x} - \boldsymbol{y}\|_1.$$

In other words, the map $\boldsymbol{x} \mapsto \boldsymbol{Q}^k\boldsymbol{x}$ is strictly contractive on S. It therefore possesses a unique fixed point $\boldsymbol{y}$, and $\lim_{m\to\infty} \boldsymbol{Q}^{mk}\boldsymbol{x} = \boldsymbol{y}$. Problem 22 asks the reader to check that the map $\boldsymbol{x} \mapsto \boldsymbol{Q}\boldsymbol{x}$ shares the fixed point $\boldsymbol{y}$ and that $\lim_{m\to\infty} \boldsymbol{Q}^m\boldsymbol{x} = \boldsymbol{y}$. ■

Proposition 15.3.2 finds applications in many branches of mathematics and statistics. Part of its value stems from the guaranteed geometric rate of convergence of the iterates $\boldsymbol{x}_{m+1} = T(\boldsymbol{x}_m)$ to the fixed point $\boldsymbol{y}$. This assertion is a straightforward consequence of the inequality

$$\|\boldsymbol{x} - \boldsymbol{y}\|_\dagger \quad \leq \quad \|\boldsymbol{x} - T(\boldsymbol{x})\|_\dagger + \|T(\boldsymbol{x}) - T(\boldsymbol{y})\|_\dagger \quad \leq \quad f(\boldsymbol{x}) + c\|\boldsymbol{x} - \boldsymbol{y}\|_\dagger$$

involving the function $f(\boldsymbol{x}) = \|T(\boldsymbol{x}) - \boldsymbol{x}\|_\dagger$. It follows that

$$\|\boldsymbol{x} - \boldsymbol{y}\|_\dagger \quad \leq \quad \frac{1}{1-c} f(\boldsymbol{x}). \tag{15.3}$$

Substituting $\boldsymbol{x}_m$ for $\boldsymbol{x}$ and applying the inequality $f[T(\boldsymbol{x})] \leq cf(\boldsymbol{x})$ repeatedly now yield the geometric bound

$$\|\boldsymbol{x}_m - \boldsymbol{y}\|_\dagger \quad \leq \quad \frac{c^m}{1-c} f(\boldsymbol{x}_0).$$

In numerical practice, inequality (15.3) gives the preferred test for declaring convergence. If one wants $\boldsymbol{x}_m$ to be within ϵ of the fixed point, then iteration should continue until $f(\boldsymbol{x}_m) \leq \epsilon(1-c)$.

15.4 Dual Functions

The Lagrange multiplier rule summarizes much of what we know about minimizing $f(\boldsymbol{x})$ subject to the constraints $g_i(\boldsymbol{x}) = 0$ for $1 \leq i \leq p$ and $h_j(\boldsymbol{x}) \leq 0$ for $1 \leq j \leq q$. Consequently, it is worth considering the standard Lagrangian function

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad = \quad f(\boldsymbol{x}) + \sum_{i=1}^{p} \lambda_i g_i(\boldsymbol{x}) + \sum_{j=1}^{q} \mu_j h_j(\boldsymbol{x})$$

in more detail. Here the multiplier vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are taken as arguments in addition to the variable $\boldsymbol{x}$. For a convex program satisfying a constraint qualification such as Slater's condition, a constrained global minimum $\hat{\boldsymbol{x}}$ of $f(\boldsymbol{x})$ is also an unconstrained global minimum of $\mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\mu}})$, where $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\mu}}$ are the corresponding Lagrange multipliers. This fact is the content of Proposition 14.7.1

The behavior of $\mathcal{L}(\hat{\boldsymbol{x}}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ as a function of $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ is also interesting. Because $g_i(\hat{\boldsymbol{x}}) = 0$ for all i and $\hat{\boldsymbol{\mu}}_j h_j(\hat{\boldsymbol{x}}) = 0$ for all j, we have

$$\begin{aligned}
\mathcal{L}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\mu}}) - \mathcal{L}(\hat{\boldsymbol{x}}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad &= \quad \sum_{j=1}^{q} (\hat{\mu}_j - \mu_j) h_j(\hat{\boldsymbol{x}}) \\
&= \quad -\sum_{j=1}^{q} \mu_j h_j(\hat{\boldsymbol{x}}) \\
&\geq \quad 0.
\end{aligned}$$

This proves the left inequality of the two saddle point inequalities

$$\mathcal{L}(\hat{\boldsymbol{x}}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \;\leq\; \mathcal{L}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\mu}}) \;\leq\; \mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\mu}}).$$

The left saddle point inequality immediately implies

$$\sup_{\boldsymbol{\lambda},\boldsymbol{\mu}\geq\mathbf{0}} \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \;\leq\; \mathcal{L}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\mu}}) \;=\; f(\hat{\boldsymbol{x}}).$$

The right saddle point inequality, valid for a convex program under Slater's constraint qualification, entails

$$\mathcal{L}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\mu}}) \;=\; \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\mu}}) \;\leq\; \sup_{\boldsymbol{\lambda},\boldsymbol{\mu}\geq\mathbf{0}} \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}).$$

Hence, we can recover the minimum value of $f(\boldsymbol{x})$ as

$$f(\hat{\boldsymbol{x}}) \;=\; \sup_{\boldsymbol{\lambda},\boldsymbol{\mu}\geq\mathbf{0}} \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}).$$

The dual function

$$\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu}) \;=\; \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$$

is jointly upper semicontinuous and concave in its arguments $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ and is well defined regardless of whether the program is convex. Maximization of the dual function subject to the constraint $\boldsymbol{\mu} \geq \mathbf{0}$ is referred to as the dual program. It trades an often simpler objective function in the original (primal) program for simpler constraints in the dual program.

In the absence of convexity or Slater's constraint qualification, we can still recover a weak form of duality based on the identity

$$f(\hat{\boldsymbol{x}}) \;=\; \inf_{\boldsymbol{x}} \sup_{\boldsymbol{\lambda},\boldsymbol{\mu}\geq\mathbf{0}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}),$$

which stems from the fact

$$\sup_{\boldsymbol{\lambda},\boldsymbol{\mu}\geq\mathbf{0}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \;=\; \begin{cases} f(\boldsymbol{x}) & \boldsymbol{x} \text{ feasible} \\ \infty & \boldsymbol{x} \text{ infeasible.} \end{cases}$$

Because $\inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \leq \mathcal{L}(\hat{\boldsymbol{x}}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \leq f(\hat{\boldsymbol{x}})$, we can assert that

$$\sup_{\boldsymbol{\lambda},\boldsymbol{\mu}\geq\mathbf{0}} \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \;\leq\; f(\hat{\boldsymbol{x}}) \;=\; \inf_{\boldsymbol{x}} \sup_{\boldsymbol{\lambda},\boldsymbol{\mu}\geq\mathbf{0}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}). \tag{15.4}$$

In other words, the minimum value of the primal problem exceeds the maximum value of the dual problem. Slater's constraint qualification guarantees for a convex program that the duality gap

$$\inf_{\boldsymbol{x}} \sup_{\boldsymbol{\lambda},\boldsymbol{\mu}\geq\mathbf{0}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) - \sup_{\boldsymbol{\lambda},\boldsymbol{\mu}\geq\mathbf{0}} \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$$

vanishes when the primal problem has a finite minimum. Weak duality also makes it evident that if the primal program is unbounded below, then the dual program has no feasible point, and that if the dual program is unbounded above, then the primal program has no feasible point.

In a convex primal program, the equality constraint functions $g_i(\boldsymbol{x})$ are affine. If the inequality constraint functions $h_j(\boldsymbol{x})$ are also affine, then we can relate the dual program to the Fenchel conjugate $f^\star(\boldsymbol{y})$ of $f(\boldsymbol{x})$. Suppose we write the constraints as $\boldsymbol{V}\boldsymbol{x} = \boldsymbol{d}$ and $\boldsymbol{W}\boldsymbol{x} \leq \boldsymbol{e}$. Then the dual function equals

$$\begin{aligned} \mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu}) &= \inf_{\boldsymbol{x}} [f(\boldsymbol{x}) + \boldsymbol{\lambda}^*(\boldsymbol{V}\boldsymbol{x} - \boldsymbol{d}) + \boldsymbol{\mu}^*(\boldsymbol{W}\boldsymbol{x} - \boldsymbol{e})] \\ &= -\boldsymbol{\lambda}^*\boldsymbol{d} - \boldsymbol{\mu}^*\boldsymbol{e} + \inf_{\boldsymbol{x}} [f(\boldsymbol{x}) + (\boldsymbol{V}^*\boldsymbol{\lambda} + \boldsymbol{W}^*\boldsymbol{\mu})^*\boldsymbol{x}] \\ &= -\boldsymbol{\lambda}^*\boldsymbol{d} - \boldsymbol{\mu}^*\boldsymbol{e} - f^\star(-\boldsymbol{V}^*\boldsymbol{\lambda} - \boldsymbol{W}^*\boldsymbol{\mu}). \end{aligned} \tag{15.5}$$

It may be that $f^\star(\boldsymbol{y})$ equals ∞ for certain values of $\boldsymbol{y}$, but we can ignore these values in maximizing $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu})$. As pointed out in Proposition 14.3.1, $f^\star(\boldsymbol{y})$ is a closed convex function.

The dual function may be differentiable even when the objective function or one of the constraints is not. Let $\boldsymbol{\omega} = (\boldsymbol{\lambda}, \boldsymbol{\mu})$, and suppose that the solution $\boldsymbol{x}(\boldsymbol{\omega})$ of $\mathcal{D}(\boldsymbol{\omega}) = \mathcal{L}(\boldsymbol{x}, \boldsymbol{\omega})$ is unique and depends continuously on $\boldsymbol{\omega}$. The inequalities

$$\begin{aligned} \mathcal{D}(\boldsymbol{\omega}_1) &= \mathcal{L}[\boldsymbol{x}(\boldsymbol{\omega}_1), \boldsymbol{\omega}_1] \leq \mathcal{L}[\boldsymbol{x}(\boldsymbol{\omega}_2), \boldsymbol{\omega}_1] \\ \mathcal{D}(\boldsymbol{\omega}_2) &= \mathcal{L}[\boldsymbol{x}(\boldsymbol{\omega}_2), \boldsymbol{\omega}_2] \leq \mathcal{L}[\boldsymbol{x}(\boldsymbol{\omega}_1), \boldsymbol{\omega}_2] \end{aligned}$$

can be re-expressed as

$$\begin{aligned} \sum_{i=1}^{p} [\lambda_{2i} - \lambda_{1i}] g_i[\boldsymbol{x}(\boldsymbol{\omega}_2)] + \sum_{j=1}^{q} [\mu_{2j} - \mu_{1j}] h_j[\boldsymbol{x}(\boldsymbol{\omega}_2)] &\leq \mathcal{D}(\boldsymbol{\omega}_2) - \mathcal{D}(\boldsymbol{\omega}_1) \\ \sum_{i=1}^{p} [\lambda_{2i} - \lambda_{1i}] g_i[\boldsymbol{x}(\boldsymbol{\omega}_1)] + \sum_{j=1}^{q} [\mu_{2j} - \mu_{1j}] h_j[\boldsymbol{x}(\boldsymbol{\omega}_1)] &\geq \mathcal{D}(\boldsymbol{\omega}_2) - \mathcal{D}(\boldsymbol{\omega}_1) \end{aligned}$$

Taking an appropriate convex combination of the left-hand sides of these two inequalities allows us to write $\mathcal{D}(\boldsymbol{\omega}_2) - \mathcal{D}(\boldsymbol{\omega}_1) = s(\boldsymbol{\omega}_2, \boldsymbol{\omega}_1)(\boldsymbol{\omega}_2 - \boldsymbol{\omega}_1)$ for a slope function $s(\boldsymbol{\omega}_2, \boldsymbol{\omega}_1)$. The slope requirement

$$\lim_{\boldsymbol{\omega}_2 \to \boldsymbol{\omega}_1} s(\boldsymbol{\omega}_2, \boldsymbol{\omega}_1) = [g_1(\boldsymbol{x}), \ldots, g_p(\boldsymbol{x}), h_1(\boldsymbol{x}), \ldots, h_q(\boldsymbol{x})]$$

with $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{\omega}_1)$ follows directly from the assumed continuity of the constraints and the solution vector $\boldsymbol{x}(\boldsymbol{\omega})$. Thus, $\mathcal{D}(\boldsymbol{\omega})$ is differentiable at $\boldsymbol{\omega}_1$.

15.5 Examples of Dual Programs

Here are some examples of dual programs that feature the close ties between duality and the Fenchel conjugate. We start with linear and quadratic programs, the simplest and most useful examples, and progress to more sophisticated examples.

Example 15.5.1 *Dual of a Linear Program*

The standard linear program minimizes $f(\boldsymbol{x}) = \boldsymbol{z}^*\boldsymbol{x}$ subject to the linear equality constraints $\boldsymbol{V}\boldsymbol{x} = \boldsymbol{d}$ and the nonnegativity constraints $\boldsymbol{x} \geq \boldsymbol{0}$. It is obvious that $f^\star(\boldsymbol{y}) = \infty$ unless $\boldsymbol{y} = \boldsymbol{z}$, in which case $f^\star(\boldsymbol{z}) = 0$. Thus in view of equation (15.5), the dual program maximizes $-\boldsymbol{\lambda}^*\boldsymbol{d}$ subject to the constraints $-\boldsymbol{V}^*\boldsymbol{\lambda} + \boldsymbol{\mu} = \boldsymbol{z}$ and $\boldsymbol{\mu} \geq \boldsymbol{0}$. Later we will demonstrate that there is no duality gap when either the primal or dual program has a finite solution [87, 183]. ■

Example 15.5.2 *Dual of a Strictly Convex Quadratic Program*

We have repeatedly visited the problem of minimizing the strictly convex function $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x} + c$ subject to the linear equality constraints $\boldsymbol{V}\boldsymbol{x} = \boldsymbol{d}$ and the linear inequality constraints $\boldsymbol{W}\boldsymbol{x} \leq \boldsymbol{e}$. Ignoring the constraints, $f(\boldsymbol{x})$ achieves its minimum value $-\frac{1}{2}\boldsymbol{b}^*\boldsymbol{A}^{-1}\boldsymbol{b} + c$ at the point $\boldsymbol{x} = -\boldsymbol{A}^{-1}\boldsymbol{b}$. In Example 14.3.1, we calculated the Fenchel conjugate

$$f^\star(\boldsymbol{y}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{b})^*\boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{b}) - c.$$

If there are no equality constraints, then the dual program maximizes the quadratic

$$\begin{aligned}
\mathcal{D}(\boldsymbol{\mu}) &= -\boldsymbol{\mu}^*\boldsymbol{e} - f^\star(-\boldsymbol{W}^*\boldsymbol{\mu}) \\
&= -\boldsymbol{\mu}^*\boldsymbol{e} - \frac{1}{2}(\boldsymbol{W}^*\boldsymbol{\mu} + \boldsymbol{b})^*\boldsymbol{A}^{-1}(\boldsymbol{W}^*\boldsymbol{\mu} + \boldsymbol{b}) + c \\
&= -\frac{1}{2}\boldsymbol{\mu}^*\boldsymbol{W}\boldsymbol{A}^{-1}\boldsymbol{W}^*\boldsymbol{\mu} - (\boldsymbol{e} + \boldsymbol{W}\boldsymbol{A}^{-1}\boldsymbol{b})^*\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{b}^*\boldsymbol{A}^{-1}\boldsymbol{b} + c
\end{aligned}$$

subject to $\boldsymbol{\mu} \geq \boldsymbol{0}$. The dual program is easier to solve than the primal program when the number of inequality constraints is small and the number of variables is large.

In the presence of equality constraints and the absence of inequality constraints, the maximum of the dual function $\mathcal{D}(\boldsymbol{\lambda}) = -\boldsymbol{\lambda}^*\boldsymbol{d} - f^\star(-\boldsymbol{V}^*\boldsymbol{\lambda})$ occurs where

$$\begin{aligned}
\boldsymbol{0} &= -\boldsymbol{d} + \boldsymbol{V}\nabla f^\star(-\boldsymbol{V}^*\boldsymbol{\lambda}) \\
&= -\boldsymbol{d} + \boldsymbol{V}\boldsymbol{A}^{-1}(-\boldsymbol{V}^*\boldsymbol{\lambda} - \boldsymbol{b}).
\end{aligned}$$

If $\boldsymbol{V}$ has full row rank, then the last equation has solution

$$\boldsymbol{\lambda} \;=\; -(\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{V}^*)^{-1}(\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{b}+\boldsymbol{d}).$$

This is just the Lagrange multiplier calculated in Example 5.2.6 and Proposition 5.2.2. Given the multiplier, it is straightforward to calculate the optimal value of the primal variable $\boldsymbol{x}$. ■

Example 15.5.3 *Dual of a Linear Semidefinite Program*

The linear semidefinite programming problem consists in minimizing the trace function $\boldsymbol{X} \mapsto \operatorname{tr}(\boldsymbol{C}\boldsymbol{X})$ over the cone of positive semidefinite matrices S_+^n subject to the linear constraints $\operatorname{tr}(\boldsymbol{A}_i\boldsymbol{X}) = b_i$ for $1 \le i \le p$. Here $\boldsymbol{C}$ and the $\boldsymbol{A}_i$ are assumed symmetric. According to Sylvester's criterion, the constraint $\boldsymbol{B} \in S_+^n$ involves a complicated system of nonlinear inequalities. It is conceptually simpler to rewrite the constraint $\boldsymbol{B} \in S_+^n$ as $-\lambda_1(\boldsymbol{X}) \le 0$, where $\lambda_1(\boldsymbol{X})$ is the minimum eigenvalue of $\boldsymbol{X}$. Example 6.3.8 proves that $\lambda_1(\boldsymbol{X})$ is concave in $\boldsymbol{X}$.

In defining the dual problem, there is a generic way of accommodating complicated convex constraints. Suppose in the standard convex program we confine $\boldsymbol{x}$ to some closed convex set S in addition to imposing explicit equality and inequality constraints. Minimizing the objective function $f(\boldsymbol{x})$ subject to all of the constraints is equivalent to minimizing $f(\boldsymbol{x}) + \delta_S(\boldsymbol{x})$ subject to the functional constraints alone. In forming the dual we therefore take the infimum of the Lagrangian over $\boldsymbol{x}$ in S rather than over all $\boldsymbol{x}$.

In linear semidefinite programming, we therefore define the dual function

$$\begin{aligned}
\mathcal{D}(\boldsymbol{\lambda}) &= \inf_{\boldsymbol{X}\in S_+^n}\Big\{\operatorname{tr}(\boldsymbol{C}\boldsymbol{X}) + \sum_{i=1}^p \lambda_i[b_i - \operatorname{tr}(\boldsymbol{A}_i\boldsymbol{X})]\Big\} \\
&= \boldsymbol{b}^*\boldsymbol{\lambda} + \inf_{\boldsymbol{X}\in S_+^n} \operatorname{tr}\Big[\Big(\boldsymbol{C} - \sum_{i=1}^p \lambda_i\boldsymbol{A}_i\Big)\boldsymbol{X}\Big] \\
&= \boldsymbol{b}^*\boldsymbol{\lambda} + \inf_{\boldsymbol{X}\in S_+^n} t(\boldsymbol{X}).
\end{aligned}$$

If $t(\boldsymbol{X}) < 0$ for some $\boldsymbol{X} \in S_+^n$, then $t(\boldsymbol{X})$ can be made to approach $-\infty$ by replacing $\boldsymbol{X}$ by $c\boldsymbol{X}$ and taking c large. It follows that we should restrict the matrix $\boldsymbol{C} - \sum_{i=1}^p \lambda_i\boldsymbol{A}_i$ to lie in the negative of the polar cone $(S_+^n)^\circ$ of S_+^n. We know from Example 14.3.7 that $(S_+^n)^\circ = -S_+^n$. When this restriction holds and $\boldsymbol{C} - \sum_{i=1}^p \lambda_i\boldsymbol{A}_i$ is positive semidefinite, the minimum of $t(\boldsymbol{X})$ is achieved by setting $\boldsymbol{X} = \boldsymbol{0}$. The dual problem thus becomes one of maximizing $\boldsymbol{b}^*\boldsymbol{\lambda}$ subject to the condition that $\boldsymbol{C} - \sum_{i=1}^p \lambda_i\boldsymbol{A}_i$ is positive semidefinite. Slater's constraint qualification guaranteeing a duality gap of 0 is just the assumption that there exists a positive definite matrix $\boldsymbol{X}$ that is feasible for the primal problem. ■

Example 15.5.4 *Regression with a Non-Euclidean Norm*

In unconstrained least squares one minimizes the criterion $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|$ with respect to $\boldsymbol{\beta}$. If we substitute a non-Euclidean norm $\|\boldsymbol{z}\|_\dagger$ for the Euclidean norm $\|\boldsymbol{z}\|$, then the problem becomes much harder. Suppose the dual norm to $\|\boldsymbol{z}\|_\dagger$ is $\|\boldsymbol{w}\|_\star$. With no Lagrange multipliers in sight, the dual function is constant. However, let us repose the problem as minimizing $\|\boldsymbol{z}\|_\dagger$ subject to the linear constraint $\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$. The Lagrangian for this version of the problem is obviously $\mathcal{L}(\boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\lambda}) = \|\boldsymbol{z}\|_\dagger + \boldsymbol{\lambda}^*(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{z})$. In view of Example 14.3.5, the Fenchel conjugate of $f(\boldsymbol{z}, \boldsymbol{\beta}) = \|\boldsymbol{z}\|_\dagger$ is

$$f^\star(\boldsymbol{w}, \boldsymbol{\gamma}) \;=\; \sup_{\boldsymbol{z}, \boldsymbol{\beta}} (\boldsymbol{w}^*\boldsymbol{z} + \boldsymbol{\gamma}^*\boldsymbol{\beta} - \|\boldsymbol{z}\|_\dagger) \;=\; \delta_B(\boldsymbol{w}) + \delta_{\{\mathbf{0}\}}(\boldsymbol{\gamma}),$$

where B is the closed unit ball associated with the dual norm. Equation (15.5) therefore produces the dual function

$$\mathcal{D}(\boldsymbol{\lambda}) \;=\; \boldsymbol{\lambda}^*\boldsymbol{y} - \delta_B(\boldsymbol{\lambda}) - \delta_{\{\mathbf{0}\}}(\boldsymbol{X}^*\boldsymbol{\lambda}).$$

The dual problem consists of minimizing $-\boldsymbol{\lambda}^*\boldsymbol{y}$ subject to the constraints $\boldsymbol{X}^*\boldsymbol{\lambda} = \mathbf{0}$ and $\|\boldsymbol{\lambda}\|_\star \leq 1$.

We can reformulate the regression problem as minimizing the criterion $\frac{1}{2}\|\boldsymbol{z}\|_\dagger^2$ subject to the linear constraint $\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$. The Lagrangian now becomes

$$\mathcal{L}(\boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\lambda}) \;=\; \frac{1}{2}\|\boldsymbol{z}\|_\dagger^2 + \boldsymbol{\lambda}^*(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{z}),$$

and in view of Example 14.3.5 the dual function reduces to

$$\mathcal{D}(\boldsymbol{\lambda}) \;=\; \boldsymbol{\lambda}^*\boldsymbol{y} - \frac{1}{2}\|\boldsymbol{\lambda}\|_\star^2 - \delta_{\{\mathbf{0}\}}(\boldsymbol{X}^*\boldsymbol{\lambda}).$$

Hence, the dual problem consists of minimizing the function $-\boldsymbol{\lambda}^*\boldsymbol{y} + \frac{1}{2}\|\boldsymbol{\lambda}\|_\star^2$ subject to the single constraint $\boldsymbol{X}^*\boldsymbol{\lambda} = \mathbf{0}$. In this case there are two different dual problems. ■

Example 15.5.5 *Dual of a Geometric Program*

In passing to the dual, it is helpful to restate geometric programming as

$$\begin{aligned} &\text{minimize} \quad \ln\left(\sum_{k=1}^{m_0} e^{\boldsymbol{a}_{0k}^*\boldsymbol{x} + b_{ok}}\right) \\ &\text{subject to} \quad \ln\left(\sum_{k=1}^{m_j} e^{\boldsymbol{a}_{jk}^*\boldsymbol{x} + b_{jk}}\right) \;\leq\; 0, \quad 1 \leq j \leq q. \end{aligned}$$

In this convex version of geometric programming, the positive constants multiplying the monomials are just the exponentials $e^{b_{jk}}$. It is hard to take the Fenchel conjugate of the associated Lagrangian so we resort to the standard trick of introducing simpler functions $f_j(\boldsymbol{z}_j)$ with distinct arguments

and compensating equality constraints. If $f_j(\boldsymbol{z}_j)$ equals $\ln\left(\sum_k e^{z_{jk}}\right)$, then Example 14.3.2 calculates the entropy conjugate

$$f_j^\star(\boldsymbol{y}_j) \;=\; \begin{cases} \sum_k y_{jk} \ln y_{jk} & \text{for } \sum_k y_{jk} = 1 \text{ and all } y_{jk} \geq 0 \\ \infty & \text{otherwise.} \end{cases}$$

The equality constraint corresponding to $f_j(\boldsymbol{z}_j)$ is $\boldsymbol{A}_j\boldsymbol{x} + \boldsymbol{b}_j - \boldsymbol{z}_j = \boldsymbol{0}$ for a matrix $\boldsymbol{A}_j$ with rows $\boldsymbol{a}_{jk}^*$. In this notation, the simplified Lagrangian becomes

$$\begin{aligned} & \mathcal{L}(\boldsymbol{x}, \boldsymbol{z}_0, \ldots, \boldsymbol{z}_q, \boldsymbol{\lambda}, \boldsymbol{\mu}) \\ = \;& f_0(\boldsymbol{z}_0) + \sum_{j=1}^q \mu_j f_j(\boldsymbol{z}_j) + \sum_{j=0}^q \boldsymbol{\lambda}_j^*(\boldsymbol{A}_j\boldsymbol{x} + \boldsymbol{b}_j - \boldsymbol{z}_j) \\ = \;& f_0(\boldsymbol{z}_0) - \boldsymbol{\lambda}_0^*\boldsymbol{z}_0 + \sum_{j=1}^q \mu_j \left[f_j(\boldsymbol{z}_j) - \mu_j^{-1}\boldsymbol{\lambda}_j^*\boldsymbol{z}_j\right] + \sum_{j=0}^q \boldsymbol{\lambda}_j^*(\boldsymbol{A}_j\boldsymbol{x} + \boldsymbol{b}_j). \end{aligned}$$

It follows that the dual function equals

$$\begin{aligned} \mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu}) \;&=\; \inf_{\boldsymbol{x}, \boldsymbol{z}_0, \ldots, \boldsymbol{z}_q} \mathcal{L}(\boldsymbol{x}, \boldsymbol{z}_0, \ldots, \boldsymbol{z}_q, \boldsymbol{\lambda}, \boldsymbol{\mu}) \\ &=\; \sum_{j=0}^q \boldsymbol{\lambda}_j^*\boldsymbol{b}_j - f_0^\star(\boldsymbol{\lambda}_0) - \sum_{j=1}^q \mu_j f_j^\star(\mu_j^{-1}\boldsymbol{\lambda}_j) + \delta_{\boldsymbol{0}}\Big(\sum_{j=0}^q \boldsymbol{A}_j^*\boldsymbol{\lambda}_j\Big). \end{aligned}$$

The exceptional cases where one or more $\mu_j = 0$ require special treatment. When $\mu_j = 0$ but $\boldsymbol{\lambda}_j \neq \boldsymbol{0}$, the dual function must be interpreted as $-\infty$. When $\mu_j = 0$ and $\boldsymbol{\lambda}_j = \boldsymbol{0}$, the expression for $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is valid provided we interpret $\mu_j f_j^\star(\mu_j^{-1}\boldsymbol{\lambda}_j)$ as 0. Because of the nature of the entropy function, the constraints $\boldsymbol{\lambda}_j \geq \boldsymbol{0}$ for $0 \leq j \leq q$ and $\|\boldsymbol{\lambda}_0\|_1 = 1$ and $\|\boldsymbol{\lambda}_j\|_1 = \mu_j$ for $1 \leq j \leq q$ are implicit in the dual function. The constraints $\mu_j \geq 0$ for $1 \leq j \leq q$ are explicit, as is the constraint $\sum_{j=0}^q \boldsymbol{A}_j^*\boldsymbol{\lambda}_j = \boldsymbol{0}$. ■

Example 15.5.6 *Dykstra's Algorithm as Block Relaxation of the Dual*

Dykstra's problem can be restated as finding the minimum of the convex function

$$g(\boldsymbol{x}) \;=\; f(\boldsymbol{x}) + \sum_{i=0}^{r-1} \delta_{C_i}(\boldsymbol{x}) \;=\; f(\boldsymbol{x}) + \delta_C(\boldsymbol{x})$$

for $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2$, C the intersection of the closed convex sets C_0 through C_{r-1}, $\delta_C(\boldsymbol{x})$ the indicator function of C as defined in Example 14.3.4, and $\boldsymbol{z}$ the external point to C. As in Example 15.5.5, it is difficult to take the Fenchel conjugate of $g(\boldsymbol{x})$. Matters simplify tremendously if we replace the argument of each $\delta_{C_i}(\boldsymbol{x})$ by $\boldsymbol{x}_i$ and impose the constraint $\boldsymbol{x}_i = \boldsymbol{x}$. Consider therefore the Lagrangian

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{\Lambda}) \;=\; f(\boldsymbol{x}) + \sum_{i=0}^{r-1} \delta_{C_i}(\boldsymbol{x}_i) + \sum_{i=0}^{r-1} \boldsymbol{\lambda}_i^*(\boldsymbol{x} - \boldsymbol{x}_i)$$

with $\boldsymbol{X} = (\boldsymbol{x}, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{r-1})$ and $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_0, \ldots, \boldsymbol{\lambda}_{r-1})$. The dual function is

$$\begin{aligned}\mathcal{D}(\boldsymbol{\Lambda}) &= -\sup_{\boldsymbol{X}} \left[-\sum_{i=0}^{r-1} \boldsymbol{\lambda}_i^* \boldsymbol{x} - f(\boldsymbol{x}) + \sum_{i=0}^{r-1} \boldsymbol{\lambda}_i^* \boldsymbol{x}_i - \sum_{i=0}^{r-1} \delta_{C_i}(\boldsymbol{x}_i) \right] \\ &= -f^\star \left(-\sum_{i=0}^{r-1} \boldsymbol{\lambda}_i \right) - \sum_{i=0}^{r-1} \delta_{C_i}^\star(\boldsymbol{\lambda}_i).\end{aligned}$$

The dual problem consists of minimizing the convex function

$$f^\star \left(-\sum_{i=0}^{r-1} \boldsymbol{\lambda}_i \right) + \sum_{i=0}^{r-1} \delta_{C_i}^\star(\boldsymbol{\lambda}_i).$$

Dykstra's algorithm solves the dual problem by block descent [9, 26].

Suppose that we fix all $\boldsymbol{\lambda}_i$ except $\boldsymbol{\lambda}_j$. The stationarity condition requires $\mathbf{0}$ to belong to the subdifferential

$$\partial \Big[f^\star \Big(-\sum_{i=0}^{r-1} \boldsymbol{\lambda}_i \Big) + \delta_{C_j}^\star(\boldsymbol{\lambda}_j) \Big] = -\partial f^\star \Big(-\sum_{i=0}^{r-1} \boldsymbol{\lambda}_i \Big) + \partial \delta_{C_j}^\star(\boldsymbol{\lambda}_j).$$

It follows that there exists a vector $\boldsymbol{x}_j$ such that $\boldsymbol{x}_j \in \partial f^\star \left(-\sum_{i=0}^{r-1} \boldsymbol{\lambda}_i \right)$ and $\boldsymbol{x}_j \in \partial \delta_{C_j}^\star(\boldsymbol{\lambda}_j)$. Propositions 14.3.1 and 14.4.4 allow us to invert these two relations. Thus, $-\boldsymbol{\lambda}_j \in \partial [f(\boldsymbol{x}_j) + \sum_{i \neq j} \boldsymbol{\lambda}_i^* \boldsymbol{x}_j]$ and $\boldsymbol{\lambda}_j \in \partial \delta_{C_j}(\boldsymbol{x}_j)$, which together are equivalent to the primal stationarity condition

$$\mathbf{0} \in \partial \Big[f(\boldsymbol{x}_j) + \sum_{i \neq j} \boldsymbol{\lambda}_i^* \boldsymbol{x}_j + \partial \delta_{C_j}(\boldsymbol{x}_j) \Big].$$

As a consequence, it suffices to minimize

$$f(\boldsymbol{x}_j) + \sum_{i \neq j} \boldsymbol{\lambda}_i^* \boldsymbol{x}_j + \delta_{C_j}(\boldsymbol{x}_j) = \frac{1}{2} \Big\| \boldsymbol{x}_j + \sum_{i \neq j} \boldsymbol{\lambda}_i - \boldsymbol{z} \Big\|^2 + \delta_{C_j}(\boldsymbol{x}_j) + c,$$

where c is an irrelevant constant. But this problem is solved by projecting $\boldsymbol{z} - \sum_{i \neq j} \boldsymbol{\lambda}_i$ onto the convex set C_j.

The update of $\boldsymbol{\lambda}_j$ satisfies

$$\boldsymbol{\lambda}_j = -\partial \Big[f(\boldsymbol{x}_j) + \sum_{i \neq j} \boldsymbol{\lambda}_i^* \boldsymbol{x}_j \Big] = \boldsymbol{z} - \boldsymbol{x}_j - \sum_{i \neq j} \boldsymbol{\lambda}_i.$$

Given the converged values of the $\boldsymbol{\lambda}_j$, the optimal $\boldsymbol{x}$ can be recovered from the stationarity condition

$$\mathbf{0} = \partial f(\boldsymbol{x}) + \sum_{i=0}^{r-1} \boldsymbol{\lambda}_i = \boldsymbol{x} - \boldsymbol{z} + \sum_{i=0}^{r-1} \boldsymbol{\lambda}_i$$

for the Lagrangian $\mathcal{L}(\boldsymbol{X}, \boldsymbol{\Lambda})$ as $\boldsymbol{x} = \boldsymbol{z} - \sum_{i=0}^{r-1} \boldsymbol{\lambda}_i$ ■

Example 15.5.7 *Duffin's Counterexample*

Consider the convex program of minimizing $f(\boldsymbol{x}) = e^{-x_2}$ subject to the inequality constraint $h(\boldsymbol{x}) = \|\boldsymbol{x}\| - x_1 \leq 0$ on R^2. This problem does not satisfy Slater's condition because all feasible $\boldsymbol{x}$ satisfy $x_2 = 0$ and consequently $h(\boldsymbol{x}) = 0$ and $f(\boldsymbol{x}) = 1$. To demonstrate that there is a duality gap, we show that the dual function

$$\mathcal{D}(\mu) \;=\; \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \mu) \;=\; \inf_{\boldsymbol{x}} [e^{-x_2} + \mu(\|\boldsymbol{x}\| - x_1)]$$

is identically 0. Because $\mathcal{L}(\boldsymbol{x}, \mu) \geq 0$ for all $\boldsymbol{x}$ and $\mu \geq 0$, it suffices to prove that $\mathcal{L}(\boldsymbol{x}, \mu)$ can be made less than any positive ϵ. Choose an x_2 so that $e^{-x_2} < \epsilon/2$. Having chosen x_2, choose x_1 so that

$$\begin{aligned}
\sqrt{x_1^2 + x_2^2} - x_1 &= x_1\sqrt{1 + \frac{x_2^2}{x_1^2}} - x_1 \\
&\leq x_1\left(1 + \frac{x_2^2}{2x_1^2}\right) - x_1 \\
&= \frac{x_2^2}{2x_1} \\
&< \frac{\epsilon}{2\mu}.
\end{aligned}$$

With these choices we have $\mathcal{L}(\boldsymbol{x}, \mu) < \epsilon$. Thus, the minimum value 1 of the primal problem is strictly greater than the maximum value 0 of the dual problem. ■

Before turning to practical applications of duality in the next section, we would like to mention the fundamental theorem of linear programming. Our proof depends on the subtle properties of polyhedral sets sketched in Appendix A.3. Readers can digest this material at their leisure.

Proposition 15.5.1 *If either the primal or dual linear program formulated in Example 15.5.1 has a solution, then the other program has a solution as well. Furthermore, there is no duality gap.*

Proof: If the primal program has a solution, then inequality (15.4) shows that the dual program has an upper bound. Proposition A.3.5 therefore implies that the dual program has a solution. Conversely, if the dual program has a solution, then inequality (15.4) shows that the primal program has a lower bound. Since a linear function is both convex and concave, a second application of Proposition A.3.5 shows that the primal program has a solution. According to Example 6.5.5, the existence of either solution forces the preferred form of the Lagrange multiplier rule and hence implies no duality gap. ■

Example 15.5.8 *Von Neumann's Minimax Theorem*

Von Neumann's minimax theorem is one of the earliest results of game theory [266]. In purely mathematical terms, it can be stated as the identity

$$\min_{\boldsymbol{y}\in S}\max_{\boldsymbol{x}\in S}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{y} \;=\; \max_{\boldsymbol{x}\in S}\min_{\boldsymbol{y}\in S}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{y}, \tag{15.6}$$

where $\boldsymbol{A} = (a_{ij})$ is an $n \times n$ matrix and S is the unit simplex

$$S \;=\; \left\{\boldsymbol{z} \in \mathsf{R}^n : z_1 \geq 0, \ldots, z_n \geq 0, \sum_{i=1}^n z_i = 1\right\}.$$

It is possible to view the identity (15.6) as a manifestation of linear programming duality. As the reader can check (Problem 28), the primal program

$$\begin{aligned} &\text{minimize} && u \\ &\text{subject to} && \sum_{j=1}^n y_j \;=\; 1, \;\; \sum_{j=1}^n a_{ij}y_j \;\leq u \;\; \forall i, \;\; y_j \geq 0 \;\; \forall j \end{aligned}$$

has dual program

$$\begin{aligned} &\text{maximize} && v \\ &\text{subject to} && \sum_{i=1}^n x_i \;=\; 1, \;\; \sum_{i=1}^n x_i a_{ij} \;\geq v \;\; \forall j, \;\; x_i \geq 0 \;\; \forall i. \end{aligned}$$

Von Neumann's identity (15.6) is true because the primal and optimal values p and d of this linear program satisfy

$$\begin{aligned} \min_{\boldsymbol{y}\in S}\max_{\boldsymbol{x}\in S}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{y} \;&=\; \min_{\boldsymbol{y}\in S}\max_{1\leq i\leq n}\boldsymbol{e}_i^*\boldsymbol{A}\boldsymbol{y} \;=\; p \\ \max_{\boldsymbol{x}\in S}\min_{\boldsymbol{y}\in S}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{y} \;&=\; \max_{\boldsymbol{x}\in S}\min_{1\leq j\leq n}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{e}_j \;=\; d. \end{aligned}$$

Von Neumann's identity is a special case of the much more general minimax principle of Sion [155, 238]. This principle implies no duality gap in convex programming given appropriate compactness assumptions. ■

15.6 Practical Applications of Duality

The two examples of this section illustrate how duality is not just a theoretical construct. It can also lead to the discovery of concrete optimization algorithms. Practitioners of optimization should always bear this in mind as well as the lower bound offered by inequality (15.4). Of course, solving the dual problem is seldom enough. To realize the potential of duality, one must convert the solution of the dual problem into a solution of the primal problem.

Example 15.6.1 *The Power Plant Problem*

The power plant production problem [226] involves minimizing

$$f(\boldsymbol{x}) \quad = \quad \sum_{i=1}^{n} f_i(x_i)$$

subject to the constraints $0 \leq x_i \leq u_i$ for each i and $\sum_{i=1}^{n} x_i \geq d$. For plant i, x_i is the power output, u_i is the capacity, and $f_i(x_i)$ is the cost. The total demand is d. The Lagrangian for this minimization problem is

$$\mathcal{L}(\boldsymbol{x}, \mu) \quad = \quad \sum_{i=1}^{n} f_i(x_i) + \mu\Big(d - \sum_{i=1}^{n} x_i\Big).$$

As a consequence of the separability of the Lagrangian, the dual function can be expressed as

$$\mathcal{D}(\mu) \quad = \quad \mu d + \sum_{i=1}^{n} \min_{0 \leq x_i \leq u_i} \Big[f_i(x_i) - \mu x_i\Big].$$

For the quadratic choices $f_i(x_i) = a_i x_i + \frac{1}{2} b_i x_i^2$ with positive cost constants a_i and b_i, the problem is a convex program, and it is possible to explicitly solve for the dual. A brief calculation shows that optimal value of x_i is

$$\hat{x}_i \quad = \quad \begin{cases} 0 & 0 \leq \mu \leq a_i \\ \frac{\mu - a_i}{b_i} & a_i \leq \mu \leq a_i + b_i u_i \\ u_i & \mu \geq a_i + b_i u_i. \end{cases}$$

These solutions translate into the dual function

$$\mathcal{D}(\mu) \quad = \quad \mu d + \sum_{i=1}^{n} \begin{cases} 0 & 0 \leq \mu \leq a_i \\ -\frac{(\mu - a_i)^2}{2b_i} & a_i \leq \mu \leq a_i + b_i u_i \\ a_i u_i + \frac{1}{2} b_i u_i^2 - \mu u_i & \mu \geq a_i + b_i u_i. \end{cases}$$

and ultimately into the derivative $\mathcal{D}'(\mu) = d - \sum_{i=1}^{n} \hat{x}_i$. Because $\mathcal{D}(\mu)$ is concave, a stationary point furnishes the global maximum. It is straightforward to implement bisection to locate a stationary point. Steepest ascent is another route to maximizing $\mathcal{D}(\mu)$. Problem 26 asks the reader to consider the impact of assuming one or more $b_i = 0$. ■

Example 15.6.2 *Linear Classification*

Classification problems are ubiquitous in statistics. Section 13.8 discusses one approach to discriminant analysis. Here we take another motivated by hyperplane separation. The binary classification problem can be phrased in terms of a training sequence of observation vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ from R^n and

an associated sequence of population indicators $s_1, \ldots, s_m$ from $\{-1, +1\}$. In favorable situations, the two different populations can be separated by a hyperplane defined by a unit vector $\boldsymbol{z}$ and constants $c_1 \leq c_2$ in the sense that

$$\begin{aligned} \boldsymbol{z}^* \boldsymbol{v}_i &\leq c_1, \quad s_i = -1 \\ \boldsymbol{z}^* \boldsymbol{v}_i &\geq c_2, \quad s_i = +1. \end{aligned}$$

The optimal separation occurs when the difference $c_2 - c_1$ is maximized.

This linear classification problem can be simplified by rewriting the separation conditions as

$$\begin{aligned} \boldsymbol{z}^* \boldsymbol{v}_i - \frac{c_1 + c_2}{2} &\leq -\frac{c_2 - c_1}{2}, \quad s_i = -1 \\ \boldsymbol{z}^* \boldsymbol{v}_i - \frac{c_1 + c_2}{2} &\geq +\frac{c_2 - c_1}{2}, \quad s_i = +1. \end{aligned}$$

If we let $a = (c_2 - c_1)/2$, $b = (c_1 + c_2)/(c_2 - c_1)$, and $\boldsymbol{y} = a^{-1}\boldsymbol{z}$, then these become

$$\begin{aligned} \boldsymbol{y}^* \boldsymbol{v}_i - b &\leq -1, \quad s_i = -1 \\ \boldsymbol{y}^* \boldsymbol{v}_i - b &\geq +1, \quad s_i = +1. \end{aligned} \tag{15.7}$$

Thus, the linear classification problem reduces to minimizing the criterion $\frac{1}{2}\|\boldsymbol{y}\|^2$ subject to the inequality constraints (15.7). Observe that the constraint functions are linear in the parameter vector $(\boldsymbol{y}^*, b)^*$ in this semidefinite quadratic programming problem. Unfortunately, because the component b does not appear in the objective function $\frac{1}{2}\|\boldsymbol{y}\|^2$, Dykstra's algorithm does not apply. Once we find $\boldsymbol{y}$ and b, we can classify a new test vector $\boldsymbol{v}$ in the $s = -1$ population when $\boldsymbol{y}^*\boldsymbol{v} - b < 0$ and in the $s = +1$ population when $\boldsymbol{y}^*\boldsymbol{v} - b > 0$.

There is no guarantee that a feasible vector $(\boldsymbol{y}^*, b)^*$ exists for the linear classification problem as stated. A more realistic version of the problem imposes the inequality constraints

$$s_i(\boldsymbol{y}^* \boldsymbol{v}_i - b) \geq 1 - \epsilon_i \tag{15.8}$$

using a slack variable $\epsilon_i \geq 0$. To penalize deviation from the ideal of perfect separation by a hyperplane, we modify the objective function to be

$$f(\boldsymbol{y}, b, \boldsymbol{\epsilon}) = \frac{1}{2}\|\boldsymbol{y}\|^2 + \delta \sum_{i=1}^{m} \epsilon_i \tag{15.9}$$

for some tuning constant $\delta > 0$. The constraints (15.8) and $\epsilon_i \geq 0$ are again linear in the parameter vector $\boldsymbol{x} = (\boldsymbol{y}^*, b, \boldsymbol{\epsilon}^*)^*$. The Lagrangian

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}, b, \boldsymbol{\epsilon}, \boldsymbol{\mu}) &= \frac{1}{2}\|\boldsymbol{y}\|^2 + \delta \sum_{i=1}^{m} \epsilon_i - \sum_{i=1}^{m} \mu_{m+i}\epsilon_i \\ &\quad + \sum_{i=1}^{m} \mu_i[-s_i(\boldsymbol{v}_i^* \boldsymbol{y} - b) + 1 - \epsilon_i]\end{aligned}$$

involves $2m$ nonnegative multipliers $\mu_1, \ldots, \mu_{2m}$.

It is simpler to solve the dual problem than the primal problem. We can formulate the dual by following the steps of Example 15.5.2. If we express the inequality constraints as $\boldsymbol{W}\boldsymbol{x} \leq \boldsymbol{e}$, then the matrix transpose $\boldsymbol{W}^*$ and the vector $\boldsymbol{e}$ are

$$\boldsymbol{W}^* = -\begin{pmatrix} s_1\boldsymbol{v}_1 & \cdots & s_m\boldsymbol{v}_m & \boldsymbol{0} \\ -s_1 & \cdots & -s_m & 0 \\ & \boldsymbol{I}_m & & \boldsymbol{I}_m \end{pmatrix}, \qquad \boldsymbol{e} = \begin{pmatrix} -\boldsymbol{1} \\ \boldsymbol{0} \end{pmatrix}$$

in the current setting. The dual problem of consists of maximizing the function $-\boldsymbol{\mu}^*\boldsymbol{e} - f^\star(-\boldsymbol{W}^*\boldsymbol{\mu})$ subject to the constraints $\mu_i \geq 0$ and restrictions imposed by the essential domain of the Fenchel conjugate $f^\star(\boldsymbol{p}, q, \boldsymbol{r})$. An easy calculation gives

$$\begin{aligned}f^\star(\boldsymbol{p}, q, \boldsymbol{r}) &= \sup_{\boldsymbol{y}, b, \boldsymbol{\epsilon}} \left[\boldsymbol{p}^*\boldsymbol{y} + qb + \boldsymbol{r}^*\boldsymbol{\epsilon} - \frac{1}{2}\|\boldsymbol{y}\|^2 - \delta \sum_{i=1}^{m} \epsilon_i\right] \\ &= \begin{cases} \infty & q \neq 0 \text{ or } r_i \neq \delta \text{ for some } i \\ \boldsymbol{p}^*\boldsymbol{p} - \frac{1}{2}\|\boldsymbol{p}\|^2 & \text{otherwise} \end{cases} \\ &= \begin{cases} \infty & q \neq 0 \text{ or } r_i \neq \delta \text{ for some } i \\ \frac{1}{2}\|\boldsymbol{p}\|^2 & \text{otherwise.} \end{cases}\end{aligned}$$

To match $-\boldsymbol{W}^*\boldsymbol{\mu}$ to the indicated essential domain, note that the restriction $q = 0$ entails the constraint $\sum_{i=1}^m s_i\mu_i = 0$, and the restriction $r_i = \delta$ entails the constraint $\mu_i + \mu_{m+i} = \delta$ and therefore the bound $\mu_i \leq \delta$. For the vector $\boldsymbol{p}$ we substitute the linear combination $\sum_{i=1}^m s_i\mu_i\boldsymbol{v}_i$. Hence, the dual problem consists in maximizing

$$\begin{aligned}-\boldsymbol{\mu}^*\boldsymbol{e} - f^\star(-\boldsymbol{W}^*\boldsymbol{\mu}) &= \sum_{i=1}^{m} \mu_i - \frac{1}{2}\left\|\sum_{i=1}^{m} s_i\mu_i\boldsymbol{v}_i\right\|^2 \qquad (15.10) \\ &= \sum_{i=1}^{m} \mu_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} s_i\mu_i\boldsymbol{v}_i^*\boldsymbol{v}_j s_j\mu_j\end{aligned}$$

subject to the constraints $\sum_{i=1}^m s_i\mu_i = 0$ and $0 \leq \mu_i \leq \delta$ for all i.

Solving the dual problem fortunately leads to a straightforward solution of the primal problem. For instance, the Lagrangian conditions

$$\frac{\partial}{\partial y_j}\mathcal{L}(\boldsymbol{y}, b, \boldsymbol{\epsilon}, \boldsymbol{\mu}) = 0$$

give $\boldsymbol{y} = \sum_{i=1}^m \mu_i s_i \boldsymbol{v}_i$. The Lagrangian condition

$$\frac{\partial}{\partial \epsilon_j} \mathcal{L}(\boldsymbol{y}, b, \boldsymbol{\epsilon}, \boldsymbol{\mu}) \quad = \quad \delta - \mu_j - \mu_{m+j} \quad = \quad 0$$

implies that $\mu_{m+j} = \delta - \mu_j$. The complementary slackness conditions

$$\begin{aligned} 0 &= \mu_i[-s_i(\boldsymbol{v}_i^* \boldsymbol{y} - b) + 1 - \epsilon_i] \\ 0 &= \mu_{m+j} \epsilon_j \end{aligned}$$

can be used to determine b and ϵ_i for $1 \leq i \leq m$. If we choose an index j such that $0 < \mu_j < \delta$, then $\mu_{m+j} > 0$ and $\epsilon_j = 0$. It follows that $-s_j(\boldsymbol{v}_j^* \boldsymbol{y} - b) + 1 = 0$ and that $b = \boldsymbol{v}_j^* \boldsymbol{y} - s_j$ since $s_j = \pm 1$. Given b, all ϵ_j with $\mu_j > 0$ are determined. If $\mu_j = 0$, then $\mu_{m+j} = \delta > 0$ and $\epsilon_j = 0$.

Despite these interesting maneuvers, we have not actually shown how to solve the dual problem. For linear classification problems with many training vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ in a high-dimensional space R^n, it is imperative to formulate an efficient algorithm. One possibility is to try a penalty method. If we subtract the square penalty $\omega(\sum_{i=1}^m s_i \mu_i)^2$ from the dual function (15.10), then the remaining box constraints are consistent with coordinate ascent. Sending ω to ∞ then produces the dual solution. Alternatively, subtracting the penalty $\omega|\sum_{i=1}^m s_i \mu_i|$ converts the dual problem into an exact penalty problem and opens up the possibility of path following as described in Chap. 16. ■

15.7 Problems

1. Let $C \subset \mathsf{R}^2$ be the cone defined by the constraint $x_1 \leq x_2$. Show that the projection operator $P_C(\boldsymbol{x})$ has components

$$P_C(\boldsymbol{x})_j \quad = \quad \begin{cases} x_j & x_1 \leq x_2 \\ \frac{1}{2}(x_1 + x_2) & x_1 > x_2. \end{cases}$$

   Let $S \subset \mathsf{R}^3$ be the cone defined by the constraint $x_2 \leq \frac{1}{2}(x_1 + x_3)$. Show that the projection operator $P_S(\boldsymbol{x})$ has components

$$P_S(\boldsymbol{x})_j \quad = \quad \begin{cases} x_j & x_2 \leq \frac{1}{2}(x_1 + x_3) \\ \frac{5}{6}x_1 + \frac{1}{3}x_2 - \frac{1}{6}x_3 & j = 1 \text{ and } x_2 > \frac{1}{2}(x_1 + x_3) \\ \frac{1}{3}x_1 + \frac{1}{3}x_2 + \frac{1}{3}x_3 & j = 2 \text{ and } x_2 > \frac{1}{2}(x_1 + x_3) \\ -\frac{1}{6}x_1 + \frac{1}{3}x_2 + \frac{5}{6}x_3 & j = 3 \text{ and } x_2 > \frac{1}{2}(x_1 + x_3). \end{cases}$$

2. If A and B are two closed convex sets, then prove that projection onto the Cartesian product $A \times B$ is effected by the Cartesian product operator $(\boldsymbol{x}, \boldsymbol{y}) \mapsto [P_A(\boldsymbol{x}), P_B(\boldsymbol{y})]$.

3. Program and test either an algorithm for projection onto the unit simplex or the pool adjacent violators algorithm.

4. If C is a closed convex set in R^n and $\boldsymbol{x} \notin C$, then demonstrate that

$$\operatorname{dist}(\boldsymbol{x}, C) = \inf_{\boldsymbol{y} \in C} \sup_{\|\boldsymbol{z}\|=1} \boldsymbol{z}^*(\boldsymbol{x} - \boldsymbol{y}) = \sup_{\|\boldsymbol{z}\|=1} \inf_{\boldsymbol{y} \in C} \boldsymbol{z}^*(\boldsymbol{x} - \boldsymbol{y}).$$

   Also prove that there exists a unit vector $\boldsymbol{z}$ with

$$\operatorname{dist}(\boldsymbol{x}, C) = \inf_{\boldsymbol{y} \in C} \boldsymbol{z}^*(\boldsymbol{x} - \boldsymbol{y}).$$

   (Hints: The first equality follows from the Cauchy-Schwarz inequality and the definition of $\operatorname{dist}(\boldsymbol{x}, C)$. The rest of the problem depends on the Cauchy-Schwarz inequality, the particular choice

$$\boldsymbol{z} = \operatorname{dist}(\boldsymbol{x}, C)^{-1}[\boldsymbol{x} - P_C(\boldsymbol{x})],$$

   and the obtuse angle criterion.)

5. Let S and T be subspaces of R^n. Demonstrate that the projections P_S and P_T satisfy $P_S P_T = P_{S \cap T}$ if and only if $P_S P_T = P_T P_S$.

6. Let C be a closed convex set in R^n. Show that

   (a) $\operatorname{dist}(\boldsymbol{x} + \boldsymbol{y}, C + \boldsymbol{y}) = \operatorname{dist}(\boldsymbol{x}, C)$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$.

   (b) $P_{C+\boldsymbol{y}}(\boldsymbol{x} + \boldsymbol{y}) = P_C(\boldsymbol{x}) + \boldsymbol{y}$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$.

   (c) $\operatorname{dist}(a\boldsymbol{x}, aC) = |a| \operatorname{dist}(\boldsymbol{x}, C)$ for all $\boldsymbol{x}$ and real a.

   (d) $P_{aC}(a\boldsymbol{x}) = aP_C(\boldsymbol{x})$ for all $\boldsymbol{x}$ and real a.

   Let S be a subspace of R^n. Show that

   (a) $\operatorname{dist}(\boldsymbol{x} + \boldsymbol{y}, S) = \operatorname{dist}(\boldsymbol{x}, S)$ for all $\boldsymbol{x} \in \mathsf{R}^n$ and $\boldsymbol{y} \in S$.

   (b) $P_S(\boldsymbol{x} + \boldsymbol{y}) = P_S(\boldsymbol{x}) + \boldsymbol{y}$ for all $\boldsymbol{x} \in \mathsf{R}^n$ and $\boldsymbol{y} \in S$.

   (c) $\operatorname{dist}(a\boldsymbol{x}, S) = |a| \operatorname{dist}(\boldsymbol{x}, S)$ for all $\boldsymbol{x} \in \mathsf{R}^n$ and real a.

   (d) $P_S(a\boldsymbol{x}) = aP_S(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathsf{R}^n$ and real a.

7. Let $\boldsymbol{M}$ be an $n \times n$ symmetric matrix with spectral decomposition $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^*$, where $\boldsymbol{U}$ is an orthogonal matrix and $\boldsymbol{D}$ is a diagonal matrix with ith diagonal entry d_i. Prove that the Frobenius norm $\|\boldsymbol{M} - P_S(\boldsymbol{M})\|_F$ is minimized over the closed convex cone S of positive semidefinite matrices by taking $P_S(\boldsymbol{M}) = \boldsymbol{U}\boldsymbol{D}_+\boldsymbol{U}^*$, where $\boldsymbol{D}_+$ is diagonal with ith diagonal entry $\max\{d_i, 0\}$.

8. Let $C = \{(\boldsymbol{x}, t) \in \mathsf{R}^{n+1} : \|\boldsymbol{x}\| \leq t\}$ denote the ice cream cone in R^{n+1}. Verify the projection formulas

$$P_C[(\boldsymbol{x}, t)] = \begin{cases} (\boldsymbol{x}, t) & \|\boldsymbol{x}\| \leq t \text{ and } t \geq 0 \\ (\boldsymbol{0}, 0) & \|\boldsymbol{x}\| \leq -t \text{ and } t \leq 0 \\ \left(\frac{\|\boldsymbol{x}\|+t}{2\|\boldsymbol{x}\|}\boldsymbol{x}, \frac{\|\boldsymbol{x}\|+t}{2}\right) & \text{otherwise.} \end{cases}$$

9. The convex regression example in Sect. 15.2 implicitly assumes that the regression function is defined on the integers $1, \ldots, n$. Consider instead the problem of finding a convex function $f(\boldsymbol{x})$ on R^m such that the sum of squares $\sum_{i=1}^n [y_i - f(\boldsymbol{x}_i)]^2$ is minimized. Demonstrate that this alternative problem can be rephrased as the quadratic programming problem of minimizing $\sum_{i=1}^n (y_i - z_i)^2$ subject to the restrictions

$$z_j \geq z_i + \boldsymbol{g}_i^*(\boldsymbol{x}_j - \boldsymbol{x}_i)$$

for all i and $j \neq i$. Here the unknown subgradients $\boldsymbol{g}_i$ must be found along with the function values z_i at the specified points $\boldsymbol{x}_i$ [19].

10. Verify the projection formula in Problem 11 of Chap. 5 by invoking the obtuse angle criterion.

11. Suppose C is a closed convex set wholly contained within an affine subspace $V = \{\boldsymbol{y} \in \mathsf{R}^n : \boldsymbol{A}\boldsymbol{y} = \boldsymbol{b}\}$. For $\boldsymbol{x} \notin V$ demonstrate the projection identity $P_C(\boldsymbol{x}) = P_C \circ P_V(\boldsymbol{x})$ [196]. (Hint: Consider the equality

$$\begin{aligned}[\boldsymbol{x} - P_C(\boldsymbol{x})]^*[\boldsymbol{y} - P_C(\boldsymbol{x})] &= [\boldsymbol{x} - P_V(\boldsymbol{x})]^*[\boldsymbol{y} - P_C(\boldsymbol{x})] \\ &\quad + [P_V(\boldsymbol{x}) - P_C(\boldsymbol{x})]^*[\boldsymbol{y} - P_C(\boldsymbol{x})]\end{aligned}$$

with the obtuse angle criterion in mind.)

12. For positive numbers $c_1, \ldots, c_n$ and nonnegative numbers $b_1, \ldots, b_n$ satisfying $\sum_{i=1}^n c_i b_i \leq 1$, define the truncated simplex

$$S = \left\{\boldsymbol{y} \in \mathsf{R}^n : \sum_{i=1}^n c_i y_i = 1,\ y_i \geq b_i,\ 1 \leq i \leq n\right\}.$$

If $\boldsymbol{x} \in \mathsf{R}^n$ has coordinate sum $\sum_{i=1}^n c_i x_i = 1$, then prove that the closest point $\boldsymbol{y}$ in S to $\boldsymbol{x}$ satisfies the Lagrange multiplier conditions

$$y_i - x_i + \lambda c_i - \mu_i = 0$$

for appropriate multipliers λ and $\mu_i \geq 0$. Further show that

$$\lambda = \frac{\boldsymbol{c}^*\boldsymbol{\mu}}{\|\boldsymbol{c}\|^2} \geq 0.$$

Why does it follow that $y_i = b_i$ whenever $x_i < b_i$? Prove that the Lagrange multiplier conditions continue to hold when $x_i < b_i$ if we replace x_i by b_i and μ_i by λc_i. Since the Lagrange multiplier conditions are sufficient as well as necessary in convex programming, this demonstrates that (a) we can replace each coordinate x_i by $\max\{x_i, b_i\}$ without changing the projection $\boldsymbol{y}$ of $\boldsymbol{x}$ onto S, and (b) $\boldsymbol{y}$ can be viewed as a point in a similar simplex in a reduced number of dimensions when one or more $x_i \leq b_i$ [196].

13. Michelot's [196] algorithm for projecting a point $\boldsymbol{x}$ onto the simplex S defined in Problem 12 cycles through the following steps:

    (a) Project onto the affine subspace $V_n = \{\boldsymbol{y} \in \mathsf{R}^n : \sum_i c_i y_i = 1\}$,

    (b) Replace each coordinate x_i by $\max\{x_i, b_i\}$,

    (c) Reduce the dimension n whenever some $x_i = b_i$.

    In view of Problems 12 and 13, demonstrate that Michelot's algorithm converges to the correct solution in at most n steps. Explicitly solve the Lagrange multiplier problem corresponding to step (a). Program and test the algorithm.

14. Consider the problem of projecting a point $\boldsymbol{x}$ onto the ℓ_1 ball

$$B \quad = \quad \{\boldsymbol{y} : \|\boldsymbol{y} - \boldsymbol{z}\|_1 \leq r\}.$$

    Show that it suffices to project $\boldsymbol{x} - \boldsymbol{z}$ onto $\{\boldsymbol{w} : \|\boldsymbol{w}\|_1 \leq r\}$ and then translate the solution $\hat{\boldsymbol{w}}$ by $\boldsymbol{z}$. Hence, assume $\boldsymbol{z} = \boldsymbol{0}$ without loss of generality. Now argue that every entry of a solution $\hat{\boldsymbol{y}}$ should have the same sign as the corresponding entry of $\boldsymbol{x}$ and that when $x_i = 0$, it does no harm to take $\hat{y}_i \geq 0$. Finally, sketch how projection onto an ℓ_1 ball can be achieved by projection onto the unit simplex.

15. The $\ell_{1,2}$ norm on R^n is useful in group penalties [230]. Suppose the sets σ_g partition $\{1, \ldots, n\}$ into groups with g as the group index. For $\boldsymbol{x} \in \mathsf{R}^n$ let $\boldsymbol{x}_{\sigma_g}$ denote the vector formed by taking the components of $\boldsymbol{x}$ derived from σ_g. The $\ell_{1,2}$ norm equals

$$\|\boldsymbol{x}\|_{1,2} \quad = \quad \sum_g \|\boldsymbol{x}_{\sigma_g}\|.$$

    Check that $\|\boldsymbol{x}\|_{1,2}$ satisfies the properties of a norm. Now consider projecting a point $\boldsymbol{x}$ onto the ball $B_r = \{\boldsymbol{y} : \|\boldsymbol{y}\|_{1,2} \leq r\}$. If we let $c_g = \|\boldsymbol{x}_{\sigma_g}\|$ and suppose that $\sum_g c_g \leq r$, then the solution is $\boldsymbol{x}$. Thus, assume the contrary. If any $c_g = \|\boldsymbol{x}_{\sigma_g}\| = 0$, argue that one should take $\boldsymbol{y}_{\sigma_g} = \boldsymbol{0}$. Assume therefore that $c_g > 0$ for every g. Collect the Euclidean distances c_g into a vector $\boldsymbol{c}$, and project $\boldsymbol{c}$ onto the closest point $\boldsymbol{d}$ in the ℓ_1 ball of radius r. This can be accomplished

by the algorithm described in Problem 14. Finally, prove that the projection $\boldsymbol{y}$ of $\boldsymbol{x}$ onto B_r satisfies $\boldsymbol{y}_{\sigma_g} = d_g c_g^{-1} \boldsymbol{x}_{\sigma_g}$. (Hints: Suppose that $r_g \geq 0$ for all g and $\sum_g r_g \leq r$. The problem of minimizing $\|\boldsymbol{y} - \boldsymbol{x}\|_{1,2}^2$ subject to $\|\boldsymbol{y}_{\sigma_g}\| \leq r_g$ for each g is separable and can be solved by projection onto the pertinent Euclidean balls. Substitution now leads to the secondary problem of minimizing

$$\sum_g (c_g - r_g)^2 1_{\{r_g \leq c_g\}} \tag{15.11}$$

subject to $r_g \geq 0$ for all g and $\sum_g r_g \leq r$. Suppose the optimal choice of the vector $\boldsymbol{r}$ involves $r_g > c_g$ for some g. There must be a corresponding g' with $r_{g'} < c_{g'}$. One can decrease the criterion (15.11) by decreasing r_g and increasing $r_{g'}$. Hence, all $r_g \leq c_g$ at the optimal point, and one can dispense with the indicators $1_{\{r_g \geq c_g\}}$ and minimize the criterion (15.11) by ordinary projection.)

16. A polyhedral set is the nonempty intersection of a finite number of halfspaces. Program Dykstra's algorithm for projection onto the closest point of an arbitrary polyhedral set. Also program cyclic projection as suggested in Proposition 15.3.1, and compare it to Dykstra's algorithm on one or more test problems.

17. Demonstrate that the map $f(x) = x + e^{-x}$ is contractive on $[0, \infty)$ but lacks a fixed point. Is $f(x)$ strictly contractive?

18. Prove that the iteration scheme

$$x_{m+1} = \frac{1}{1 + x_m}$$

with domain $[0, \infty)$ has one fixed point y and that y is globally attractive. Is the function $f(x) = (1 + x)^{-1}$ contractive or strictly contractive?

19. Consider the map

$$T(\boldsymbol{x}) = \begin{pmatrix} \frac{1}{2(1+x_2)} \\ \frac{1}{2} e^{-x_1} \end{pmatrix}$$

from $\mathsf{R}_+^2 = \{\boldsymbol{x} \in \mathsf{R}^2 : \boldsymbol{x} \geq \boldsymbol{0}\}$ to itself. Calculate the differential

$$dT(\boldsymbol{x}) = \begin{pmatrix} 0 & -\frac{1}{2(1+x_2)^2} \\ -\frac{1}{2} e^{-x_1} & 0 \end{pmatrix},$$

and show that $\|dT(\boldsymbol{x})\| \leq \frac{1}{2}$ for all $\boldsymbol{x}$. Deduce the mean value inequality

$$\|T(\boldsymbol{y}) - T(\boldsymbol{z})\| \leq \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{z}\|$$

implying that $T(\boldsymbol{x})$ is a strict contraction with a unique fixed point.

20. Let $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{-1}$ be an $n \times n$ diagonalizable matrix, where $\boldsymbol{D}$ is diagonal and $\boldsymbol{U}$ is invertible. Here the ith column $\boldsymbol{u}_i$ of $\boldsymbol{U}$ is an eigenvector of $\boldsymbol{M}$ with eigenvalue d_i equal to the ith diagonal entry of $\boldsymbol{D}$. For a vector $\boldsymbol{x}$ with expansion $\sum_{i=1}^n a_i \boldsymbol{u}_i$, define the vector norm $\|\boldsymbol{x}\|_\dagger = \sum_{i=1}^n |a_i|$. Verify that this defines a norm with induced matrix norm $\|\boldsymbol{M}\|_\dagger = \max_{1 \le i \le n} |d_i|$. Show that the affine map $\boldsymbol{x} \mapsto \boldsymbol{M}\boldsymbol{x} + \boldsymbol{v}$ is a contraction under $\|\boldsymbol{x}\|_\dagger$ whenever the spectral radius of $\boldsymbol{M}$ is strictly less than 1. Note that the d_i and $\boldsymbol{u}_i$ may be complex. What is the fixed point of the map?

21. Suppose the maps $T_0, \ldots, T_{r-1}$ are paracontractive with fixed point sets $F_0, \ldots, F_{r-1}$. If $F = \cap_{i=0}^{r-1} F_i$ is nonempty, then show that the map $S = T_{r-1} \circ \cdots \circ T_0$ is paracontractive with fixed point set $F = \cap_{i=0}^{r-1} F_i$.

22. Suppose the continuous map T from a closed convex set D to itself has a k-fold composition $S = T \circ \cdots \circ T$ that is a strict contraction. Demonstrate that T and S share a unique fixed point, and that $\boldsymbol{x}_{m+1} = T(\boldsymbol{x}_m)$ converges to it.

23. Let $T(\boldsymbol{x})$ map the compact convex set C into itself. If there is a norm $\|\boldsymbol{x}\|_\dagger$ under which $\|T(\boldsymbol{y}) - T(\boldsymbol{x})\|_\dagger \le \|\boldsymbol{y} - \boldsymbol{x}\|_\dagger$ for all $\boldsymbol{y}$ and $\boldsymbol{x}$, then show that $T(\boldsymbol{x})$ has a fixed point. Use this result to prove that every finite state Markov chain possesses a stationary distribution. (Hints: Choose any $\boldsymbol{z} \in C$ and $\epsilon \in (0, 1)$ and define

$$T_\epsilon(\boldsymbol{x}) \quad = \quad (1 - \epsilon) T(\boldsymbol{x}) + \epsilon \boldsymbol{z}.$$

Argue that $T_\epsilon(\boldsymbol{x})$ is a strict contraction and send ϵ to 0.)

24. Under the hypotheses of Problem 23, demonstrate that the set of fixed points is nonempty, compact, and convex. (Hint: To prove convexity, suppose $\boldsymbol{x}$ and $\boldsymbol{y}$ are fixed points. For $\lambda \in [0, 1]$, argue that the point $\boldsymbol{z} = \lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y}$ satisfies

$$\begin{aligned} \|\boldsymbol{y} - \boldsymbol{x}\|_\dagger &\le \|T(\boldsymbol{y}) - T(\boldsymbol{z})\|_\dagger + \|T(\boldsymbol{z}) - T(\boldsymbol{x})\|_\dagger \\ &\le \|\boldsymbol{y} - \boldsymbol{x}\|_\dagger. \end{aligned}$$

Deduce from this result that $T(\boldsymbol{z}) = \boldsymbol{z}$.)

25. Consider the problem of minimizing the convex function $f(\boldsymbol{x})$ subject to the affine equality constraints $g_i(\boldsymbol{x}) = 0$ for $1 \le i \le p$ and the convex inequality constraints $h_j(\boldsymbol{x}) \le c_j$ for $1 \le j \le q$. Let $v(\boldsymbol{c})$ be the optimal value of $f(\boldsymbol{x})$ subject to the constraints. Show that the function $v(\boldsymbol{c})$ is convex in $\boldsymbol{c}$.

26. Describe the dual function for the power plan problem when one or more of the cost functions $f_i(x_i) = a_i x_i$ is linear instead of quadratic. Does this change affect the proposed solution by bisection or steepest ascent?

27. Calculate the dual function for the problem of minimizing $|x|$ subject to $x \leq -1$. Show that the optimal values of the primal and dual problems agree.

28. Verify that the two linear programs in Example 15.5.8 are dual programs.

29. Demonstrate that the dual function for Example 5.5.3 is

$$\mathcal{D}(\mu) = \begin{cases} 0 & \mu = 0 \\ 2\sum_{i=1}^n \sqrt{\mu a_i c_i} - \mu b & \mu > 0. \end{cases}$$

Check that this problem is a convex program and that Slater's condition is satisfied.

30. Derive the dual function

$$\mathcal{D}(\lambda, \boldsymbol{\mu}) = -\lambda - \boldsymbol{\mu}^* \boldsymbol{e} - e^{-\lambda - 1} \sum_{i=1}^n e^{-\boldsymbol{w}_i^* \boldsymbol{\mu}}$$

for the problem of minimizing the negative entropy $\sum_{i=1}^n x_i \ln x_i$ subject to the constraints $\sum_{i=1}^n x_i = 1$, $\boldsymbol{W}\boldsymbol{x} \leq \boldsymbol{e}$, and all $x_i \geq 0$. Here the vector $\boldsymbol{w}_i$ is the ith column of $\boldsymbol{W}$. (Hint: See Table 14.1.)

31. Consider the problem of minimizing the convex function

$$f(\boldsymbol{x}) = x_1 \ln x_1 - x_1 + x_2 \ln x_2 - x_2$$

subject to the constraints $x_1 + 2x_2 \leq 1$, $x_1 \geq 0$, and $x_2 \geq 0$. Show that the primal and dual optimal values coincide [17].

32. In the analytic centering program one minimizes the objective function $f(\boldsymbol{x}) = -\sum_{i=1}^n \ln x_i$ subject to the linear equality constraints $\boldsymbol{V}\boldsymbol{x} = \boldsymbol{d}$ and the positivity constraints $x_i > 0$ for all i. We can incorporate the positivity constraints into the objective function by defining $f(\boldsymbol{x}) = \infty$ if any $x_i \leq 0$. With this essential domain in mind, calculate the Fenchel conjugate

$$f^\star(\boldsymbol{y}) = \begin{cases} -n - \sum_{i=1}^n \ln(-y_i) & \text{all } y_i < 0 \\ \infty & \text{any } y_i \geq 0\,. \end{cases}$$

Show that the dual problem consists in maximizing

$$-\boldsymbol{\lambda}^* \boldsymbol{d} + n + \sum_{i=1}^n \ln(\boldsymbol{V}^* \boldsymbol{\lambda})_i$$

subject to the constraints $(\boldsymbol{V}^* \boldsymbol{\lambda})_i > 0$ for all i. The primal problem is easy to solve if we know the Lagrange multipliers. Indeed, straightforward differentiation shows that $x_i = 1/(\boldsymbol{V}^* \boldsymbol{\lambda})_i$. The moral here is

that if the number of rows of $\boldsymbol{V}$ is small, then it is advantageous to solve the dual problem for $\boldsymbol{\lambda}$ and insert this value into the explicit solution for $\boldsymbol{x}$.

33. In the trust region problem of Sect. 11.7, one minimizes a quadratic $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x} + c$ subject to the constraint $\|\boldsymbol{x}\|^2 \leq r^2$. Why does the constraint satisfy Slater's condition? If $\boldsymbol{A}$ is positive definite, then the problem is convex, and the primal and dual programs have the same optimal values. Calculate the dual function $\mathcal{D}(\mu)$ depending on the Lagrange multiplier μ. If $\boldsymbol{A}$ is not positive definite, then show that

$$\frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x} + c \;=\; -\alpha r^2 + \frac{1}{2}\boldsymbol{x}^*(\boldsymbol{A} + \alpha\boldsymbol{I})\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x} + c$$

subject to the equality constraint $\|\boldsymbol{x}\|^2 = r^2$. How does this permit one to handle indefinite programs? The article [245] treats this problem in detail.

34. In the experimental design literature, the problems of D-optimal and A-optimal designs are well studied [225]. These involve minimizing the objective functions

$$\begin{aligned} f(\boldsymbol{x}) &= \ln\det\Big(\sum_{i=1}^{p} x_i\boldsymbol{v}_i\boldsymbol{v}_i^*\Big)^{-1} \\ g(\boldsymbol{x}) &= \operatorname{tr}\Big(\sum_{i=1}^{p} x_i\boldsymbol{v}_i\boldsymbol{v}_i^*\Big)^{-1} \end{aligned}$$

subject to the explicit constraints $\sum_{i=1}^{p} x_i = 1$ and $x_i \geq 0$ and the implicit constraint that the matrix $\sum_{i=1}^{p} x_i\boldsymbol{v}_i\boldsymbol{v}_i^*$ is positive definite. The vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ in R^n are given in advance. To formulate the dual problems, set $\boldsymbol{W} = \sum_{i=1}^{p} x_i\boldsymbol{v}_i\boldsymbol{v}_i^*$ and require $\boldsymbol{W}$ to be positive definite. Derive the dual problems with the constraint $\boldsymbol{W} = \sum_{i=1}^{p} x_i\boldsymbol{v}_i\boldsymbol{v}_i^*$ included. Show that the dual problems involve maximizing $\ln\det\boldsymbol{X}$ and $\operatorname{tr}(\boldsymbol{X}^{1/2})^2$, respectively, subject to the constraints that $\boldsymbol{X}$ is positive definite and $\boldsymbol{v}_i^*\boldsymbol{X}\boldsymbol{v}_i \leq 1$ for all i. See the reference [262] for this formulation and techniques for maximizing the dual function. (Hints: The dual function for the objective function $f(\boldsymbol{x})$ is

$$\begin{aligned} \mathcal{D}(\boldsymbol{X}, \lambda) \;=\; & \inf_{\boldsymbol{W}, \boldsymbol{x} \geq \boldsymbol{0}} \Big\{ \ln\det\boldsymbol{W}^{-1} + \operatorname{tr}\Big[\boldsymbol{X}\Big(\boldsymbol{W} - \sum_{i=1}^{p} x_i\boldsymbol{v}_i\boldsymbol{v}_i^*\Big)\Big]\Big\} \\ & + \lambda\Big(\sum_{i=1}^{p} x_i - 1\Big). \end{aligned}$$

Eliminate λ by reparameterizing $\boldsymbol{X}$. For the objective function $g(\boldsymbol{x})$, take $p = 1$ in the next exercise.)

35. Let $\boldsymbol{X}$ be a positive definite matrix and p a positive scalar. Demonstrate that the function $f(\boldsymbol{W}) = \operatorname{tr}(\boldsymbol{W}^{-p}) + p \operatorname{tr}(\boldsymbol{X}\boldsymbol{W})$ achieves its minimum of $(p+1)\operatorname{tr}[\boldsymbol{X}^{p/(p+1)}]$ at the matrix $\boldsymbol{W} = \boldsymbol{X}^{-1/(p+1)}$. Here the argument $\boldsymbol{W}$ is also assumed positive definite. (Hint: Reduce the problem to one of calculating a Fenchel conjugate via Proposition 14.6.1.)

36. We have repeatedly visited the problem of projecting an exterior point onto a closed convex set C. Consider a non-Euclidean norm $\|\boldsymbol{x}\|_\dagger$ with dual norm $\|\boldsymbol{y}\|_\star$. Let $\boldsymbol{u}$ be a point exterior to C. Argue that the projection of $\boldsymbol{u}$ onto C under this alternative norm can be found by minimizing the criterion $\|\boldsymbol{z}\|_\dagger + \delta_C(\boldsymbol{x})$ subject to $\boldsymbol{u} - \boldsymbol{x} = \boldsymbol{z}$. Demonstrate that the dual problem can be phrased as minimizing

$$\mathcal{D}(\boldsymbol{\lambda}) \quad = \quad \boldsymbol{\lambda}^* \boldsymbol{u} - \sup_{\boldsymbol{x} \in C} \boldsymbol{\lambda}^* \boldsymbol{x} \quad \text{for} \quad \|\boldsymbol{\lambda}\|_\star \leq 1.$$

16
Convex Minimization Algorithms

16.1 Introduction

This chapter delves into three advanced algorithms for convex minimization. The projected gradient algorithm is useful in minimizing a strictly convex quadratic over a closed convex set. Although the algorithm extends to more general convex functions, the best theoretical results are available in this limited setting. We rely on the MM principle to motivate and extend the algorithm. The connections to Dykstra's algorithm and the contraction mapping principle add to the charm of the subject. On the minus side of the ledger, the projected gradient method can be very slow to converge. This defect is partially offset by ease of coding in many problems.

The second algorithm, path following in the exact penalty method, requires a fairly sophisticated understanding of convex calculus. As described in Chap. 13, classical penalty methods for solving constrained optimization problems exploit smooth penalties and send the tuning constant to infinity. If one substitutes absolute value and hinge penalties for square penalties, then there is no need to pass to the limit. Taking the penalty tuning constant sufficiently large generates a penalized problem with the same minimum as the constrained problem. In path following we track the minimum point of the penalized objective function as the tuning constant increases. Invocation of the implicit function theorem reduces path following to an exercise in numerically solving an ordinary differential equation [283, 284].

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_16,

Our third algorithm, Bregman iteration [120, 208, 279], has found the majority of its applications in image processing. In ℓ_1 penalized image restoration, it gives sparser, better fitting signals. In total variation penalized image reconstruction, it gives higher contrast images with decent smoothing. The basis pursuit problem [38, 74] of minimizing $\|\boldsymbol{u}\|_1$ subject to $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}$ readily succumbs to Bregman iteration. In many cases the basis pursuit solution is the sparest consistent with the constraint. One can solve the basis pursuit problem by linear programming, but conventional solvers are not tailored to dense matrices $\boldsymbol{A}$ and sparse solutions. Many applications require substitution of $\|\boldsymbol{D}\boldsymbol{u}\|_1$ for $\|\boldsymbol{u}\|_1$ for a smoothing matrix $\boldsymbol{D}$. This complication motivated the introduction of split Bregman iteration [106], which we briefly cover. Solution techniques continue to evolve rapidly in Bregman iteration. The whole field is driven by the realization that well-controlled sparsity gives better statistical inference and faster, more reliable algorithms than competing models and methods.

16.2 Projected Gradient Algorithm

For an $n \times n$ positive definite matrix $\boldsymbol{A}$ and a closed convex set $S \subset \mathsf{R}^n$, consider the problem of minimizing the quadratic function

$$f(\boldsymbol{x}) \;=\; \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x}$$

over S. We have studied this problem in depth in the special case $\boldsymbol{A} = \boldsymbol{I}$, where it reduces to projection onto S. Given that projection is relatively easy for a variety of convex sets, it is worth asking when the more general problem can be reduced to this special case. One way of answering the question is through the majorization

$$\begin{aligned}
\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} &= (\boldsymbol{x} - \boldsymbol{x}_m + \boldsymbol{x}_m)^*\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_m + \boldsymbol{x}_m) \\
&= (\boldsymbol{x} - \boldsymbol{x}_m)^*\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_m) + 2\boldsymbol{x}_m^*\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_m) + \boldsymbol{x}_m^*\boldsymbol{A}\boldsymbol{x}_m \\
&\leq \|\boldsymbol{A}\|_2\|\boldsymbol{x} - \boldsymbol{x}_m\|^2 + 2\boldsymbol{x}_m^*\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_m) + \boldsymbol{x}_m^*\boldsymbol{A}\boldsymbol{x}_m.
\end{aligned}$$

This majorization leads to the function

$$g(\boldsymbol{x} \mid \boldsymbol{x}_m) \;=\; \frac{\|\boldsymbol{A}\|_2}{2}\|\boldsymbol{x} - \boldsymbol{x}_m\|^2 + \boldsymbol{x}_m^*\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_m) + \boldsymbol{b}^*(\boldsymbol{x} - \boldsymbol{x}_m) + c$$

majorizing $f(\boldsymbol{x})$, where c is an irrelevant constant. Completing the square allows one to rewrite the surrogate function as

$$g(\boldsymbol{x} \mid \boldsymbol{x}_m) \;=\; \frac{\|\boldsymbol{A}\|_2}{2}\left\|\boldsymbol{x} - \boldsymbol{x}_m + \frac{1}{\|\boldsymbol{A}\|_2}[\boldsymbol{A}\boldsymbol{x}_m + \boldsymbol{b}]\right\|^2 + d$$

for another irrelevant constant d. The majorization of $f(\boldsymbol{x})$ persists if we replace the coefficient $\|\boldsymbol{A}\|_2$ appearing in $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ by a larger constant.

The surrogate function $g(\boldsymbol{x} \mid \boldsymbol{x}_m)$ is essentially a Euclidean distance, and minimizing it over S is accomplished by projection onto S. If $P_S(\boldsymbol{y})$ is the projection operator, then the algorithm map boils down to

$$M(\boldsymbol{x}) \;=\; P_S\left[\boldsymbol{x} - \frac{1}{\|\boldsymbol{A}\|_2}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})\right] \;=\; P_S\left[\boldsymbol{x} - \frac{1}{\|\boldsymbol{A}\|_2}\nabla f(\boldsymbol{x})\right].$$

According to the MM principle, this projected gradient satisfies the descent property for the objective function $f(\boldsymbol{x})$. More generally, the algorithm map

$$M_\rho(\boldsymbol{x}) \;=\; P_S\left[\boldsymbol{x} - \frac{\rho}{\|\boldsymbol{A}\|_2}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})\right] \;=\; P_S\left[\boldsymbol{x} - \frac{\rho}{\|\boldsymbol{A}\|_2}\nabla f(\boldsymbol{x})\right].$$

also possesses the descent property for all $\rho \in (0,1]$. Convergence of the projected gradient algorithm is guaranteed by the contraction mapping theorem stated in Proposition 15.3.2. Indeed, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ denote the eigenvalues of $\boldsymbol{A}$. Since $\|P_S(\boldsymbol{u}) - P_S(\boldsymbol{v})\| \leq \|\boldsymbol{u}-\boldsymbol{v}\|$ for all points $\boldsymbol{u}$ and $\boldsymbol{v}$ and the matrix $\boldsymbol{I}-\alpha\boldsymbol{A}$ has ith eigenvalue $1-\alpha\lambda_i$ for any constant α, we have

$$\begin{aligned}
\|M_\rho(\boldsymbol{x}) - M_\rho(\boldsymbol{y})\| &\leq \left\|\boldsymbol{x} - \frac{\rho}{\|\boldsymbol{A}\|_2}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}) - \boldsymbol{y} + \frac{\rho}{\|\boldsymbol{A}\|_2}(\boldsymbol{A}\boldsymbol{y}+\boldsymbol{b})\right\| \\
&= \left\|\left(\boldsymbol{I} - \frac{\rho}{\|\boldsymbol{A}\|_2}\boldsymbol{A}\right)(\boldsymbol{x}-\boldsymbol{y})\right\| \\
&\leq \max_i \left|1 - \frac{\rho\lambda_i}{\lambda_1}\right| \|\boldsymbol{x}-\boldsymbol{y}\|.
\end{aligned}$$

Provided $\lambda_n > 0$ and $\rho \in (0,2)$, it follows that the map $M_\rho(\boldsymbol{x})$ is a strict contraction on S. Except for a detail, this proves the second claim of the next proposition.

Proposition 16.2.1 *The projected gradient algorithm*

$$\boldsymbol{x}_{m+1} \;=\; P_S\left[\boldsymbol{x}_m - \frac{\rho}{\|\boldsymbol{A}\|_2}(\boldsymbol{A}\boldsymbol{x}_m+\boldsymbol{b})\right] \tag{16.1}$$

for the convex quadratic $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x}$ is a descent algorithm whenever $\rho \in (0,1]$. The algorithm converges to the unique minimum of $f(\boldsymbol{x})$ on the convex set S whenever $f(\boldsymbol{x})$ is strictly convex and $\rho \in (0,2)$.

Proof: Because the iteration map is a strict contraction, the iterates converge at a linear rate to its unique fixed point $\boldsymbol{y}$. It suffices to prove that $\boldsymbol{y}$ furnishes the minimum. In view of the obtuse angle criterion, if $\boldsymbol{z}$ is any point of S, we have

$$\left[\boldsymbol{y} - \frac{\rho}{\|\boldsymbol{A}\|_2}\nabla f(\boldsymbol{y}) - \boldsymbol{y}\right]^*\left[\boldsymbol{z}-\boldsymbol{y}\right] \;\leq\; 0.$$

However, this is equivalent to the condition $df(\boldsymbol{y})(\boldsymbol{z}-\boldsymbol{y}) \geq 0$, which is both necessary and sufficient for a minimum. ■

The special case of least squares estimation is important in Proposition 16.2.1. If

$$f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{Q}\boldsymbol{x}\|^2 = \frac{1}{2}\boldsymbol{x}^*\boldsymbol{Q}^*\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{y}^*\boldsymbol{Q}\boldsymbol{x} + \frac{1}{2}\|\boldsymbol{y}\|^2, \quad (16.2)$$

then in our previous notation $\boldsymbol{A} = \boldsymbol{Q}^*\boldsymbol{Q}$ and $\boldsymbol{b} = -\boldsymbol{Q}^*\boldsymbol{y}$. Furthermore, the matrix norm $\|\boldsymbol{A}\|_2 = \|\boldsymbol{Q}^*\boldsymbol{Q}\|_2 = \|\boldsymbol{Q}\|_2^2$. In practice finding the norm $\|\boldsymbol{A}\|_2$ is an issue. One can substitute the larger norm $\|\boldsymbol{A}\|_F$ for $\|\boldsymbol{A}\|_2$ as noted in Proposition 2.2.1. Alternatively, one can backtrack in the projected gradient algorithm (16.1). This involves making an initial guess a_0 of $\|\boldsymbol{A}\|_2$ and selecting a constant $c > 1$. If the point $\boldsymbol{x}_{m+1}$ does not decrease $f(\boldsymbol{x})$, then replace a_0 by $a_1 = ca_0$ and recompute $\boldsymbol{x}_{m+1}$. If this new $\boldsymbol{x}_{m+1}$ does not decrease $f(\boldsymbol{x})$, then replace a_1 by $a_2 = c^2a_0$, and so forth. For k large enough, a_k must exceed $\|\boldsymbol{A}\|_2$. Problem 15 of Chap. 11 explores a simple algorithm of Hestenes and Karush [126] that efficiently produces the largest eigenvalue $\|\boldsymbol{A}\|_2$ of $\boldsymbol{A}$ for n large.

Projected gradient algorithms are not limited to quadratic functions. Consider an arbitrary differentiable function $f(\boldsymbol{x})$ whose gradient satisfies the Lipschitz inequality

$$\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| \leq b\|\boldsymbol{y}-\boldsymbol{x}\|$$

for some b and all $\boldsymbol{x}$ and $\boldsymbol{y}$. The projected gradient algorithm

$$\boldsymbol{x}_{m+1} = P_S\left[\boldsymbol{x}_m - \frac{\rho}{b}\nabla f(\boldsymbol{x}_m)\right]$$

with $\rho \in (0,2)$ is designed to minimize $f(\boldsymbol{x})$ over a closed convex set S. Problems 1 and 2 sketch a few convergence results in this context [226]. See also Problem 31 of Chap. 4. For some functions the Lipschitz condition is only valid for a ball B centered at the current point $\boldsymbol{x}_m$. Then the algorithm that projects $\boldsymbol{x}_m - \frac{\rho}{b}\nabla f(\boldsymbol{x}_m)$ onto the intersection $B \cap S$ also retains the descent property. This amendment to the projected gradient method is inspired by the trust region strategy.

Example 16.2.1 *Projection onto the Image of a Convex Set*

Suppose projection onto the convex set S is easy. Given a compatible matrix $\boldsymbol{Q}$, the projected gradient algorithm allows us to project a point $\boldsymbol{y}$ onto the image set $\boldsymbol{Q}S$. One merely minimizes the criterion (16.2) over S. For example, let S be the closed convex set $\{\boldsymbol{x} : \boldsymbol{x}_i \geq 0 \ \forall\ i > 1\}$, and let $\boldsymbol{Q}$ be the lower-triangular matrix whose nonzero entries equal 1. The set $\boldsymbol{Q}S$ is the set $\{\boldsymbol{w} : w_1 \leq w_2 \leq \cdots \leq w_n\}$ whose entries are nondecreasing.

The two transformations $\boldsymbol{w} = \boldsymbol{Q}\boldsymbol{x}$ and $\boldsymbol{v} = \boldsymbol{Q}^*\boldsymbol{u}$ can be implemented via the fast recurrences

$$\begin{aligned} w_1 &= x_1, \quad w_{k+1} = w_k + x_{k+1} \\ v_n &= u_n, \quad v_k = v_{k+1} + u_k. \end{aligned}$$

The first of these recurrences operates in a forward direction; the second operates in a backward direction. Projection onto $\boldsymbol{Q}S$ is quick because the recurrences and projection onto S are both quick. In practice, the pool adjacent violators algorithm of Example 15.2.3 is faster overall. ■

Example 16.2.2 *Logistic Regression with a Group Lasso*

Data from the National Opinion Research Center offers the opportunity to perform logistic regression with grouped categorical predictors [57]. Here we consider the dichotomy happy (very happy plus pretty happy) versus unhappy surveyed on $n = 1,566$ people. In addition to the primary response, each participant registered a level in five predictor categories: gender, marital status, education, financial status, and health. Within a category (or group) the first level is taken as baseline and omitted in analysis. In logistic regression we seek to maximize the loglikelihood

$$\begin{aligned} \ln L(\boldsymbol{\theta}) &= \sum_{i=1}^{n} [y_i \ln p_i + (1 - y_i) \ln(1 - p_i)] \\ y_i &\in \{0, 1\}, \quad p_i = \frac{e^{\boldsymbol{x}_i^* \boldsymbol{\theta}}}{1 + e^{\boldsymbol{x}_i^* \boldsymbol{\theta}}}. \end{aligned}$$

Here $\boldsymbol{x}_i$ is the predictor vector for person i omitting the first level of each category; all entries of $\boldsymbol{x}_i$ equal 0 or 1. The vector $\boldsymbol{\theta}$ encodes the corresponding regression coefficients. The ball $S = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_{1,2} \leq r\}$ is defined via the $\ell_{1,2}$ norm mentioned in Problem 15 of Chap. 15. This problem also outlines an effective algorithm for projection onto an $\ell_{1,2}$ ball.

Constrained maximization performs continuous model selection. The $\ell_{1,2}$ constraint groups the various parameters by category. As explained in Example 8.7, one can majorize $-\ln L(\boldsymbol{\theta})$ by the convex quadratic

$$f(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m) = -\ln L(\boldsymbol{\theta}_m) - d \ln L(\boldsymbol{\theta}_m)(\boldsymbol{\theta} - \boldsymbol{\theta}_m) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_m)^* \boldsymbol{A}(\boldsymbol{\theta} - \boldsymbol{\theta}_m),$$

where $\boldsymbol{A} = \frac{1}{4}\boldsymbol{X}^*\boldsymbol{X}$ and $\boldsymbol{X}$ is the matrix whose ith row is $\boldsymbol{x}_i^*$. The projected gradient algorithm (16.1) with $\rho = 1$ therefore drives $f(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$ downhill and $\ln L(\boldsymbol{\theta})$ uphill. Alternation of majorization and projection is effective in producing constrained maximum likelihood estimates in the limit.

Table 16.1 lists the estimated regression coefficients within each category as a function of the radius r. The most interesting findings in the table are the low impact of education on happiness and the parameter reversals

TABLE 16.1. Logistic regression with a Group-Lasso constraint

Radius r	0.00	0.20	0.40	0.60	0.80	1.00
Iterations	16	10	12	13	13	13
Female	–	–	–	–	–	–
Male	0.00	0.00	0.00	0.00	−0.00	−0.04
Married	–	–	–	–	–	–
Never married	0.00	0.00	0.00	0.00	−0.03	−0.06
Divorced	0.00	0.00	0.00	0.00	−0.05	−0.12
Widowed	0.00	0.00	0.00	0.00	−0.05	−0.12
Separated	0.00	0.00	0.00	0.00	−0.05	−0.12
Some high school	–	–	–	–	–	
High school	0.00	0.00	0.00	0.00	0.00	0.00
Junior college	0.00	0.00	0.00	0.00	0.00	0.00
Bachelor	0.00	0.00	0.00	0.00	0.00	0.00
Graduate	0.00	0.00	0.00	0.00	0.00	0.00
Poor	–	–	–	–	–	–
Below average	0.00	−0.10	−0.16	−0.20	−0.21	−0.21
Average	0.00	0.08	0.13	0.18	0.20	0.21
Above average	0.00	0.07	0.13	0.20	0.23	0.25
Rich	0.00	0.01	0.01	0.02	0.03	0.03
Poor health	–	–	–	–	–	–
Fair health	0.00	−0.02	−0.06	−0.09	−0.09	−0.09
Good health	0.00	0.00	0.01	0.04	0.06	0.07
Excellent health	0.00	0.05	0.14	0.25	0.31	0.34

between poor and below-average financial status and between poor health and fair health. The number of iterations until convergence displayed in Table 16.1 suggest good numerical performance on this typical problem. ■

One can generalize the projected gradient algorithm in various ways. For instance, in the projected Newton method, one projects the partial Newton step $\boldsymbol{x}_m - \tau d^2 f(\boldsymbol{x}_m)^{-1} \nabla f(\boldsymbol{x}_m)$ onto the constraint set S [12]. Here the step length τ is usually taken to be 1, and the second differential $d^2 f(\boldsymbol{x}_m)$ is assumed positive definite. The projected Newton strategy tends to reduce the number of iterations while increasing the complexity per iteration. To minimize generic convex functions, one can substitute subgradients for gradients. Thus, one projects $\boldsymbol{x}_m - \tau \boldsymbol{g}_m$ onto S for $\boldsymbol{g}_m \in \partial f(\boldsymbol{x}_m)$ and some optimal choice of the constant τ. Unfortunately, there exist subgradients whose negatives are not descent directions. For instance, $-g$ is an ascent direction of $f(x) = |x|$ for any nontrivial $g \in \partial f(0) = [-1, 1]$. The next proposition partially salvages the situation.

Proposition 16.2.2 *Suppose* $\boldsymbol{y}$ *minimizes the convex function* $f(\boldsymbol{x})$ *over the closed convex set* S*. If* $\boldsymbol{g}_m \in \partial f(\boldsymbol{x}_m)$*,*

$$\boldsymbol{x}_{m+1} = P_S\left[\boldsymbol{x}_m - \tau \boldsymbol{g}_m\right],$$

and $\boldsymbol{x}_m$ *is not optimal, then the choice*

$$0 < \tau < \frac{2[f(\boldsymbol{x}_m) - f(\boldsymbol{y})]}{\|\boldsymbol{g}_m\|^2}$$

produces $\|\boldsymbol{x}_{m+1} - \boldsymbol{y}\| < \|\boldsymbol{x}_m - \boldsymbol{y}\|$.

Proof: Because projection is nonexpansive,

$$\begin{aligned} \|\boldsymbol{x}_{m+1} - \boldsymbol{y}\|^2 &= \|P_S(\boldsymbol{x}_m - \tau \boldsymbol{g}_m) - P_S(\boldsymbol{y})\|^2 \\ &\leq \|\boldsymbol{x}_m - \tau \boldsymbol{g}_m - \boldsymbol{y}\|^2 \\ &= \|\boldsymbol{x}_m - \boldsymbol{y}\|^2 - 2\tau \boldsymbol{g}_m^*(\boldsymbol{x}_m - \boldsymbol{y}) + \tau^2 \|\boldsymbol{g}_m\|^2. \end{aligned}$$

Hence, the inequality $f(\boldsymbol{y}) \geq f(\boldsymbol{x}_m) + \boldsymbol{g}_m^*(\boldsymbol{y} - \boldsymbol{g}_m)$ implies

$$\begin{aligned} \|\boldsymbol{x}_{m+1} - \boldsymbol{y}\|^2 &\leq \|\boldsymbol{x}_m - \boldsymbol{y}\|^2 + 2\tau[f(\boldsymbol{y}) - f(\boldsymbol{x}_m)] + \tau^2 \|\boldsymbol{g}_m\|^2 \\ &= \|\boldsymbol{x}_m - \boldsymbol{y}\|^2 + h(\tau) \end{aligned}$$

for the obvious quadratic $h(\tau)$, which satisfies $h(0) = 0$ and attains its minimum value

$$\min_{\tau} h(\tau) = -\frac{[f(\boldsymbol{x}_m) - f(\boldsymbol{y})]^2}{\|\boldsymbol{g}_m\|^2}$$

at the positive point

$$\hat{\tau} = \frac{f(\boldsymbol{x}_m) - f(\boldsymbol{y})}{\|\boldsymbol{g}_m\|^2}.$$

The claim now follows from the symmetry of $h(\tau)$ around the point $\hat{\tau}$. ■

In practice the value $f(\boldsymbol{y})$ is not known beforehand. This necessitates some strategy for choosing the step-length constants τ_m. For the sake of brevity, we refer the reader to the book [226] for further discussion.

16.3 Exact Penalties and Lagrangians

In nonlinear programming, exact penalty methods minimize the function

$$\mathcal{E}_\rho(\boldsymbol{y}) = f(\boldsymbol{y}) + \rho \sum_{i=1}^{p} |g_i(\boldsymbol{y})| + \rho \sum_{j=1}^{q} \max\{0, h_j(\boldsymbol{y})\},$$

where $f(\boldsymbol{y})$ is the objective function, $g_i(\boldsymbol{y})$ is an equality constraint, and $h_j(\boldsymbol{y})$ is an inequality constraint. It is interesting to compare this function to the Lagrangian function

$$\mathcal{L}(\boldsymbol{y}) \;=\; f(\boldsymbol{y}) + \sum_{i=1}^{p} \lambda_i g_i(\boldsymbol{y}) + \sum_{j=1}^{q} \mu_j h_j(\boldsymbol{y})$$

capturing the behavior of $f(\boldsymbol{y})$ near a constrained local minimum $\boldsymbol{x}$. Proposition 5.2.1 demonstrates that the Lagrangian satisfies the stationarity condition $\nabla \mathcal{L}(\boldsymbol{x}) = \boldsymbol{0}$; its inequality multipliers are nonnegative and obey the complementary slackness requirements $\mu_j h_j(\boldsymbol{x}) = 0$. In an exact penalty method we take

$$\rho \;>\; \max\{|\lambda_1|, \ldots, |\lambda_p|, \mu_1, \ldots, \mu_q\}. \tag{16.3}$$

This choice creates the favorable circumstances

$$\begin{aligned}
\mathcal{L}(\boldsymbol{y}) &< \mathcal{E}_\rho(\boldsymbol{y}) \quad \text{for all infeasible } \boldsymbol{y} \\
\mathcal{L}(\boldsymbol{z}) &\leq f(\boldsymbol{z}) = \mathcal{E}_\rho(\boldsymbol{z}) \quad \text{for all feasible } \boldsymbol{z} \\
\mathcal{L}(\boldsymbol{x}) &= f(\boldsymbol{x}) = \mathcal{E}_\rho(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \text{ optimal}
\end{aligned}$$

with profound consequences. As the next proposition proves, minimizing $\mathcal{E}_\rho(\boldsymbol{y})$ is effective in minimizing $f(\boldsymbol{y})$ subject to the constraints.

In most problems the Lagrange multipliers are unknown, so that a certain amount of trial and error is necessary to ensure that ρ is large enough. Alternatively, one can simply follow the exact solution path until it merges with the constrained solution. Before we study path following in detail, it is helpful to prove necessary and sufficient conditions for the two solutions to coincide. We start with the sufficient conditions expressed in Proposition 5.5.2.

Proposition 16.3.1 *Under the assumptions of Proposition 5.5.2, a constrained local minimum $\boldsymbol{x}$ of $f(\boldsymbol{y})$ is an unconstrained local minimum of $\mathcal{E}_\rho(\boldsymbol{y})$ given inequality (16.3).*

Proof: Suppose the contrary is true, and let $\boldsymbol{x}_m$ be a sequence of points that converge to $\boldsymbol{x}$ and satisfy $\mathcal{E}_\rho(\boldsymbol{x}_m) < \mathcal{E}_\rho(\boldsymbol{x})$. Without loss of generality, assume that the unit vectors

$$\boldsymbol{v}_m \;=\; \frac{1}{\|\boldsymbol{x}_m - \boldsymbol{x}\|}(\boldsymbol{x}_m - \boldsymbol{x})$$

converge to a unit vector $\boldsymbol{v}$. Now consider the difference quotients

$$\frac{\mathcal{L}(\boldsymbol{x}_m) - \mathcal{L}(\boldsymbol{x})}{\|\boldsymbol{x}_m - \boldsymbol{x}\|} \;\leq\; \frac{\mathcal{E}_\rho(\boldsymbol{x}_m) - \mathcal{E}_\rho(\boldsymbol{x})}{\|\boldsymbol{x}_m - \boldsymbol{x}\|} \;<\; 0. \tag{16.4}$$

A brief calculation shows that

$$\begin{aligned}\lim_{m\to\infty}\frac{\mathcal{L}(\boldsymbol{x}_m)-\mathcal{L}(\boldsymbol{x})}{\|\boldsymbol{x}_m-\boldsymbol{x}\|} &= \lim_{m\to\infty} s_{\mathcal{L}}(\boldsymbol{x}_m,\boldsymbol{x})\boldsymbol{v}_m\\ &= d\mathcal{L}(\boldsymbol{x})\boldsymbol{v}\\ &= df(\boldsymbol{x})\boldsymbol{v}+\sum_{i=1}^{p}\lambda_i dg_i(\boldsymbol{x})\boldsymbol{v}+\sum_{j=1}^{q}\mu_j dh_j(\boldsymbol{x})\boldsymbol{v},\end{aligned}$$

where $s_{\mathcal{L}}(\boldsymbol{y},\boldsymbol{x})$ is the slope function of $\mathcal{L}(\boldsymbol{y})$ around $\boldsymbol{x}$. (See Sect. 4.4 for a discussion of slope functions.) The stationarity condition $\nabla\mathcal{L}(\boldsymbol{x})=\boldsymbol{0}$ implies $d\mathcal{L}(\boldsymbol{x})\boldsymbol{v}=0$.

The limit of the difference quotient for $\mathcal{E}_\rho(\boldsymbol{y})$ is more subtle to calculate. Under the subscript convention for slope functions, the equality $g_i(\boldsymbol{x})=0$ implies

$$\lim_{m\to\infty}\frac{|g_i(\boldsymbol{x}_m)|-|g_i(\boldsymbol{x})|}{\|\boldsymbol{x}_m-\boldsymbol{x}\|} = \lim_{m\to\infty}|s_{g_i}(\boldsymbol{x}_m,\boldsymbol{x})\boldsymbol{v}_m| = |dg_i(\boldsymbol{x})\boldsymbol{v}|.$$

The equality $h_j(\boldsymbol{x})=0$ for an active inequality constraint entails

$$\lim_{m\to\infty}\frac{\max\{0,h_j(\boldsymbol{x}_m)\}-\max\{0,h_j(\boldsymbol{x})\}}{\|\boldsymbol{x}_m-\boldsymbol{x}\|} = \max\{0,dh_j(\boldsymbol{x})\boldsymbol{v}\}.$$

The inequalities $h_j(\boldsymbol{x})<0$ and $h_j(\boldsymbol{x}_m)<0$ for an inactive inequality constraint likewise entail

$$\lim_{m\to\infty}\frac{\max\{0,h_j(\boldsymbol{x}_m)\}-\max\{0,h_j(\boldsymbol{x})\}}{\|\boldsymbol{x}_m-\boldsymbol{x}\|} = 0.$$

Hence, if the first r inequality constraints are active, then inequality (16.4) yields

$$\begin{aligned}0 &\geq \lim_{m\to\infty}\frac{\mathcal{E}_\rho(\boldsymbol{x}_m)-\mathcal{E}_\rho(\boldsymbol{x})}{\|\boldsymbol{x}_m-\boldsymbol{x}\|}\\ &= df(\boldsymbol{x})\boldsymbol{v}+\rho\sum_{i=1}^{p}|dg_i(\boldsymbol{x})\boldsymbol{v}|+\rho\sum_{j=1}^{r}\max\{0,dh_j(\boldsymbol{x})\boldsymbol{v}\}.\end{aligned}$$

Subtracting $d\mathcal{L}(\boldsymbol{x})\boldsymbol{v}=0$ from the last inequality gives

$$\begin{aligned}0 &\geq \sum_{i=1}^{p}[\rho|dg_i(\boldsymbol{x})\boldsymbol{v}|-\lambda_i dg_i(\boldsymbol{x})\boldsymbol{v}]\\ &\quad+\sum_{j=1}^{r}[\rho\max\{0,dh_j(\boldsymbol{x})\boldsymbol{v}\}-\mu_j dh_j(\boldsymbol{x})\boldsymbol{v}]\\ &\geq \sum_{i=1}^{p}[\rho|dg_i(\boldsymbol{x})\boldsymbol{v}|-\lambda_i dg_i(\boldsymbol{x})\boldsymbol{v}]\\ &\quad+\sum_{j=1}^{r}(\rho-\mu_j)\max\{0,dh_j(\boldsymbol{x})\boldsymbol{v}\}.\end{aligned}$$

Because all of the terms in the last two sums are nonnegative, they in fact vanish. But these are precisely the tangency conditions $dg_i(\boldsymbol{x})\boldsymbol{v} = 0$ and $dh_j(\boldsymbol{x})\boldsymbol{v} \leq 0$ for $h_j(\boldsymbol{x})$ active.

To finish the proof, we now pass to the limit in the second-order Taylor expansion

$$\frac{1}{2}\boldsymbol{v}_m^* s_{\mathcal{L}}^2(\boldsymbol{x}_m, \boldsymbol{x})\boldsymbol{v}_m \quad = \quad \frac{1}{\|\boldsymbol{x}_m - \boldsymbol{x}\|^2}\left[\mathcal{L}(\boldsymbol{x}_m) - \mathcal{L}(\boldsymbol{x})\right] \quad < \quad 0$$

and conclude that $\boldsymbol{v}^* d^2\mathcal{L}(\boldsymbol{x})\boldsymbol{v} \leq 0$, contrary to the assumption of Proposition 5.5.2 that no such unit tangent vector exists. Thus, the supposition that $\boldsymbol{x}$ is not a local minimum of $\mathcal{E}_\rho(\boldsymbol{y})$ is untenable. ■

Further theoretical progress can be made can be made by assuming that the equality constraints are affine and that the objective and inequality constraint functions are convex in addition to being differentiable. In these circumstances $\mathcal{E}_\rho(\boldsymbol{x})$ is convex owing to the closure properties of convex functions described in Proposition 6.3.3. Furthermore, at a feasible point $\boldsymbol{x}$ with the first r inequality constraints active, the sum and chain rules yield

$$\partial\mathcal{E}_\rho(\boldsymbol{x}) \quad = \quad \nabla f(\boldsymbol{x}) + \rho\sum_{i=1}^{p}[-1,1]\nabla g_i(\boldsymbol{x}) + \rho\sum_{j=1}^{r}[0,1]\nabla h_j(\boldsymbol{x}).$$

Hence, if

$$\mathbf{0} \quad = \quad \nabla f(\boldsymbol{x}) + \sum_{i=1}^{p}\lambda_i\nabla g_i(\boldsymbol{x}) + \sum_{j=1}^{r}\mu_j\nabla h_j(\boldsymbol{x}), \tag{16.5}$$

then certainly $\mathbf{0} \in \partial\mathcal{E}_\rho(\boldsymbol{x})$. Thus, a constrained minimum point of $f(\boldsymbol{y})$ satisfying the multiplier rule (16.5) corresponds to an unconstrained minimum point of $\mathcal{E}_\rho(\boldsymbol{y})$. This is just the content of Proposition 16.3.1 specialized to convex programming.

Conversely, assume that $\boldsymbol{z}$ is an unconstrained minimum point of $\mathcal{E}_\rho(\boldsymbol{y})$ and that $\boldsymbol{x}$ is a constrained minimum point of $f(\boldsymbol{y})$ satisfying the multiplier rule (16.5). If $\boldsymbol{z}$ is feasible, then $\boldsymbol{z}$ also furnishes a constrained minimum of $f(\boldsymbol{y})$ because $f(\boldsymbol{y})$ and $\mathcal{E}_\rho(\boldsymbol{y})$ coincide on the feasible region. If $\boldsymbol{z}$ is infeasible, then

$$\mathcal{E}_\rho(\boldsymbol{x}) \quad = \quad f(\boldsymbol{x}) \quad = \quad \mathcal{L}(\boldsymbol{x}) \quad \leq \quad \mathcal{L}(\boldsymbol{z}) \quad < \quad \mathcal{E}_\rho(\boldsymbol{z}).$$

Here the inequality $\mathcal{L}(\boldsymbol{x}) \leq \mathcal{L}(\boldsymbol{z})$ reflects the fact that the convex function $\mathcal{L}(\boldsymbol{y})$ attains its minimum at the stationary point $\boldsymbol{x}$. The contradiction $\mathcal{E}_\rho(\boldsymbol{x}) < \mathcal{E}_\rho(\boldsymbol{z})$ now shows that $\boldsymbol{z}$ is feasible and consequently furnishes a constrained minimum of $f(\boldsymbol{y})$. For the sake of completeness, we restate this result as a formal proposition.

Proposition 16.3.2 *Suppose in a convex program with differentiable objective and constraint functions that there exists a constrained minimum $\boldsymbol{x}$ of $f(\boldsymbol{y})$ satisfying the multiplier rule. Under condition (16.3), a point $\boldsymbol{z}$ furnishes a constrained minimum of $f(\boldsymbol{y})$ if and only if it furnishes an unconstrained minimum of $\mathcal{E}_\rho(\boldsymbol{y})$.*

Proof: See the forgoing discussion. ■

In path following, one tracks the postulated minimum point $\boldsymbol{x}(\rho)$ of $\mathcal{E}_\rho(\boldsymbol{y})$ as a function of ρ until ρ exceeds the Lagrange multiplier threshold. Thus, specifying a stationarity condition for $\mathcal{E}_\rho(\boldsymbol{y})$ is crucial. Unfortunately, our previous derivations of stationarity conditions assumed either differentiability or convexity. In general nonlinear programs, $\mathcal{E}_\rho(\boldsymbol{y})$ is neither. Here we tackle the stationarity condition via forward directional derivatives. Once again assume that $\boldsymbol{x}$ is a local minimum of $\mathcal{E}_\rho(\boldsymbol{y})$. If the objective and constraint functions are differentiable at $\boldsymbol{x}$, then they possess directional derivatives at $\boldsymbol{x}$, and

$$d_{\boldsymbol{v}}\mathcal{E}_\rho(\boldsymbol{x}) = df(\boldsymbol{x})\boldsymbol{v} + \rho\sum_{i=1}^{p} d_{\boldsymbol{v}}|g_i(\boldsymbol{x})| + \rho\sum_{j=1}^{q} d_{\boldsymbol{v}}\max\{0, h_j(\boldsymbol{x})\}.$$

In proving Proposition 16.3.1, we calculated the obscure pieces making up $d_{\boldsymbol{v}}\mathcal{E}_\rho(\boldsymbol{y})$. Based on the notation

$$\begin{aligned} \mathcal{N}_{\mathrm{E}} &= \{i : g_i(\boldsymbol{x}) < 0\} & \mathcal{N}_{\mathrm{I}} &= \{j : h_j(\boldsymbol{x}) < 0\} \\ \mathcal{Z}_{\mathrm{E}} &= \{i : g_i(\boldsymbol{x}) = 0\} & \mathcal{Z}_{\mathrm{I}} &= \{j : h_j(\boldsymbol{x}) = 0\} \\ \mathcal{P}_{\mathrm{E}} &= \{i : g_i(\boldsymbol{x}) > 0\} & \mathcal{P}_{\mathrm{I}} &= \{j : h_j(\boldsymbol{x}) > 0\} \end{aligned} \tag{16.6}$$

we have

$$\begin{aligned} d_{\boldsymbol{v}}\mathcal{E}_\rho(\boldsymbol{x}) &= df(\boldsymbol{x})\boldsymbol{v} - \rho\sum_{i\in\mathcal{N}_{\mathrm{E}}} dg_i(\boldsymbol{x})\boldsymbol{v} + \rho\sum_{i\in\mathcal{P}_{\mathrm{E}}} dg_i(\boldsymbol{x})\boldsymbol{v} + \rho\sum_{j\in\mathcal{P}_{\mathrm{I}}} dh_j(\boldsymbol{x})\boldsymbol{v} \\ &\quad +\rho\sum_{i\in\mathcal{Z}_{\mathrm{E}}} |dg_i(\boldsymbol{x})\boldsymbol{v}| + \rho\sum_{j\in\mathcal{Z}_{\mathrm{I}}} \max\{0, dh_j(\boldsymbol{x})\boldsymbol{v}\} \\ &= \boldsymbol{w}^*\boldsymbol{v} + \rho\sum_{i\in\mathcal{Z}_{\mathrm{E}}} |dg_i(\boldsymbol{x})\boldsymbol{v}| + \rho\sum_{j\in\mathcal{Z}_{\mathrm{I}}} \max\{0, dh_j(\boldsymbol{x})\boldsymbol{v}\} \end{aligned}$$

for the obvious choice of $\boldsymbol{w}$. At a local minimum $\boldsymbol{x}$ of $\mathcal{E}_\rho(\boldsymbol{y})$, all directional derivatives satisfy $d_{\boldsymbol{v}}\mathcal{E}_\rho(\boldsymbol{x}) \geq 0$.

We now focus on the function $\mathcal{K}(\boldsymbol{v}) = d_{\boldsymbol{v}}\mathcal{E}_\rho(\boldsymbol{x})$ and derive an appropriate stationarity condition. Since the composition of a convex function with a linear function is convex, $\mathcal{K}(\boldsymbol{v})$ is convex even when $\mathcal{E}_\rho(\boldsymbol{x})$ is not. Hence, in dealing with $\mathcal{K}(\boldsymbol{v})$, we can invoke the rules of the convex calculus developed in Sects. 14.4 and 14.5. Because $\mathcal{K}(\boldsymbol{v})$ achieves its minimum value of 0 at the origin $\mathbf{0}$, Proposition 14.4.3 implies the containment $\mathbf{0} \in \partial\mathcal{K}(\mathbf{0})$. Applying

the rules of convex differentiation to $\mathcal{K}(\boldsymbol{v})$ gives the subdifferential

$$\partial\mathcal{K}(\mathbf{0}) \;=\; \boldsymbol{w} + \rho \sum_{i\in\mathcal{Z}_{\mathrm{E}}} [-1,1]\nabla g_i(\boldsymbol{x}) + \rho \sum_{j\in\mathcal{Z}_{\mathrm{I}}} [0,1]\nabla h_j(\boldsymbol{x}).$$

The stationarity condition

$$\begin{aligned}\mathbf{0} \;\in\; & \nabla f(\boldsymbol{x}) - \rho \sum_{i\in\mathcal{N}_{\mathrm{E}}} \nabla g_i(\boldsymbol{x}) + \rho \sum_{i\in\mathcal{P}_{\mathrm{E}}} \nabla g_i(\boldsymbol{x}) + \rho \sum_{j\in\mathcal{P}_{\mathrm{I}}} \nabla h_j(\boldsymbol{x})\boldsymbol{v} \\ & +\rho \sum_{i\in\mathcal{Z}_{\mathrm{E}}} [-1,1]\nabla g_i(\boldsymbol{x}) + \rho \sum_{j\in\mathcal{Z}_{\mathrm{I}}} [0,1]\nabla dh_j(\boldsymbol{x}) \qquad (16.7)\end{aligned}$$

for $\mathcal{E}_\rho(\boldsymbol{x})$ is a consequence of these considerations.

16.4 Mechanics of Path Following

Throughout this section we restrict our attention to convex programs. As a prelude to our derivation of the path following algorithm, we record several properties of $\mathcal{E}_\rho(\boldsymbol{x})$ that mitigate the failure of differentiability.

Proposition 16.4.1 *The surrogate function $\mathcal{E}_\rho(\boldsymbol{x})$ is increasing in ρ. Furthermore, $\mathcal{E}_\rho(\boldsymbol{x})$ is strictly convex for one $\rho > 0$ if and only if it is strictly convex for all $\rho > 0$. Likewise, when $f(\boldsymbol{x})$ is bounded below, $\mathcal{E}_\rho(\boldsymbol{x})$ is coercive for one $\rho > 0$ if and only if is coercive for all $\rho > 0$. Finally, if $f(\boldsymbol{x})$ is strictly convex (or coercive), then all $\mathcal{E}_\rho(\boldsymbol{x})$ are strictly convex (or coercive).*

Proof: The first assertion is obvious. For the second assertion, consider more generally a finite family $u_1(\boldsymbol{x}), \ldots, u_q(\boldsymbol{x})$ of convex functions, and suppose a linear combination $\sum_{k=1}^q c_k u_k(\boldsymbol{x})$ with positive coefficients is strictly convex. It suffices to prove that any other linear combination $\sum_{k=1}^q b_k u_k(\boldsymbol{x})$ with positive coefficients is strictly convex. For any two points $\boldsymbol{x} \neq \boldsymbol{y}$ and any scalar $\alpha \in (0,1)$, we have

$$u_k[\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}] \leq \alpha u_k(\boldsymbol{x}) + (1-\alpha)u_k(\boldsymbol{y}). \qquad (16.8)$$

Since $\sum_{k=1}^q c_k u_k(\boldsymbol{x})$ is strictly convex, strict inequality must hold for at least one k. Hence, multiplying inequality (16.8) by b_k and adding gives

$$\sum_{k=1}^q b_k u_k[\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}] \;<\; \alpha \sum_{k=1}^q b_k u_k(\boldsymbol{x}) + (1-\alpha) \sum_{k=1}^q b_k u_k(\boldsymbol{y}).$$

The third assertion follows from the criterion given in Proposition 12.3.1. Indeed, suppose $\mathcal{E}_\rho(\boldsymbol{x})$ is coercive, but $\mathcal{E}_{\rho^*}(\boldsymbol{x})$ is not coercive. Then there exists a point $\boldsymbol{x}$, a direction $\boldsymbol{v}$, and a sequence of scalars t_n tending to

∞ such that $\mathcal{E}_{\rho^*}(\boldsymbol{x} + t_n\boldsymbol{v})$ is bounded above. This requires the sequence $f(\boldsymbol{x}+t_n\boldsymbol{v})$ and each of the sequences $|g_i(\boldsymbol{x}+t_n\boldsymbol{v})|$ and $\max\{0, h_j(\boldsymbol{x}+t_n\boldsymbol{v})\}$ to remain bounded above. But in this circumstance the sequence $\mathcal{E}_\rho(\boldsymbol{x}+t_n\boldsymbol{v})$ also remains bounded above. The final assertion is also obvious. ■

To speak coherently of solution paths, one must validate the existence, uniqueness, and continuity of the solution $\boldsymbol{x}(\rho)$ to the stationarity condition (16.7). Uniqueness follows from assuming that $f(\boldsymbol{x})$ is strictly convex or more generally by assuming that $\mathcal{E}_\rho(\boldsymbol{x})$ is strictly convex for a single positive ρ. Existence and continuity are more subtle. Let us restate the stationarity condition as

$$\mathbf{0} \;=\; \nabla f(\boldsymbol{x}) + \rho \sum_{i=1}^{r} s_i \nabla g_i(\boldsymbol{x}) + \rho \sum_{j=1}^{s} t_j \nabla h_j(\boldsymbol{x}) \tag{16.9}$$

for coefficient sets $\{s_i\}_{i=1}^r$ and $\{t_j\}_{j=1}^s$ that satisfy

$$s_i \;\in\; \begin{cases} \{-1\} & g_i(\boldsymbol{x}) < 0 \\ [-1,1] & g_i(\boldsymbol{x}) = 0 \\ \{1\} & g_i(\boldsymbol{x}) > 0 \end{cases} \quad \text{and} \quad t_j \;\in\; \begin{cases} \{0\} & h_j(\boldsymbol{x}) < 0 \\ [0,1] & h_j(\boldsymbol{x}) = 0 \\ \{1\} & h_j(\boldsymbol{x}) > 0. \end{cases} \tag{16.10}$$

This notation puts us into position to state and prove some basic facts.

Proposition 16.4.2 *If $\mathcal{E}_\rho(\boldsymbol{y})$ is strictly convex and coercive, then the solution path $\boldsymbol{x}(\rho)$ of equation (16.7) exists and is continuous in ρ. If the gradient vectors $\{\nabla g_i(\boldsymbol{x}) : g_i(\boldsymbol{x}) = 0\} \cup \{\nabla h_j(\boldsymbol{x}) : h_j(\boldsymbol{x}) = 0\}$ of the active constraints are linearly independent at $\boldsymbol{x}(\rho)$ for $\rho > 0$, then in addition the coefficients $s_i(\rho)$ and $t_j(\rho)$ are unique and continuous near ρ.*

Proof: In accord with Proposition 16.4.1, we assume that either $f(\boldsymbol{x})$ is strictly convex and coercive or restrict our attention to the open interval $(0, \infty)$. Consider a subinterval $[a, b]$ containing ρ and fix a point $\boldsymbol{x}$ in the common domain of the functions $\mathcal{E}_\rho(\boldsymbol{y})$. The coercivity of $\mathcal{E}_a(\boldsymbol{y})$ and the inequalities

$$\mathcal{E}_a[\boldsymbol{x}(\rho)] \;\leq\; \mathcal{E}_\rho[\boldsymbol{x}(\rho)] \;\leq\; \mathcal{E}_\rho(\boldsymbol{x}) \;\leq\; \mathcal{E}_b(\boldsymbol{x})$$

demonstrate that the solution vector $\boldsymbol{x}(\rho)$ is bounded over $[a, b]$. To prove continuity, suppose that it fails for a given $\rho \in [a, b]$. Then there exists an $\epsilon > 0$ and a sequence ρ_n tending to ρ such $\|\boldsymbol{x}(\rho_n) - \boldsymbol{x}(\rho)\| \geq \epsilon$ for all n. Since $\boldsymbol{x}(\rho_n)$ is bounded, we can pass to a subsequence if necessary and assume that $\boldsymbol{x}(\rho_n)$ converges to some point $\boldsymbol{y}$. Taking limits in the inequality $\mathcal{E}_{\rho_n}[\boldsymbol{x}(\rho_n)] \leq \mathcal{E}_{\rho_n}(\boldsymbol{x})$ shows that $\mathcal{E}_\rho(\boldsymbol{y}) \leq \mathcal{E}_\rho(\boldsymbol{x})$ for all $\boldsymbol{x}$. Because $\boldsymbol{x}(\rho)$ is unique, we reach the contradictory conclusions $\|\boldsymbol{y} - \boldsymbol{x}(\rho)\| \geq \epsilon$ and $\boldsymbol{y} = \boldsymbol{x}(\rho)$.

Verification of the second claim is deferred to permit further discussion of path following. The claim says that an active constraint ($g_i(\boldsymbol{x}) = 0$ or $h_j(\boldsymbol{x}) = 0$) remains active until its coefficient hits an endpoint of its

subdifferential. Because the solution path is, in fact, piecewise smooth, one can follow the coefficient path by numerically solving an ordinary differential equation (ODE). ■

Along the solution path we keep track of the index sets defined in equation (16.6) and determined by the signs of the constraint functions. For the sake of simplicity, assume that at the beginning of the current segment s_i does not equal -1 or 1 when $i \in \mathcal{Z}_{\mathrm{E}}$ and t_j does not equal 0 or 1 when $j \in \mathcal{Z}_{\mathrm{I}}$. In other words, the coefficients of the active constraints occur on the interiors, either $(-1, 1)$ or $(0, 1)$, of their subdifferentials. Let us show in this circumstance that the solution path can be extended in a smooth fashion. Our plan of attack is to reparameterize by the Lagrange multipliers of the active constraints. Thus, set $\lambda_i = \rho s_i$ for $i \in \mathcal{Z}_{\mathrm{E}}$ and $\omega_j = \rho t_j$ for $j \in \mathcal{Z}_{\mathrm{I}}$. These multipliers satisfy $-\rho < \lambda_i < \rho$ and $0 < \omega_j < \rho$. The stationarity condition now reads

$$\begin{aligned} \mathbf{0} &= \nabla f(\boldsymbol{x}) - \rho \sum_{i \in \mathcal{N}_{\mathrm{E}}} \nabla g_i(\boldsymbol{x}) + \rho \sum_{i \in \mathcal{P}_{\mathrm{E}}} \nabla g_i(\boldsymbol{x}) + \rho \sum_{j \in \mathcal{P}_{\mathrm{I}}} \nabla h_j(\boldsymbol{x}) \\ &\quad + \sum_{i \in \mathcal{Z}_{\mathrm{E}}} \lambda_i \nabla g_i(\boldsymbol{x}) + \sum_{j \in \mathcal{Z}_{\mathrm{I}}} \omega_j \nabla h_j(\boldsymbol{x}). \end{aligned}$$

For convenience now define

$$\begin{aligned} \boldsymbol{U}_{\mathcal{Z}}(\boldsymbol{x}) &= \begin{bmatrix} dg_{\mathcal{Z}_{\mathrm{E}}}(\boldsymbol{x}) \\ dh_{\mathcal{Z}_{\mathrm{I}}}(\boldsymbol{x}) \end{bmatrix} \\ \boldsymbol{u}_{\bar{\mathcal{Z}}}(\boldsymbol{x}) &= -\sum_{i \in \mathcal{N}_{\mathrm{E}}} \nabla g_i(\boldsymbol{x}) + \sum_{i \in \mathcal{P}_{\mathrm{E}}} \nabla g_i(\boldsymbol{x}) + \sum_{j \in \mathcal{P}_{\mathrm{I}}} \nabla h_j(\boldsymbol{x}). \end{aligned}$$

In this notation the stationarity equation can be recast as

$$\mathbf{0} = \nabla f(\boldsymbol{x}) + \rho \boldsymbol{u}_{\bar{\mathcal{Z}}}(\boldsymbol{x}) + \boldsymbol{U}_{\mathcal{Z}}^*(\boldsymbol{x}) \begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\omega} \end{pmatrix}.$$

Under the assumption that the matrix $\boldsymbol{U}_{\mathcal{Z}}(\boldsymbol{x})$ has full row rank, one can solve for the Lagrange multipliers in the form

$$\begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\omega} \end{pmatrix} = -[\boldsymbol{U}_{\mathcal{Z}}(\boldsymbol{x}) \boldsymbol{U}_{\mathcal{Z}}^*(\boldsymbol{x})]^{-1} \boldsymbol{U}_{\mathcal{Z}}(\boldsymbol{x}) \left[\nabla f(\boldsymbol{x}) + \rho \boldsymbol{u}_{\bar{\mathcal{Z}}}(\boldsymbol{x})\right]. \quad (16.11)$$

Hence, the multipliers are unique. Continuity of the multipliers is a consequence of the continuity of the solution path $\boldsymbol{x}(\rho)$ and the continuity of all functions in sight on the right-hand side of equation (16.11). This observation completes the proof of Proposition 16.4.2.

In addition to the stationarity condition, one must enforce the constraint equations $0 = g_i(\boldsymbol{x})$ for $i \in \mathcal{Z}_{\mathrm{E}}$ and $0 = h_j(\boldsymbol{x})$ for $j \in \mathcal{Z}_{\mathrm{I}}$. Collectively the stationarity and constraint equations can be written as a vector equation $\mathbf{0} = k(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \rho)$ with the active constraints appended below the stationarity condition. To solve for $\boldsymbol{x}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$ in terms of ρ, we apply the implicit

function theorem. This requires calculating the differential of $k(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \rho)$ with respect to the underlying dependent variables $\boldsymbol{x}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\omega}$ and the independent variable ρ. Because the equality constraints are affine, a brief calculation gives

$$\begin{aligned} \partial_{\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\omega}} k &= \begin{bmatrix} d^2 f(\boldsymbol{x}) + \rho \sum_{j \in \mathcal{P}_{\mathrm{I}}} d^2 h_j(\boldsymbol{x}) + \sum_{j \in \mathcal{Z}_{\mathrm{I}}} \omega_j d^2 h_j(\boldsymbol{x}) & \boldsymbol{U}_{\mathcal{Z}}^*(\boldsymbol{x}) \\ \boldsymbol{U}_{\mathcal{Z}}(\boldsymbol{x}) & \mathbf{0} \end{bmatrix} \\ \partial_\rho k &= \begin{bmatrix} \boldsymbol{u}_{\bar{\mathcal{Z}}}(\boldsymbol{x}) \\ \mathbf{0} \end{bmatrix}. \end{aligned}$$

In view of Proposition 5.2.2, the matrix $\partial_{\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\omega}} k(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \rho)$ is nonsingular when its upper-left block is positive definite and its lower-left block has full row rank. Given that it is nonsingular, the implicit function theorem applies, and we can in principle solve for $\boldsymbol{x}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$ in terms of ρ. More importantly, the implicit function theorem supplies the derivative

$$\frac{d}{d\rho} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{\lambda} \\ \boldsymbol{\omega} \end{bmatrix} = -(\partial_{\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\omega}} k)^{-1} \partial_\rho k, \tag{16.12}$$

which is the key to path following. We summarize our findings in the next proposition.

Proposition 16.4.3 *Suppose the surrogate function $\mathcal{E}_\rho(\boldsymbol{y})$ is strictly convex and coercive. If at the point ρ_0 the matrix $\partial_{\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\omega}} k(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \rho)$ is nonsingular and the coefficient of each active constraint occurs on the interior of its subdifferential, then the solution path $\boldsymbol{x}(\rho)$ and Lagrange multipliers $\boldsymbol{\lambda}(\rho)$ and $\boldsymbol{\omega}(\rho)$ satisfy the differential equation (16.12) in the vicinity of ρ_0.*

If one views ρ as time, then one can trace the solution path along the current time segment until either an inactive constraint becomes active or the coefficient of an active constraint hits the boundary of its subdifferential. The earliest hitting time or escape time over all constraints determines the duration of the current segment. When the hitting time for an inactive constraint occurs first, we move the constraint to the appropriate active set $\mathcal{Z}_{\mathrm{E}}$ or $\mathcal{Z}_{\mathrm{I}}$ and keep the other constraints in place. Similarly, when the escape time for an active constraint occurs first, we move the constraint to the appropriate inactive set and keep the other constraints in place. In the second scenario, if s_i hits the value -1, then we move i to $\mathcal{N}_{\mathrm{E}}$; If s_i hits the value 1, then we move i to $\mathcal{P}_{\mathrm{E}}$. Similar comments apply when a coefficient t_j hits 0 or 1. Once this move is executed, we commence path following along the new segment. Path following continues until for sufficiently large ρ, the sets $\mathcal{N}_{\mathrm{E}}$, $\mathcal{P}_{\mathrm{E}}$, and $\mathcal{P}_{\mathrm{I}}$ are exhausted, $\boldsymbol{u}_{\bar{\mathcal{Z}}} = \mathbf{0}$, and the solution vector $\boldsymbol{x}(\rho)$ stabilizes.

Path following simplifies considerably in convex quadratic programming with objective function $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^* A \boldsymbol{x} + \boldsymbol{b}^* \boldsymbol{x}$ and equality constraints

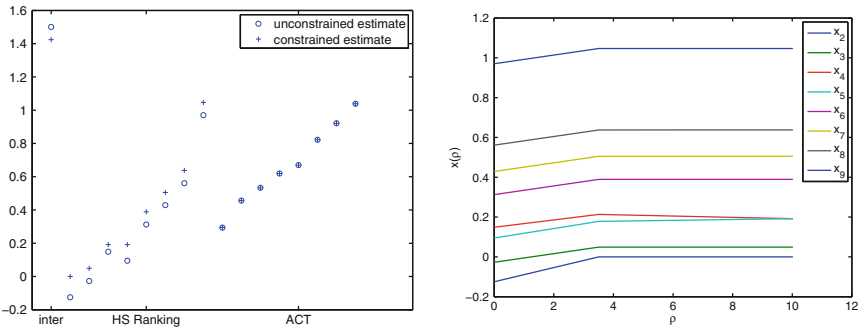

FIGURE 16.1. *Left*: Unconstrained and constrained estimates for the Iowa GPA data. *Right*: Solution paths for the high school rank regression coefficients

$\boldsymbol{V}\boldsymbol{x} = \boldsymbol{d}$ and inequality constraints $\boldsymbol{W}\boldsymbol{x} \leq \boldsymbol{e}$, where $\boldsymbol{A}$ is positive semidefinite. The exact penalized objective function becomes

$$\mathcal{E}_\rho(\boldsymbol{x}) \quad = \quad \frac{1}{2}\boldsymbol{x}^* A\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x} + \rho\sum_{i=1}^{s}|\boldsymbol{v}_i^*\boldsymbol{x} - d_i| + \rho\sum_{j=1}^{*}(\boldsymbol{w}_j^*\boldsymbol{x} - e_j)_+.$$

Since both the equality and inequality constraints are affine, their second derivatives vanish. Both $\boldsymbol{U}_{\mathcal{Z}}$ and $\boldsymbol{u}_{\bar{\mathcal{Z}}}$ are constant on the current path segment, and the path $\boldsymbol{x}(\rho)$ satisfies

$$\frac{d}{d\rho}\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{\lambda}\\ \boldsymbol{\omega}\end{bmatrix} \quad = \quad -\begin{pmatrix}\boldsymbol{A} & \boldsymbol{U}_{\bar{\mathcal{Z}}}^*\\ \boldsymbol{U}_{\mathcal{Z}} & \boldsymbol{0}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{u}_{\bar{\mathcal{Z}}}\\ \boldsymbol{0}\end{pmatrix}. \tag{16.13}$$

Because the solution path $\boldsymbol{x}(\rho)$ is piecewise linear, it is possible to anticipate the next hitting or exit time and take a large jump. The matrix inverse appearing in equation (16.13) can be efficiently updated by the sweep operator of computational statistics [283].

Example 16.4.1 *Partial Isotone Regression*

Order-constrained regression is now widely accepted as an important modeling tool in statistics [219, 236]. If $\boldsymbol{x}$ is the parameter vector, monotone regression includes isotone constraints $x_1 \leq x_2 \leq \cdots \leq x_m$ and antitone constraints $x_1 \geq x_2 \geq \cdots \geq x_m$. In partially ordered regression, subsets of the parameters are subject to isotone or antitone constraints. As an example of partial isotone regression, consider the data from Table 1.3.1 of the reference [219] on the first-year grade point averages (GPA) of 2397 University of Iowa freshmen. These data can be downloaded as part of the `R` package `ic.infer`. The ordinal predictors high school rank (as a percentile) and ACT (a standard aptitude test) score are discretized into nine ordered categories each. It is rational to assume that college performance is isotone separately within each predictor set. Figure 16.1 shows

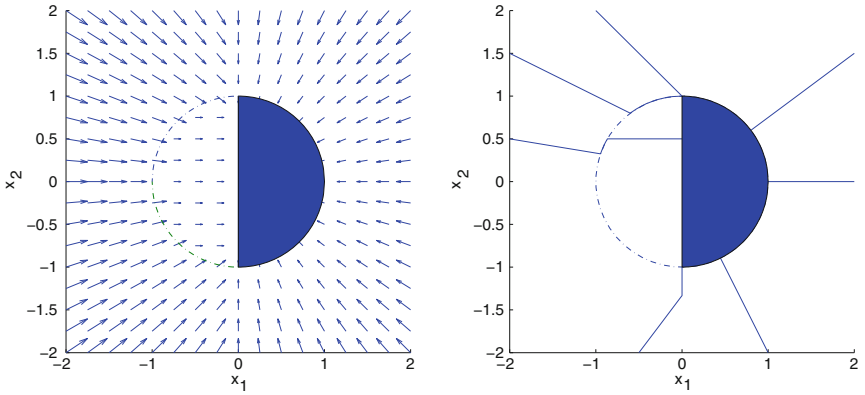

FIGURE 16.2. Projection to the positive half-disc. *Left*: Derivatives at $\rho = 0$ for projection onto the half-disc. *Right*: Projection trajectories from various initial points

the unconstrained and constrained solutions for the intercept and the two predictor sets and the solution path of the regression coefficients for the high school rank predictor. In this quadratic programming problem, the solution path is piecewise linear. In contrast the next example involves nonlinear path segments. ■

Example 16.4.2 *Projection onto the Half-Disc*

Dykstra's algorithm as explained in Sect. 15.2 projects an exterior point onto the intersection of a finite number of closed convex sets. The projection problem also yields to path following. Consider our previous toy example of projecting a point $\boldsymbol{b} \in \mathsf{R}^2$ onto the intersection of the closed unit ball and the closed half space $x_1 \geq 0$. This is equivalent to solving

$$\begin{array}{lrcl} \text{minimize} & f(\boldsymbol{x}) & = & \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{b}\|^2 \\ \text{subject to} & h_1(\boldsymbol{x}) & = & \frac{1}{2}\|\boldsymbol{x}\|^2 - \frac{1}{2} \leq 0, \quad h_2(\boldsymbol{x}) = -x_1 \leq 0 \end{array}$$

with gradients and second differentials

$$\begin{aligned} \nabla f(\boldsymbol{x}) &= \boldsymbol{x}-\boldsymbol{b}, \quad \nabla h_1(\boldsymbol{x}) = \boldsymbol{x}, \quad \nabla h_2(\boldsymbol{x}) = -\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \\ d^2 f(\boldsymbol{x}) &= d^2 h_1(\boldsymbol{x}) = \boldsymbol{I}_2, \quad d^2 h_2(\boldsymbol{x}) = \boldsymbol{0}. \end{aligned}$$

Path following starts from the unconstrained solution $\boldsymbol{x}(0) = \boldsymbol{b}$. The left panel of Fig. 16.2 plots the vector field $\frac{d}{d\rho}\boldsymbol{x}$ at the time $\rho = 0$. The right panel shows the solution path for projection from the points $(-2, 0.5)$,

$(-2, 1.5)$, $(-1, 2)$, $(2, 1.5)$, $(2, 0)$, $(1, 2)$, and $(-0.5, -2)$ onto the feasible region. In contrast to the previous example, small steps are taken. In projecting the point $\boldsymbol{b}^* = (-1, 2)$ onto $(0, 1)$, our software exploits the ODE45 solver of MATLAB. Following the solution path requires derivatives at 19 different time points. Dykstra's algorithm by comparison takes about 30 iterations to converge. ■

16.5 Bregman Iteration

We met Bregman functions previously in Sect. 13.3 in the study of adaptive barrier methods. If $J(\boldsymbol{u})$ is a convex function and $\boldsymbol{p} \in \partial J(\boldsymbol{v})$ is any subgradient, then the associated Bregman function

$$D_J^{\boldsymbol{p}}(\boldsymbol{u} \mid \boldsymbol{v}) \quad = \quad J(\boldsymbol{u}) - J(\boldsymbol{v}) - \boldsymbol{p}^*(\boldsymbol{u} - \boldsymbol{v})$$

defines a kind of distance anchored at $\boldsymbol{v}$ [24, 25]. When $J(\boldsymbol{u})$ is differentiable, the superscript $\boldsymbol{p}$ is redundant, and we omit it. The exercises at the end of the chapter list some properties of Bregman distances. In general, the symmetry and triangle properties of a metric fail. Figure 16.3 illustrates graphically a Bregman distance for a smooth function in one dimension. Problem 9 lists some commonly encountered Bregman distances.

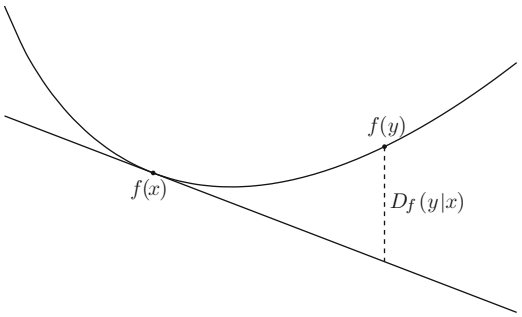

FIGURE 16.3. The Bregman distance generated by a smooth function $f(\boldsymbol{x})$

Example 16.5.1 *Kullback–Leibler Divergence*

The density function of a random vector $\boldsymbol{Y}$ in a natural exponential family can be written as

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}) \quad = \quad g(\boldsymbol{y})e^{-h(\boldsymbol{\theta})+\boldsymbol{y}^*\boldsymbol{\theta}}$$

for a parameter vector $\boldsymbol{\theta}$. Here $\boldsymbol{Y}$ itself is the sufficient statistic. The five most important univariate examples are: the normal distribution with known variance, the Poisson distribution, the gamma distribution with

known shape parameter, the binomial distribution with known number of trials, and the negative binomial with known number of required successes. The analysis of Sect. 10.6 shows that

$$\mathrm{E}_{\boldsymbol{\theta}}(\boldsymbol{Y}) \;=\; \nabla h(\boldsymbol{\theta}), \quad \mathrm{Var}_{\boldsymbol{\theta}}(\boldsymbol{Y}) \;=\; d^2 h(\boldsymbol{\theta}).$$

It follows that $h(\boldsymbol{\theta})$ is convex. The Fenchel conjugate

$$h^\star(\boldsymbol{y}) \;=\; \sup_{\boldsymbol{\theta}} \left[\boldsymbol{y}^*\boldsymbol{\theta} - h(\boldsymbol{\theta})\right] \;=\; \sup_{\boldsymbol{\theta}} \ln f(\boldsymbol{y} \mid \boldsymbol{\theta})$$

determines the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ through the likelihood equation $\boldsymbol{y} = \nabla h(\boldsymbol{\theta})$. The Bregman distance

$$\begin{aligned} D_h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0) &= h(\boldsymbol{\theta}_1) - h(\boldsymbol{\theta}_0) - dh(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0) \\ &= \mathrm{E}_{\boldsymbol{\theta}_0}\left[h(\boldsymbol{\theta}_1) - h(\boldsymbol{\theta}_0) - \boldsymbol{Y}^*(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)\right] \\ &= \mathrm{E}_{\boldsymbol{\theta}_0}\left[\ln \frac{f(\boldsymbol{Y} \mid \boldsymbol{\theta}_0)}{f(\boldsymbol{Y} \mid \boldsymbol{\theta}_1)}\right] \end{aligned}$$

coincides with the Kullback–Leibler divergence of the densities $f(\boldsymbol{y} \mid \boldsymbol{\theta}_0)$ and $f(\boldsymbol{y} \mid \boldsymbol{\theta}_1)$. ■

Osher [120, 208, 279] and colleagues have pioneered the application of Bregman iteration in compressed sensing. One of their motivating examples is to minimize the convex function $J(\boldsymbol{u}) = \|\boldsymbol{u}\|_1$ subject to the equality constraint $H(\boldsymbol{u}) = \min_{\boldsymbol{v}} H(\boldsymbol{v})$ for $H(\boldsymbol{u})$ smooth and convex. In the simple case of basis pursuit, $H(\boldsymbol{u})$ equals the sum of squares criterion $\frac{1}{2}\|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f}\|^2$. The next Bregman iterate $\boldsymbol{u}_{k+1}$ minimizes the surrogate function

$$G(\boldsymbol{u} \mid \boldsymbol{u}_k) \;=\; \lambda H(\boldsymbol{u}) + D_J^{\boldsymbol{p}_k}(\boldsymbol{u} \mid \boldsymbol{u}_k).$$

Here $\lambda > 0$ is a scaling constant determining the relative contribution of $H(\boldsymbol{u})$. For the sake of simplicity, we will take $\lambda = 1$ by absorbing its value in the definition of $H(\boldsymbol{u})$. We will also shift $H(\boldsymbol{u})$ so that its minimum value is 0. This action does not affect the choice of the update $\boldsymbol{u}_{k+1}$. Bregman iteration tries to drive $H(\boldsymbol{u})$ to 0 and simultaneously minimize $J(\boldsymbol{u})$ subject to the constraint $H(\boldsymbol{u}) = 0$. There are four obvious questions raised by Bregman iteration. First, does the next iterate $\boldsymbol{u}_{k+1}$ exist? This is certainly the case when $G(\boldsymbol{u} \mid \boldsymbol{u}_k)$ is coercive. Second, is $\boldsymbol{u}_{k+1}$ uniquely determined? Strict convexity of $G(\boldsymbol{u} \mid \boldsymbol{u}_k)$ suffices. Third, how can one find $\boldsymbol{u}_{k+1}$? In sparse problems with $J(\boldsymbol{u}) = \|\boldsymbol{u}\|_1$, coordinate descent works well. The fourth question of how to choose the subgradient $\boldsymbol{p}_k$ is the most subtle of all.

The MM principle is the key to understanding Bregman iteration. Both the objective function $H(\boldsymbol{u})$ and the Bregman function $D_J^{\boldsymbol{p}}(\boldsymbol{u} \mid \boldsymbol{u}_k)$ are convex. Because $D_J^{\boldsymbol{p}}(\boldsymbol{u} \mid \boldsymbol{u}_k)$ is anchored at $\boldsymbol{u}_k$ and majorizes the constant

0, the surrogate function $G(\boldsymbol{u} \mid \boldsymbol{u}_k)$ majorizes $H(\boldsymbol{u})$ at $\boldsymbol{u}_k$. Minimizing the surrogate therefore drives $H(\boldsymbol{u})$ downhill. The preferred update of the subgradient is straightforward to explain in this context. The convex stationarity condition $\mathbf{0} \in \nabla H(\boldsymbol{u}_{k+1}) + \partial J(\boldsymbol{u}_{k+1}) - \boldsymbol{p}_k$ shows that

$$\boldsymbol{p}_{k+1} \;=\; \boldsymbol{p}_k - \nabla H(\boldsymbol{u}_{k+1}) \;\in\; \partial J(\boldsymbol{u}_{k+1}).$$

Thus, the MM update furnishes a candidate subgradient. Treating the last equation recursively gives the further identity

$$\boldsymbol{p}_k \;=\; \boldsymbol{p}_0 - \sum_{i=1}^{k} \nabla H(\boldsymbol{u}_i). \tag{16.14}$$

If $J(\boldsymbol{u}) = \|\boldsymbol{u}\|_1$, then $\boldsymbol{u}_0 = \boldsymbol{p}_0 = \mathbf{0}$ is an obvious starting point for Bregman iteration.

Fortunately, there is a simple proof that the Bregman iterates as just defined send $H(\boldsymbol{u})$ to 0. The argument invokes the identity

$$D_J^{\boldsymbol{p}}(\boldsymbol{u} \mid \boldsymbol{v}) + D_J^{\boldsymbol{q}}(\boldsymbol{v} \mid \boldsymbol{w}) - D_J^{\boldsymbol{q}}(\boldsymbol{u} \mid \boldsymbol{w}) \;=\; (\boldsymbol{p}-\boldsymbol{q})^*(\boldsymbol{v}-\boldsymbol{u}), \tag{16.15}$$

which the reader can readily verify. If we suppose that $H(\boldsymbol{u})$ is coercive, then it achieves its minimum of 0 at some point $\hat{\boldsymbol{u}}$. The identity (16.15) and the convexity of $H(\boldsymbol{u})$ therefore imply that

$$\begin{aligned}
& D_J^{\boldsymbol{p}_k}(\hat{\boldsymbol{u}} \mid \boldsymbol{u}_k) - D_J^{\boldsymbol{p}_{k-1}}(\hat{\boldsymbol{u}} \mid \boldsymbol{u}_{k-1}) \\
\leq\; & D_J^{\boldsymbol{p}_k}(\hat{\boldsymbol{u}} \mid \boldsymbol{u}_k) + D_J^{\boldsymbol{p}_{k-1}}(\boldsymbol{u}_k \mid \boldsymbol{u}_{k-1}) - D_J^{\boldsymbol{p}_{k-1}}(\hat{\boldsymbol{u}} \mid \boldsymbol{u}_{k-1}) \\
=\; & (\boldsymbol{p}_k - \boldsymbol{p}_{k-1})^*(\boldsymbol{u}_k - \hat{\boldsymbol{u}}) \\
=\; & dH(\boldsymbol{u}_k)(\hat{\boldsymbol{u}} - \boldsymbol{u}_k) \\
\leq\; & H(\hat{\boldsymbol{u}}) - H(\boldsymbol{u}_k) \\
=\; & -H(\boldsymbol{u}_k).
\end{aligned} \tag{16.16}$$

Summing the extremes of inequality (16.16) from 1 to m produces

$$\begin{aligned}
\sum_{k=1}^{m} [D_J^{\boldsymbol{p}_k}(\hat{\boldsymbol{u}} \mid \boldsymbol{u}_k) - D_J^{\boldsymbol{p}_{k-1}}(\hat{\boldsymbol{u}} \mid \boldsymbol{u}_{k-1})] \;&=\; D_J^{\boldsymbol{p}_m}(\hat{\boldsymbol{u}} \mid \boldsymbol{u}_m) - D_J^{\boldsymbol{p}_0}(\hat{\boldsymbol{u}} \mid \boldsymbol{u}_0) \\
&\leq\; -\sum_{k=1}^{m} H(\boldsymbol{u}_k) \\
&\leq\; -mH(\boldsymbol{u}_m).
\end{aligned}$$

Rearranging this now yields

$$H(\boldsymbol{u}_m) \;\leq\; \frac{1}{m}\Big[D_J^{\boldsymbol{p}_0}(\hat{\boldsymbol{u}} \mid \boldsymbol{u}_0) - D_J^{\boldsymbol{p}_m}(\hat{\boldsymbol{u}} \mid \boldsymbol{u}_m)\Big] \;\leq\; \frac{1}{m} D_J^{\boldsymbol{p}_0}(\hat{\boldsymbol{u}} \mid \boldsymbol{u}_0).$$

The convergence of $H(\boldsymbol{u}_m)$ to 0 is an immediate consequence.

For the sum of squares criterion $H(\boldsymbol{u}) = \frac{\lambda}{2}\|\boldsymbol{A}\boldsymbol{u}-\boldsymbol{f}\|^2$, Bregman iteration simplifies considerably. Given the initial conditions $\boldsymbol{u}_0 = \boldsymbol{p}_0 = \boldsymbol{0}$, define $\boldsymbol{f}_0 = \boldsymbol{f}$ and $\boldsymbol{f}_k = \boldsymbol{f}_{k-1} + (\boldsymbol{f} - \boldsymbol{A}\boldsymbol{u}_k)$. It then follows from equation (16.14) and telescoping that

$$\begin{aligned}
H(\boldsymbol{u}) - \boldsymbol{p}_k^* \boldsymbol{u} &= H(\boldsymbol{u}) + \Big[\sum_{i=1}^{k} \nabla H(\boldsymbol{u}_i)\Big]^* \boldsymbol{u} \\
&= \frac{\lambda}{2}\|\boldsymbol{A}\boldsymbol{u}\|^2 + \frac{\lambda}{2}\|\boldsymbol{f}\|^2 - \lambda \boldsymbol{f}^* \boldsymbol{A}\boldsymbol{u} + \lambda\Big[\sum_{i=1}^{k} \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{u}_i - \boldsymbol{f})\Big]^* \boldsymbol{u} \\
&= \frac{\lambda}{2}\|\boldsymbol{A}\boldsymbol{u}\|^2 + \frac{\lambda}{2}\|\boldsymbol{f}\|^2 - \lambda\Big[\boldsymbol{f} + \sum_{i=1}^{k}(\boldsymbol{f} - \boldsymbol{A}\boldsymbol{u}_i)\Big]^* \boldsymbol{A}\boldsymbol{u} \\
&= \frac{\lambda}{2}\|\boldsymbol{A}\boldsymbol{u}\|^2 + \frac{\lambda}{2}\|\boldsymbol{f}\|^2 - \lambda \boldsymbol{f}_k^* \boldsymbol{A}\boldsymbol{u} \\
&= \frac{\lambda}{2}\|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f}_k\|^2 - \frac{\lambda}{2}\|\boldsymbol{f}_k\|^2 + \frac{\lambda}{2}\|\boldsymbol{f}\|^2.
\end{aligned}$$

Thus, the Bregman surrogate function becomes

$$G(\boldsymbol{u} \mid \boldsymbol{u}_k) = \frac{\lambda}{2}\|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f}_k\|^2 + J(\boldsymbol{u}) + c_k,$$

where c_k is an irrelevant constant. Section 13.5 sketches how to find $\boldsymbol{u}_{k+1}$ by coordinate descent when $J(\boldsymbol{u}) = \|\boldsymbol{u}\|_1$.

The linearized version of Bregman iteration [209] relies on the approximation

$$H(\boldsymbol{u}) \approx H(\boldsymbol{u}_k) + dH(\boldsymbol{u}_k)(\boldsymbol{u} - \boldsymbol{u}_k).$$

Since this Taylor expansion is apt to be accurate only when $\|\boldsymbol{u} - \boldsymbol{u}_k\|$ is small, the surrogate function is re-defined as

$$G(\boldsymbol{u} \mid \boldsymbol{u}_k) = H(\boldsymbol{u}_k) + dH(\boldsymbol{u}_k)(\boldsymbol{u} - \boldsymbol{u}_k) + D_J^{\boldsymbol{p}_k}(\boldsymbol{u} \mid \boldsymbol{u}_k) + \frac{1}{2\delta}\|\boldsymbol{u} - \boldsymbol{u}_k\|^2$$

for $\delta > 0$. The quadratic penalty majorizes 0 and shrinks the next iterate $\boldsymbol{u}_{k+1}$ toward $\boldsymbol{u}_k$. Separation of parameters in the surrogate function when $J(\boldsymbol{u}) = \|\boldsymbol{u}\|_1$ is a huge advantage in solving for $\boldsymbol{u}_{k+1}$. Examination of the convex stationarity conditions then shows that

$$u_{k+1,j} = \begin{cases} u_{k+1,j}^+ = u_{kj} + \delta\Big[p_{kj} - \frac{\partial}{\partial u_j} H(\boldsymbol{u}_k) - 1\Big] & u_{k+1,j}^+ > 0 \\ u_{k+1,j}^- = u_{kj} + \delta\Big[p_{kj} - \frac{\partial}{\partial u_j} H(\boldsymbol{u}_k) + 1\Big] & u_{k+1,j}^- < 0 \\ 0 & \text{otherwise.} \end{cases}$$

Given that the surrogate function approximately majorizes $H(\boldsymbol{u})$ in a neighborhood of $\boldsymbol{u}_k$, one can also confidently expect the linearized Bregman iterates to drive $H(\boldsymbol{u})$ downhill.

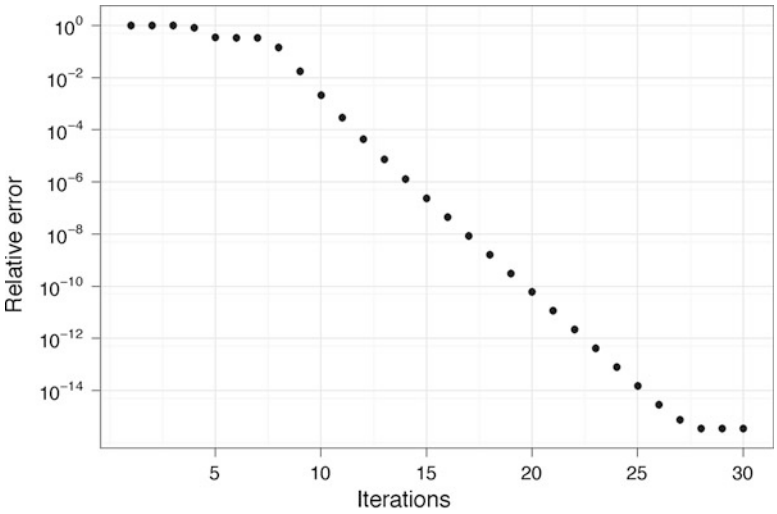

FIGURE 16.4. Relative error versus iteration number for basis pursuit

Figure 16.4 plots the results of linearized Bregman iteration for a simple numerical example. Here the entries of the $10^2 \times 10^5$ matrix $\boldsymbol{A}$ are populated with independent standard normal deviates. The generating sparse vector $\boldsymbol{u}$ has all entries equal to 0 except for

$$u_{45373} = -1.162589, \quad u_{57442} = 2.436616, \quad u_{81515} = 1.876241.$$

The response vector $\boldsymbol{f}$ equals $\boldsymbol{Au}$, and the Bregman constants are $\delta = 0.01$ and $\lambda = 1000$. The figure plots the relative error $\|\boldsymbol{u}-\boldsymbol{u}_k\|/\|\boldsymbol{u}\|$ as a function of iteration number k. It is remarkable how quickly the true $\boldsymbol{u}$ is recovered in the absence of noise. As one might expect, lasso penalized linear regression, also known as basis pursuit denoising, produces solutions very similar to basis pursuit.

16.6 Split Bregman Iteration

Our focus on the penalty $J(\boldsymbol{u}) = \|\boldsymbol{u}\|_1$ obscures the fact that many applications really require $J(\boldsymbol{u}) = \|\boldsymbol{Du}\|_1$ for a constant matrix $\boldsymbol{D}$. For instance, the well-known image denoising model of Rudin, Osher, and Fatemi [224] minimizes the total variation regularized sum of squares criterion

$$\frac{1}{2}\sum_{i,j}(w_{ij} - u_{ij})^2 + \rho \sum_{i,j} \sqrt{(u_{i+1,j} - u_{ij})^2 + (u_{i,j+1} - u_{ij})^2},$$

where w_{ij} is the corrupted intensity and u_{ij} is the true intensity for pixel (i, j) of an image. The total variation penalty represented by the second

sum is intended to smooth the reconstructed image while preserving its edges. A similar effect can be achieved by adopting the anisotropic penalty

$$\rho \sum_{i,j} \Big(|u_{i+1,j} - u_{ij}| + |u_{i,j+1} - u_{ij}| \Big),$$

which has the form $J(\boldsymbol{u}) = \rho \|\boldsymbol{D}\boldsymbol{u}\|_1$ for $\boldsymbol{D}\boldsymbol{u}$ linear.

Split Bregman iteration is intended to handle this kind of situation [106]. Consider minimization of the criterion $E(\boldsymbol{u}) + \|\boldsymbol{D}\boldsymbol{u}\|_1$, where $\boldsymbol{D}$ is linear and $E(\boldsymbol{u})$ is convex and differentiable. In split Bregman iteration the general idea is to introduce a new variable $\boldsymbol{d} = \boldsymbol{D}\boldsymbol{u}$ and carry out Bregman iteration with the objective function $H(\boldsymbol{u}, \boldsymbol{d}) = \frac{1}{2\delta}\|\boldsymbol{d} - \boldsymbol{D}\boldsymbol{u}\|^2$ modified by the Bregman function derived from $J(\boldsymbol{u}, \boldsymbol{d}) = E(\boldsymbol{u}) + \|\boldsymbol{d}\|_1$. Thus, one selects the pair $(\boldsymbol{u}_{k+1}, \boldsymbol{d}_{k+1})$ to minimize the criterion

$$\frac{1}{2\delta}\|\boldsymbol{d} - \boldsymbol{D}\boldsymbol{u}\|^2 + E(\boldsymbol{u}) + \|\boldsymbol{d}\|_1 - \boldsymbol{p}_k^*(\boldsymbol{u} - \boldsymbol{u}_k) - \boldsymbol{q}_k^*(\boldsymbol{d} - \boldsymbol{d}_k) \tag{16.17}$$

for subgradients $\boldsymbol{p}_k \in \partial E(\boldsymbol{u}_k)$ and $\boldsymbol{q}_k \in \partial \|\boldsymbol{d}_k\|_1$. Once the next iterate $(\boldsymbol{u}_{k+1}, \boldsymbol{d}_{k+1})$ is determined, the new subgradients

$$\begin{aligned} \boldsymbol{p}_{k+1} &= \boldsymbol{p}_k - \nabla_{\boldsymbol{u}} H(\boldsymbol{u}_{k+1}, \boldsymbol{d}_{k+1}) \\ \boldsymbol{q}_{k+1} &= \boldsymbol{q}_k - \nabla_{\boldsymbol{d}} H(\boldsymbol{u}_{k+1}, \boldsymbol{d}_{k+1}) \end{aligned}$$

are defined. Block relaxation is the natural method of minimizing the criterion (16.17). If $E(\boldsymbol{u})$ is well behaved, then one can minimize

$$\frac{1}{2\delta}\|\boldsymbol{d} - \boldsymbol{D}\boldsymbol{u}\|^2 + E(\boldsymbol{u}) - \boldsymbol{p}_k^*(\boldsymbol{u} - \boldsymbol{u}_k)$$

with respect to $\boldsymbol{u}$ by Newton's method. Minimizing

$$\frac{1}{2\delta}\|\boldsymbol{d} - \boldsymbol{D}\boldsymbol{u}\|^2 + \|\boldsymbol{d}\|_1 - \boldsymbol{q}_k^*(\boldsymbol{d} - \boldsymbol{d}_k)$$

with respect to $\boldsymbol{d}$ can be achieved in a single iteration by the shrinkage rule

$$d_j = \begin{cases} d_j^+ = (\boldsymbol{D}\boldsymbol{u})_j + \delta(q_{kj} - 1) & d_j^+ > 0 \\ d_j^- = (\boldsymbol{D}\boldsymbol{u})_j + \delta(q_{kj} + 1) & d_j^- < 0 \\ 0 & \text{otherwise} \end{cases}$$

suggested in our discussion of linearized Bregman iteration.

Example 16.6.1 *The Fused Lasso*

The fused lasso [257] penalty is the ℓ_1 norm of the successive differences between parameters. The simplest possible fused lasso problem minimizes the penalized sum of squares criterion

$$E(\boldsymbol{u}) + \|\boldsymbol{D}\boldsymbol{u}\|_1 = \frac{\lambda}{2}\sum_{i=1}^{n}(y_i - u_i)^2 + \sum_{i=1}^{n-1}|u_{i+1} - u_i|. \tag{16.18}$$

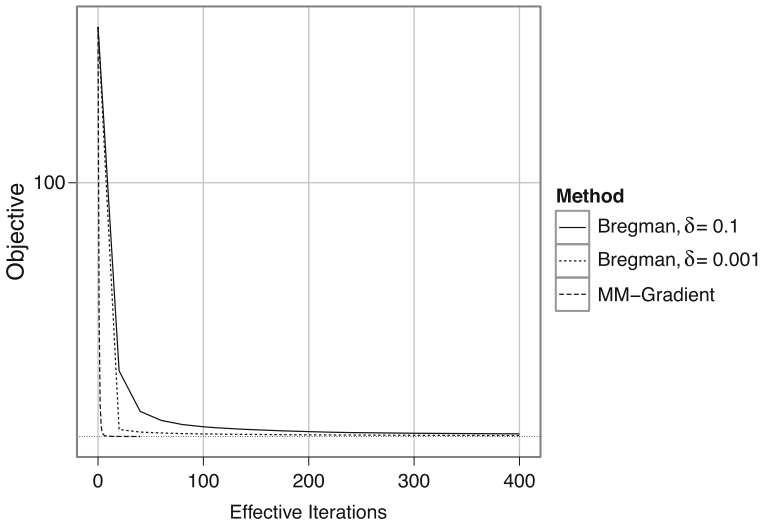

FIGURE 16.5. Convergence to the objective (16.18) for the fused lasso problem

Note that the matrix $\boldsymbol{D}$ in the penalty $\|\boldsymbol{D}\boldsymbol{u}\|_1$ is bidiagonal. The differences $u_{i+1} - u_i$ under the absolute value signs make it difficult to implement coordinate descent, so split Bregman iteration is an attractive possibility. An alternative [280] is to change the penalty slightly and minimize the revised objective function

$$f(\boldsymbol{u}) \;=\; \frac{\lambda}{2}\sum_{i=1}^{n}(y_i - u_i)^2 + \sum_{i=1}^{n-1}[(u_{i+1} - u_i)^2 + \epsilon]^{1/2},$$

which smoothly approximates the original objective function for small positive ϵ. The majorization (8.12) translates into the quadratic majorization

$$g(\boldsymbol{u} \mid \boldsymbol{u}_k) \;=\; \frac{\lambda}{2}\sum_{i=1}^{n}(y_i - u_i)^2 + \sum_{i=1}^{n-1}\frac{(u_{i+1} - u_i)^2}{2[u_{k,i+1} - u_{ki})^2 + \epsilon]^{1/2}} + c$$

for an irrelevant constant c. The MM gradient algorithm for updating $\boldsymbol{u}$ is easy to implement because the second differential $d^2 g(\boldsymbol{u} \mid \boldsymbol{u}_k)$ is tridiagonal and efficiently inverted at $\boldsymbol{u} = \boldsymbol{u}_k$ by Thomas's algorithm [52]. In fact, one step of Newton's method minimizes the quadratic $g(\boldsymbol{u} \mid \boldsymbol{u}_k)$. Updating $\boldsymbol{u}$ in split Bregman iteration also benefits from Thomas's algorithm. Observe, however, that the multiple inner iterations of block descent puts split Bregman iteration at a computational disadvantage.

Figure 16.5 compares the performance of split Bregman iteration and the MM gradient algorithm on simulated data with independently generated responses $y_1, \ldots, y_{1000}$ sampled from two Gaussian densities. These densities share a common standard deviation of 0.2 but differ in their means of

0.6 for $501 \leq i \leq 550$ and 0 otherwise. Thus, there is a brief elevation in the signal for a short segment in the middle. Both algorithms commence from the starting values $\boldsymbol{u}_0 = \boldsymbol{y}$ and $\boldsymbol{d}_0 = \boldsymbol{0}$. For split Bregman iteration $\boldsymbol{p}_0 = \boldsymbol{q}_0 = \boldsymbol{0}$; for the MM gradient algorithm $\epsilon = 10^{-10}$. The tuning constant $\lambda = 2.5/\sqrt{\ln 1000}$. The effective number of iterations plotted in Fig. 16.5 counts the total inner iterations for split Bregman iteration. ■

16.7 Convergence of Bregman Iteration

Proving convergence of Bregman iteration is difficult [147, 279]. The crux of the problem is that minimizing $J(\boldsymbol{u})$ is secondary to minimizing $H(\boldsymbol{u})$. Owing to the difficulties, we will only tackle convergence for the special choices $H(\boldsymbol{u}) = \frac{\lambda}{2}\|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f}\|^2$ and $J(\boldsymbol{u}) = \|\boldsymbol{D}\boldsymbol{u}\|_1$ made in image denoising. If $\boldsymbol{A}$ does not have full column rank, then the purpose of Bregman iteration is to minimize the secondary criterion $J(\boldsymbol{u}) = \|\boldsymbol{D}\boldsymbol{u}\|_1$ subject to $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}$. The first question that comes to mind is whether the minimum exists. Let K be the kernel of the matrix $\boldsymbol{A}$ and $\boldsymbol{x}$ be a particular solution of the equation $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}$. This equation's solution space is simply $\boldsymbol{x} + K$. If $\boldsymbol{D}\boldsymbol{x} = \boldsymbol{0}$, then the minimum value of $J(\boldsymbol{u})$ is achieved at $\boldsymbol{x}$. If in addition $\boldsymbol{D}\boldsymbol{y} = \boldsymbol{0}$ for some $\boldsymbol{y} \in K$, then the minimum of 0 is also achieved along the entire line through $\boldsymbol{x}$ along the direction $\boldsymbol{y}$.

Now consider whether the function $\boldsymbol{y} \mapsto \|\boldsymbol{D}(\boldsymbol{x} + \boldsymbol{y})\|_1$ is coercive on K. Coerciveness implies that the minimum exists. In view of the equivalence of norms, it suffices to decide whether the function $\boldsymbol{y} \mapsto \|\boldsymbol{D}(\boldsymbol{x} + \boldsymbol{y})\|$ is coercive. Proposition 12.3.1 implies that the Euclidean norm is coercive if and only if

$$\lim_{t\to\infty} \|\boldsymbol{D}(\boldsymbol{x} + t\boldsymbol{y})\|^2 = \lim_{t\to\infty} \|\boldsymbol{D}\boldsymbol{x}\|^2 + 2t\boldsymbol{y}^*\boldsymbol{D}^*\boldsymbol{D}\boldsymbol{x} + t^2\boldsymbol{y}^*\boldsymbol{D}^*\boldsymbol{D}\boldsymbol{y} = \infty$$

for all $\boldsymbol{y} \in K$. This is clearly equivalent to the condition $\boldsymbol{D}\boldsymbol{y} \neq \boldsymbol{0}$ for all $\boldsymbol{y} \in K$. Hence, the condition $\boldsymbol{D}\boldsymbol{y} \neq \boldsymbol{0}$ for every $\boldsymbol{y} \in K$ is necessary and sufficient for coerciveness. According to the discussion following Example 5.5.3, another equivalent condition is that the matrix $\boldsymbol{A}^*\boldsymbol{A} + \gamma\boldsymbol{D}^*\boldsymbol{D}$ is positive definite for some $\gamma > 0$.

The boundedness of the iterates and subdifferentials plays a key role in proving convergence. The latter are easy to handle because the chain rule entails $\partial J(\boldsymbol{u}) = \boldsymbol{D}^*\partial\|\boldsymbol{D}\boldsymbol{u}\|_1$. For any $\boldsymbol{v} \in \mathsf{R}^n$, $\partial\|\boldsymbol{v}\|_1$ is contained in the n-dimensional cube $[-1, 1]^n$. Hence, $\partial J(\boldsymbol{u})$ is contained in the compact set $\boldsymbol{D}^*[-1, 1]^n$ for all $\boldsymbol{u}$. One can guarantee that the iterates $\boldsymbol{u}_k$ are also bounded whenever $H(\boldsymbol{u})$ is coercive. Unfortunately, this is not the case for an under-determined system $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}$. We will simply postulate that the iteration sequence $\boldsymbol{u}_k$ is bounded. We will also assume that $J(\boldsymbol{u})$ achieves its constrained minimum at some point $\boldsymbol{w}$.

Now suppose some subsequence of the Bregman iterates $\boldsymbol{u}_{k+1}$ converges to a cluster point $\boldsymbol{y}$. To show that $\boldsymbol{y}$ minimizes the criterion $J(\boldsymbol{u})$ subject to $H(\boldsymbol{u}) = 0$, we examine the inequality

$$\begin{aligned} & H(\boldsymbol{u}_{k+1}) + J(\boldsymbol{u}_{k+1}) - J(\boldsymbol{u}_k) - \boldsymbol{p}_k^*(\boldsymbol{u}_{k+1} - \boldsymbol{u}_k) \\ & \leq H(\boldsymbol{w}) + J(\boldsymbol{w}) - J(\boldsymbol{u}_k) - \boldsymbol{p}_k^*(\boldsymbol{w} - \boldsymbol{u}_k). \end{aligned}$$

By passing to a subsequence if necessary, we can assume that $\boldsymbol{u}_k$ converges to $\boldsymbol{x}$, $\boldsymbol{u}_{k+1}$ converges to $\boldsymbol{y}$, and $\boldsymbol{p}_k$ converges to $\boldsymbol{p}$. If we can show that

$$\boldsymbol{p}^*(\boldsymbol{y} - \boldsymbol{x}) \quad = \quad \boldsymbol{p}^*(\boldsymbol{w} - \boldsymbol{x}) \quad = \quad 0,$$

then in the limit the descent property of $H(\boldsymbol{u})$ implies $J(\boldsymbol{y}) \leq J(\boldsymbol{w})$. Thus, the cluster point $\boldsymbol{y}$ is also optimal.

We now check that $\boldsymbol{p}^*(\boldsymbol{y} - \boldsymbol{x}) = 0$. The other inner product vanishes for similar reasons. Recall that we commence with $\boldsymbol{u}_0 = \boldsymbol{p}_0 = \boldsymbol{0}$. Thus, $\boldsymbol{p}_0$ belongs to the range of $\boldsymbol{A}^*$. In general, this assertion is true for all $\boldsymbol{p}_k$ because $\boldsymbol{p}_{k+1} = \boldsymbol{p}_k - \lambda \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{u}_{k+1} - \boldsymbol{f})$. Given that the range of $\boldsymbol{A}^*$ is closed, there exists a vector $\boldsymbol{z}$ with $\boldsymbol{p} = \boldsymbol{A}^*\boldsymbol{z}$. Furthermore, $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} = \boldsymbol{f}$ since $H(\boldsymbol{y}) = H(\boldsymbol{x}) = 0$. It follows that

$$\boldsymbol{p}^*(\boldsymbol{y} - \boldsymbol{x}) \quad = \quad \boldsymbol{z}^*\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{x}) \quad = \quad 0.$$

Let us summarize this discussion in a proposition.

Proposition 16.7.1 *Consider Bregman iteration for the function choices* $H(\boldsymbol{u}) = \frac{\lambda}{2}\|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f}\|^2$ *and* $J(\boldsymbol{u}) = \|\boldsymbol{D}\boldsymbol{u}\|_1$. *The minimum value of* $J(\boldsymbol{u})$ *subject to* $H(\boldsymbol{u}) = 0$ *is attained provided* $\boldsymbol{D}\boldsymbol{u} \neq \boldsymbol{0}$ *whenever* $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{0}$. *When the minimum is attained and the Bregman iterates* $\boldsymbol{u}_k$ *remain bounded, every cluster point of the iterates* $\boldsymbol{u}_k$ *achieves the minimum.*

Proof: See the foregoing comments. ■

For the basis pursuit problem, it is known that Bregman iteration converges in a finite number of steps [279]. Problems 15 and 16 sketch a proof of this important fact.

16.8 Problems

1. Suppose that the real-valued function $f(\boldsymbol{x})$ is twice differentiable and that $b = \sup_{\boldsymbol{z}} \|d^2 f(\boldsymbol{z})\|$ is finite. Prove the Lipschitz inequality

$$\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| \quad \leq \quad b\|\boldsymbol{y} - \boldsymbol{x}\|$$

for all $\boldsymbol{x}$ and $\boldsymbol{y}$. If $f(\boldsymbol{x})$ satisfies the Lipschitz inequality, then prove that

$$g(\boldsymbol{x} \mid \boldsymbol{x}_m) \quad = \quad f(\boldsymbol{x}_m) + df(\boldsymbol{x}_m)(\boldsymbol{x} - \boldsymbol{x}_m) + \frac{b}{2}\|\boldsymbol{x} - \boldsymbol{x}_m\|^2$$

majorizes $f(\boldsymbol{x})$. (Hint: Expand $f(\boldsymbol{x})$ to first order around $\boldsymbol{x}_m$. Rearrange the integral remainder and bound.)

2. As described in Problem 1, assume the gradient of $f(\boldsymbol{x})$ is Lipschitz. Consider the projected gradient algorithm

$$\boldsymbol{x}_{m+1} \quad = \quad P_S\left[\boldsymbol{x}_m - \frac{\rho}{b}\nabla f(\boldsymbol{x}_m)\right]$$

for $\rho \in (0,2)$. From the obtuse angle criterion deduce

$$df(\boldsymbol{x}_m)(\boldsymbol{x}_{m+1} - \boldsymbol{x}_m) \quad \leq \quad -\frac{b}{\rho}\|\boldsymbol{x}_{m+1} - \boldsymbol{x}_m\|^2.$$

Add this inequality to the majorizing inequality and further deduce

$$f(\boldsymbol{x}_{m+1}) \quad \leq \quad f(\boldsymbol{x}_m) - \left[\frac{b}{\rho} - \frac{b}{2}\right]\|\boldsymbol{x}_{m+1} - \boldsymbol{x}_m\|^2.$$

It follows that the sequence $f(\boldsymbol{x}_m)$ is decreasing. If $f(\boldsymbol{x})$ is coercive or S is compact, then argue that $\lim_{m\to\infty} f(\boldsymbol{x}_m)$ exists and that $\lim_{m\to\infty} \|\boldsymbol{x}_{m+1} - \boldsymbol{x}_m\| = 0$. If we suppose $\boldsymbol{y}$ is a cluster point of the sequence $\boldsymbol{x}_m$, then the second of these limits shows that

$$P_S\left[\boldsymbol{y} - \frac{\rho}{b}\nabla f(\boldsymbol{y})\right] \quad = \quad \boldsymbol{y}.$$

Apply the obtuse angle criterion to any $\boldsymbol{z} \in S$, and deduce the necessary condition $df(\boldsymbol{y})(\boldsymbol{z} - \boldsymbol{y}) \geq 0$ for optimality. When $f(\boldsymbol{x})$ is also convex, conclude that $\boldsymbol{y}$ provides a global minimum. If $f(\boldsymbol{x})$ is strictly convex, then $\lim_{m\to\infty} \boldsymbol{x}_m$ exists and furnishes the unique global minimum point.

3. Prove that the function $\mathcal{E}_\rho(\boldsymbol{y})$ appearing in the exact penalty method is convex whenever $f(\boldsymbol{y})$ and the inequality constraints $h_j(\boldsymbol{y})$ are convex and the equality constraints $g_i(\boldsymbol{y})$ are affine.

4. In the exact penalty method for projection onto the half-disc, show that the solution path initially: (a) heads toward the origin when $\boldsymbol{x} \in \{\boldsymbol{x} : \|\boldsymbol{x}\|^2 > 1, x_1 > 0\}$ or $\boldsymbol{x} \in \{\boldsymbol{x} : |x_2| > 1, x_1 = 0\}$, (b) heads toward the point $(1,0)^*$ when $\boldsymbol{x} \in \{\boldsymbol{x} : \|\boldsymbol{x}\|^2 > 1, x_1 < 0\}$, (c) follows the unit circle when $\boldsymbol{x} \in \{\boldsymbol{x} : \|\boldsymbol{x}\|^2 = 1, x_1 < 0\}$, and (d) heads toward the x_2-axis when $\boldsymbol{x} \in \{\boldsymbol{x} : \|\boldsymbol{x}\|^2 < 1, x_1 < 0\}$. (Hint: See the left panel of Fig. 16.2.)

5. Path following can be conducted without invoking the exact penalty method [31, 46]. Consider minimizing the convex differentiable function $f(\boldsymbol{x})$ subject to the standard linear programming constraints $\boldsymbol{x} \geq \boldsymbol{0}$ and $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Given an initial feasible point $\boldsymbol{x}_0 > \boldsymbol{0}$, one can devise a differential equation $\frac{d}{dt}x(t) = G(\boldsymbol{x})$ whose solution $x(t)$ is likely

to converge to the optimal point. Simply take $G(\boldsymbol{x}) = D(\boldsymbol{x})P(\boldsymbol{x})v(\boldsymbol{x})$, where $D(\boldsymbol{x}) = \operatorname{diag}(\boldsymbol{x})$ is a diagonal matrix with diagonal entries given by the vector $\boldsymbol{x}$ and $P(\boldsymbol{x})$ is orthogonal projection onto the null space of $\boldsymbol{A}D(\boldsymbol{x})$. The matrix $D(\boldsymbol{x})$ slows the trajectory down as it approaches a boundary $x_i = 0$. The matrix $P(\boldsymbol{x})$ ensures that the value of $\boldsymbol{A}x(t)$ remains fixed at the constant $\boldsymbol{b}$. Check this fact. Show that the choice $v(\boldsymbol{x}) = -Q(\boldsymbol{x})P(\boldsymbol{x})D(\boldsymbol{x})\nabla f(\boldsymbol{x})$ for $Q(\boldsymbol{x})$ positive semidefinite yields $\frac{d}{dt}f[x(t)] \leq 0$. In other words, $f(\boldsymbol{x})$ is a Liapunov function for the solution path. For an analogue of steepest descent useful in linear programming, the choice $Q(\boldsymbol{x}) = \boldsymbol{I}$ is obvious. For an analogue of Newton's method, the choice $Q(\boldsymbol{x}) = d^2 f(\boldsymbol{x})^{-1}$ has merit. In a statistical setting where $-f(\boldsymbol{x})$ is a loglikelihood, $\boldsymbol{x}$ is a parameter vector, and $-\nabla f(\boldsymbol{x})$ is the score, substitution of the expected information matrix for the observed information matrix is also reasonable.

6. Implement the path following algorithm of Problem 5 for linear programming, linear regression, or linear logistic regression. You may use the differential equation solver of Matlab or Euler's method. Cheney [46] mentions some tactics for linear programming that ease the computational burden and make the overall algorithm more stable.

7. Homotopy methods follow solution paths. Suppose you want to solve the equation $f(\boldsymbol{x}) = \boldsymbol{0}$. Choose any point $\boldsymbol{x}_0$ and define the homotopy

$$h(t, \boldsymbol{x}) \;=\; tf(\boldsymbol{x}) + (1-t)[f(\boldsymbol{x}) - f(\boldsymbol{x}_0)] \;=\; f(\boldsymbol{x}) + (t-1)f(\boldsymbol{x}_0).$$

Note that $h(0, \boldsymbol{x}) = f(\boldsymbol{x}) - f(\boldsymbol{x}_0)$ and $h(1, \boldsymbol{x}) = f(\boldsymbol{x})$. Furthermore, $h(0, \boldsymbol{x}_0) = \boldsymbol{0}$. Path following starts at time $t = 0$ and $\boldsymbol{x} = \boldsymbol{x}_0$. Apply the implicit function theorem to the equation $h(t, \boldsymbol{x}) = \boldsymbol{0}$, and demonstrate that a continuously differentiable solution path $x(t)$ exists satisfying the differential equation

$$\frac{d}{dt}x(t) \;=\; -df[x(t)]^{-1} f(\boldsymbol{x}_0). \tag{16.19}$$

8. Continuing Problem 7, suppose

$$f(\boldsymbol{x}) \;=\; \begin{pmatrix} \sin x_1 + e^{x_2} - 3 \\ (x_2+3)^2 - x_1 - 4 \end{pmatrix}.$$

Numerically integrate the differential equation (16.19) starting at the point $\boldsymbol{x}_0 = (5, 3)^*$. You should have an approximate zero at the time $t = 1$. If necessary, polish this approximate solution by Newton's method. The purpose of path following is to reliably reach a neighborhood of a solution.

9. Show that the four functions $J_1(\boldsymbol{u}) = \|\boldsymbol{u}\|^2$, $J_2(\boldsymbol{u}) = -\sum_i \log u_i$, $J_3(\boldsymbol{u}) = \sum_i u_i \ln u_i$, and $J_4(\boldsymbol{u} \mid \boldsymbol{v}) = \boldsymbol{u}^* \boldsymbol{A} \boldsymbol{u}$ generate the Bregman distances

$$\begin{aligned} D_{J_1}(\boldsymbol{u} \mid \boldsymbol{v}) &= \|\boldsymbol{u} - \boldsymbol{v}\|^2 \\ D_{J_2}(\boldsymbol{u} \mid \boldsymbol{v}) &= \sum_i \left[\frac{u_i}{v_i} - \log\left(\frac{u_i}{v_i}\right) - 1 \right] \\ D_{J_3}(\boldsymbol{u} \mid \boldsymbol{v}) &= \sum_i u_i \ln\left(\frac{u_i}{v_i}\right) - \sum_i (u_i - v_i) \\ D_{J_4}(\boldsymbol{u} \mid \boldsymbol{v}) &= (\boldsymbol{u} - \boldsymbol{v})^* \boldsymbol{A} (\boldsymbol{u} - \boldsymbol{v}). \end{aligned}$$

   For $J_4(\boldsymbol{u})$ assume $\boldsymbol{A}$ is positive semidefinite.

10. Prove the generalized Pythagorean identity (16.15).

11. The Bregman function $D_J^{\boldsymbol{p}}(\boldsymbol{u} \mid \boldsymbol{v})$ has some of the properties of a distance. Show that it is 0 when $\boldsymbol{u} = \boldsymbol{v}$ and nonnegative when $\boldsymbol{u} \neq \boldsymbol{v}$. If $J(\boldsymbol{u})$ is strictly convex, then show that $D_J^{\boldsymbol{p}}(\boldsymbol{u} \mid \boldsymbol{v})$ is positive when $\boldsymbol{u} \neq \boldsymbol{v}$. Also prove that $D_J^{\boldsymbol{p}}(\boldsymbol{u} \mid \boldsymbol{v})$ is convex in its argument $\boldsymbol{u}$. Finally, prove that $D_J^{\boldsymbol{p}}(\boldsymbol{u} \mid \boldsymbol{v}) \geq D_J^{\boldsymbol{p}}(\boldsymbol{w} \mid \boldsymbol{v})$ when $\boldsymbol{w}$ lies on the line segment between $\boldsymbol{u}$ and $\boldsymbol{v}$.

12. For differentiable functions, demonstrate the Bregman identities

$$\begin{aligned} D_{cJ}(\boldsymbol{u} \mid \boldsymbol{v}) &= c D_J(\boldsymbol{u} \mid \boldsymbol{v}) \quad \text{for} \quad c \geq 0 \\ D_{J_1+J_2}(\boldsymbol{u} \mid \boldsymbol{v}) &= D_{J_1}(\boldsymbol{u} \mid \boldsymbol{v}) + D_{J_2}(\boldsymbol{u} \mid \boldsymbol{v}) \\ D_J(\boldsymbol{u} \mid \boldsymbol{v}) &= 0 \quad \text{for} \quad J(\boldsymbol{u}) \text{ affine.} \end{aligned}$$

13. Suppose X is a random variable with mean μ and $f(x)$ is a convex function defined on R. Prove Jensen's inequality in the form

$$\mathrm{E}[f(X)] - f(\mu) \quad = \quad \mathrm{E}[D_f^p(X \mid \mu)] \quad \geq \quad 0,$$

   where p is any subgradient of $f(x)$ at μ.

14. Suppose the convex function $f(\boldsymbol{x})$ and its Fenchel conjugate $f^\star(\boldsymbol{y})$ are both differentiable. Let $\boldsymbol{u} = \nabla f^\star(\boldsymbol{u}^\star)$ and $\boldsymbol{v}^\star = \nabla f(\boldsymbol{v})$. Prove the duality result

$$D_f(\boldsymbol{u} \mid \boldsymbol{v}) \quad = \quad D_{f^\star}(\boldsymbol{v}^\star \mid \boldsymbol{u}^\star).$$

   (Hint: Recall when equality holds in the Fenchel-Young inequality.)

15. Assume the Bregman iterate $\boldsymbol{u}_k$ satisfies $H(\boldsymbol{u}_k) = \frac{\lambda}{2} \|\boldsymbol{A}\boldsymbol{u}_k - \boldsymbol{f}\|^2 = 0$. Prove that $\boldsymbol{u}_k$ also minimizes $J(\boldsymbol{u})$ subject to $H(\boldsymbol{u}) = 0$. (Hint: Let $\hat{\boldsymbol{u}}$ be optimal for $J(\boldsymbol{u})$ given the constraint $H(\boldsymbol{u}) = 0$. Note that $J(\boldsymbol{u}_k) \leq J(\hat{\boldsymbol{u}}) - \boldsymbol{p}_k^*(\hat{\boldsymbol{u}} - \boldsymbol{u}_k)$ and $\boldsymbol{p}_k$ belongs to the range of $\boldsymbol{A}^*$.)

16. For the basis pursuit problem, let (I_+^j, I_-^j, E^j) be a partition of the index set $\{1, 2, \ldots, n\}$, and define

$$\begin{aligned} U^j &= \{\boldsymbol{u} \in \mathsf{R}^n : u_i \geq 0, i \in I_+^j; u_i \leq 0, i \in I_-^j; u_i = 0, i \in E^j\} \\ H^j &= \inf\left\{\frac{1}{2}\|\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f}\|^2 : \boldsymbol{u} \in U^j\right\}. \end{aligned}$$

Show that there are a finite number of distinct partitions U^j and that their union equals R^n. At iteration k define

$$\begin{aligned} I_+^k &= \{i : p_{ki} = 1\} \\ I_-^k &= \{i : p_{ki} = -1\} \\ E^k &= \{i : p_{ki} \in (-1, 1)\}. \end{aligned}$$

Demonstrate that $\boldsymbol{p}_k \in \partial J(\boldsymbol{u}_k)$ implies $\boldsymbol{u}_k \in U^k$ and that $\boldsymbol{u}_k \in U^j$ with $H^j > 0$ can happen for only finitely many k. Hence, for some k we have $\boldsymbol{u}_k \in U^j$ with $H^j = 0$. Show that this entails $\boldsymbol{A}\boldsymbol{u}_{k+1} = \boldsymbol{f}$. Now invoke problem 15 to prove that Bregman iteration converges in a finite number of steps.

17. In Bregman iteration suppose $H(\boldsymbol{u})$ achieves its minimum of 0 at the point $\hat{\boldsymbol{u}}$. If $H(\boldsymbol{u}_{k-1}) > 0$, then prove that

$$D_J^{\boldsymbol{p}_k}(\hat{\boldsymbol{u}} \mid \boldsymbol{u}_k) \quad < \quad D_J^{\boldsymbol{p}_{k-1}}(\hat{\boldsymbol{u}} \mid \boldsymbol{u}_{k-1}).$$

(Hint: Consider inequality (16.16).)

18. Suppose the convex functions $H(\boldsymbol{u})$ and $J(\boldsymbol{u})$ in our discussion of Bregman iteration are differentiable. Prove that the surrogate function $G(\boldsymbol{u} \mid \boldsymbol{u}_k) = H(\boldsymbol{u}) + D_J(\boldsymbol{u} \mid \boldsymbol{u}_k)$ satisfies

$$G(\boldsymbol{u} \mid \boldsymbol{u}_{k-1}) \quad = \quad G(\boldsymbol{u}_k \mid \boldsymbol{u}_{k-1}) + D_H(\boldsymbol{u} \mid \boldsymbol{u}_k) + D_J(\boldsymbol{u} \mid \boldsymbol{u}_k).$$

(Hint: $\boldsymbol{0} = \nabla H(\boldsymbol{u}_k) + \nabla J(\boldsymbol{u}_k) - \nabla J(\boldsymbol{u}_{k-1})$.)

17
The Calculus of Variations

17.1 Introduction

The calculus of variations deals with infinite dimensional optimization problems. Seventeenth century mathematicians and physicists such as Newton, Galileo, Huygens, John and James Bernoulli, L'Hôpital, and Leibniz posed and solved many variational problems. In the eighteenth century Euler made more definitive strides that were clarified and extended by Lagrange. In the nineteenth and twentieth centuries, the intellectual stimulus offered by the calculus of variations was instrumental in the development of functional analysis and control theory. Some of this rich history is explored in the books [240, 258].

The current chapter surveys the classical and more elementary parts of the calculus of variations. The subject matter is optimization of functionals such as

$$F(\boldsymbol{x}) \quad = \quad \int_a^b f[t, x(t), x'(t)]\, dt \tag{17.1}$$

depending on a continuously differentiable function $x(t)$ over an interval $[a, b]$. Euler and Lagrange were able to deduce that the solution satisfies the differential equation

$$\frac{d}{dt}\frac{\partial}{\partial x'} f[t, x(t), x'(t)] \quad = \quad \frac{\partial}{\partial x} f[t, x(t), x'(t)]. \tag{17.2}$$

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8_17,

If constraints are imposed on the solution, then these constraints enter the Euler-Lagrange equation via multipliers. Thus, the theory parallels the finite-dimensional case.

However, as one might expect, the theory is harder. Proof strategies based on compactness often fail while strategies based on convexity usually succeed. Much of the theory involving differentials fortunately generalizes. We tackle this theory for normed vector spaces. Most other introductions to the calculus of variations substitute a weaker version of differentiation that relies entirely on directional derivatives. In our view this forfeits the chance to bridge the gap between advanced calculus and functional analysis. Having expended so much energy on developing Carathéodory's version of the differential, we continue to pursue that definition here. Readers interested in the more traditional perspective can consult the references [46, 102, 228, 233, 240]

17.2 Classical Problems

As motivating examples, we briefly discuss some of the classical problems of the calculus of variations. Finding a solution to one of these problems is often helped by a judicious choice of a coordinate system.

Example 17.2.1 *Geodesics*

A geodesic is the shortest path between two points. In the absence of constraints, a geodesic is a straight line. Suppose the points in question are $\boldsymbol{p} = \mathbf{0}$ and $\boldsymbol{q}$ in R^n. A path is a differentiable curve $\boldsymbol{x}$ from $[0,1]$ starting at $\mathbf{0}$ and ending at $\boldsymbol{q}$. To prove that the optimal path is a straight line, we must show that $y(t) = t\boldsymbol{q}$ minimizes the functional

$$G(\boldsymbol{x}) \;=\; \int_0^1 \|x'(t)\|\,dt.$$

But this is obvious from the inequality

$$\|\boldsymbol{q}\| \;=\; \left\|\int_0^1 x'(t)\,dt\right\| \;\leq\; \int_0^1 \|x'(t)\|\,dt.$$

A somewhat harder problem is to show that a geodesic on a sphere follows a great circle. If the sphere has radius r, then a feasible path $x(t)$ must satisfy $\|x(t)\| = r$ for all t. It is convenient to pass to spherical coordinates and assume that the sphere resides in ordinary space R^3 with its center at the origin. It is also convenient to take the initial point $\boldsymbol{p}$ as the north pole and parameterize the azimuthal angle $\theta(\phi)$ of a path by the polar angle ϕ, where $\phi \in [0,\pi]$ and $\theta \in [0,2\pi]$. The path $y(\phi) = (r,\theta(\phi),\phi)^*$

in spherical coordinates automatically satisfies the radial constraint. The usual arguments from elementary calculus show that $y(\phi)$ has arclength

$$G(\boldsymbol{y}) \quad = \quad r \int_0^{\phi_1} \sqrt{1 + [\theta'(\phi) \sin \phi]^2}\, d\phi$$

between the north pole and $\boldsymbol{q} = (r, \theta_1, \phi_1)$. It is clear that $G(\boldsymbol{y})$ is minimized by taking $\theta(\phi)$ equal to the constant θ_1. This implies that the solution is an arc of a great circle. ■

Example 17.2.2 *Minimal Surface Area of Revolution*

In the plane R^2, imagine rotating a curve $y(x)$ about the x-axis. This generates a surface of revolution with area

$$S(\boldsymbol{y}) \quad = \quad 2\pi \int_{x_0}^{x_1} y(x) \sqrt{1 + y'(x)^2}\, dx.$$

Here the curve begins at $y(x_0) = \boldsymbol{y}_0$ and ends at $y(x_1) = \boldsymbol{y}_1$. If it is possible to pass a catenary curve through these points, then it describes the surface with minimum area. The calculus of variations offers an easy route to this conclusion. ■

Example 17.2.3 *Passage of Light through an Inhomogeneous Medium*

If we look at an object close to the horizon but well above the earth's surface, the light from it will bend as it passes through the atmosphere. This is a consequence of the fact that the speed of light decreases as it passes through an increasingly dense medium. If we assume the earth is flat and the speed of light $v(y)$ varies with the distance y above the earth, then the ray will take the path $y(x)$ of least time. The total travel time is

$$T(\boldsymbol{y}) \quad = \quad \int_{x_0}^{x_1} \frac{\sqrt{1 + y'(x)^2}}{v[y(x)]}\, dx$$

when the source is situated at (x_0, y_0) and we are situated at (x_1, y_1). The calculus of variations provides theoretical insight into this generalization of Snell's problem. ■

Example 17.2.4 *Lagrange's versus Newton's Version of Mechanics*

The calculus of variations offers an alternative approach to classical mechanics. For a particle with kinetic energy $T(\boldsymbol{v}) = \frac{1}{2} m \|\boldsymbol{v}\|^2$ in a conservative force field with potential $U(x)$, we define the action integral

$$A(\boldsymbol{x}) \quad = \quad \int_{t_0}^{t_1} \{T[x'(t)] - U[x(t)]\} dt \tag{17.3}$$

on the path $x(t)$ of the particle from time t_0 to time t_1. The path actually taken furnishes a stationary value of $A(\boldsymbol{x})$. One can demonstrate this fact by showing that the Euler-Lagrange equations coincide with Newton's equations of motion. ■

Example 17.2.5 *Isoperimetric Problem*

This classical Greek problem involves finding the plane curve of given length ℓ enclosing the greatest area. The circle of perimeter ℓ is the obvious solution. If we let a horizontal line segment form one side of the figure, then the solution is a circular arc. This version of the problem can be formalized by writing the enclosed area as

$$A(\boldsymbol{y}) = \int_{x_0}^{x_1} y(x)\,dx$$

and its length as

$$L(\boldsymbol{y}) = \int_{x_0}^{x_1} \sqrt{1+y'(x)^2}\,dx.$$

The constrained problem of minimizing $A(\boldsymbol{y})$ subject to $L(\boldsymbol{y}) = \ell$ and $y(x_0) = y(x_1) = 0$ can be solved by introducing a Lagrange multiplier. ■

Example 17.2.6 *Splines*

A spline is a smooth curve interpolating a given function at specified points. From the variational perspective, we would like to find the function $x(t)$ minimizing the curvature

$$C(\boldsymbol{x}) = \int_a^b x''(t)^2 dt \tag{17.4}$$

subject to the constraints $x(s_i) = x_i$ at $n+1$ points in the interval $[a, b]$. In this situation the solution has limited smoothness. The interpolation points are called nodes and are numbered so that $s_0 = a < s_1 < \cdots < s_n = b$. ■

17.3 Normed Vector Spaces

In the calculus of variations, functions are viewed as vectors belonging to normed vector spaces of infinite dimension. The vector space and attached norm vary from problem to problem. Many of the concepts from finite-dimensional linear algebra extend without comment to the infinite-dimensional setting. The new concepts that crop up are fairly subtle, so it is worth spending some time on concrete examples.

For instance, consider the collection $\mathcal{C}^0[a, b]$ of continuous real-valued functions defined on a compact interval $[a, b]$. It is clear that $\mathcal{C}^0[a, b]$ is closed

under addition and scalar multiplication and that the function $x(t) \equiv 0$ serves as the zero vector or origin. The choice of norm is less obvious. Three natural possibilities are:

$$\begin{aligned} \|\boldsymbol{x}\|_\infty &= \sup_{t\in[a,b]} |x(t)| \\ \|\boldsymbol{x}\|_1 &= \int_a^b |x(t)|dt \\ \|\boldsymbol{x}\|_2 &= \left[\int_a^b |x(t)|^2 dt\right]^{1/2}. \end{aligned}$$

It is straightforward to check that $\|\boldsymbol{x}\|_\infty$ and $\|\boldsymbol{x}\|_1$ satisfy the requirements of a norm as set down in Chap. 2. To prove that $\|\boldsymbol{x}\|_2$ qualifies as a norm, it is best to define it in terms of the inner product

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \int_a^b x(t)y(t)dt$$

as

$$\|\boldsymbol{x}\|_2 = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x}\rangle}$$

and invoke the Cauchy-Schwarz inequality.

In contrast to finite-dimensional spaces, not all norms are equivalent on infinite-dimensional spaces. Consider the sequence of continuous functions

$$x_n(t) = \begin{cases} 1 - nt & t \in [0, 1/n] \\ 0 & t \notin [0, 1/n] \end{cases}$$

on the unit interval. It is clear that $\|\boldsymbol{x}_n\|_\infty = 1$ for all n while

$$\begin{aligned} \|\boldsymbol{x}_n\|_1 &= \int_0^{1/n} (1 - nt)\, dt = \frac{1}{2n} \\ \|\boldsymbol{x}_n\|_2^2 &= \int_0^{1/n} (1 - nt)^2\, dt = \frac{1}{3n}. \end{aligned}$$

Thus, $x_n(t)$ converges to the origin in two of these norms but not in the third.

Besides continuous functions, it is often useful to consider continuously differentiable functions. The vector space of such functions over the interval $[a, b]$ is denoted by $\mathcal{C}^1[a, b]$. The norm

$$\|\boldsymbol{x}\|_{\infty 1} = \sup_{t\in[a,b]} |x(t)| + \sup_{t\in[a,b]} |x'(t)| \tag{17.5}$$

is fairly natural when one is interested in uniform convergence of a sequence $x_n(t)$ together with its derivatives $x_n'(t)$. On the vector space $\mathcal{C}^m[a,b]$ of functions with m continuous derivatives, one can define the similar norm

$$\|\boldsymbol{x}\|_{\infty m} = \sum_{k=0}^{m} \sup_{t\in[a,b]} |x^{(k)}(t)|. \tag{17.6}$$

In most applications, the completeness of a normed vector space is an issue. A sequence $x_n(t)$ is said to be Cauchy if for every $\epsilon > 0$ there exists an integer n such that $\|\boldsymbol{x}_j - \boldsymbol{x}_k\| < \epsilon$ whenever $j \geq n$ and $k \geq n$. A normed vector space is complete if every Cauchy sequence possesses a limit in the space. For example, the vector space $\mathcal{C}^0[a,b]$ is complete under the uniform norm $\|\boldsymbol{x}\|_\infty$. For a proof of this fact, observe that

$$|x_j(t) - x_k(t)| \leq \|\boldsymbol{x}_j - \boldsymbol{x}_k\|_\infty$$

for every $t \in [a,b]$. This implies that the sequence $x_n(t)$ is Cauchy on the real line and possesses a limit $x(t)$. Since the convergence to $x(t)$ is uniform in t, Proposition 2.8.1 can be invoked to finish the proof.

The space $\mathcal{C}^0[a,b]$ is not complete under either of the norms $\|\boldsymbol{x}\|_1$ or $\|\boldsymbol{x}\|_2$. It is possible to extend $\mathcal{C}^0[a,b]$ to larger normed vector spaces $\mathrm{L}_1[a,b]$ and $\mathrm{L}_2[a,b]$ that are complete under these norms. The process of completion is one of the most fascinating parts of the theory of integration. Unfortunately, the work involved is far more than we can undertake here. Complete normed vector spaces are called Banach spaces; complete inner product spaces are called Hilbert spaces.

The space $\mathcal{C}^1[a,b]$ under the norm (17.5) is also complete. A little thought makes it clear that a Cauchy sequence $x_n(t)$ in $\mathcal{C}^1[a,b]$ not only converges uniformly to a continuous function $x(t)$, but its sequence of derivatives $x_n'(t)$ also converges uniformly to a continuous function $y(t)$. Applying the dominated convergence theorem to the sequence

$$x_n(t) = \int_a^t x_n'(s)ds$$

shows that

$$x(t) = \int_a^t y(s)ds.$$

It now follows from the fundamental theorem of calculus that $x'(t)$ exists and equals $y(t)$. A slight extension of this argument demonstrates that $\mathcal{C}^m[a,b]$ is complete under the norm (17.6). For this reason, we will tacitly assume in the sequel that $\mathcal{C}^m[a,b]$ is equipped with the $\|\boldsymbol{x}\|_{\infty m}$ norm.

17.4 Linear Operators and Functionals

Linear algebra focuses on linear maps and their matrix representations. In infinite-dimensional spaces, linear maps are referred to as linear operators. When the range of a linear operator is the real line, the operator is said to be a linear functional. Unfortunately, linear operators on normed vector spaces are no longer automatically continuous. Continuity is intimately tied to boundedness. A linear operator A from a normed linear space X with norm $\|\boldsymbol{x}\|_p$ to a normed linear space Y with norm $\|\boldsymbol{y}\|_q$ is bounded if there exists a constant c such that $\|A(\boldsymbol{x})\|_q \leq c\|\boldsymbol{x}\|_p$ for all $\boldsymbol{x} \in X$. The least such constant c determines the induced operator norm $\|A\|$. This verbal description just recapitulates the definition given in equation (2.2) of Chap. 2. For linear functionals, mathematicians invariably use the norm $\|y\|_q = |y|$ derived from the absolute value function. It is trivial to check that the collection of bounded linear operators between two normed vector spaces is closed under pointwise addition and scalar multiplication. Thus, this collection is a normed vector space in its own right.

Here are three typical bounded linear functionals:

$$A_1(\boldsymbol{x}) = x(d), \quad A_2(\boldsymbol{x}) = \int_a^b x(t)dt, \quad A_3(\boldsymbol{x}) = x'(d). \tag{17.7}$$

The first two of these are defined on the space $\mathcal{C}^0[a,b]$ and the third on the space $\mathcal{C}^1[a,b]$. The evaluation point d can be any point from $[a,b]$. Straightforward arguments show that the induced norms satisfy the inequalities $\|A_1\| \leq 1$, $\|A_2\| \leq b-a$, and $\|A_3\| \leq 1$.

Bounded linear operators are a little more exotic. If $y(t)$ is a monotone function mapping $[a,b]$ into itself, then the linear operator

$$A_4(\boldsymbol{x}) = \boldsymbol{x} \circ \boldsymbol{y}$$

composing $\boldsymbol{x}$ and $\boldsymbol{y}$ maps $\mathcal{C}^0[a,b]$ into itself. For example, if $[a,b] = [0,1]$ and $y(t) = t^2$, then $A_4(\boldsymbol{x})(t) = x(t^2)$. The operator A_4 has norm $\|A_4\| \leq 1$. Because integration turns one continuous function into another, the linear operator

$$A_5(\boldsymbol{x})(s) = \int_a^s x(t)dt$$

also maps $\mathcal{C}^0[a,b]$ into itself. This operator has norm $\|A_5\| \leq b-a$. It is a special case of the linear operator

$$A_6(\boldsymbol{x})(s) = \int_a^b K(s,t)x(t)dt \tag{17.8}$$

defined by a bounded function $K(s,t)$ with domain the square $[a,b] \times [a,b]$. If $|K(s,t)| \leq c$ for all s and t, then the inequality

$$\begin{aligned}\left|\int_a^b K(s,t)x(t)dt\right| &\leq \int_a^b |K(s,t)||x(t)|dt \\ &\leq c\|x\|_\infty(b-a),\end{aligned}$$

shows that $\|A_6\| \leq c(b-a)$.

For a final example, let us define the injection operator $\text{Inj}_1(x)$ by the formula

$$\text{Inj}_1(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{x}' \end{pmatrix}.$$

This is a linear operator from the normed vector space $\mathcal{C}^1[a,b]$ to the vector space $\mathcal{C}_2^0[a,b]$ of continuous vector-valued functions with two components. It is easy to check that

$$\left\|\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}\right\| = \|\boldsymbol{x}\|_\infty + \|\boldsymbol{y}\|_\infty$$

defines a norm on $\mathcal{C}_2^0[a,b]$. Furthermore, under this norm and the standard norm $\|\cdot\|_{\infty 1}$ on $\mathcal{C}^1[a,b]$, the linear operator $\text{Inj}_1(x)$ has induced operator norm 1. ■

The next proposition clarifies the relationship between boundedness and continuity.

Proposition 17.4.1 *The following three assertions concerning a linear operator A are equivalent:*

(a) A is continuous,

(b) A is continuous at the origin $\mathbf{0}$,

(c) A is bounded.

Proof: Assertion (a) clearly implies assertion (b). Let the domain of A have norm $\|\cdot\|_p$ and the range norm $\|\cdot\|_q$. Assertion (c) implies assertion (a) because of the inequality

$$\|A(\boldsymbol{y}) - A(\boldsymbol{x})\|_q = \|A(\boldsymbol{y}-\boldsymbol{x})\|_q \leq \|A\| \cdot \|\boldsymbol{y}-\boldsymbol{x}\|_p.$$

To complete the proof, we must show that assertion (b) implies assertion (c). If A is unbounded, then there exists a sequence $\boldsymbol{x}_n \neq \mathbf{0}$ with

$$\|A(\boldsymbol{x}_n)\|_q \geq n\|\boldsymbol{x}_n\|_p.$$

If we set

$$\boldsymbol{y}_n = \frac{1}{n\|\boldsymbol{x}_n\|_p} x_n,$$

then $\boldsymbol{y}_n$ converges to $\mathbf{0}$, but $\|A(\boldsymbol{y}_n)\|_q \geq 1$ does not converge to 0. ■

17.5 Differentials

Our approach to differentiation is to replace slope matrices by slope operators. Let $F(\boldsymbol{y})$ be a nonlinear operator from a normed vector space U with norm $\|\cdot\|_p$ to a normed vector space V with norm $\|\cdot\|_q$. The right-hand side of the slope equation

$$F(\boldsymbol{y}) - F(\boldsymbol{x}) \quad = \quad S(\boldsymbol{y}, \boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \tag{17.9}$$

now involves a bounded linear operator $S(\boldsymbol{y}, \boldsymbol{x})$ from U to V operating on the vector $\boldsymbol{y} - \boldsymbol{x}$ in U. The operator $S(\boldsymbol{y}, \boldsymbol{x})$ has induced norm

$$\|S(\boldsymbol{y}, \boldsymbol{x})\| \quad = \quad \sup_{\|\boldsymbol{u}\|_p = 1} \|S(\boldsymbol{y}, \boldsymbol{x})\boldsymbol{u}\|_q.$$

The operator $F(\boldsymbol{y})$ is said to have differential $dF(\boldsymbol{x})$ at $\boldsymbol{x}$ provided the slope equation (17.9) holds for all $\boldsymbol{y}$ sufficiently close to $\boldsymbol{x}$ and

$$\lim_{y \to x} \|S(\boldsymbol{y}, \boldsymbol{x}) - dF(\boldsymbol{x})\| \quad = \quad 0.$$

This last equation implicitly requires $dF(\boldsymbol{x})$ to be a bounded linear operator from U to V. Furthermore, the affine map $\boldsymbol{y} \mapsto F(\boldsymbol{x}) + dF(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})$ uniformly approximates $F(\boldsymbol{y})$ in the sense that

$$\begin{aligned} \|F(\boldsymbol{y}) - F(\boldsymbol{x}) - dF(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})\|_q &= \|[S(\boldsymbol{y}, \boldsymbol{x}) - dF(\boldsymbol{x})](\boldsymbol{y} - \boldsymbol{x})\|_q \\ &\leq \|S(\boldsymbol{y}, \boldsymbol{x}) - dF(\boldsymbol{x})\| \cdot \|\boldsymbol{y} - \boldsymbol{x}\|_p \end{aligned}$$

for $\boldsymbol{y}$ close to $\boldsymbol{x}$. Thus, we arrive at an appropriate extension of Carathéodory's definition of the differential. It is also possible to define Fréchet's differential in this setting. As Proposition 4.4.1 of Chap. 4 shows, the two definitions are equivalent on inner product spaces. On more general normed vector spaces, it is unclear whether a Fréchet differentiable function is necessarily Carathéodory differentiable.

The various rules for combining differentials remain in force, and versions of the inverse and implicit function theorems continue to hold. The proofs of these theorems must be modified to avoid an appeal to compactness [46]. Rather than concentrate on extending our earlier proofs, we prefer to offer some concrete examples. The simplest example is a bounded linear operator $F(\boldsymbol{x})$. In this case, application of the definition shows that $dF(\boldsymbol{x})\boldsymbol{u} = F(\boldsymbol{u})$. Here are some more subtle examples.

Example 17.5.1 *The Squaring Operator*

Consider the operator $F(\boldsymbol{y}) = \boldsymbol{y}^2$ on $\mathcal{C}^0[a, b]$. The relation

$$F(\boldsymbol{y}) - F(\boldsymbol{x}) \quad = \quad (\boldsymbol{y} + \boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})$$

suggests that we take $S(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{y}+\boldsymbol{x}$. For this to make sense, we reinterpret each of the symbols $\boldsymbol{y}$ and $\boldsymbol{x}$ as multiplication by the function in question. Thus, the operator $\boldsymbol{y}$ takes $\boldsymbol{u} \in \mathcal{C}^0[a, b]$ to the function $y(t)u(t)$. The obvious bound $\|\boldsymbol{y}\boldsymbol{u}\|_\infty \leq \|\boldsymbol{y}\|_\infty \|\boldsymbol{u}\|_\infty$ says $\boldsymbol{y}$ viewed as a linear operator has induced norm $\|\boldsymbol{y}\| \leq \|\boldsymbol{y}\|_\infty$. Once we apply $\boldsymbol{y}$ to itself, then it becomes clear that $\|\boldsymbol{y}\| = \|\boldsymbol{y}\|_\infty$. Given these preliminaries, the identities

$$\|S(\boldsymbol{y}, \boldsymbol{x}) - 2\boldsymbol{x}\| \quad = \quad \|\boldsymbol{y} - \boldsymbol{x}\| \quad = \quad \|\boldsymbol{y} - \boldsymbol{x}\|_\infty$$

now demonstrate that $dF(\boldsymbol{x})$ exists and equals the multiplication operator $2\boldsymbol{x}$. ■

Example 17.5.2 *An Integration Functional*

Suppose the continuous function $f(t, x)$ has a continuous partial derivative $\partial_2 f(t, x) = \frac{\partial}{\partial x} f(t, x)$. If we fix t and choose the canonical slope function

$$s(t, y, x) \quad = \quad \int_0^1 \partial_2 f[t, x + s(y - x)]\, ds$$

of $f(t, x)$ as a function of x, then

$$f(t, y) - f(t, x) \quad = \quad s(t, y, x)(y - x).$$

Furthermore, this choice ensures that $s(t, y, x)$ is jointly continuous in its three arguments. Now consider the functional

$$F(\boldsymbol{x}) \quad = \quad \int_a^b f[t, x(t)]\, dt$$

on $\mathcal{C}^0[a, b]$. The equation

$$F(\boldsymbol{y}) - F(\boldsymbol{x}) \quad = \quad \int_a^b s[t, y(t), x(t)][y(t) - x(t)]\, dt$$

suggests the candidate differential

$$dF(\boldsymbol{x})\boldsymbol{u} \quad = \quad \int_a^b \partial_2 f[t, x(t)]u(t)\, dt. \tag{17.10}$$

To prove this contention, we first show that the linear functional

$$S(\boldsymbol{y}, \boldsymbol{x})\boldsymbol{u} \quad = \quad \int_a^b s[t, y(t), x(t)]u(t)\, dt$$

is bounded. Because the interval $[a, b]$ is compact, $x(t)$ is bounded. If we fix $\delta > 0$ and limit attention to those $\boldsymbol{y}$ with $\|\boldsymbol{y} - \boldsymbol{x}\|_\infty < \delta$, then all values of $x(t)$ and $y(t)$ occur within an interval $[c, d]$. For such $\boldsymbol{y}$ we have

$$\begin{aligned} \left| \int_a^b s[t, y(t), x(t)]u(t)\, dt \right| \quad &\leq \quad \int_a^b |s[t, y(t), x(t)]| \cdot |u(t)|\, dt \\ &\leq \quad k(b - a)\|\boldsymbol{u}\|_\infty, \end{aligned}$$

where k is the supremum of $|s(t,y,x)|$ on $J = [a,b]\times[c,d]\times[c,d]$. This settles the question of boundedness. Given the uniform continuity of $s(t,y,x)$ on J, if we choose $\|\boldsymbol{y}-\boldsymbol{x}\|_\infty$ small enough, then $|s[t,y(t),x(t)]-s[t,x(t),x(t)]|$ can be made uniformly smaller than a preassigned $\epsilon > 0$. For those $\boldsymbol{y}$ it follows that

$$\left|\int_a^b \{s[t,y(t),x(t)]-s[t,x(t),x(t)]\}u(t)\,dt\right| \quad \leq \quad \epsilon(b-a)\|\boldsymbol{u}\|_\infty,$$

and this implies that $S(\boldsymbol{y},\boldsymbol{x})$ converges to $S(\boldsymbol{x},\boldsymbol{x})$ in the relevant operator norm.

In the classical theory, equation (17.10) is viewed as a directional derivation. Fixing the "direction" $u(t)$ in function space, one calculates the directional derivative

$$\lim_{\epsilon\to 0}\frac{F(\boldsymbol{x}+\epsilon\boldsymbol{u})-F(\boldsymbol{x})}{\epsilon} \quad = \quad \int_a^b \partial_2 f[t,x(t)]u(t)\,dt$$

by differentiation under the integral sign and the chain rule. This is rigorous as far as it goes, but it does not prove the existence of the differential. Having the full apparatus of differentials at our disposal unifies the theory and eases the process of generalization. ■

Example 17.5.3 *Differentials of More General Functionals*

Many of the classical examples of the calculus of variations are slight elaborations of the last example. For instance, we can replace the argument $x(t)$ of $F(\boldsymbol{x})$ by a continuous vector-valued function on $[a,b]$. The differential (17.10) is still valid provided we interpret $\partial_2 f(t,\boldsymbol{x})$ as the differential of $f(t,\boldsymbol{x})$ with respect to $\boldsymbol{x}$ and assume $x(t)$ belongs to the normed vector space $\mathcal{C}_m^0[a,b]$ of continuous functions with m components for some m. Another profitable extension is to consider functionals of the form

$$G(\boldsymbol{x}) \quad = \quad \int_a^b f[t,x(t),x'(t)]\,dt$$

depending on $x'(t)$ as well as $x(t)$. Straightforward extension of our previous arguments yield

$$dG(\boldsymbol{x})\boldsymbol{u} \;=\; \int_a^b \{\partial_2 f[t,x(t),x'(t)]u(t)+\partial_3 f[t,x(t),x'(t)]u'(t)\}\,dt, \quad (17.11)$$

where $\partial_3 f(t,x,x') = \frac{\partial}{\partial x'}f(t,x,x')$. Now we must assume that both partial derivatives $\partial_2 f(t,x,x')$ and $\partial_3 f(t,x,x')$ are jointly continuous in all variables. Alternatively, we can derive this result by noting that $G(\boldsymbol{x})$ is the composition of the functional $F(\boldsymbol{x})$ of the last example with the injection operator $\mathrm{Inj}_1(\boldsymbol{x})$. The differential (17.11) then reduces to the chain rule

$$dG(\boldsymbol{x}) \quad = \quad dF[\mathrm{Inj}_1(\boldsymbol{x})]d\,\mathrm{Inj}_1(\boldsymbol{x}).$$

This argument remains valid when $x(t)$ is vector-valued provided we interpret the partial derivatives $\partial_2 f(t, \boldsymbol{x}, \boldsymbol{x}')$ and $\partial_3 f(t, \boldsymbol{x}, \boldsymbol{x}')$ as partial differentials. Even this formula can be generalized by considering functionals

$$G(\boldsymbol{x}) = \int_a^b f[t, x(t), x'(t), \ldots, x^{(k)}(t)]\, dt$$

depending on $x(t)$ and its first k derivatives. In this case, the formula

$$dG(\boldsymbol{x})\boldsymbol{u} = \int_a^b \sum_{j=0}^{k} \partial_{j+1} f[t, x(t), \ldots, x^{(k)}(t)] u^{(j)}(t)\, dt \quad (17.12)$$

just summarizes the chain rule

$$dG(\boldsymbol{x}) = dF[\mathrm{Inj}_k(\boldsymbol{x})] d\,\mathrm{Inj}_k(\boldsymbol{x})$$

involving the injection operator $\mathrm{Inj}_k(\boldsymbol{x}) = (x, x', \ldots, x^{(k)})^*$. ■

17.6 The Euler-Lagrange Equation

In proving the Euler-Lagrange equation (17.2), we assume that the function $x(t)$ minimizes the functional (17.1) among all competing functions in the space $\mathcal{C}^1[a, b]$. When the boundary values $x(a) = c$ and $x(b) = d$ are fixed, we consider the revised functional

$$F(\boldsymbol{x} + \boldsymbol{u}) = \int_a^b f[t, x(t) + u(t), x'(t) + u'(t)]\, dt \quad (17.13)$$

defined on the set of continuously differentiable functions $u(t)$ with $u(a) = 0$ and $u(b) = 0$. This closed subspace $\mathcal{D}^1[a, b]$ of $\mathcal{C}^1[a, b]$ qualifies as a Banach space in its own right, and if $x(t)$ minimizes the original functional, then $\boldsymbol{u} = \boldsymbol{0}$ minimizes the revised functional (17.13).

Fermat's principle, which is valid on any normed vector space, requires the differential of $F(\boldsymbol{x}+\boldsymbol{u})$ to vanish at $\boldsymbol{u} = \boldsymbol{0}$. In view of Example (17.5.3), this means

$$dF(\boldsymbol{x})\boldsymbol{u} = \int_a^b \{\partial_2 f[t, x(t), x'(t)] u(t) + \partial_3 f[t, x(t), x'(t)] u'(t)\}\, dt, \quad (17.14)$$

must vanish for every $\boldsymbol{u} \in \mathcal{D}^1[a, b]$. The appearance of $u'(t)$ in addition to $u(t)$ in this last equation is awkward. We can eliminate it invoking the integration by parts formula

$$\int_a^b \partial_3 f[t, x(t), x'(t)] u'(t)\, dt = -\int_a^b \frac{d}{dt} \partial_3 f[t, x(t), x'(t)] u(t)\, dt \quad (17.15)$$

using the boundary conditions $u(a) = u(b) = 0$. If we put

$$v(t) \quad = \quad \partial_2 f[t, x(t), x'(t)] - \frac{d}{dt}\partial_3 f[t, x(t), x'(t)],$$

then Fermat's principle reads

$$\int_a^b v(t)u(t)\,dt \quad = \quad 0$$

for every $\boldsymbol{u} \in \mathcal{D}^1[a, b]$. The special case $k = 0$ of the next lemma completes the proof of the Euler-Lagrange equation (17.2).

Proposition 17.6.1 (Du Bois-Reymond) *Let the function $v(t)$ be continuous on $[a, b]$ and satisfy*

$$\int_a^b v(t)u(t)\,dt \quad = \quad 0$$

for every $u(t)$ in $\mathcal{C}^k[a, b]$ with

$$u^{(j)}(a) \quad = \quad u^{(j)}(b) \quad = \quad 0\,, \quad 0 \le j \le k.$$

Then $v(t) = 0$ for all t.

Proof: Suppose $v(t)$ is not identically 0. By continuity there exists an interval $[c, d] \subset [a, b]$ on which $v(t)$ is either strictly positive or strictly negative. Without loss of generality, we assume the former case and take $a < c < d < b$. It is straightforward to prove by induction that the function

$$u(t) \quad = \quad \begin{cases} (t-c)^{k+1}(d-t)^{k+1} & t \in [c, d] \\ 0 & t \notin [c, d] \end{cases}$$

is continuously differentiable up to order k. Because the continuous function $v(t)u(t)$ is positive throughout the open interval (c, d) and vanishes outside it, the integral

$$\int_a^b v(t)u(t)\,dt \quad > \quad 0.$$

This contradiction proves the claim. ■

Two cases of the Euler-Lagrange equation (17.2) merit special mention. First, if $f(t, x, x')$ does not depend on x, then

$$\frac{d}{dt}\frac{\partial}{\partial x'} f[t, x(t), x'(t)] \quad = \quad 0,$$

and

$$\frac{\partial}{\partial x'} f[t, x(t), x'(t)] \quad = \quad c \tag{17.16}$$

for some constant c. Second, if $f(t, x, x')$ does not depend on t, then

$$x'(t)\frac{\partial}{\partial x'}f[t, x(t), x'(t)] - f[t, x(t), x'(t)] \quad = \quad c \tag{17.17}$$

is constant. This assertion follows from the identity

$$\begin{aligned}\frac{d}{dt}\Big[x'\partial_3 f - f\Big] &= x''\partial_3 f + x'\frac{d}{dt}\partial_3 f - \partial_1 f - \partial_2 f x' - \partial_3 f x'' \\ &= x'\Big[\frac{d}{dt}\partial_3 f - \partial_2 f\Big] \\ &= 0.\end{aligned}$$

In the absence of fixed boundary values $x(a) = c$ and $x(b) = d$, the integration by parts argument invoked in deriving the Euler-Lagrange equations is still valid. However, perturbations with $u(a) \neq 0$ or $u(b) \neq 0$ are now pertinent. To ensure that the boundary contributions

$$\partial_3 f[b, x(b), x'(b)]u(b) - \partial_3 f[a, x(a), x'(a)]u(a)$$

to $dF(\boldsymbol{x})\boldsymbol{u}$ vanish, we must assume that the multiplier $\partial_3 f[a, x(a), x'(a)]$ of $u(a)$ vanishes when $x(a)$ is not fixed and the multiplier $\partial_3 f[b, x(b), x'(b)]$ of $u(b)$ vanishes when $x(b)$ is not fixed. These constraints are referred to as free or natural boundary conditions.

Some applications involve optimization over multivariate functions $x(t)$. In this case the Euler-Lagrange equation (17.2) holds component by component. Derivation of this result relies on considering multivariate perturbations $u(t)$ with all but one component identically 0. Our discussion of Lagrangian mechanics in the next section requires multivariate functions. Functionals depending on derivatives beyond the first derivative can also be treated by the methods described in this section as noted in Problem 15. To derive the appropriate generalization of the Euler-Lagrange equations, we again use integration by parts and Proposition 17.6.1. Our consideration of splines provides insight into how one deals with functionals depending on second derivatives.

The classical theory of the calculus of variations involves only necessary conditions for an optimum. The modern theory takes up sufficient conditions as well [102, 139]. Although it is impossible to do justice to the modern theory in a brief exposition, it is fair to point out the important role of convexity. A functional $F(\boldsymbol{x})$ is convex if Jensen's inequality

$$F[\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}] \quad \leq \quad \alpha F(\boldsymbol{x}) + (1-\alpha)F(\boldsymbol{y})$$

holds for all $\boldsymbol{x}$, $\boldsymbol{y}$, and $\alpha \in [0, 1]$. Proposition 6.5.2 forges the most helpful connection between convexity and optimization. When $\boldsymbol{x}$ is a stationary point of $F(\boldsymbol{x})$, then the operative inequality $dF(\boldsymbol{x})\boldsymbol{u} \geq 0$ of the proposition automatically holds for all admissible u. A more crucial point is to

recognize convexity when it occurs. Fortunately, one simple test works. If the integrand $f(t, x, x')$ in the functional (17.1) is convex in its last two variables x and x' for each fixed t, then it is trivial to show that $F(\boldsymbol{x})$ is convex in $\boldsymbol{x}$. As pointed out momentarily, this is the case with geodesic problems.

17.7 Applications of the Euler-Lagrange Equation

To illustrate the solution of the Euler-Lagrange equations, we revisit the first four examples of Sect. 17.2. None of these examples involves constraints beyond fixed boundary conditions.

Example 17.7.1 *Geodesic on a Cylinder*

Without loss of generality, we suppose that the z axis and the central axis of the cylinder coincide. If the cylinder has radius r, then a curve on its surface can be represented by the triple $[r\cos\theta, r\sin\theta, z(\theta)]$ using cylindrical coordinates (θ, z). The distance connecting the point (θ_0, z_0) to the point (θ_1, z_1) is

$$G(\boldsymbol{z}) \;=\; \int_{\theta_0}^{\theta_1} \sqrt{r^2 + z'(\theta)^2}\, d\theta.$$

Here we assume $\theta_0 < \theta_1$ and $z_0 \neq z_1$. If $\theta_0 = \theta_1$, then it is clear that the geodesic is a vertical line on the surface. If $z_0 = z_1$, then the geodesic is a circular arc in a plane perpendicular to the central axis.

Because the function $f(\theta, z, z') = \sqrt{r^2 + z'^2}$ does not depend on z, the simplified Euler-Lagrange equation (17.16) applies. This requires

$$\frac{z'}{\sqrt{r^2 + z'^2}}$$

to be constant, which is achieved by taking z' to be constant. In view of the constraints $z(\theta_0) = z_0$ and $z(\theta_1) = z_1$, we have

$$z(\theta) \;=\; \frac{z_1 - z_0}{\theta_1 - \theta_0}\theta + \frac{\theta_1 z_0 - \theta_0 z_1}{\theta_1 - \theta_0}.$$

To prove that this solution provides the minimum of $G(\boldsymbol{z})$, it suffices to note that $f(\theta, z, z') = \sqrt{r^2 + z'^2}$ is convex in (z, z'). Convexity follows from the relationship between $\sqrt{r^2 + z'^2}$ and the Euclidean norm of R^2. ■

Example 17.7.2 *Minimal Surface Area of Revolution*

Because $f(x, y, y') = 2\pi y\sqrt{1 + y'^2}$ does not depend on x, the simplified Euler-Lagrange equation (17.17) requires

$$\frac{y'^2 y}{\sqrt{1 + y'^2}} - y\sqrt{1 + y'^2} \;=\; c$$

to be constant. Straightforward algebra now gives $y^2 = c^2(1 + y'^2)$ and

$$y' = \left[\left(\frac{y}{c}\right)^2 - 1\right]^{1/2} = \frac{1}{c}\sqrt{y^2 - c^2}.$$

If we rewrite this as

$$\frac{dy}{\sqrt{y^2 - c^2}} = \frac{dx}{c},$$

then Table 4.1 shows that

$$\operatorname{arccosh}^{-1}\frac{y}{c} = \frac{x}{c} + d$$

for some constant d. It follows that

$$y(x) = c\cosh\left(\frac{x}{c} + d\right).$$

If we take $(x_0, y_0) = (0, a)$, then we can eliminate the constant c and express

$$y(x) = \frac{a}{\cosh d}\cosh\left(\frac{\cosh d}{a}x + d\right).$$

In some cases the adjustable parameter d can be used to match the other end of the catenary curve $y(x)$ to (x_1, y_1). In other cases, there is no minimal surface of revolution in the ordinary sense [15]. ■

Example 17.7.3 *Passage of Light through an Inhomogeneous Medium*

This is another example of a integrand $f(x, y, y') = \sqrt{1 + y'^2}/v(y)$ that does not depend on x. The quantity (17.17)

$$\frac{y'^2}{v(y)\sqrt{1 + y'^2}} - \frac{\sqrt{1 + y'^2}}{v(y)} = -\frac{1}{v(y)\sqrt{1 + y'^2}}$$

is constant along the optimal path. If we differentiate the identity

$$-\ln v[y(x)] - \ln\sqrt{1 + y'^2(x)} = \ln(-c)$$

with respect to x, then the formula

$$y''(x) = -\left[1 + y'(x)^2\right]\frac{v'[y(x)]}{v[y(x)]}$$

emerges. Because the speed of light increases with increasing altitude, $v'[y(x)] > 0$, and hence $y''(x) < 0$. In other words, the path that light follows toward an observer on the earth is concave and bends downward. Hence, the setting sun appears higher in the sky than it actually is [240]. Problem 20 considers the special case $v(y) = y$, where one can explicitly solve for $y(x)$. ■

Example 17.7.4 *Lagrange's versus Newton's Version of Mechanics*

Let $x(t)$ be the path taken in R^3 by a single particle with initial and final values $x(t_0) = \boldsymbol{a}$ and $x(t_1) = \boldsymbol{b}$. The Euler-Lagrange equations at the stationary value of the action integral (17.3) are

$$\begin{aligned} 0 &= -\frac{\partial}{\partial x_i} U[x(t)] - \frac{d}{dt}\frac{\partial}{\partial x_i'} T[x'(t)] \\ &= -\frac{\partial}{\partial x_i} U[x(t)] - m x_i''(t). \end{aligned}$$

These are clearly Newton's equations of motions in a conservative force field. The total energy $E = T(\boldsymbol{x}') + U(\boldsymbol{x})$ of the particle is conserved because

$$\frac{d}{dt} E(t) = m \sum_{i=1}^{3} x_i'(t) x_i''(t) + \sum_{i=1}^{3} \frac{\partial}{\partial x_i} U[x(t)] x_i'(t) = 0.$$

In many applications we follow several particles. For instance in the n-body problem, n particles move under the influence of their mutual gravitational fields. Let the ith particle have mass m_i and position $x_i(t)$ in R^3 at time t. Assuming the gravitational constant is 1, the potential energy of the system is given by

$$U(\boldsymbol{x}) = - \sum_{\{i,j\}} \frac{m_i m_j}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|},$$

where the sum ranges over all pairs $\{i, j\}$ of particles and $x(t)$ is the vector constructed by stacking the $x_i(t)$. The total kinetic energy of the particles is

$$T(\boldsymbol{x}') = \frac{1}{2} \sum_{i=1}^{n} m_i \|\boldsymbol{x}_i'\|^2.$$

According to Lagrangian mechanics, the motion of the system yields the stationary value of the action integral (17.3). The Euler-Lagrange equations are

$$\begin{aligned} 0 &= -\frac{\partial}{\partial x_{ik}} U[x(t)] - \frac{d}{dt}\frac{\partial}{\partial x_{ik}'} T[x'(t)] \\ &= -\sum_{j \neq i} \frac{m_i m_j}{\|x_i(t) - x_j(t)\|^2} \cdot \frac{x_{ik}(t) - x_{jk}(t)}{\|x_i(t) - x_j(t)\|} - m_i x_{ik}''(t). \end{aligned}$$

In other words, the force exerted on particle i by particle j is given by Newton's inverse square law. The total energy $E = T(\boldsymbol{x}') + U(\boldsymbol{x})$ of the system is conserved because

$$\begin{aligned} \frac{d}{dt} E(t) &= \sum_{i=1}^{n} m_i \sum_{k=1}^{3} x_{ik}'(t) x_{ik}''(t) + \sum_{i=1}^{n} \sum_{k=1}^{3} \frac{\partial}{\partial x_{ik}} U[x(t)] x_{ik}'(t) \\ &= 0. \end{aligned}$$

This result holds for any conservative dynamical system governed by the Euler-Lagrange equations with potential $U(\boldsymbol{x})$. Conservation of linear and angular momentum also hold in the n-body framework as documented in Problem 21. Dynamical systems with more general potential functions do not necessarily preserve linear and angular momentum. ■

17.8 Lagrange's Lacuna

Let us now expose and correct a deception foisted on the reader in deriving the Euler-Lagrange equations. At a crucial point in our derivation we invoked the integration-by-parts formula (17.15) without proving that the factor $\partial_3 f[t, x(t), x'(t)]$ is continuously differentiable. This gap in Lagrange's original treatment was noted by Du Bois-Reymond. His correction is based on the following variant of Proposition 17.6.1.

Proposition 17.8.1 *Let the function $v(t)$ be continuous on $[a, b]$ and satisfy*

$$\int_a^b v(t)u^{(k)}(t)\,dt \;=\; 0$$

for every test function $u(t)$ in $\mathcal{C}^k[a, b]$ with

$$u^{(j)}(a) \;=\; u^{(j)}(b) \;=\; 0\,, \quad 0 \le j \le k-1.$$

Then $v(t)$ is a polynomial of degree $k-1$ on $[a, b]$.

Proof: Integration by parts shows that

$$\int_a^b p(t)u^{(k)}(t)\,dt \;=\; 0$$

and therefore that

$$\int_a^b [v(t) - p(t)]u^{(k)}(t)\,dt \;=\; 0$$

for every test function $u(t)$ and polynomial $p(t)$ of degree $k-1$. If we can construct a test function $u(t)$ and a polynomial $p(t)$ of degree $k-1$ such that $u^{(k)}(t) = v(t) - p(t)$, then the identity

$$\int_a^b [v(t) - p(t)]^2\,dt \;=\; 0$$

implies $v(t) = p(t)$. Construction of $u(t)$ involves some technicalities that we will avoid by focusing on the case $k = 2$. The proof of the case $k = 1$ is

similar but simpler. If we put $p(t) = c + d(t-a)$, then it seems sensible to define

$$\begin{aligned} u'(t) &= \int_a^t [v(s) - c - d(s-a)]\,ds \\ &= \int_a^t v(s)\,ds - c(t-a) - \frac{d}{2}(t-a)^2 \\ u(t) &= \int_a^t \int_a^s v(r)\,dr\,ds - c\int_a^t (s-a)\,ds - \frac{d}{2}\int_a^t (s-a)^2\,ds \\ &= \int_a^t \int_a^s v(r)\,dr\,ds - \frac{c}{2}(t-a)^2 - \frac{d}{6}(t-a)^3. \end{aligned}$$

Because $u'(a) = u(a) = 0$ is clearly true, the only remaining issue is whether we can choose c and d so that $u'(b) = u(b) = 0$. However, this is possible since the matrix implementing the linear system

$$\begin{aligned} 0 &= \int_a^b v(s)\,ds - c(b-a) - \frac{d}{2}(b-a)^2 \\ 0 &= \int_a^b \int_a^s v(r)\,dr\,ds - \frac{c}{2}(b-a)^2 - \frac{d}{6}(b-a)^3 \end{aligned}$$

is invertible. Indeed, invertibility follows from the identity

$$\det \begin{pmatrix} (b-a) & (b-a)^2/2 \\ (b-a)^2/2 & (b-a)^3/6 \end{pmatrix} = -(b-a)^4/12.$$

Thus, c and d can be chosen so that $u(t)$ satisfies the requisite boundary conditions. ■

To apply Proposition 17.8.1 in the derivation of the Euler-Lagrange equation, we define the function

$$g(t) = \int_a^t \partial_2 f[s, x(s), x'(s)]\,ds.$$

The fundamental theorem of calculus implies that $g(t)$ is continuously differentiable. Hence, integration by parts gives the alternative

$$dF(\boldsymbol{x})\boldsymbol{u} = \int_a^b \{-g(t) + \partial_3 f[t, x(t), x'(t)]\}u'(t)\,dt$$

to equation (17.14). According to Proposition 17.8.1 with $k = 1$, the function

$$-g(t) + \partial_3 f[t, x(t), x'(t)] = c$$

for some constant c. It follows that $\partial_3 f[t, x(t), x'(t)] = g(t) + c$ is continuously differentiable as required.

17.9 Variational Problems with Constraints

The isoperimetric problem requires maximizing the area enclosed by a curve of fixed perimeter. More generally suppose we wish to maximize a functional $F(\boldsymbol{y})$ subject to the equality constraints $G_i(\boldsymbol{y}) = 0$ for $1 \leq i \leq p$. If $\boldsymbol{x}$ furnishes a local maximum, then we can hope that the Lagrange multiplier rule

$$dF(\boldsymbol{x})\boldsymbol{u} + \sum_{i=1}^{p} \lambda_i dG_i(\boldsymbol{x})\boldsymbol{u} = 0 \tag{17.18}$$

will be valid for all admissible functions u. This turns out to be the case if the differentials $dG_1(\boldsymbol{x}), \ldots, dG_p(\boldsymbol{x})$ are linearly independent. Linear independence fails whenever there exists a nontrivial vector $\boldsymbol{c}$ with components $c_1, \ldots, c_p$ such that

$$\sum_{i=1}^{p} c_i dG_i(\boldsymbol{x})\boldsymbol{u} = 0$$

for all possible $\boldsymbol{u}$. Such a failure is impossible if there exists a finite sequence $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ of admissible functions such that the square matrix $[dG_i(\boldsymbol{x})\boldsymbol{u}_j]$ is invertible. For the sake of simplicity, we will take linear independence to mean that for some choice of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ the matrix $[dG_i(\boldsymbol{x})\boldsymbol{u}_j]$ is invertible. With this stipulation in mind, we now prove the multiplier rule (17.18) under independence.

Our strategy will be to examine map

$$H(\boldsymbol{\alpha}) = \begin{pmatrix} F(\boldsymbol{x} + \alpha_0\boldsymbol{u} + \alpha_1\boldsymbol{u}_1 + \cdots + \alpha_p\boldsymbol{u}_p) \\ G_1(\boldsymbol{x} + \alpha_0\boldsymbol{u} + \alpha_1\boldsymbol{u}_1 + \cdots + \alpha_p\boldsymbol{u}_p) \\ \vdots \\ G_p(\boldsymbol{x} + \alpha_0\boldsymbol{u} + \alpha_1\boldsymbol{u}_1 + \cdots + \alpha_p\boldsymbol{u}_p) \end{pmatrix}$$

defined for the given functions $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ and an arbitrary admissible function $\boldsymbol{u}$. Note that $H(\boldsymbol{\alpha})$ maps R^{p+1} into itself. The differential of $H(\boldsymbol{\alpha})$ at $\mathbf{0}$ is the Jacobian matrix

$$dH(\mathbf{0}) = \begin{pmatrix} dF(\boldsymbol{x})\boldsymbol{u} & dF(\boldsymbol{x})\boldsymbol{u}_1 & \cdots & dF(\boldsymbol{x})\boldsymbol{u}_p \\ dG_1(\boldsymbol{x})\boldsymbol{u} & dG_1(\boldsymbol{x})\boldsymbol{u}_1 & \cdots & dG_1(\boldsymbol{x})\boldsymbol{u}_p \\ \vdots & \vdots & \vdots & \vdots \\ dG_p(\boldsymbol{x})\boldsymbol{u} & dG_p(\boldsymbol{x})\boldsymbol{u}_1 & \cdots & dG_p(\boldsymbol{x})\boldsymbol{u}_p \end{pmatrix}.$$

Assuming the objective functional $F(\boldsymbol{x})$ and the constraints $G_i(\boldsymbol{x})$ are continuously differentiable in a neighborhood of $\boldsymbol{x}$, the function $H(\boldsymbol{\alpha})$ is continuously differentiable in a neighborhood of $\mathbf{0}$. Hence, we can invoke the inverse function theorem, Proposition 4.6.1, provided $dH(\mathbf{0})$ is invertible.

Assume for the sake of argument that this is the case. Then $H(\boldsymbol{\alpha})$ maps a neighborhood of $\mathbf{0}$ onto a neighborhood of the image

$$[F(\boldsymbol{x}), G_1(\boldsymbol{x}), \ldots, G_p(\boldsymbol{x})]^* \quad = \quad [F(\boldsymbol{x}), 0, \ldots, 0]^*.$$

Taking the neighborhood of $\mathbf{0}$ to be arbitrarily small implies that there are functions $\boldsymbol{v} = \boldsymbol{x} + \alpha_0 \boldsymbol{u} + \alpha_1 \boldsymbol{u}_1 + \cdots + c_p \boldsymbol{u}_p$ arbitrarily close to $\boldsymbol{x}$ satisfying $F(\boldsymbol{v}) > F(\boldsymbol{x})$ and $G_i(\boldsymbol{v}) = 0$ for all i. This contradicts the assumption that $\boldsymbol{x}$ furnishes a local maximum subject to the constraints. It follows that the matrix $dH(\mathbf{0})$ must be singular.

The connection of this condition to the multiplier rule (17.18) becomes less obscure when we exploit the fact that $\det dH(\mathbf{0}) = 0$. Expanding this determinant on the first column of $dH(\mathbf{0})$ leads to the equation

$$0 \quad = \quad \mu_0 dF(\boldsymbol{x})\boldsymbol{u} + \sum_{i=1}^{p} \mu_i dG_i(\boldsymbol{x})\boldsymbol{u}, \tag{17.19}$$

where the cofactor $\mu_0 = \det[dG_i(\boldsymbol{x})\boldsymbol{u}_j]$ is nonzero by assumption. If we divide equation (17.19) by μ_0, then we arrive at the multiplier rule (17.18) with $\lambda_i = \mu_i/\mu_0$.

Example 17.9.1 *Isoperimetric Problem*

If we assume that the perimeter constraint has a nontrivial differential, then the combination of the multiplier rule and the Euler-Lagrange equation (17.2) implies

$$1 - \frac{d}{dx}\left[\lambda \frac{y'(x)}{\sqrt{1 + y'(x)^2}}\right] \quad = \quad 0.$$

This forces λ to be nonzero, and integration on x produces

$$\frac{y'(x)}{\sqrt{1 + y'(x)^2}} \quad = \quad cx + d$$

for $c = 1/\lambda$ and d a constant of integration. This last equation can be solved for $y'(x)$ in the form

$$y'(x) \quad = \quad \frac{cx + d}{\sqrt{1 - (cx + d)^2}}$$

and integrated to give

$$y(x) \quad = \quad -\frac{1}{c}\sqrt{1 - (cx + d)^2} + e$$

for some additional constant e. If we take $x_0 = -1$ and $x_1 = 1$, then application of the boundary conditions $y(-1) = y(1) = 0$ yields

$$\begin{aligned} e^2 &= \frac{1}{c^2}\left[1 - (-c+d)^2\right] \\ e^2 &= \frac{1}{c^2}\left[1 - (c+d)^2\right]. \end{aligned}$$

It follows that $d = 0$ and $e = \sqrt{c^{-2} - 1}$. Finally, the constraint

$$\begin{aligned} \ell &= \int_{-1}^{1} \sqrt{1 + y'(x)^2}\, dx \\ &= \int_{-1}^{1} \frac{1}{\sqrt{1 - c^2x^2}}\, dx \\ &= \frac{2}{c} \arcsin c \end{aligned}$$

determines the constant c. More precisely, the line $c \mapsto \ell c$ and the convex increasing curve $c \mapsto 2 \arcsin c$ intersect in a single point in $(0, 1]$ whenever $2 < l \leq \pi$. Because

$$[y(x) - e]^2 + x^2 = \frac{1}{c^2},$$

the function $y(x)$ traces out a circular arc. If $\ell = \pi$, then the arc is a semicircle with center at the origin $\mathbf{0}$.

This analysis is predicated on the assumption that there exists a continuously differentiable function $u(x)$ with

$$\begin{aligned} dL(\boldsymbol{y})\boldsymbol{u} &= \int_{-1}^{1} \frac{y'(x)}{\sqrt{1 + y'(x)^2}} u'(x)\, dx \\ &= \int_{-1}^{1} cxu'(x)\, dx \\ &\neq 0 \end{aligned}$$

and $u(-1) = u(1) = 0$. One obvious choice is $u(x) = (1 - x^2)^2$. ■

17.10 Natural Cubic Splines

Our treatment of splines involves functions from the Banach space $\mathcal{C}^2[a, b]$. If $x(t)$ minimizes the spline functional (17.4), then the differential

$$dC(\boldsymbol{x})\boldsymbol{u} = 2 \int_a^b x''(t)u''(t)\, dt$$

is a special case of equation (17.12) and vanishes for all admissible $u(t)$. Such functions $u(t)$ satisfy $u(s_i) = 0$ at every node s_i. If we focus on test functions $u(t)$ that vanish outside a node interval $[s_{i-1}, s_i]$, then we can invoke Proposition 17.8.1 and infer that $x''(t)$ is linear on $[s_{i-1}, s_i]$. This compels $x(t)$ to be a piecewise cubic polynomial throughout $[a, b]$.

On node interval $[s_{i-1}, s_i]$, integration by parts and the cubic nature of $x(t)$ yield

$$\int_{s_{i-1}}^{s_i} x''(t)u''(t)\,dt \quad = \quad x''(t)u'(t)\Big|_{s_{i-1}}^{s_i} - x'''(t)u(t)\Big|_{s_{i-1}}^{s_i}. \quad (17.20)$$

The constraint $u(s_{i-1}) = u(s_i) = 0$ forces the boundary terms involving $x'''(t)$ to vanish. When we add the contributions (17.20) to form the full integral $\int_a^b x''(t)u''(t)\,dt$, all of the boundary terms involving $x''(t)$ cancel except for $x''(s_n)u'(s_n) - x''(s_0)u'(s_0)$. If we impose the natural boundary conditions $x''(s_n) = x''(s_0) = 0$, then all terms vanish, and we recover the necessary condition $dC(\boldsymbol{x})\boldsymbol{u} = 2\int_a^b x''(t)u''(t)\,dt = 0$ for optimality.

It is possible to demonstrate that there is precisely one piecewise cubic polynomial $x(t)$ from $\mathcal{C}^2[a, b]$ that interpolates the given values x_i at the nodes s_i and satisfies the natural boundary conditions $x''(s_0) = x''(s_n) = 0$. This exercise in linear algebra is carried out in the reference [166]. It is shown there that $x(t)$ minimizes $C(\boldsymbol{y})$ subject to the interpolation constraints. The fact that the minimum is unique is hardly surprising given the convexity of the functional $C(\boldsymbol{y})$ on $\mathcal{C}^2[a, b]$.

17.11 Problems

1. Prove that the space $\mathcal{C}^0[a, b]$ is not finite dimensional. (Hint: The functions $1, t, \ldots, t^n$ are linearly independent.)

2. Prove that the set $P_n[0, 1]$ of polynomials of degree n or less on $[0, 1]$ is a closed subspace of $\mathcal{C}^0[0, 1]$.

3. Suppose X is an inner product space. Prove that the induced norm satisfies the parallelogram identity

$$\|\boldsymbol{u} + \boldsymbol{v}\|^2 + \|\boldsymbol{u} - \boldsymbol{v}\|^2 \quad = \quad 2(\|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2)$$

   for all vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. Show that the norm of $\mathcal{C}^0[0, 1]$ is not induced by an inner product by producing $\boldsymbol{u}$ and $\boldsymbol{v}$ that fail the parallelogram identity.

4. Demonstrate that the closed unit ball $B = \{\boldsymbol{x} : \|\boldsymbol{x}\| \leq 1\}$ in $\mathcal{C}^0[0, 1]$ is not compact. (Hint: The sequence $1, t, t^2, t^3, \ldots$ has no uniformly convergent subsequence.)

5. Demonstrate that a closed subset of a Banach space is complete.

6. Let X and Y be normed vector spaces with norms $\|\boldsymbol{x}\|_p$ and $\|\boldsymbol{y}\|_q$. The product space $X \times Y$ is a vector space if addition and scalar multiplication are defined coordinatewise. Show that the following are equivalent norms on $X \times Y$:

$$\begin{aligned} \|(\boldsymbol{x}, \boldsymbol{y})\|_\infty &= \max\{\|\boldsymbol{x}\|_p, \|\boldsymbol{y}\|_q\} \\ \|(\boldsymbol{x}, \boldsymbol{y})\|_1 &= \|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_q \\ \|(\boldsymbol{x}, \boldsymbol{y})\|_2 &= \sqrt{\|\boldsymbol{x}\|_p^2 + \|\boldsymbol{y}\|_q^2}. \end{aligned}$$

   See Example 2.5.6 for the notion of equivalent norms.

7. Continuing Problem 6, show that $X \times Y$ is a Banach space under any of three proposed norms whenever X and Y are both Banach spaces.

8. Demonstrate that the three linear functionals A_1, A_2, and A_3 defined in (17.7) actually have the operator norms 1, $b - a$, and 1 on their respective normed linear spaces.

9. Suppose the function $K(s,t)$ in equation (17.8) is square-integrable over $[a,b] \times [a,b]$. Prove that the corresponding operator A_6 maps the Hilbert space $\mathrm{L}_2[a,b]$ into itself in such a way that

$$\|A_6\|^2 \leq \int_a^b \int_a^b K(s,t)^2 ds\, dt.$$

10. Let $L(\boldsymbol{y})$ be a continuous linear functional on a normed vector space. Under what circumstances does $L(\boldsymbol{y})$ have a minimum or a maximum?

11. On an inner product space prove that the functional $\|\boldsymbol{x}\|^2$ is differentiable with differential $dF(\boldsymbol{x})\boldsymbol{y} = 2\langle \boldsymbol{x}, \boldsymbol{y} \rangle$. Deduce from this fact that the norm $G(\boldsymbol{x}) = \|\boldsymbol{x}\|$ is differentiable with differential $dG(\boldsymbol{x})\boldsymbol{y} = \langle \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}, \boldsymbol{y} \rangle$ whenever $\boldsymbol{x} \neq \boldsymbol{0}$.

12. Consider the linear functional $F(\boldsymbol{y}) = \langle \boldsymbol{v}, \boldsymbol{y} \rangle$ on an inner product space. Here the vector $\boldsymbol{v}$ is fixed. Let $g(t)$ be a differentiable function on the real line, and define $G(\boldsymbol{y}) = g \circ F(\boldsymbol{y})$. Show that $G(\boldsymbol{y})$ has differential $dG(\boldsymbol{x})\boldsymbol{u} = g'(\langle \boldsymbol{v}, \boldsymbol{x} \rangle)\langle \boldsymbol{v}, \boldsymbol{u} \rangle$ at the vector $\boldsymbol{x}$.

13. Some variational problems have no solution. For instance, show that the following two functionals have no minimum subject to the given constraints:

    (a) The integral $\int_0^1 \sqrt{1 + y'(x)^2}\, dx$ subject to the four constraints $y(0) = y(1) = 0$ and $y'(0) = y'(1) = 1$.

(b) Weierstrass's integral $\int_{-1}^{1} x^2 y'(x)^2\,dx$ subject to the constraints $y(-1) = -1$ and $y(1) = 1$.

In each case $y(x)$ can be any piecewise differentiable function on the given interval.

14. Consider the closed subspace of $\mathcal{C}^1[0, 2\pi]$ defined by functions that are periodic. Wirtinger's inequality says

$$\int_0^{2\pi} f'(t)^2 dt \quad \geq \quad \int_0^{2\pi} f(t)^2 dt$$

for every such function $f(t)$ with equality if and only if

$$f(t) \quad = \quad a\cos t + b\sin t.$$

Prove this result by considering the functional

$$W(\boldsymbol{f}) \quad = \quad \int_0^{2\pi} \left[f'(t)^2 - f(t)^2\right] dt$$

For the purpose of this problem your may assume that the minimum of $W(\boldsymbol{f})$ is attained.

15. Consider a functional

$$F(\boldsymbol{x}) \quad = \quad \int_a^b f[t, x(t), x'(t), \ldots, x^{(k)}(t)]\,dt$$

defined on $\mathcal{C}^k[a, b]$. Derive the Euler-Lagrange equation

$$\frac{\partial}{\partial x} f + \sum_{j=1}^{k} (-1)^j \frac{d^j}{dt^j}\left[\frac{\partial}{\partial x^{(j)}} f\right] \quad = \quad 0$$

for the function $x(t)$ optimizing $F(\boldsymbol{x})$. What assumptions are necessary to justify this derivation?

16. Let $y(x)$ denote a continuously differentiable curve over the interval $[-1, 1]$ satisfying $y(-1) = y(1) = 0$. Find $y(x)$ that minimizes the length

$$L(\boldsymbol{y}) \quad = \quad \int_{-1}^{1} \sqrt{1 + y'(x)^2}\,dx.$$

while enclosing a fixed area

$$\int_{-1}^{1} y(x)\,dx \quad = \quad A \quad \leq \quad \frac{\pi}{2}.$$

17. If possible, find the minimum and maximum of the functional

$$F(\boldsymbol{y}) \quad = \quad \int_0^1 [y'(x)^2 + x^2]\, dx$$

over $\mathcal{C}^1[0,1]$ subject to the constraints $y(0) = 0$, $y(1) = 0$, and $\int_0^1 y(x)^2 dx = 1$. Verify that your minimum solution yields the minimum value.

18. Find the minimum value of the functional

$$M(\boldsymbol{y}) \quad = \quad \int_0^1 xy(x)\, dx$$

over $\mathcal{C}^0[0,1]$ subject to the constraint

$$V(\boldsymbol{y}) \quad = \quad \int_0^1 y(x)^2\, dx \quad = \quad \frac{1}{12}.$$

Verify that your solution gives the minimum.

19. Find the minimum value of the functional

$$F(\boldsymbol{y}) \quad = \quad \int_0^\pi [y'(x)^2 + 2y(x)\sin x]\, dx$$

subject to the single constraint $y(0) = 0$. How does this differ from the solution when the constraint $y(\pi) = 0$ is added? Verify in each case that the minimum is achieved.

20. Suppose the speed of light in the upper-half xy-plane is $v(y) = y$. Find the path of light connecting (x_0, y_0) to (x_1, y_1). Show that the travel time is infinite if and only if either y_0 or y_1 equals 0.

21. In the n-body problem, linear and angular momentum are defined by

$$\begin{aligned} L(t) \quad &= \quad \sum_{i=1}^n m_i\, x_i'(t) \\ A(t) \quad &= \quad \sum_{i=1}^n m_i\, x_i(t) \times x_i'(t). \end{aligned}$$

Prove that these quantities are conserved. (Hints: The force exerted by i on j is equal in magnitude and opposite in direction to the force exerted by j on i. The cross product $\boldsymbol{v} \times \boldsymbol{v}$ of a vector with itself vanishes.)

22. Suppose the integrand $\mathcal{L}(t, \boldsymbol{x}, \boldsymbol{v})$ of the functional

$$F(\boldsymbol{x}) = \int_{t_0}^{t_1} \mathcal{L}[t, x(t), x'(t)]\, dt$$

is strictly convex in the variable $\boldsymbol{v}$ with continuous second differential $\partial_3^2 \mathcal{L}(t, \boldsymbol{x}, \boldsymbol{v})$. In mechanics, $\mathcal{L}(t, \boldsymbol{x}, \boldsymbol{v})$ is called the Lagrangian, and its Fenchel conjugate

$$\mathcal{H}(t, \boldsymbol{x}, \boldsymbol{p}) = \sup_{\boldsymbol{v}} [\boldsymbol{p}^* \boldsymbol{v} - \mathcal{L}(t, \boldsymbol{x}, \boldsymbol{v})]$$

is called the Hamiltonian. According to Proposition 14.3.2, we recover $\mathcal{L}(t, \boldsymbol{x}, \boldsymbol{v})$ as

$$\mathcal{L}(t, \boldsymbol{x}, \boldsymbol{v}) = \sup_{\boldsymbol{p}} [\boldsymbol{v}^* \boldsymbol{p} - \mathcal{H}(t, \boldsymbol{x}, \boldsymbol{p})].$$

Show that $\boldsymbol{v}$ and $\boldsymbol{p}$ determine each other through the equations

$$\begin{aligned} \partial_3 \mathcal{L}(t, \boldsymbol{x}, \boldsymbol{v}) &= \boldsymbol{p}^* \\ \partial_3 \mathcal{H}(t, \boldsymbol{x}, \boldsymbol{p}) &= \boldsymbol{v}^* \end{aligned}$$

and that

$$\mathcal{H}(t, \boldsymbol{x}, \boldsymbol{p}) = \boldsymbol{p}^* \boldsymbol{v} - \mathcal{L}(t, \boldsymbol{x}, \boldsymbol{v}) \tag{17.21}$$

when these equations are satisfied. From the expression (17.21) deduce that

$$\begin{aligned} d\mathcal{H}(t, \boldsymbol{x}, \boldsymbol{p}) &= -d\mathcal{L}(t, \boldsymbol{x}, \boldsymbol{v}) + \boldsymbol{p}^* d\boldsymbol{v} + \boldsymbol{v}^* d\boldsymbol{p} \\ &= -\partial_1 \mathcal{L}(t, \boldsymbol{x}, \boldsymbol{v}) dt - \partial_2 \mathcal{L}(t, \boldsymbol{x}, \boldsymbol{v}) d\boldsymbol{x} + \partial_3 \mathcal{H}(t, \boldsymbol{x}, \boldsymbol{p}) d\boldsymbol{p}. \end{aligned}$$

Compare this to the differential

$$d\mathcal{H}(t, \boldsymbol{x}, \boldsymbol{p}) = \partial_1 \mathcal{H}(t, \boldsymbol{x}, \boldsymbol{p}) dt + \partial_2 \mathcal{H}(t, \boldsymbol{x}, \boldsymbol{p}) d\boldsymbol{x} + \partial_3 \mathcal{H}(t, \boldsymbol{x}, \boldsymbol{p}) d\boldsymbol{p}$$

and conclude that

$$\partial_2 \mathcal{L}(t, \boldsymbol{x}, \boldsymbol{v}) = -\partial_2 \mathcal{H}(t, \boldsymbol{x}, \boldsymbol{p}).$$

Use this result to prove that the Euler-Lagrange equation

$$\partial_2 \mathcal{L}[t, x(t), x'(t)] = \frac{d}{dt} \partial_3 \mathcal{L}[t, x(t), x'(t)]$$

is equivalent to the two Hamiltonian equations

$$\begin{aligned} \frac{d}{dt} x(t) &= \partial_3 \mathcal{H}[t, x(t), p(t)]^* \\ \frac{d}{dt} p(t) &= -\partial_2 \mathcal{H}[t, x(t), p(t)]^*. \end{aligned}$$

Thus, the first-order Hamiltonian differential equations can serve as a substitute for the second-order Euler-Lagrange differential equation in these circumstances.

23. Continuing Problem 22, show that the Hamiltonian is constant over time t whenever it does not depend explicitly on time, that is

$$\mathcal{H}(t, \boldsymbol{x}, \boldsymbol{p}) \quad = \quad \mathcal{H}(\boldsymbol{x}, \boldsymbol{p}).$$

24. Continuing Problem 22, suppose $\mathcal{L}(t, \boldsymbol{x}, \boldsymbol{v})$ can be expressed as the difference between the kinetic energy $T(\boldsymbol{v}) = \frac{1}{2}\sum_{i=1}^{n} m_i \|\boldsymbol{v}_i\|^2$ and the potential energy $U(\boldsymbol{x})$. Demonstrate that $\mathcal{H}(t, \boldsymbol{x}, \boldsymbol{p})$ is the sum $T(\boldsymbol{v}) + U(\boldsymbol{x})$ of the kinetic and potential energies. (Hint: See Problem 10 of Chap. 14.)

Appendix: Mathematical Notes

A.1 Univariate Normal Random Variables

A random variable X is said to be standard normal if it possesses the density function

$$\psi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

To find the characteristic function $\hat{\psi}(s) = \mathrm{E}(e^{isX})$ of X, we derive and solve a differential equation. Differentiation under the integral sign and integration by parts together imply that

$$\begin{aligned}
\frac{d}{ds}\hat{\psi}(s) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{isx} ix e^{-\frac{x^2}{2}} dx \\
&= -\frac{i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{isx} \frac{d}{dx} e^{-\frac{x^2}{2}} dx \\
&= -\frac{i}{\sqrt{2\pi}} e^{isx} e^{-\frac{x^2}{2}} \Big|_{-\infty}^{\infty} - \frac{s}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{isx} e^{-\frac{x^2}{2}} dx \\
&= -s\hat{\psi}(s).
\end{aligned}$$

The unique solution to this differential equation with initial value $\hat{\psi}(0) = 1$ is $\hat{\psi}(s) = e^{-s^2/2}$. The differential equation also yields the moments

$$\mathrm{E}(X) = \frac{1}{i} \frac{d}{ds} \hat{\psi}(0) = 0$$

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8,

and

$$\begin{aligned} \mathrm{E}(X^2) &= \frac{1}{i^2}\frac{d^2}{ds^2}\hat{\psi}(0) \\ &= \frac{1}{i^2}\Big[-\hat{\psi}(s) + s^2\hat{\psi}(s)\Big]_{s=0} \\ &= 1. \end{aligned}$$

An affine transformation $Y = \sigma X + \mu$ of X is normally distributed with density

$$\frac{1}{\sigma}\psi\Big(\frac{y-\mu}{\sigma}\Big) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

Here we take $\sigma > 0$. The general identity $\mathrm{E}\left[e^{is(\mu+\sigma X)}\right] = e^{is\mu}E\left[e^{i(\sigma s)X}\right]$ permits us to write the characteristic function of Y as

$$e^{is\mu}\hat{\psi}(\sigma s) = e^{is\mu-\frac{\sigma^2 s^2}{2}}.$$

The mean and variance of Y are μ and σ^2.

One of the most useful properties of normally distributed random variables is that they are closed under the formation of independent linear combinations. Thus, if Y and Z are independent and normally distributed, then $aY + bZ$ is normally distributed for any choice of the constants a and b. To prove this result, it suffices to assume that Y and Z are standard normal. In view of the form of $\hat{\psi}(s)$, we then have

$$\mathrm{E}\left[e^{is(aY+bZ)}\right] = \mathrm{E}\left[e^{i(as)Y}\right]\mathrm{E}\left[e^{i(bs)Z}\right] = \hat{\psi}\left(\sqrt{a^2+b^2}s\right).$$

Thus, if we accept the fact that a distribution function is uniquely defined by its characteristic function, $aY + bZ$ is normally distributed with mean 0 and variance $a^2 + b^2$.

Doubtless the reader is also familiar with the central limit theorem. For the record, recall that if X_n is a sequence of independent identically distributed random variables with common mean μ and common variance σ^2, then

$$\lim_{n\to\infty}\Pr\left[\frac{\sum_{j=1}^n(X_j-\mu)}{\sqrt{n\sigma^2}} \le x\right] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{u^2}{2}}du.$$

Of course, there is a certain inevitability to the limit being standard normal; namely, if the X_n are standard normal to begin with, then the standardized sum $n^{-1/2}\sum_{j=1}^n X_j$ is also standard normal.

A.2 Multivariate Normal Random Vectors

We now extend the univariate normal distribution to the multivariate normal distribution. Among the many possible definitions, we adopt the one most widely used in stochastic simulation. Our point of departure will be random vectors with independent standard normal components. If such a random vector $\boldsymbol{X}$ has n components, then its density is

$$\prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_j^2/2} \quad = \quad \left(\frac{1}{2\pi}\right)^{n/2} e^{-\boldsymbol{x}^*\boldsymbol{x}/2}.$$

Because the standard normal distribution has mean 0, variance 1, and characteristic function $e^{-s^2/2}$, it follows that $\boldsymbol{X}$ has mean vector $\boldsymbol{0}$, variance matrix $\boldsymbol{I}$, and characteristic function

$$\mathrm{E}(e^{i\boldsymbol{s}^*\boldsymbol{X}}) \quad = \quad \prod_{j=1}^{n} e^{-s_j^2/2} \quad = \quad e^{-\boldsymbol{s}^*\boldsymbol{s}/2}.$$

We now define any affine transformation $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\mu}$ of $\boldsymbol{X}$ to be multivariate normal [218]. This definition has several practical consequences. First, it is clear that $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{Y}) = \boldsymbol{A}\,\mathrm{Var}(\boldsymbol{X})\boldsymbol{A}^* = \boldsymbol{A}\boldsymbol{A}^* = \boldsymbol{\Omega}$. Second, any affine transformation $\boldsymbol{B}\boldsymbol{Y} + \boldsymbol{\nu} = \boldsymbol{B}\boldsymbol{A}\boldsymbol{X} + \boldsymbol{B}\boldsymbol{\mu} + \boldsymbol{\nu}$ of $\boldsymbol{Y}$ is also multivariate normal. Third, any subvector of $\boldsymbol{Y}$ is multivariate normal. Fourth, the characteristic function of $\boldsymbol{Y}$ is

$$\mathrm{E}(e^{i\boldsymbol{s}^*\boldsymbol{Y}}) \quad = \quad e^{i\boldsymbol{s}^*\boldsymbol{\mu}}\,\mathrm{E}(e^{i\boldsymbol{s}^*\boldsymbol{A}\boldsymbol{X}}) \quad = \quad e^{i\boldsymbol{s}^*\boldsymbol{\mu} - \boldsymbol{s}^*\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{s}/2} \quad = \quad e^{i\boldsymbol{s}^*\boldsymbol{\mu} - \boldsymbol{s}^*\boldsymbol{\Omega}\boldsymbol{s}/2}.$$

This enumeration omits two more subtle issues. One is whether $\boldsymbol{Y}$ possesses a density. Observe that $\boldsymbol{Y}$ lives in an affine subspace of dimension equal to or less than the rank of $\boldsymbol{A}$. Thus, if $\boldsymbol{Y}$ has m components, then $n \geq m$ must hold in order for $\boldsymbol{Y}$ to possess a density. A second issue is the existence and nature of the conditional density of a set of components of $\boldsymbol{Y}$ given the remaining components. We can clarify both of these issues by making canonical choices of $\boldsymbol{X}$ and $\boldsymbol{A}$ based on the classical QR decomposition of a matrix, which follows directly from the Gram-Schmidt orthogonalization procedure [48].

Assuming that $n \geq m$, we can write

$$\boldsymbol{A}^* \quad = \quad \boldsymbol{Q}\begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{pmatrix},$$

where $\boldsymbol{Q}$ is an $n \times n$ orthogonal matrix and $\boldsymbol{R} = \boldsymbol{L}^*$ is an $m \times m$ upper-triangular matrix with nonnegative diagonal entries. (If $n = m$, we omit the zero matrix in the QR decomposition.) It follows that

$$\boldsymbol{A}\boldsymbol{X} \quad = \quad (\,\boldsymbol{L} \quad \boldsymbol{0}^*\,)\,\boldsymbol{Q}^*\boldsymbol{X} \quad = \quad (\,\boldsymbol{L} \quad \boldsymbol{0}^*\,)\,\boldsymbol{Z}.$$

In view of the usual change-of-variables formula for probability densities and the facts that the orthogonal matrix $\boldsymbol{Q}^*$ preserves inner products and has determinant ± 1, the random vector $\boldsymbol{Z}$ has n independent standard normal components and serves as a substitute for $\boldsymbol{X}$. Not only is this true, but we can dispense with the last $n - m$ components of $\boldsymbol{Z}$ because they are multiplied by the matrix $\boldsymbol{0}^*$. Thus, we can safely assume $n = m$ and calculate the density of $\boldsymbol{Y} = \boldsymbol{L}\boldsymbol{Z} + \boldsymbol{\mu}$ when $\boldsymbol{L}$ is invertible. In this situation, $\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{L}^*$ is termed the Cholesky decomposition, and the usual change-of-variables formula shows that $\boldsymbol{Y}$ has density

$$\begin{aligned} f(\boldsymbol{y}) &= \left(\frac{1}{2\pi}\right)^{n/2} |\det \boldsymbol{L}^{-1}| e^{-(\boldsymbol{y}-\boldsymbol{\mu})^*(\boldsymbol{L}^{-1})^*\boldsymbol{L}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})/2} \\ &= \left(\frac{1}{2\pi}\right)^{n/2} |\det \boldsymbol{\Omega}|^{-1/2} e^{-(\boldsymbol{y}-\boldsymbol{\mu})^*\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})/2}, \end{aligned}$$

where $\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{L}^*$ is the variance matrix of $\boldsymbol{Y}$.

To address the issue of conditional densities, consider the compatibly partitioned vectors $\boldsymbol{Y}^* = (\boldsymbol{Y}_1^*, \boldsymbol{Y}_2^*)$, $\boldsymbol{X}^* = (\boldsymbol{X}_1^*, \boldsymbol{X}_2^*)$, and $\boldsymbol{\mu}^* = (\boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*)$ and matrices

$$\boldsymbol{L} = \begin{pmatrix} \boldsymbol{L}_{11} & \boldsymbol{0} \\ \boldsymbol{L}_{21} & \boldsymbol{L}_{22} \end{pmatrix}, \qquad \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix}.$$

Now suppose that $\boldsymbol{X}$ is standard normal, that $\boldsymbol{Y} = \boldsymbol{L}\boldsymbol{X} + \boldsymbol{\mu}$, and that $\boldsymbol{L}_{11}$ has full rank. For $\boldsymbol{Y}_1 = \boldsymbol{y}_1$ fixed, the equation $\boldsymbol{y}_1 = \boldsymbol{L}_{11}\boldsymbol{X}_1 + \boldsymbol{\mu}_1$ shows that $\boldsymbol{X}_1$ is fixed at the value $\boldsymbol{x}_1 = \boldsymbol{L}_{11}^{-1}(\boldsymbol{y}_1 - \boldsymbol{\mu}_1)$. Because no restrictions apply to $\boldsymbol{X}_2$, we have

$$\boldsymbol{Y}_2 = \boldsymbol{L}_{22}\boldsymbol{X}_2 + \boldsymbol{L}_{21}\boldsymbol{L}_{11}^{-1}(\boldsymbol{y}_1 - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_2.$$

Thus, $\boldsymbol{Y}_2$ given $\boldsymbol{Y}_1$ is normal with mean $\boldsymbol{L}_{21}\boldsymbol{L}_{11}^{-1}(\boldsymbol{y}_1 - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_2$ and variance $\boldsymbol{L}_{22}\boldsymbol{L}_{22}^*$. To express these in terms of the blocks of $\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{L}^*$, observe that

$$\begin{aligned} \boldsymbol{\Omega}_{11} &= \boldsymbol{L}_{11}\boldsymbol{L}_{11}^* \\ \boldsymbol{\Omega}_{21} &= \boldsymbol{L}_{21}\boldsymbol{L}_{11}^* \\ \boldsymbol{\Omega}_{22} &= \boldsymbol{L}_{21}\boldsymbol{L}_{21}^* + \boldsymbol{L}_{22}\boldsymbol{L}_{22}^*. \end{aligned}$$

The first two of these equations imply that $\boldsymbol{L}_{21}\boldsymbol{L}_{11}^{-1} = \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}$. The last equation then gives

$$\begin{aligned} \boldsymbol{L}_{22}\boldsymbol{L}_{22}^* &= \boldsymbol{\Omega}_{22} - \boldsymbol{L}_{21}\boldsymbol{L}_{21}^* \\ &= \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}(\boldsymbol{L}_{11}^*)^{-1}\boldsymbol{L}_{11}^{-1}\boldsymbol{\Omega}_{12} \\ &= \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}. \end{aligned}$$

These calculations do not require that $\boldsymbol{Y}_2$ possess a density. In summary, the conditional distribution of $\boldsymbol{Y}_2$ given $\boldsymbol{Y}_1$ is normal with mean and variance

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y}_2 \mid \boldsymbol{Y}_1) &= \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}(\boldsymbol{Y}_1 - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_2 \\ \mathrm{Var}(\boldsymbol{Y}_2 \mid \boldsymbol{Y}_1) &= \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}. \end{aligned} \tag{A.1}$$

A.3 Polyhedral Sets

A polyhedral set S is the nonempty intersection of a finite number of half-spaces. Symbolically, S can be represented as

$$S = \{\boldsymbol{x} \in \mathsf{R}^m : \boldsymbol{v}_i^* \boldsymbol{x} \leq c_i \text{ for } 1 \leq i \leq p\}. \tag{A.2}$$

As previously noted, S is closed and convex. If all $c_i = 0$, then S is said to be a polyhedral cone. We now prove a sequence of propositions that lay out the basic facts about polyhedral cones and sets. Consult Examples 2.4.1 and 14.3.7 for background material on convex cones.

Proposition A.3.1 *The polar cone of a polyhedral cone is a finitely generated cone and vice versa. In matrix notation, the cones* $\{\boldsymbol{x} : \boldsymbol{V}^* \boldsymbol{x} \leq \boldsymbol{0}\}$ *and* $\{\boldsymbol{V}\boldsymbol{a} : \boldsymbol{a} \geq \boldsymbol{0}\}$ *constitute a polar pair.*

Proof: Assume the polyhedral set (A.2) is a cone. Let us show that its polar cone is

$$T = \left\{\boldsymbol{y} \in \mathsf{R}^m : \boldsymbol{y} = \sum_{i=1}^{p} a_i \boldsymbol{v}_i,\ a_i \geq 0 \text{ for all } i\right\}.$$

According to Example 2.4.1, the finitely generated cone T is closed and convex. In view of Example 14.3.7, it therefore suffices to prove that the polar cone T° coincides with S. But this follows from the simple observation that $\boldsymbol{x}^* \boldsymbol{v}_i \leq 0$ for all $\boldsymbol{v}_i$ if and only if $\boldsymbol{x}^* (\sum_{i=1}^{p} a_i \boldsymbol{v}_i) \leq 0$ for all conical combinations of the $\boldsymbol{v}_i$. ■

Proposition A.3.2 *A cone is polyhedral if and only if it is finitely generated.*

Proof: Consider a cone C generated by the vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ in R^n. Let us prove by induction on p that C is a polyhedral cone. Suppose $p = 1$ and $\boldsymbol{u}_1 = \boldsymbol{0}$. If $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ is the standard basis of R^n, then

$$C = \{\boldsymbol{0}\} = \{\boldsymbol{x} : \boldsymbol{e}_i^* \boldsymbol{x} \leq 0,\ (-\boldsymbol{e}_i)^* \boldsymbol{x} \leq 0, \text{ for all } i\}.$$

If $p = 1$ and $\boldsymbol{u}_1 \neq \boldsymbol{0}$, then choose an orthogonal basis $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ with $\boldsymbol{w}_1 = -\boldsymbol{u}_1$. In this basis, we have

$$C = \{\boldsymbol{x} : \boldsymbol{w}_1^* \boldsymbol{x} \leq 0,\ \boldsymbol{w}_i^* \boldsymbol{x} \leq 0,\ (-\boldsymbol{w}_i)^* \boldsymbol{x} \leq 0, \text{ for all } i \geq 2\}.$$

Now assume $p > 1$, and let K be the cone generated by $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{p-1}$. By the induction hypothesis, K has polyhedral representation

$$K = \{\boldsymbol{x} : \boldsymbol{v}_i^* \boldsymbol{x} \leq 0,\ 1 \leq i \leq q\}.$$

Furthermore, the cone C generated by $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ satisfies

$$\begin{aligned} C &= \{\boldsymbol{x} : \boldsymbol{x} - t\boldsymbol{u}_p \in K \text{ for some } t \geq 0\} \\ &= \{\boldsymbol{x} : \boldsymbol{v}_i^*(\boldsymbol{x} - t\boldsymbol{u}_p) \leq 0 \text{ for all } i \text{ and some } t \geq 0\} \\ &= \{\boldsymbol{x} : \boldsymbol{v}_i^*\boldsymbol{x} \leq t\boldsymbol{v}_i^*\boldsymbol{u}_p \text{ for all } i \text{ and some } t \geq 0\}. \end{aligned}$$

To represent C as a polyhedral cone, we eliminate the variable t by the Fourier–Motzkin maneuver. Accordingly, define the index sets

$$I_- = \{i : \boldsymbol{v}_i^*\boldsymbol{u}_p < 0\}, \; I_0 = \{i : \boldsymbol{v}_i^*\boldsymbol{u}_p = 0\}, \; I_+ = \{i : \boldsymbol{v}_i^*\boldsymbol{u}_p > 0\}.$$

In this notation $\boldsymbol{x} \in C$ if and only if there exists $t \geq 0$ with

$$\frac{\boldsymbol{v}_i^*\boldsymbol{x}}{\boldsymbol{v}_i^*\boldsymbol{u}_p} \leq t \leq \frac{\boldsymbol{v}_j^*\boldsymbol{x}}{\boldsymbol{v}_j^*\boldsymbol{u}_p}, \qquad \boldsymbol{v}_k^*\boldsymbol{x} \leq 0$$

for all $i \in I_+$, $j \in I_-$, and $k \in I_0$. Evidently, an appropriate $t \geq 0$ exists if and only if

$$\max_{i \in I_+} \frac{\boldsymbol{v}_i^*\boldsymbol{x}}{\boldsymbol{v}_i^*\boldsymbol{u}_p} \leq \min_{j \in I_-} \frac{\boldsymbol{v}_j^*\boldsymbol{x}}{\boldsymbol{v}_j^*\boldsymbol{u}_p} \quad \text{and} \quad \min_{j \in I_-} \frac{\boldsymbol{v}_j^*\boldsymbol{x}}{\boldsymbol{v}_j^*\boldsymbol{u}_p} \geq 0.$$

Because $\boldsymbol{v}_j^*\boldsymbol{u}_p < 0$ for $j \in I_-$, the second of the last two inequalities is equivalent to $\boldsymbol{v}_j^*\boldsymbol{x} \leq 0$ for all $j \in I_-$. It follows that C can be represented as the polyhedral cone

$$C = \left\{\boldsymbol{x} : \boldsymbol{v}_k^*\boldsymbol{x} \leq 0, \; \frac{\boldsymbol{v}_i^*\boldsymbol{x}}{\boldsymbol{v}_i^*\boldsymbol{u}_p} \leq \frac{\boldsymbol{v}_j^*\boldsymbol{x}}{\boldsymbol{v}_j^*\boldsymbol{u}_p}, \; k \in I_- \cup I_0, i \in I_+, j \in I_-\right\}$$

involving no mention of the scalar t.

Conversely, if C is a polyhedral cone, then Proposition A.3.1 implies that C° is a finitely generated cone. By the argument just given, C° is also a polyhedral cone. A second application of Proposition A.3.1 shows that $C^{\circ\circ} = C$ is a finitely generated cone. ■

Proposition A.3.3 (Minkowski–Weyl) *A nonempty set S is polyhedral if and only if it can be represented as*

$$S = \left\{\boldsymbol{x} : \boldsymbol{x} = \sum_{i=1}^{q} a_i\boldsymbol{u}_i + \sum_{j=1}^{r} b_j\boldsymbol{w}_j, \; \sum_{j=1}^{r} b_j = 1, \text{ all } a_i \geq 0, \; b_j \geq 0\right\}. \quad \text{(A.3)}$$

In other words, S is the algebraic sum of a finitely generated cone and the convex hull of a finite set of points.

Proof: Consider the polyhedral set S appearing in equation (A.2), and define the polyhedral cone

$$\begin{aligned} T &= \{(\boldsymbol{x}, t) : t \geq 0 \text{ and } \boldsymbol{v}_i^*\boldsymbol{x} \leq c_i t \text{ for } 1 \leq i \leq p\} \\ &= \{(\boldsymbol{x}, t) : t \geq 0 \text{ and } \boldsymbol{v}_i^*\boldsymbol{x} - c_i t \leq 0 \text{ for } 1 \leq i \leq p\}. \end{aligned}$$

Proposition A.3.2 implies that T is a finitely generated cone with generators $(\boldsymbol{u}_i, 0)$ and $(\boldsymbol{w}_j, 1)$ for $1 \le i \le q$ and $1 \le j \le r$. The representation (A.3) is now a consequence of the fact that $S = \{\boldsymbol{x} : (\boldsymbol{x}, 1) \in T\}$.

For the converse, suppose S is a set of the form (A.3). Define T to be the cone generated by the vectors $(\boldsymbol{u}_i, 0)$ and $(\boldsymbol{w}_j, 1)$ for $1 \le i \le q$ and $1 \le j \le r$. Proposition A.3.2 identifies T as a polyhedral cone satisfying

$$T = \{(\boldsymbol{x}, t) : t \ge 0 \text{ and } \boldsymbol{v}_i^* \boldsymbol{x} \le c_i t \text{ for } 1 \le i \le p\}$$

for appropriate vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ and corresponding scalars $c_1, \ldots, c_p$. But this means that $S = \{\boldsymbol{x} : (\boldsymbol{x}, 1) \in T\}$ is also a polyhedral set. ■

Proposition A.3.4 *The collection of polyhedral sets enjoys the following closure properties:*

(a) *The nonempty intersection of two polyhedral sets is polyhedral.*

(b) *The inverse image of a polyhedral set under a linear transformation is polyhedral.*

(c) *The Cartesian product of two polyhedral sets is polyhedral.*

(d) *The image of a polyhedral set under a linear transformation is polyhedral.*

(e) *The vector sum of two polyhedral sets is polyhedral.*

Proof: Assertion (a) is obvious. Consider the polyhedral set (A.2). If $\boldsymbol{M}$ is a linear transformation from R^n to R^m, then the set equality

$$\begin{aligned} \boldsymbol{M}^{-1}(S) &= \{\boldsymbol{x} \in \mathsf{R}^n : \boldsymbol{v}_i^*(\boldsymbol{M}\boldsymbol{x}) \le c_i \text{ for } 1 \le i \le p\} \\ &= \{\boldsymbol{x} \in \mathsf{R}^n : (\boldsymbol{M}^* \boldsymbol{v}_i)^* \boldsymbol{x} \le c_i \text{ for } 1 \le i \le p\} \end{aligned}$$

proves assertion (b). To verify assertion (c), consider a second polyhedral set

$$T = \left\{\boldsymbol{y} \in \mathsf{R}^n : \boldsymbol{w}_j^* \boldsymbol{y} \le d_j \text{ for } 1 \le j \le q\right\}.$$

The Cartesian product amounts to

$$S \times T = \{(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{v}_i^* \boldsymbol{x} \le c_i,\ \boldsymbol{w}_j^* \boldsymbol{y} \le d_j \text{ for all possible } i \text{ and } j\}.$$

To prove assertion (d), consider the Minkowski–Weyl representation

$$S = \left\{\boldsymbol{x} \in \mathsf{R}^m : \boldsymbol{x} = \sum_{i=1}^{p} c_i \boldsymbol{v}_i + \sum_{j=1}^{q} d_j \boldsymbol{w}_j\right\},$$

where the first sum ranges over all convex combinations and the second sum ranges over all conical combinations. If $\boldsymbol{N}$ is a linear transformation from R^m to R^n, then

$$\boldsymbol{N}(S) \quad = \quad \Big\{\boldsymbol{y} \in \mathsf{R}^n : \boldsymbol{y} = \sum_{i=1}^{p} c_i \boldsymbol{N}\boldsymbol{v}_i + \sum_{j=1}^{q} d_j \boldsymbol{N}\boldsymbol{w}_j\Big\}$$

is a Minkowski–Weyl representation of the image. Finally, to prove assertion (e), note that $S+T$ is the image of the Cartesian product $S \times T$ under the linear map $(\boldsymbol{v}, \boldsymbol{w}) \mapsto \boldsymbol{v} + \boldsymbol{w}$. Application of properties (c) and (d) complete the proof. ■

Proposition A.3.5 *Let $f(\boldsymbol{x})$ be a convex function with domain a polyhedral set S. If $\sup_{\boldsymbol{x}\in S} f(\boldsymbol{x}) < \infty$, then $f(\boldsymbol{x})$ attains its maximum over S at one of the points $\boldsymbol{w}_j$ defining the convex hull part of S.*

Proof: Assume that S has Minkowski–Weyl representation (A.3) and that $M = \sup_{\boldsymbol{x}\in S} f(\boldsymbol{x})$. If $\boldsymbol{x} \in S$ and $\boldsymbol{u}_i$ is one of the vectors defining the conical part of S, then $\boldsymbol{x} + a_i\boldsymbol{u}_i \in S$ for all $a_i \geq 0$. Furthermore, for $t \geq 1$ we have

$$\begin{aligned} f(\boldsymbol{x} + a_i\boldsymbol{u}_i) &\leq \Big(1 - \frac{1}{t}\Big) f(\boldsymbol{x}) + \frac{1}{t} f(\boldsymbol{x} + t a_i \boldsymbol{u}_i) \\ &\leq \Big(1 - \frac{1}{t}\Big) f(\boldsymbol{x}) + \frac{M}{t}. \end{aligned}$$

Sending t to ∞ shows that $f(\boldsymbol{x} + a_i\boldsymbol{u}_i) \leq f(\boldsymbol{x})$. Hence, we may confine our attention to those points in the representation (A.3) with all $a_i = 0$. With this understanding, Jensen's inequality

$$f\Big(\sum_{j=1}^{r} b_j \boldsymbol{w}_j\Big) \quad \leq \quad \sum_{j=1}^{r} b_j f(\boldsymbol{w}_j) \quad \leq \quad \max\{f(\boldsymbol{w}_1), \ldots, f(\boldsymbol{w}_r)\}$$

demonstrates that $f(\boldsymbol{x})$ attains its maximum over S at one of the points $\boldsymbol{w}_j$ defining the convex hull part of S. ■

A.4 Birkhoff's Theorem and Fan's Inequality

Birkhoff's theorem deals with the set Γ^n of $n \times n$ doubly stochastic matrices. Every matrix $\boldsymbol{M} = (m_{ij})$ in Γ^n has nonnegative entries and row and column sums equal to 1. The affine constraints defining Γ^n compel it to be a compact polyhedral set with Minkowski–Weyl representation

$$\Gamma^n \quad = \quad \Big\{\boldsymbol{x} : \boldsymbol{x} = \sum_{j=1}^{r} b_j \boldsymbol{w}_j,\ \sum_{j=1}^{r} b_j = 1,\ b_j \geq 0\Big\} \tag{A.4}$$

lacking a conical part. See Proposition A.3.3. Compact polyhedral sets are called convex polytopes and are characterized by their extreme points.

A point $\boldsymbol{x}$ in a convex set S is said to be extreme if it cannot be written as a nontrivial convex combination of two points from S. It turns out that the Minkowski–Weyl vectors $\boldsymbol{w}_i$ in a convex polytope can be taken to be extreme points. Indeed, suppose $\boldsymbol{w}_i$ can be represented as

$$\boldsymbol{w}_i \;=\; \alpha \sum_{j=1}^{r} a_j \boldsymbol{w}_j + (1-\alpha) \sum_{j=1}^{r} b_j \boldsymbol{w}_j$$

with $\alpha \in (0,1)$. Either $a_i < 1$ or $b_i < 1$; otherwise, the two points on the right of the equation coincide with $\boldsymbol{w}_i$. Subtracting $[\alpha a_i + (1-\alpha) b_i]\boldsymbol{w}_i$ from both sides of the equation and rescaling give $\boldsymbol{w}_i$ as a convex combination $\sum_{j \neq i} c_j \boldsymbol{w}_j$. In any convex combination $\boldsymbol{v}$ of the vectors $\{\boldsymbol{w}_j\}_{j=1}^r$, one can replace $\boldsymbol{w}_i$ by this convex combination and represent $\boldsymbol{v}$ by a convex combination of the remaining vectors $\boldsymbol{w}_j \neq \boldsymbol{w}_i$. If any non-extreme points remain after deletion of $\boldsymbol{w}_i$, then this substitution and reduction process can be repeated. Ultimately it halts with a set of vectors $\boldsymbol{w}_j$ composed entirely of extreme points. In fact, these are the only extreme points of the convex polytope.

Birkhoff's theorem identifies the permutation matrices as the extreme points of Γ^n. A permutation matrix $\boldsymbol{P} = (p_{ij})$ has entries drawn from the set $\{0,1\}$. Each of its rows and columns has exactly one entry equal to 1. The permutation matrices do not exhaust Γ^n. For instance, the matrix $\frac{1}{n}\mathbf{1}\mathbf{1}^*$ belongs to Γ^n. For another example, take any orthogonal matrix $\boldsymbol{U} = (u_{ij})$ and form the matrix $\boldsymbol{M}$ with entries $m_{ij} = u_{ij}^2$. This matrix resides in Γ^n as well.

Proposition A.4.1 (Birkhoff) *Every doubly stochastic matrix can be represented as a convex combination of permutation matrices.*

Proof: It suffices to prove that the permutation matrices are the extreme points of Γ^n. Suppose the permutation matrix $\boldsymbol{P}$ satisfies

$$\boldsymbol{P} \;=\; \alpha \boldsymbol{Q} + (1-\alpha)\boldsymbol{R}$$

for two doubly stochastic matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ and $\alpha \in (0,1)$. If an entry p_{ij} equals 1, then the two corresponding entries q_{ij} and r_{ij} of $\boldsymbol{Q}$ and $\boldsymbol{R}$ must also equal 1. Likewise, if an entry p_{ij} equals 0, then the corresponding entries q_{ij} and r_{ij} of $\boldsymbol{Q}$ and $\boldsymbol{R}$ must also equal 0. Thus, both $\boldsymbol{Q}$ and $\boldsymbol{R}$ coincide with $\boldsymbol{P}$. As a consequence, $\boldsymbol{P}$ is an extreme point.

Conversely, suppose $\boldsymbol{M} = (m_{ij})$ is an extreme point that is not a permutation matrix. Take any index pair (i_0, j_0) with $0 < m_{i_0 j_0} < 1$. Because every row sum equals 1, there is an index $j_1 \neq j_0$ with $0 < m_{i_0 j_1} < 1$. Similarly, there is index i_1 with $0 < m_{i_1 j_1} < 1$. This process of jumping

along a row and then along a column creates a path from index pair to index pair. Eventually, the path intersects itself. Take a closed circuit

$$(i_k, j_k) \to (i_{k+1}, j_k) \to \cdots \to (i_l, j_l) = (i_k, j_k)$$

or

$$(i_{k+1}, j_k) \to (i_{k+1}, j_{k+1}) \to \cdots \to (i_l, j_l) = (i_{k+1}, j_k)$$

and construct a matrix $\boldsymbol{N}$ whose entries are 0 except for entries along the path. For these special entries alternate the values 1 and -1. It is clear that this construction forces $\boldsymbol{N}$ to have row and column sums equal to 0. Because the entries of $\boldsymbol{M}$ along the path occur in the open interval $(0, 1)$, there exists a positive constant ϵ such that $\boldsymbol{A} = \boldsymbol{M} + \epsilon\boldsymbol{N}$ and $\boldsymbol{B} = \boldsymbol{M} - \epsilon\boldsymbol{N}$ are both doubly stochastic. The representation

$$\boldsymbol{M} \quad = \quad \frac{1}{2}(\boldsymbol{A} + \boldsymbol{B})$$

now demonstrates that $\boldsymbol{M}$ is not an extreme point. Hence, only permutation matrices can be extreme points. ■

As a prelude to stating Fan's inequality, we first prove a classic rearrangement theorem of Hardy, Littlewood, and Pólya [118]. Consider two increasing sequences $a_1 \leq a_2 \leq \cdots \leq a_n$ and $b_1 \leq b_2 \leq \cdots \leq b_n$. If σ is any permutation of $\{1, \ldots, n\}$, then the theorem says

$$\sum_{i=1}^{n} a_i b_{\sigma(i)} \quad \leq \quad \sum_{i=1}^{n} a_i b_i. \tag{A.5}$$

To quote the celebrated trio:

> The theorem becomes obvious if we interpret the a_i as distances along a rod to hooks and the b_i as weights suspended from the hooks. To get the maximum statical moment with respect to the end of the rod, we hang the heaviest weights on the hooks farthest from the end.

To prove the result, suppose the a_i are in ascending order, but the b_i are not. Then there are indices $j < k$ with $a_j \leq a_k$ and $b_j > b_k$. Because

$$a_j b_k + a_k b_j - (a_j b_j + a_k b_k) \quad = \quad (a_k - a_j)(b_j - b_k) \quad \geq \quad 0,$$

we can increase the sum by exchanging b_j and b_k. A finite number of such exchanges (transpositions) puts the b_i into ascending order.

Proposition A.4.2 (von Neumann–Fan) *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $n \times n$ symmetric matrices with ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and $\mu_1 \geq \cdots \geq \mu_n$. Then*

$$\operatorname{tr}(\boldsymbol{AB}) \quad \leq \quad \lambda_1\mu_1 + \cdots + \lambda_n\mu_n. \tag{A.6}$$

Equality holds in inequality (A.6) if and only if

$$\boldsymbol{A} \;=\; \boldsymbol{W}\boldsymbol{D}_A\boldsymbol{W}^* \quad \textit{and} \quad \boldsymbol{B} \;=\; \boldsymbol{W}\boldsymbol{D}_B\boldsymbol{W}^*$$

are simultaneously diagonalizable by an orthogonal matrix $\boldsymbol{W}$ and diagonal matrices $\boldsymbol{D}_A$ and $\boldsymbol{D}_B$ whose entries are ordered from largest to smallest.

Proof: There exists a pair of orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ with

$$\boldsymbol{A} \;=\; \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^* \quad \text{and} \quad \boldsymbol{B} \;=\; \sum_{i=1}^{n} \mu_i \boldsymbol{v}_i \boldsymbol{v}_i^*.$$

It follows that

$$\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) \;=\; \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \mu_j \operatorname{tr}(\boldsymbol{u}_i \boldsymbol{u}_i^* \boldsymbol{v}_j \boldsymbol{v}_j^*).$$

The matrix $\boldsymbol{C}$ with entries $c_{ij} = \operatorname{tr}(\boldsymbol{u}_i \boldsymbol{u}_i^* \boldsymbol{v}_j \boldsymbol{v}_j^*)$ is doubly stochastic because its entries $c_{ij} = (\boldsymbol{u}_i^* \boldsymbol{v}_j)^2$ are nonnegative and the column sums satisfy

$$\sum_{i=1}^{n} \operatorname{tr}(\boldsymbol{u}_i \boldsymbol{u}_i^* \boldsymbol{v}_j \boldsymbol{v}_j^*) \;=\; \operatorname{tr}\Big(\sum_{i=1}^{n} \boldsymbol{u}_i \boldsymbol{u}_i^* \boldsymbol{v}_j \boldsymbol{v}_j^*\Big) \;=\; \operatorname{tr}(\boldsymbol{I}_n \boldsymbol{v}_j \boldsymbol{v}_j^*) \;=\; \boldsymbol{v}_j^* \boldsymbol{v}_j \;=\; 1.$$

Virtually the same argument shows that the row sums equal 1. Proposition A.3.3 therefore implies that $\boldsymbol{C}$ is a convex combination $\sum_k \alpha_k \boldsymbol{P}_k$ of permutation matrices. This representation gives

$$\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) \;=\; \boldsymbol{\lambda}^* \boldsymbol{C} \boldsymbol{\mu} \;=\; \sum_k \alpha_k \boldsymbol{\lambda}^* \boldsymbol{P}_k \boldsymbol{\mu} \;\le\; \sum_k \alpha_k \sum_{i=1}^{n} \lambda_i \mu_i \;=\; \sum_{i=1}^{n} \lambda_i \mu_i$$

in view of the Hardy–Littlewood–Pólya rearrangement inequality (A.5).

Equality holds in inequality (A.6) under the stated conditions because

$$\operatorname{tr}(\boldsymbol{W}\boldsymbol{D}_A\boldsymbol{W}^*\boldsymbol{W}\boldsymbol{D}_B\boldsymbol{W}^*) \;=\; \operatorname{tr}(\boldsymbol{D}_A\boldsymbol{D}_B).$$

Here we apply the cyclic permutation property of the trace function and the identity $\boldsymbol{W}^*\boldsymbol{W} = \boldsymbol{I}_n$.

Conversely, suppose inequality (A.6) is an equality. Following Theobald [254], let $\boldsymbol{E} = \boldsymbol{A} + \boldsymbol{B}$ have ordered spectral decomposition $\boldsymbol{E} = \boldsymbol{W}\boldsymbol{D}_E\boldsymbol{W}^*$ with ρ_i the ith diagonal entry of $\boldsymbol{D}_E$. We now show that $\boldsymbol{A}$ and $\boldsymbol{B}$ have ordered spectral decompositions $\boldsymbol{W}\boldsymbol{D}_A\boldsymbol{W}^*$ and $\boldsymbol{W}\boldsymbol{D}_B\boldsymbol{W}^*$, respectively. The first half of the proposition and the hypothesis imply

$$\begin{aligned} \operatorname{tr}(\boldsymbol{W}\boldsymbol{D}_A\boldsymbol{W}^*\boldsymbol{B}) \;&\le\; \sum_{i=1}^{n} \lambda_i \mu_i \;=\; \operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) \\ \operatorname{tr}(\boldsymbol{A}\boldsymbol{E}) \;&\le\; \sum_{i=1}^{n} \lambda_i \rho_i \;=\; \operatorname{tr}(\boldsymbol{D}_A\boldsymbol{D}_E). \end{aligned}$$

It follows that

$$\begin{aligned}
\operatorname{tr}(\boldsymbol{W}\boldsymbol{D}_A\boldsymbol{W}^*\boldsymbol{A}) &= \operatorname{tr}(\boldsymbol{W}\boldsymbol{D}_A\boldsymbol{W}^*\boldsymbol{E}) - \operatorname{tr}(\boldsymbol{W}\boldsymbol{D}_A\boldsymbol{W}^*\boldsymbol{B}) \\
&\geq \operatorname{tr}(\boldsymbol{D}_A\boldsymbol{W}^*\boldsymbol{E}\boldsymbol{W}) - \operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) \\
&= \operatorname{tr}(\boldsymbol{D}_A\boldsymbol{D}_E) - \operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) \\
&\geq \operatorname{tr}(\boldsymbol{A}\boldsymbol{E}) - \operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) \\
&= \operatorname{tr}(\boldsymbol{A}^2).
\end{aligned}$$

Since $\operatorname{tr}(\boldsymbol{A}^2) = \|\boldsymbol{A}\|_F^2$ and $\|\boldsymbol{W}\boldsymbol{D}_A\boldsymbol{W}^*\|_F = \|\boldsymbol{D}_A\|_F = \|\boldsymbol{A}\|_F$, the Cauchy-Schwarz inequality for the Frobenius inner product now gives

$$\|\boldsymbol{A}\|_F^2 = \|\boldsymbol{W}\boldsymbol{D}_A\boldsymbol{W}^*\|_F\|\boldsymbol{A}\|_F \geq \operatorname{tr}(\boldsymbol{W}\boldsymbol{D}_A\boldsymbol{W}^*\boldsymbol{A}) \geq \|\boldsymbol{A}\|_F^2.$$

Because equality must hold in this inequality, the standard necessary condition for equality in the Cauchy-Schwarz inequality forces $\boldsymbol{W}\boldsymbol{D}_A\boldsymbol{W}^*$ to equal $c\boldsymbol{A}$ for some constant c. In fact, $c = 1$ because $\boldsymbol{W}\boldsymbol{D}_A\boldsymbol{W}^*$ has the same norm as $\boldsymbol{A}$. A similar argument implies that $\boldsymbol{B} = \boldsymbol{W}\boldsymbol{D}_B\boldsymbol{W}^*$. ■

Another proof of the sufficiency half of Proposition A.4.2 is possibly more illuminating [172]. Again let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $n \times n$ symmetric matrices with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and $\mu_1 \geq \cdots \geq \mu_n$. One can easily show that the set of symmetric matrices with the same eigenvalues as $\boldsymbol{A}$ is compact. Therefore the continuous function $\boldsymbol{M} \mapsto \operatorname{tr}(\boldsymbol{M}\boldsymbol{B})$ achieves its maximum over the set at some matrix $\boldsymbol{A}_{max}$. Now take any skew symmetric matrix $\boldsymbol{C}$ and consider the one-parameter family of matrices $\boldsymbol{A}(t) = e^{t\boldsymbol{C}}\boldsymbol{A}_{max}e^{-t\boldsymbol{C}}$ similar to $\boldsymbol{A}_{max}$. Since $\boldsymbol{C}^* = -\boldsymbol{C}$, the matrix exponential $e^{t\boldsymbol{C}}$ is orthogonal with transpose $e^{-t\boldsymbol{C}}$. The optimality of $\boldsymbol{A}_{max}$ implies that

$$\begin{aligned}
\frac{d}{dt}\operatorname{tr}[\boldsymbol{A}(t)\boldsymbol{B}]\Big|_{t=o} &= \operatorname{tr}\{[\boldsymbol{C}\boldsymbol{A}_{max} - \boldsymbol{A}_{max}\boldsymbol{C}]\boldsymbol{B}\} \\
&= \operatorname{tr}[\boldsymbol{C}(\boldsymbol{A}_{max}\boldsymbol{B} - \boldsymbol{B}\boldsymbol{A}_{max})]
\end{aligned}$$

vanishes. This suggests taking $\boldsymbol{C}$ equal to the skew symmetric commutator matrix $\boldsymbol{A}_{max}\boldsymbol{B} - \boldsymbol{B}\boldsymbol{A}_{max}$. It then follows that

$$0 = \operatorname{tr}(\boldsymbol{C}\boldsymbol{C}) = -\operatorname{tr}(\boldsymbol{C}\boldsymbol{C}^*) = -\|\boldsymbol{C}\|_F^2.$$

In other words $\boldsymbol{C}$ vanishes, and $\boldsymbol{A}_{max}$ and $\boldsymbol{B}$ commute. Commuting symmetric matrices can be simultaneously diagonalized by a common orthogonal matrix. Hence,

$$\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) \leq \operatorname{tr}(\boldsymbol{A}_{max}\boldsymbol{B}) = \sum_{i=1}^{n} \lambda_{\sigma(i)}\mu_i$$

for some permutation σ. Application of inequality (A.5) finishes the proof. The next proposition summarizes the foregoing discussion.

Proposition A.4.3 *Consider an $n \times n$ symmetric matrix $\boldsymbol{B}$ with ordered spectral decomposition $\boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\mu})\boldsymbol{U}^*$. On the compact set of $n \times n$ symmetric matrices with ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$, the linear function*

$$\boldsymbol{A} \quad \mapsto \quad \mathrm{tr}(\boldsymbol{AB})$$

attains its maximum of $\sum_{i=1}^n \lambda_i \mu_i$ at the point $\boldsymbol{A} = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{U}^*$.

A.5 Singular Value Decomposition

In many statistical applications involving large data sets, statisticians are confronted with a large $m \times n$ matrix $\boldsymbol{X} = (x_{ij})$ that encodes n features on each of m objects. For instance, in gene microarray studies x_{ij} represents the expression level of the ith gene under the jth experimental condition [190]. In information retrieval, x_{ij} represents the frequency of the jth word or term in the ith document [11]. The singular value decomposition (svd) captures the structure of such matrices. In many applications there are alternatives to the svd, but these are seldom as informative. From the huge literature on the svd, the books [64, 105, 107, 136, 137, 232, 249, 260] are especially recommended.

The spectral theorem for symmetric matrices discussed in Example 1.4.3 states that an $m \times m$ symmetric matrix $\boldsymbol{M}$ can be written as $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^*$ for an orthogonal matrix $\boldsymbol{U}$ and a diagonal matrix $\boldsymbol{\Lambda}$ with diagonal entries λ_i. If $\boldsymbol{U}$ has columns $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$, then the matrix product $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^*$ unfolds into the sum of outer products

$$\boldsymbol{M} \quad = \quad \sum_{j=1}^m \lambda_j \boldsymbol{u}_j \boldsymbol{u}_j^*.$$

When $\lambda_j \neq 0$ for $j \leq k$ and $\lambda_j = 0$ for $j > k$, $\boldsymbol{M}$ has rank k and only the first k terms of the sum are relevant. The svd seeks to generalize the spectral theorem to nonsymmetric matrices. In this case there are two orthonormal sets of vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ instead of one, and we write

$$\boldsymbol{M} \quad = \quad \sum_{j=1}^k \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^* \quad = \quad \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* \tag{A.7}$$

for matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ with orthonormal columns $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$, respectively. If $\sigma_j < 0$, one can exchange $-\boldsymbol{u}_j$ for $\boldsymbol{u}_j$ and $-\sigma_j$ for σ_j. Hence, it is possible to take the σ_j to be nonnegative in the representation (A.7).

For some purposes, it is better to fill out the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ to full orthogonal matrices. If $\boldsymbol{M}$ is $m \times n$, then $\boldsymbol{U}$ is viewed as $m \times m$, $\boldsymbol{\Sigma}$ as

$m \times n$, and $\boldsymbol{V}$ as $n \times n$. The svd then becomes

$$\boldsymbol{M} \;=\; (\boldsymbol{u}_1 \ldots \boldsymbol{u}_k\, \boldsymbol{u}_{k+1} \ldots \boldsymbol{u}_m) \begin{pmatrix} \sigma_1^2 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_k^2 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & 0 & \ldots & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_1^* \\ \vdots \\ \boldsymbol{v}_k^* \\ \boldsymbol{v}_{k+1}^* \\ \vdots \\ \boldsymbol{v}_n^* \end{pmatrix},$$

assuming $k < \min\{m, n\}$. The scalars $\sigma_1, \ldots, \sigma_k$ are said to be singular values and conventionally are listed in decreasing order. The vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ are known as left singular vectors and the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ as right singular vectors.

To prove that the svd of an $m \times n$ matrix $\boldsymbol{M}$ exists, consider the symmetric matrix product $\boldsymbol{M}^*\boldsymbol{M}$. Let $\boldsymbol{M}^*\boldsymbol{M}$ have spectral decomposition $\boldsymbol{V}\boldsymbol{\Omega}\boldsymbol{V}^*$ with the eigenvalues σ_i^2 arranged from greatest to least along the diagonal of $\boldsymbol{\Omega}$. The calculation

$$(\boldsymbol{M}\boldsymbol{v}_i)^*\boldsymbol{M}\boldsymbol{v}_j \;=\; \boldsymbol{v}_i^*\boldsymbol{M}^*\boldsymbol{M}\boldsymbol{v}_j \;=\; \sigma_j^2\boldsymbol{v}_i^*\boldsymbol{v}_j \;=\; \begin{cases} 0 & i \neq j \\ \sigma_i^2 & i = j \end{cases}$$

shows that the vectors $\boldsymbol{M}\boldsymbol{v}_i$ are orthogonal. Furthermore, when $\sigma_i^2 > 0$, the normalized vector $\boldsymbol{u}_i = \sigma_i^{-1}\boldsymbol{M}\boldsymbol{v}_i$ is a unit vector. If we suppose that $\sigma_k > 0$ but $\sigma_{k+1} = 0$, then the representation (A.7) is valid because

$$\left(\sum_{j=1}^{k} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^*\right)\boldsymbol{v}_i \;=\; \boldsymbol{M}\boldsymbol{v}_i$$

for all vectors in the orthonormal basis $\{\boldsymbol{v}_i\}_{i=1}^n$.

Fan's inequality generalizes to nonsymmetric matrices provided one substitutes ordered singular values for ordered eigenvalues. Let us first reduce the problem to square matrices. If $\boldsymbol{M}$ is $m \times n$ with $n < m$, then the svd $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ translates into the svd

$$(\,\boldsymbol{M} \quad \boldsymbol{0}\,) \;=\; \boldsymbol{U}\,(\,\boldsymbol{\Sigma} \quad \boldsymbol{0}\,) \begin{pmatrix} \boldsymbol{V}^* & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{m-n} \end{pmatrix}$$

preserving the nontrivial singular values. Furthermore, the trace of two such expanded matrices satisfies

$$\operatorname{tr}\left[\,(\,\boldsymbol{M} \quad \boldsymbol{0}\,)^*\,(\,\boldsymbol{N} \quad \boldsymbol{0}\,)\right] \;=\; \operatorname{tr}(\boldsymbol{M}^*\boldsymbol{N}).$$

A similar representation applies when $m < n$.

We now reduce the problem to symmetric matrices. Suppose $\boldsymbol{M}$ is an $m \times m$ square matrix with svd $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. The matrix

$$\boldsymbol{K} \;=\; \begin{pmatrix} \boldsymbol{0} & \boldsymbol{M} \\ \boldsymbol{M}^* & \boldsymbol{0} \end{pmatrix}$$

is a $2m \times 2m$ symmetric matrix, and the $2m$ vectors

$$\boldsymbol{w}_i = \frac{1}{\sqrt{2}} \begin{pmatrix} \boldsymbol{u}_i \\ \boldsymbol{v}_i \end{pmatrix} \quad \text{and} \quad \boldsymbol{z}_i = \frac{1}{\sqrt{2}} \begin{pmatrix} -\boldsymbol{u}_i \\ \boldsymbol{v}_i \end{pmatrix}$$

are orthogonal unit vectors. Because $\boldsymbol{M}^*\boldsymbol{u}_i = \sigma_i \boldsymbol{v}_i$ and $\boldsymbol{M}\boldsymbol{v}_i = \sigma_i \boldsymbol{u}_i$, the vector $\boldsymbol{w}_i$ is an eigenvector of $\boldsymbol{K}$ with eigenvalue σ_i, and the vector $\boldsymbol{z}_i$ is an eigenvector of $\boldsymbol{K}$ with eigenvalue $-\sigma_i$. Hence, we have found the spectral decomposition of $\boldsymbol{K}$.

Now let $\boldsymbol{N}$ be a second $m \times m$ matrix with svd $\boldsymbol{P}\boldsymbol{\Omega}\boldsymbol{Q}^*$ and a similar expansion to a $2m \times 2m$ symmetric matrix $\boldsymbol{L}$. Fan's inequality for symmetric matrices gives

$$\begin{aligned} 2\,\mathrm{tr}(\boldsymbol{M}^*\boldsymbol{N}) &= \mathrm{tr}(\boldsymbol{M}^*\boldsymbol{N}) + \mathrm{tr}(\boldsymbol{N}\boldsymbol{M}^*) \\ &= \mathrm{tr}(\boldsymbol{M}^*\boldsymbol{N}) + \mathrm{tr}(\boldsymbol{M}\boldsymbol{N}^*) \\ &= \mathrm{tr}(\boldsymbol{K}^*\boldsymbol{L}) \\ &\leq \sum_{i=1}^{m} \sigma_i \omega_i + \sum_{i=1}^{m} (-\sigma_i)(-\omega_i) \\ &= 2 \sum_{i=1}^{m} \sigma_i \omega_i. \end{aligned}$$

Equality occurs in this inequality if and only if $\boldsymbol{P} = \boldsymbol{U}$ and $\boldsymbol{Q} = \boldsymbol{V}$. Thus, Fan's inequality extends to arbitrary matrices if we substitute ordered singular values for ordered eigenvalues.

A.6 Hadamard Semidifferentials

A function $f(\boldsymbol{y})$ mapping an open set U of R^p into R^q is said to be Hadamard semidifferentiable at $\boldsymbol{x} \in U$ if for every vector $\boldsymbol{v}$ the uniform limit

$$\lim_{t \downarrow 0, \boldsymbol{w} \to \boldsymbol{v}} \frac{f(\boldsymbol{x} + t\boldsymbol{w}) - f(\boldsymbol{x})}{t} = d_{\boldsymbol{v}} f(\boldsymbol{x})$$

exists [63]. Taking $\boldsymbol{w}$ identically equal to $\boldsymbol{v}$ shows that $d_{\boldsymbol{v}} f(\boldsymbol{x})$ coincides with the forward directional derivative of $f(\boldsymbol{y})$ at $\boldsymbol{x}$ in the direction $\boldsymbol{v}$. Some authors equate semidifferentiability at $\boldsymbol{x}$ to the existence of all possible forward directional derivatives. Hadamard's definition is more restrictive and yields a richer theory. For the sake of brevity, we will omit the prefix Hadamard in discussing semidifferentials. It is also convenient to restate the definition in terms of sequences. Thus, semidifferentiability requires the limit

$$\lim_{n \to \infty} \frac{f(\boldsymbol{x} + t_n \boldsymbol{w}_n) - f(\boldsymbol{x})}{t_n} = d_{\boldsymbol{v}} f(\boldsymbol{x})$$

to exist and to be independent of the particular sequences $t_n \downarrow 0$ and $\boldsymbol{w}_n \to \boldsymbol{v}$. The relation between differentiability and semidifferentiability is spelled out in our first proposition.

Proposition A.6.1 *A function $f(\boldsymbol{y})$ differentiable at $\boldsymbol{x}$ is also semidifferentiable at $\boldsymbol{x}$. If $f(\boldsymbol{y})$ is semidifferentiable at $\boldsymbol{x}$, and the map $\boldsymbol{v} \mapsto d_{\boldsymbol{v}} f(\boldsymbol{x})$ is linear, then $f(\boldsymbol{y})$ is differentiable at $\boldsymbol{x}$.*

Proof: If $f(\boldsymbol{y})$ differentiable at $\boldsymbol{x}$, then

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \quad = \quad df(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) + o(\|\boldsymbol{y} - \boldsymbol{x}\|)$$

as $\boldsymbol{y}$ approaches $\boldsymbol{x}$. Choosing $\boldsymbol{y} = \boldsymbol{x} + t\boldsymbol{w}$ for $t > 0$ and $\|\boldsymbol{w} - \boldsymbol{v}\|$ small shows that the Hadamard difference quotient approaches $d_{\boldsymbol{v}} f(x) = df(\boldsymbol{x})\boldsymbol{v}$. To prove the partial converse, define

$$g(\boldsymbol{u}) \quad = \quad \frac{f(\boldsymbol{x} + \boldsymbol{u}) - f(\boldsymbol{x}) - d_{\boldsymbol{u}} f(x)}{\|\boldsymbol{u}\|},$$

and set $c = \limsup_{\|\boldsymbol{u}\| \to 0} \|g(\boldsymbol{u})\|$. It suffices to prove that $c = 0$. Choose a sequence $\boldsymbol{u}_n \neq \boldsymbol{0}$ such that $t_n = \|\boldsymbol{u}_n\|$ converges to 0 and $\|g(\boldsymbol{u}_n)\|$ converges to c. Because the unit sphere is compact, some subsequence of the sequence $\boldsymbol{w}_n = \|\boldsymbol{u}_n\|^{-1}\boldsymbol{u}_n$ converges to a unit vector $\boldsymbol{v}$. Without loss of generality, take the subsequence to be the original sequence. It then follows that

$$\begin{aligned}
\lim_{n \to \infty} g(\boldsymbol{u}_n) &= \lim_{n \to \infty} \frac{f(\boldsymbol{x} + t_n \boldsymbol{w}_n) - f(\boldsymbol{x}) - d_{t_n \boldsymbol{w}_n} f(x)}{t_n} \\
&= \lim_{n \to \infty} \left[\frac{f(\boldsymbol{x} + t_n \boldsymbol{w}_n) - f(\boldsymbol{x})}{t_n} - d_{\boldsymbol{w}_n} f(x) \right] \\
&= d_{\boldsymbol{v}} f(\boldsymbol{x}) - d_{\boldsymbol{v}} f(\boldsymbol{x}).
\end{aligned}$$

The second equality here invokes the homogeneity of the map $\boldsymbol{v} \mapsto d_{\boldsymbol{v}} f(\boldsymbol{x})$. The third equality invokes the continuity of the map, which is a consequence of linearity. This calculation proves that $c = 0$. ■

Our second proposition demonstrates the utility of Hadamard's definition of semidifferentiability.

Proposition A.6.2 *A function $f(\boldsymbol{y})$ semidifferentiable at $\boldsymbol{x}$ is continuous at $\boldsymbol{x}$.*

Proof: Suppose $\boldsymbol{x}_n$ tends to $\boldsymbol{x}$ but $f(\boldsymbol{x}_n)$ does not tend to $f(\boldsymbol{x})$. Then there exists an $\epsilon > 0$ such that $\|f(\boldsymbol{x}_n) - f(\boldsymbol{x})\| \geq \epsilon$ for infinitely many n. Without loss of generality, we may assume that the entire sequence possesses this property. Now write

$$\boldsymbol{x}_n \quad = \quad \boldsymbol{x} + \boldsymbol{x}_n - \boldsymbol{x} \quad = \quad \boldsymbol{x} + t_n \frac{\boldsymbol{x}_n - \boldsymbol{x}_n}{\|\boldsymbol{x}_n - \boldsymbol{x}\|}$$

by taking $t_n = \|\boldsymbol{x}_n - \boldsymbol{x}\|$. Again some subsequence of the sequence

$$\boldsymbol{w}_n \quad = \quad \frac{\boldsymbol{x}_n - \boldsymbol{x}_n}{\|\boldsymbol{x}_n - \boldsymbol{x}\|}$$

of unit vectors converges to a unit vector $\boldsymbol{v}$. Passing to the subsequence where this occurs if necessary, we have

$$f(\boldsymbol{x} + t_n \boldsymbol{w}_n) - f(\boldsymbol{x}) \quad = \quad t_n d_{\boldsymbol{v}} f(\boldsymbol{x}) + o(t_n), \tag{A.8}$$

contrary to the assumption that $\|f(\boldsymbol{x}_n) - f(\boldsymbol{x})\| \geq \epsilon$ for all large n. ■

The approximate equality (A.8) has an important consequence in minimization of real-valued functions. Suppose we can find a direction $\boldsymbol{v}$ with $d_{\boldsymbol{v}} f(\boldsymbol{x}) < 0$. Then $\boldsymbol{v}$ is a descent direction from $\boldsymbol{x}$ in the sense that $f(\boldsymbol{x} + t\boldsymbol{v}) < f(\boldsymbol{x})$ for all sufficiently small $t > 0$. Thus, back-tracking is bound to produce a decrease in $f(\boldsymbol{y})$.

Here is a simple test for establishing semidifferentiability.

Proposition A.6.3 *Suppose $f(\boldsymbol{y})$ is locally Lipschitz around $\boldsymbol{x}$ and possesses all possible forward directional derivatives there. Then $f(\boldsymbol{y})$ is semidifferentiable at $\boldsymbol{x}$.*

Proof: Consider the expansion

$$\frac{f(\boldsymbol{x} + t\boldsymbol{w}) - f(\boldsymbol{x})}{t} \quad = \quad \frac{f(\boldsymbol{x} + t\boldsymbol{w}) - f(\boldsymbol{x} + t\boldsymbol{v})}{t} + \frac{f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x})}{t}.$$

As $t \downarrow 0$, the second fraction on the right of this equation tends to the forward directional derivative of $f(\boldsymbol{y})$ at $\boldsymbol{x}$ in the direction $\boldsymbol{v}$. If c is a Lipschitz constant for $f(\boldsymbol{y})$ in a neighborhood of $\boldsymbol{x}$, then the first fraction on the right of the equation is locally bounded in norm by $c\|\boldsymbol{w} - \boldsymbol{v}\|$, which can be made arbitrarily small by taking $\boldsymbol{w}$ close to $\boldsymbol{v}$. ■

Example A.6.1 *Convex Functions*

Any convex function $f(\boldsymbol{y})$ is locally Lipschitz and possesses all possible forward directional derivatives at an interior point $\boldsymbol{x}$ of its essential domain. Proposition A.6.3 implies that $f(\boldsymbol{y})$ is semidifferentiable at such points. ■

Example A.6.2 *Norms*

A norm $\|\boldsymbol{x}\|_\dagger$ on R^p is convex and therefore semidifferentiable in $\boldsymbol{x}$. In fact, a norm is globally Lipschitz because

$$\big|\|\boldsymbol{y}\|_\dagger - \|\boldsymbol{x}\|_\dagger\big| \quad \leq \quad \|\boldsymbol{y} - \boldsymbol{x}\|_\dagger \quad \leq \quad c\|\boldsymbol{y} - \boldsymbol{x}\|$$

for some constant $c > 0$. At the origin the semidifferential reduces to

$$\lim_{t \downarrow 0, \boldsymbol{w} \to \boldsymbol{v}} \frac{\|t\boldsymbol{w}\|_\dagger - 0}{t} \quad = \quad \|\boldsymbol{v}\|_\dagger.$$

At other points the semidifferential is more complicated to calculate. ■

Semidifferentiable functions obey most of the classical rules of differentiation.

Proposition A.6.4 *Let $f(\boldsymbol{y})$ and $g(\boldsymbol{y})$ be two functions semidifferentiable at the point $\boldsymbol{x}$. Then the homogeneity, sum, product, inverse, and chain rules*

$$\begin{aligned} d_{\boldsymbol{v}}[cf(\boldsymbol{x})] &= cd_{\boldsymbol{v}}f(\boldsymbol{x}) \\ d_{\boldsymbol{v}}[f(\boldsymbol{x})+g(\boldsymbol{x})] &= d_{\boldsymbol{v}}f(\boldsymbol{x})+d_{\boldsymbol{v}}g(\boldsymbol{x}) \\ d_{\boldsymbol{v}}[f(\boldsymbol{x})g(\boldsymbol{x})] &= d_{\boldsymbol{v}}f(\boldsymbol{x})g(\boldsymbol{x})+f(\boldsymbol{x})d_{\boldsymbol{v}}g(\boldsymbol{x}) \\ d_{\boldsymbol{v}}f(\boldsymbol{x})^{-1} &= -f(\boldsymbol{x})^{-1}d_{\boldsymbol{v}}f(\boldsymbol{x})f(\boldsymbol{x})^{-1} \\ d_{\boldsymbol{v}}f\circ g(\boldsymbol{x}) &= d_{d_{\boldsymbol{v}}g(\boldsymbol{x})}f[g(\boldsymbol{x})] \end{aligned}$$

are valid under the usual compatibility assumptions for vector and matrix-valued functions.

Proof: These claims follow directly from the definition of semidifferentiability. Consider for instance the quotient rule. We simply write

$$\frac{1}{t}\left[\frac{1}{f(\boldsymbol{x}+t\boldsymbol{w})}-\frac{1}{f(\boldsymbol{x})}\right] = -f(\boldsymbol{x}+t\boldsymbol{w})^{-1}\cdot\frac{f(\boldsymbol{x}+t\boldsymbol{w})-f(\boldsymbol{x})}{t}\cdot f(\boldsymbol{x})^{-1}$$

and take limits, invoking the continuity of $f(\boldsymbol{y})$ at $\boldsymbol{x}$ in the process. For the chain rule, set

$$\boldsymbol{u} = \frac{g(\boldsymbol{x}+t\boldsymbol{w})-g(\boldsymbol{x})}{t}$$

and rewrite the defining difference quotient as

$$\frac{f[g(\boldsymbol{x}+t\boldsymbol{w})]-f[g(\boldsymbol{x})]}{t} = \frac{f[g(\boldsymbol{x})+t\boldsymbol{u}]-f[g(\boldsymbol{x})]}{t}.$$

Since $\boldsymbol{u}$ tends to $d_{\boldsymbol{v}}g(\boldsymbol{x})$, the limit $d_{d_{\boldsymbol{v}}g(\boldsymbol{x})}f[g(\boldsymbol{x})]$ emerges. ■

Example A.6.3 *Semidifferential of $\|\boldsymbol{x}\|_r$ for $1 \le r < \infty$*

The sum rule and a brief calculation give

$$d_{\boldsymbol{v}}\|\boldsymbol{x}\|_1 = \sum_{i=1}^{p} d_{\boldsymbol{v}}|x_i| = \sum_{i=1}^{p}\begin{cases} v_i & x_i > 0 \\ |v_i| & x_i = 0 \\ -v_i & x_i < 0. \end{cases}$$

Application of the chain and rules shows that the norm

$$\|\boldsymbol{x}\|_r = \left(\sum_{i=1}^{p}|x_i|^r\right)^{1/r}$$

for $1 < r < \infty$ has semidifferential

$$\begin{aligned} d_{\boldsymbol{v}}\|\boldsymbol{x}\|_r &= \frac{1}{r}\left(\sum_{i=1}^p |x_i|^r\right)^{\frac{1}{r}-1} \sum_{i=1}^p r|x_i|^{r-1} d_{\boldsymbol{v}}|x_i| \\ &= \|\boldsymbol{x}\|_r^{1-r} \sum_{i=1}^p |x_i|^{r-1}\operatorname{sgn}(x_i)v_i. \end{aligned}$$

Note that the semidifferential $d_{\boldsymbol{v}}\|\boldsymbol{x}\|_r$ does not necessarily converge to the semidifferential $d_{\boldsymbol{v}}\|\boldsymbol{x}\|_1$ as r tends to 1. ■

Example A.6.4 *Differences of Convex Functions*

If $f(\boldsymbol{y})$ and $g(\boldsymbol{y})$ are convex, then the difference $h(\boldsymbol{y}) = f(\boldsymbol{y}) - g(\boldsymbol{y})$ may not be convex. However, it is semidifferentiable throughout the intersection of the interiors of the essential domains of $f(\boldsymbol{y})$ and $g(\boldsymbol{y})$. ■

More surprising than the classical rules are the maxima and minima rules of the next proposition.

Proposition A.6.5 *Assume the real-valued functions $f_1(\boldsymbol{y}), \ldots, f_m(\boldsymbol{y})$ are semi-differentiable at the point $\boldsymbol{x}$. If*

$$I(\boldsymbol{x}) = \{i : f_i(\boldsymbol{x}) = \max_i f_i(\boldsymbol{x})\} \text{ and } J(\boldsymbol{x}) = \{i : f_i(\boldsymbol{x}) = \min_i f_i(\boldsymbol{x})\},$$

then $\max_i f_i(\boldsymbol{y})$ *and* $\min_i f_i(\boldsymbol{y})$ *are semidifferentiable at* $\boldsymbol{x}$ *and*

$$d_{\boldsymbol{v}} \max_i f_i(\boldsymbol{x}) = \max_{i \in I(\boldsymbol{x})} d_{\boldsymbol{v}} f_i(\boldsymbol{x}) \quad \text{and} \quad d_{\boldsymbol{v}} \min_i f_i(\boldsymbol{x}) = \min_{i \in J(\boldsymbol{x})} d_{\boldsymbol{v}} f_i(\boldsymbol{x}).$$

Proof: The general rules follow from the case $m = 2$ and induction. Consider the minima rule. If $f_1(\boldsymbol{x}) < f_2(\boldsymbol{x})$, then this inequality persists in a neighborhood of $\boldsymbol{x}$. Hence, the rule follows by taking the limit of the difference quotient

$$\frac{\min\{f_1(\boldsymbol{x}+t\boldsymbol{w}), f_2(\boldsymbol{x}+t\boldsymbol{w})\} - \min\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\}}{t} = \frac{f_1(\boldsymbol{x}+t\boldsymbol{w}) - f_1(\boldsymbol{x})}{t}.$$

The case $f_2(\boldsymbol{x}) < f_1(\boldsymbol{x})$ is handled similarly. For the case $f_2(\boldsymbol{x}) = f_1(\boldsymbol{x})$, we have

$$\frac{\min_{1\le i\le 2} f_i(\boldsymbol{x}+t\boldsymbol{w}) - \min_{1\le i\le 2} f_i(\boldsymbol{x})}{t} = \min_{1\le i\le 2} \frac{f_i(\boldsymbol{x}+t\boldsymbol{w}) - f_i(\boldsymbol{x})}{t}.$$

Taking limits again validates the rule. ■

Example A.6.5 *Semidifferential of* $\|\boldsymbol{x}\|_\infty$

The maxima rule implies

$$d_{\boldsymbol{v}}\|\boldsymbol{x}\|_\infty = \max_{i\in I(\boldsymbol{x})} d_{\boldsymbol{v}}|x_i| = \max_{i\in I(\boldsymbol{x})} \begin{cases} v_i & x_i > 0 \\ |v_i| & x_i = 0 \\ -v_i & x_i < 0 \end{cases}$$

for $I(\boldsymbol{x}) = \{i : |x_i| = \|\boldsymbol{x}\|_\infty\}$. ■

Proposition A.6.5 was generalized by Danskin [54]. Here is a convex version of Danskin's result.

Proposition A.6.6 *Consider a continuous function* $f(\boldsymbol{x}, \boldsymbol{y})$ *of two variables* $\boldsymbol{x} \in \mathsf{R}^p$ *and* $\boldsymbol{y} \in C$ *for some compact set* $C \subset \mathsf{R}^q$. *Suppose* $f(\boldsymbol{x}, \boldsymbol{y})$ *is convex in* $\boldsymbol{x}$ *for each fixed* $\boldsymbol{y}$. *Then the function* $g(\boldsymbol{x}) = \sup_{\boldsymbol{y}\in C} f(\boldsymbol{x}, \boldsymbol{y})$ *is convex and has semidifferential*

$$d_{\boldsymbol{v}}g(\boldsymbol{x}) = \sup_{\boldsymbol{y}\in S(\boldsymbol{x})} d_{\boldsymbol{v}}f(\boldsymbol{x}, \boldsymbol{y}),$$

where $S(\boldsymbol{x})$ *denotes the solution set of* $\boldsymbol{y} \in C$ *satisfying* $f(\boldsymbol{x}, \boldsymbol{y}) = g(\boldsymbol{x})$. *Finally, if* $S(\boldsymbol{x})$ *consists of a single point* $\boldsymbol{y}$ *and* $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})$ *exists, then* $\nabla g(\boldsymbol{x})$ *exists and equals* $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})$.

Proof: The function $g(\boldsymbol{x})$ is finite by virtue of the continuity of $f(\boldsymbol{x}, \boldsymbol{y})$ in $\boldsymbol{y}$ and the compactness of C. It is convex because convexity is preserved under suprema. It therefore possesses a semidifferential, whose value equals the forward directional derivative $d_{\boldsymbol{v}}g(\boldsymbol{x})$. Now select any $\boldsymbol{y}_n \in S(\boldsymbol{x} + t_n\boldsymbol{v})$ and any $\boldsymbol{y} \in S(\boldsymbol{x})$. The inequality

$$\begin{aligned} \frac{g(\boldsymbol{x} + t_n\boldsymbol{v}) - g(\boldsymbol{x})}{t_n} &= \frac{f(\boldsymbol{x} + t_n\boldsymbol{v}, \boldsymbol{y}_n) - f(\boldsymbol{x}, \boldsymbol{y})}{t_n} \\ &= \frac{f(\boldsymbol{x} + t_n\boldsymbol{v}, \boldsymbol{y}_n) - f(\boldsymbol{x} + t_n\boldsymbol{v}, \boldsymbol{y})}{t_n} \\ &\quad + \frac{f(\boldsymbol{x} + t_n\boldsymbol{v}, \boldsymbol{y}) - f(\boldsymbol{x}, \boldsymbol{y})}{t_n} \\ &\geq \frac{f(\boldsymbol{x} + t_n\boldsymbol{v}, \boldsymbol{y}) - f(\boldsymbol{x}, \boldsymbol{y})}{t_n} \end{aligned}$$

implies in the limit that $d_{\boldsymbol{v}}g(\boldsymbol{x}) \geq d_{\boldsymbol{v}}f(\boldsymbol{x}, \boldsymbol{y})$ and consequently that

$$d_{\boldsymbol{v}}g(\boldsymbol{x}) \geq \sup_{\boldsymbol{y}\in S(\boldsymbol{x})} d_{\boldsymbol{v}}f(\boldsymbol{x}, \boldsymbol{y}).$$

In the process this logic reveals that $\sup_{\boldsymbol{y}\in S(\boldsymbol{x})} d_{\boldsymbol{v}}f(\boldsymbol{x}, \boldsymbol{y})$ is finite.

To prove the reverse inequality, observe that the inequalities

$$d_{\boldsymbol{v}}g(\boldsymbol{x}) \leq \frac{g(\boldsymbol{x} + t_n\boldsymbol{v}) - g(\boldsymbol{x})}{t_n}$$

$$\begin{aligned} &\leq \frac{f(\boldsymbol{x}+t_n\boldsymbol{v},\boldsymbol{y}_n)-f(\boldsymbol{x},\boldsymbol{y}_n)}{t_n} \\ &\leq d_{\boldsymbol{v}}f(\boldsymbol{x}+t_n\boldsymbol{v},\boldsymbol{y}_n) \end{aligned}$$

for $\boldsymbol{y}_n \in S(\boldsymbol{x}+t_n\boldsymbol{v})$ simply reflect the monotonicity relations between difference quotients and directional derivatives for a convex function discussed in Sect. 6.4. To complete the proof, it suffices to argue that

$$\limsup_{n\to\infty} d_{\boldsymbol{v}}f(\boldsymbol{x}+t_n\boldsymbol{v},\boldsymbol{y}_n) \leq d_{\boldsymbol{v}}f(\boldsymbol{x},\boldsymbol{y})$$

for some point $\boldsymbol{y} \in S(\boldsymbol{x})$. Fortunately, we can identify $\boldsymbol{y}$ as the limit of any convergent subsequence of the original sequence $\boldsymbol{y}_n$. Without loss of generality, assume that this subsequence coincides with the original sequence. Taking limits in the inequality $f(\boldsymbol{x}+t_n\boldsymbol{v},\boldsymbol{y}_n) \geq f(\boldsymbol{x}+t_n\boldsymbol{v},\boldsymbol{z})$ implies that $f(\boldsymbol{x},\boldsymbol{y}) \geq f(\boldsymbol{x},\boldsymbol{z})$ for all $\boldsymbol{z}$; hence, $\boldsymbol{y} \in S(\boldsymbol{x})$. Now for any $\epsilon > 0$, all sufficiently small $t > 0$ satisfy

$$\frac{f(\boldsymbol{x}+t\boldsymbol{v},\boldsymbol{y})-f(\boldsymbol{x},\boldsymbol{y})}{t} \leq d_{\boldsymbol{v}}f(\boldsymbol{x},\boldsymbol{y})+\frac{\epsilon}{2}.$$

Hence, for such a t and sufficiently large n, joint continuity and monotonicity imply

$$\begin{aligned} d_{\boldsymbol{v}}f(\boldsymbol{x}+t_n\boldsymbol{v},\boldsymbol{y}_n) &\leq \frac{f(\boldsymbol{x}+t_n\boldsymbol{v}+t\boldsymbol{v},\boldsymbol{y}_n)-f(\boldsymbol{x}+t_n\boldsymbol{v},\boldsymbol{y}_n)}{t} \\ &\leq \frac{f(\boldsymbol{x}+t\boldsymbol{v},\boldsymbol{y})-f(\boldsymbol{x},\boldsymbol{y})}{t}+\frac{\epsilon}{2} \\ &\leq d_{\boldsymbol{v}}f(\boldsymbol{x},\boldsymbol{y})+\epsilon \end{aligned}$$

Since ϵ can be taken arbitrarily small in the inequality

$$d_{\boldsymbol{v}}f(\boldsymbol{x}+t_n\boldsymbol{v},\boldsymbol{y}_n) \leq d_{\boldsymbol{v}}f(\boldsymbol{x},\boldsymbol{y})+\epsilon,$$

this completes the derivation of the semidifferential. To verify the last claim of the proposition, note that $d_{\boldsymbol{v}}g(\boldsymbol{x}) = d_{\boldsymbol{v}}f(\boldsymbol{x},\boldsymbol{y}) = \nabla_{\boldsymbol{x}}f(\boldsymbol{x},\boldsymbol{y})^*\boldsymbol{v}$. ■

Danskin's original argument dispenses with convexity and relies on the existence and continuity of the gradient $\nabla_{\boldsymbol{x}}f(\boldsymbol{x},\boldsymbol{y})$. For our purposes the convex version is more convenient.

Example A.6.6 *Orthogonally Invariant Matrix Norms*

Let $\|\boldsymbol{A}\|_\dagger$ be a matrix norm on $m \times n$ matrices. As pointed out in Example 14.3.6, every matrix norm has a dual norm $\|\boldsymbol{B}\|_\star$ in terms of which

$$\|\boldsymbol{A}\|_\dagger = \sup_{\|\boldsymbol{B}\|_\star=1} \operatorname{tr}(\boldsymbol{B}^*\boldsymbol{A}). \tag{A.9}$$

The semidifferential of the linear map $\boldsymbol{A} \mapsto \operatorname{tr}(\boldsymbol{B}^*\boldsymbol{A})$ is

$$d_{\boldsymbol{C}} \operatorname{tr}(\boldsymbol{B}^*\boldsymbol{A}) = \operatorname{tr}(\boldsymbol{B}^*\boldsymbol{C}),$$

and the set $\{\boldsymbol{B} : \|\boldsymbol{B}\|_\star = 1\}$ is compact. Thus, the hypotheses of Proposition A.6.6 are met. Furthermore, representation (A.9) and duality show that $\|\cdot\|_\dagger$ is orthogonally invariant if and only if $\|\cdot\|_\star$ is orthogonally invariant. For an orthogonally invariant pair, the restriction $\|\boldsymbol{B}\|_\star = 1$ can be re-expressed as $\|\boldsymbol{D}_{\boldsymbol{B}}\|_\star = 1$ using the diagonal matrix $\boldsymbol{D}_{\boldsymbol{B}}$ whose diagonal entries are the singular values of $\boldsymbol{B}$.

According to Fan's inequality, $\mathrm{tr}(\boldsymbol{B}^*\boldsymbol{A}) \le \sum_i \beta_i\alpha_i$, where the α_i are the ordered singular values of $\boldsymbol{A}$ and the β_i are the ordered singular values of $\boldsymbol{B}$. Equality is attained in Fan's inequality if and only if $\boldsymbol{A}$ and $\boldsymbol{B}$ have ordered singular value decompositions $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}_{\boldsymbol{A}}\boldsymbol{V}^*$ and $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}_{\boldsymbol{B}}\boldsymbol{V}^*$ with shared singular vectors. Let $S_{\boldsymbol{A}}$ be the set of diagonal matrices $\boldsymbol{D}_{\boldsymbol{B}}$ with ordered diagonal entries β_i, $\|\boldsymbol{D}_{\boldsymbol{B}}\|_\star = 1$, and $\sum_i \beta_i\alpha_i = \|\boldsymbol{A}\|_\dagger$. If the matrix $\boldsymbol{U}$ has columns $\boldsymbol{u}_i$ and the matrix $\boldsymbol{V}$ has columns $\boldsymbol{v}_i$, then it follows that

$$d_{\boldsymbol{C}}\|\boldsymbol{A}\|_\dagger = \sup_{\substack{\boldsymbol{D}_{\boldsymbol{B}} \in S_{\boldsymbol{A}} \\ \boldsymbol{A}=\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{A}}\boldsymbol{V}^*}} \mathrm{tr}(\boldsymbol{V}\boldsymbol{D}_{\boldsymbol{B}}\boldsymbol{U}^*\boldsymbol{C}) = \sup_{\substack{\boldsymbol{D}_{\boldsymbol{B}} \in S_{\boldsymbol{A}} \\ \boldsymbol{A}=\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{A}}\boldsymbol{V}^*}} \sum_i \beta_i \boldsymbol{u}_i^*\boldsymbol{C}\boldsymbol{v}_i,$$

where the suprema extend over all singular value decompositions of $\boldsymbol{A}$. When the singular values are distinct, the singular vectors of $\boldsymbol{A}$ are unique up to sign. The singular values are always unique. As an example, consider the spectral norm $\|\boldsymbol{B}\| = \beta_1$ and its dual the nuclear norm $\|\boldsymbol{B}\|_\star = \sum_i \beta_i$. The forward directional derivatives $d_{\boldsymbol{C}}\|\boldsymbol{A}\|$ and $d_{\boldsymbol{C}}\|\boldsymbol{A}\|_\star$ amount to

$$\begin{aligned} d_{\boldsymbol{C}}\|\boldsymbol{A}\| &= \sup_{\boldsymbol{A}=\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{A}}\boldsymbol{V}^*} \boldsymbol{u}_1^*\boldsymbol{C}\boldsymbol{v}_1 \\ d_{\boldsymbol{C}}\|\boldsymbol{A}\|_\star &= \sup_{\boldsymbol{A}=\boldsymbol{U}\boldsymbol{D}_{\boldsymbol{A}}\boldsymbol{V}^*} \left(\sum_{\alpha_i>0} \boldsymbol{u}_i^*\boldsymbol{C}\boldsymbol{v}_i + \sum_{\alpha_i=0} \sup_{\beta_i \in [0,1]} \beta_i \boldsymbol{u}_i^*\boldsymbol{C}\boldsymbol{v}_i \right). \end{aligned}$$

In the first case we take $\beta_1 = 1$ and all remaining $\beta_i = 0$, and in the second case we take $\beta_i = 1$ when $\alpha_i > 0$ and $\beta_i \in [0,1]$ otherwise. These directional derivatives are consistent with the rule (14.5) and the subdifferentials found in Example 14.5.7 and Problem 38 of Chap. 14. ■

Example A.6.7 *Induced Matrix Norms*

Let $\|\boldsymbol{x}\|_a$ and $\|\boldsymbol{y}\|_b$ be vector norms on R^m and R^n, respectively. These induce a norm on $m \times n$ matrices $\boldsymbol{M}$ via

$$\|\boldsymbol{M}\|_{a,b} = \sup_{\|\boldsymbol{x}\|_a=1} \|\boldsymbol{M}\boldsymbol{x}\|_b.$$

The chain rule implies $d_{\boldsymbol{N}}\|\boldsymbol{M}\boldsymbol{x}\|_b = d_{\boldsymbol{N}\boldsymbol{x}}\|\boldsymbol{y}\|_b$ with $\boldsymbol{y} = \boldsymbol{M}\boldsymbol{x}$. Hence, Proposition A.6.6 gives

$$d_{\boldsymbol{N}}\|\boldsymbol{M}\|_{a,b} = \sup_{\substack{\|\boldsymbol{x}\|_a=1 \\ \|\boldsymbol{M}\boldsymbol{x}\|_b=\|\boldsymbol{M}\|_{a,b}}} d_{\boldsymbol{N}\boldsymbol{x}}\|\boldsymbol{M}\boldsymbol{x}\|_b.$$

In the special case where the two vectors norms are both ℓ_1 norms, one has $\|\boldsymbol{M}\|_{1,1} = \max_j \sum_i |m_{ij}|$. Suppose this maximum is attained for a unique index $j = k$. Then the best vector $\boldsymbol{x}$ is the standard unit vector $\boldsymbol{e}_k$, and one can show that

$$d_{\boldsymbol{N}}\|\boldsymbol{M}\|_{1,1} = \sum_i \begin{cases} n_{ik} & m_{ik} > 0 \\ |n_{ik}| & m_{ik} = 0 \\ -n_{ik} & m_{ik} < 0 \end{cases}.$$

The reader might like to consider the case of two ℓ_∞ vector norms. ■

Example A.6.8 *Differentiability of a Fenchel Conjugate*

Let $f(\boldsymbol{x})$ be a strictly convex function satisfying the growth condition $\liminf_{\|\boldsymbol{x}\|\to\infty} \|\boldsymbol{x}\|^{-1} f(\boldsymbol{x}) = \infty$. The Fenchel conjugate

$$f^\star(\boldsymbol{y}) = \sup_{\boldsymbol{x}} [\boldsymbol{y}^* \boldsymbol{x} - f(\boldsymbol{x})]$$

is finite for all $\boldsymbol{y}$, and the supremum is attained at a unique point $\boldsymbol{x}$. In general for any compact set S of points $\boldsymbol{y}$, the corresponding set C of optimal points $\boldsymbol{x}$ is compact. If on the contrary C is unbounded, then there exist paired sequences $\boldsymbol{y}_n$ and $\boldsymbol{x}_n$ with $\lim_{n\to\infty} \|\boldsymbol{x}_n\| = \infty$. Given the boundedness of S and the inequalities

$$-f(\boldsymbol{0}) \leq f^\star(\boldsymbol{y}_n) = \boldsymbol{y}_n^* \boldsymbol{x}_n - f(\boldsymbol{x}_n) \leq \|\boldsymbol{y}_n\|\|\boldsymbol{x}_n\| - f(\boldsymbol{x}_n),$$

this contradicts the growth condition. The reader can check that C is closed. Propositions A.6.6 and A.6.1 now imply that $d_{\boldsymbol{v}} f^\star(\boldsymbol{y}) = \boldsymbol{x}^* \boldsymbol{v}$ for the optimal $\boldsymbol{x}$ and that $f^\star(\boldsymbol{y})$ is differentiable with $\nabla f^\star(\boldsymbol{y}) = \boldsymbol{x}$. ■

Example A.6.9 *Distance to a Closed Convex Set C*

The distance function $\operatorname{dist}(\boldsymbol{x}, C)$ is convex and locally constant at an interior point of C. All directional derivatives consequently vanish there. At an exterior point the identity

$$\nabla \operatorname{dist}(\boldsymbol{x}, C)^2 = 2[\boldsymbol{x} - P_C(\boldsymbol{x})] \tag{A.10}$$

involving the projection operator $P_C(\boldsymbol{x})$ shows that $\operatorname{dist}(\boldsymbol{x}, C)$ is differentiable. Indeed, because $\operatorname{dist}(\boldsymbol{x}, C)^2 > 0$, the chain rule yields

$$\nabla \operatorname{dist}(\boldsymbol{x}, C) = \nabla \sqrt{\operatorname{dist}(\boldsymbol{x}, C)^2} = \frac{\boldsymbol{x} - P_C(\boldsymbol{x})}{\operatorname{dist}(\boldsymbol{x}, C)}.$$

To prove formula (A.10), set

$$\Delta = \operatorname{dist}(\boldsymbol{x} + \boldsymbol{y}, C)^2 - \operatorname{dist}(\boldsymbol{x}, C)^2.$$

In view of the inequality $\operatorname{dist}(\boldsymbol{x}, C)^2 \leq \|\boldsymbol{x} - P_C(\boldsymbol{x}+\boldsymbol{y})\|^2$, one can construct the lower bound

$$\begin{aligned}
\Delta &\geq \|\boldsymbol{x}+\boldsymbol{y}-P_C(\boldsymbol{x}+\boldsymbol{y})\|^2 - \|\boldsymbol{x}-P_C(\boldsymbol{x}+\boldsymbol{y})\|^2 \\
&= \|\boldsymbol{y}\|^2 + 2\boldsymbol{y}^*[\boldsymbol{x}-P_C(\boldsymbol{x}+\boldsymbol{y})] \\
&= \|\boldsymbol{y}\|^2 + 2\boldsymbol{y}^*[\boldsymbol{x}-P_C(\boldsymbol{x})] + 2\boldsymbol{y}^*[P_C(\boldsymbol{x})-P_C(\boldsymbol{x}+\boldsymbol{y})] \qquad \text{(A.11)} \\
&\geq \|\boldsymbol{y}\|^2 + 2\boldsymbol{y}^*[\boldsymbol{x}-P_C(\boldsymbol{x})] - 2\|\boldsymbol{y}\| \cdot \|P_C(\boldsymbol{x})-P_C(\boldsymbol{x}+\boldsymbol{y})\| \\
&\geq \|\boldsymbol{y}\|^2 + 2\boldsymbol{y}^*[\boldsymbol{x}-P_C(\boldsymbol{x})] - 2\|\boldsymbol{y}\|^2.
\end{aligned}$$

The penultimate inequality here is just the Cauchy–Schwartz inequality. The final inequality is a consequence of the non-expansiveness of the projection operator. The analogous inequality $\operatorname{dist}(\boldsymbol{x}+\boldsymbol{y}, C)^2 \leq \|\boldsymbol{x}+\boldsymbol{y}-P_C(\boldsymbol{x})\|^2$ gives the upper bound

$$\Delta \leq \|\boldsymbol{x}+\boldsymbol{y}-P_C(\boldsymbol{x})\|^2 - \|\boldsymbol{x}-P_C(\boldsymbol{x})\|^2 = \|\boldsymbol{y}\|^2 + 2\boldsymbol{y}^*[\boldsymbol{x}-P_C(\boldsymbol{x})]. \quad \text{(A.12)}$$

The two bounds (A.11) and (A.12) together imply that

$$\Delta = 2\boldsymbol{y}^*[\boldsymbol{x}-P_C(\boldsymbol{x})] + o(\|\boldsymbol{y}\|)$$

and consequently that $\nabla \operatorname{dist}(\boldsymbol{x}, C)^2 = 2[\boldsymbol{x}-P_C(\boldsymbol{x})]$ according to Fréchet's definition of the differential.

In contrast differentiability of $\operatorname{dist}(\boldsymbol{x}, C)$ at boundary points of C is not assured. To calculate $d_{\boldsymbol{v}} \operatorname{dist}(\boldsymbol{x}, C)$, we first observe that $\operatorname{dist}(\boldsymbol{x}, C)$ equals the composition of the functions $f(\boldsymbol{y}) = \|\boldsymbol{y}\|$ and $g(\boldsymbol{x}) = \boldsymbol{x} - P_C(\boldsymbol{x})$. Given that $g(\boldsymbol{x}) = \boldsymbol{0}$, the chain rule therefore implies that

$$d_{\boldsymbol{v}} \operatorname{dist}(\boldsymbol{x}, C) = \|\boldsymbol{v} - d_{\boldsymbol{v}} P_C(\boldsymbol{x})\|.$$

Thus, we need to evaluate $d_{\boldsymbol{v}} P_C(\boldsymbol{x})$. Without loss of generality, assume that $\boldsymbol{x} = \boldsymbol{0}$, and define $T(\boldsymbol{x})$ to be the closure of the convex cone $\cup_{t>0} \frac{1}{t} C$. Formally, $T(\boldsymbol{x})$ is the tangent space of C at the point $\boldsymbol{0}$. We now demonstrate that $d_{\boldsymbol{v}} P_C(\boldsymbol{x}) = P_{T(\boldsymbol{x})}(\boldsymbol{v})$ is the projection of $\boldsymbol{v}$ onto $T(\boldsymbol{x})$. The easily checked identity

$$\frac{P_C(\boldsymbol{0}+t\boldsymbol{v}) - P_C(\boldsymbol{0})}{t} = P_{t^{-1}C}(\boldsymbol{v})$$

is our point of departure. We must demonstrate that $P_{t^{-1}C}(\boldsymbol{v})$ converges to $P_{T(\boldsymbol{x})}(\boldsymbol{v})$. Because $\boldsymbol{0} \in C$, the sets $t^{-1}C$ increase as t decreases. Hence, the trajectory $P_{t^{-1}C}(\boldsymbol{v})$ remains bounded. Let $\boldsymbol{x}$ be any cluster point of $\boldsymbol{x}(t) = P_{t^{-1}C}(\boldsymbol{v})$. The obtuse angle criterion along the converging sequence $\boldsymbol{x}(t_n)$ requires

$$[\boldsymbol{v} - \boldsymbol{x}(t_n)]^*[\boldsymbol{z} - \boldsymbol{x}(t_n)] \leq 0$$

for every $\boldsymbol{z} \in t_n^{-1} C$. In the limit the inequality

$$(\boldsymbol{v} - \boldsymbol{x})^*(\boldsymbol{z} - \boldsymbol{x}) \leq 0$$

holds for every $\boldsymbol{z} \in \cup_{t>0} \frac{1}{t} C$ and therefore for every $\boldsymbol{z}$ in the closure of this set. The only point in this closed set that qualifies is $P_{T(\boldsymbol{x})}(\boldsymbol{v})$. Thus, all cluster points of $\boldsymbol{x}(t)$ reduce to $P_{T(\boldsymbol{x})}(\boldsymbol{v})$. ■

A.7 Problems

1. Show that a polyhedral set S is compact if and only if it is the convex hull of a finite set of points.

2. Prove that a polyhedral set S possesses at most finitely many extreme points. (Hint: In the representation (A.2), a constraint $\boldsymbol{v}_i^* \boldsymbol{x} \leq c_i$ is said to be active at $\boldsymbol{x}$ whenever equality holds there. If two points have the same active constraints, then neither of them is extreme.)

3. Demonstrate that every compact convex set possesses at least one extreme point. (Hint: The point farthest from the origin is extreme.)

4. Let $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ be a set of points in R^n. If the pair $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ attains the maximum Euclidean distance between two points in S, then show that $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are both extreme points of $\operatorname{conv}(S)$.

5. Let C_1 and C_2 be two closed convex sets. Demonstrate that

$$\operatorname{dist}(C_1, C_2) \quad = \quad \inf_{\boldsymbol{x} \in C_1, \boldsymbol{y} \in C_2} \|\boldsymbol{x} - \boldsymbol{y}\|$$

   is attained whenever (a) $C_1 \cap C_2 \neq \emptyset$, (b) either C_1 or C_2 is compact, or (c) both C_1 and C_2 are polyhedral.

6. For two sequences $a_1 \leq a_2 \leq \cdots \leq a_n$ and $b_1 \leq b_2 \leq \cdots \leq b_n$, demonstrate that

$$\sum_{i=1}^n a_i b_{\sigma(i)} \quad \geq \quad \sum_{i=1}^n a_i b_{n-i+1}$$

   for every permutation σ.

7. Based on the previous exercise and the proof of Proposition A.4.2, devise a lower bound on $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B})$ for symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

8. Under the Frobenius inner product $\langle \boldsymbol{M}, \boldsymbol{N} \rangle = \operatorname{tr}(\boldsymbol{M}^* \boldsymbol{N})$ on square matrices, show that the subspaces S and A of symmetric and skew-symmetric matrices are orthogonal complements. Here $\boldsymbol{M} \in A$ if and only if $\boldsymbol{M} = -\boldsymbol{M}^*$. Find the projection operators P_S and P_A.

9. Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ are two $n \times n$ symmetric matrices with ordered eigenvalues $\{\lambda_i\}_{i=1}^n$ and $\{\mu_i\}_{i=1}^n$. Prove that

$$\sum_{i=1}^n (\lambda_i - \mu_i)^2 \quad \leq \quad \|\boldsymbol{A} - \boldsymbol{B}\|_F^2.$$

(Hint: $\sum_{i=1}^n \lambda_i^2 = \|\boldsymbol{A}\|_F^2$ and similarly for $\boldsymbol{B}$.)

10. The function

$$f(\boldsymbol{x}) \quad = \quad \begin{cases} \frac{x_1^6}{(x_2 - x_1^2)^2 + x_1^8} & \boldsymbol{x} \neq \boldsymbol{0} \\ 0 & \boldsymbol{x} = \boldsymbol{0} \end{cases}$$

illustrates the difference between Hadamard semidifferentiability at a point and the mere existence of all forward directional derivatives at the point. Prove the following assertions:

(a) The forward directional derivative $d_{\boldsymbol{v}} f(\boldsymbol{0}) = 0$ for all $\boldsymbol{v}$. (Hint: Treat the cases $v_2 = 0$ and $v_2 \neq 0$ separately.)

(b) $f(\boldsymbol{x})$ is discontinuous at $\boldsymbol{0}$. (Hint: Take the limit of $f(t, t^2)$ as $t \to 0$.)

(c) The Hadamard semidifferential $d_{(1,0)} f(\boldsymbol{0})$ does not exist. (Hint: Contrast the convergence of the relevant difference quotient for the scalar sequence $t_n = \frac{1}{n} \to 0$ and the two vector sequences $\boldsymbol{w}_n = (1, 0)^*$ and $\boldsymbol{w}_n = (1, \frac{1}{n})^* \to (1, 0)^*$.)

11. Demonstrate that a real-valued semidifferentiable function $f(\boldsymbol{x})$ satisfies

$$d_{\boldsymbol{v}} |f(\boldsymbol{x})| \quad = \quad \begin{cases} d_{\boldsymbol{v}} f(\boldsymbol{x}) & f(\boldsymbol{x}) > 0 \\ |d_{\boldsymbol{v}} f(\boldsymbol{x})| & f(\boldsymbol{x}) = 0 \\ -d_{\boldsymbol{v}} f(\boldsymbol{x}) & f(\boldsymbol{x}) < 0 \end{cases}$$

and

$$d_{\boldsymbol{v}} \max\{f(\boldsymbol{x}), 0\} \quad = \quad \begin{cases} d_{\boldsymbol{v}} f(\boldsymbol{x}) & f(\boldsymbol{x}) > 0 \\ \max\{d_{\boldsymbol{v}} f(\boldsymbol{x}), 0\} & f(\boldsymbol{x}) = 0 \\ 0 & f(\boldsymbol{x}) < 0. \end{cases}$$

These formulas are pertinent in calculating the directional derivatives of the exact penalty function $\mathcal{E}_\rho(\boldsymbol{y})$ discussed in Sect. 16.3.

12. Suppose the function $f(\boldsymbol{x})$ is Lipschitz in a neighborhood of the point $\boldsymbol{y}$ with Lipschitz constant c. Prove the inequality

$$\|d_{\boldsymbol{v}} f(\boldsymbol{y}) - d_{\boldsymbol{w}} f(\boldsymbol{y})\| \quad \leq \quad c\|\boldsymbol{v} - \boldsymbol{w}\|$$

assuming the indicated forward directional derivatives exist. The special case $\boldsymbol{w} = \boldsymbol{0}$ gives $\|d_{\boldsymbol{v}} f(\boldsymbol{x})\| \leq c\|\boldsymbol{v}\|$.

References

[1] Acosta E, Delgado C (1994) Fréchet versus Carathéodory. Am Math Mon 101:332–338

[2] Acton FS (1990) Numerical methods that work. Mathematical Association of America, Washington, DC

[3] Anderson TW (2003) An introduction to multivariate statistical analysis, 3rd edn. Wiley, Hoboken

[4] Armstrong RD, Kung MT (1978) Algorithm AS 132: least absolute value estimates for a simple linear regression problem. Appl Stat 27:363–366

[5] Arthur D, Vassilvitskii S (2007) k-means++: the advantages of careful seeding. In: 2007 symposium on discrete algorithms (SODA). Society for Industrial and Applied Mathematics, Philadelphia, 2007

[6] Barlow RE, Bartholomew DJ, Bremner JM, Brunk HD (1972) Statistical inference under order restrictions; the theory and application of isotonic regression. Wiley, New York

[7] Bartle RG (1996) Return to the Riemann integral. Am Math Mon 103:625–632

[8] Baum LE (1972) An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3:1–8

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8,

[9] Bauschke HH, Lewis AS (2000) Dykstra's algorithm with Bregman projections: a convergence proof. Optimization 48:409–427

[10] Beltrami EJ (1970) An algorithmic approach to nonlinear analysis and optimization. Academic, New York

[11] Berry MW, Drmac Z, Jessup ER (1999) Matrices, vector spaces, and information retrieval. SIAM Rev 41:335–362

[12] Bertsekas DP (1999) Nonlinear programming, 2nd edn. Athena Scientific, Belmont

[13] Bertsekas DP (2009) Convex optimization theory. Athena Scientific, Belmont

[14] Bishop YMM, Feinberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT, Cambridge

[15] Bliss GA (1925) Calculus of variations. Mathematical Society of America, Washington, DC

[16] Böhning D, Lindsay BG (1988) Monotonicity of quadratic approximation algorithms. Ann Inst Stat Math 40:641–663

[17] Borwein JM, Lewis AS (2000) Convex analysis and nonlinear optimization: theory and examples. Springer, New York

[18] Botsko MW, Gosser RA (1985) On the differentiability of functions of several variables. Am Math Mon 92:663–665

[19] Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University Press, Cambridge

[20] Boyd S, Kim SJ, Vandenberghe L, Hassibi A (2007) A tutorial on geometric programming. Optim Eng 8:67–127

[21] Boyle JP, Dykstra RL (1985) A method for finding projections onto the intersection of convex sets in Hilbert space. In: Advances in order restricted statistical inference. Lecture notes in statistics. Springer, New York, pp 28–47

[22] Bradley EL (1973) The equivalence of maximum likelihood and weighted least squares estimates in the exponential family. J Am Stat Assoc 68:199–200

[23] Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs. Biometrika 39:324–345

[24] Bregman LM (1965) The method of successive projection for finding a common point of convex sets. Sov Math Dokl 6:688–692

[25] Bregman LM (1967) The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. USSR Comput Math Math Phy 7:200–217

[26] Bregman LM, Censor Y, Reich S (2000) Dykstra's algorithm as the nonlinear extension of Bregman's optimization method. J Convex Anal 6:319–333

[27] Brent RP (1973) Some efficient algorithms for solving systems of nonlinear equations. SIAM J Numer Anal 10:327–344

[28] Brezhneva OA, Tret'yakov AA, Wright SE (2010) A simple and elementary proof of the Karush-Kuhn-Tucker theorem for inequality-constrained optimization. Optim Lett 3:7–10

[29] Bridger M, Stolzenberg G (1999) Uniform calculus and the law of bounded change. Am Math Mon 106:628–635

[30] Brinkhuis J, Tikhomirov V (2005) Optimization: insights and applications. Princeton University Press, Princeton

[31] Brophy JF, Smith PW (1988) Prototyping Karmarkar's algorithm using MATH-PROTRAN. IMSL Dir 5:2–3

[32] Broyden CG (1965) A class of methods for solving nonlinear simultaneous equations. Math Comput 19:577–593

[33] Byrd RH, Nocedal J (1989) A tool for the analysis of quasi-Newton methods with application to unconstrained minimization. SIAM J Numer Anal 26:727–739

[34] Byrne CL (2009) A first course in optimization. Department of Mathematical Sciences, University of Massachusetts Lowell, Lowell

[35] Cai J-F, Candés EJ, Shen Z (2008) A singular value thresholding algorithm for matrix completion. SIAM J Optim 20:1956–1982

[36] Candés EJ, Tao T (2007) The Danzig selector: statistical estimation when p is much larger than n. Ann Stat 35:2313–2351

[37] Candés EJ, Tao T (2009) The power of convex relaxation: near-optimal matrix completion. IEEE Trans Inform Theor 56:2053–2080

[38] Candés EJ, Romberg J, Tao T (2006) Stable signal recovery from incomplete and inaccurate measurements. Comm Pure Appl Math 59:1207–1223

[39] Candés EJ, Wakin M, Boyd S (2007) Enhancing sparsity by reweighted ℓ_1 minimization. J Fourier Anal Appl 14:877–905

[40] Carathéodory C (1954) Theory of functions of a complex variable, vol 1. Chelsea, New York

[41] Censor Y, Zenios SA (1992) Proximal minimization with D-functions. J Optim Theor Appl 73:451–464

[42] Censor Y, Chen W, Combettes PL, Davidi R, Herman GT (2012) On the effectiveness of projection methods for convex feasibility problems with linear inequality constraints. Comput Optim Appl 51:1065–1088

[43] Charnes A, Frome EL, Yu PL (1976) The equivalence of generalized least squares and maximum likelihood in the exponential family. J Am Stat Assoc 71:169–171

[44] Chen J, Tan X (2009) Inference for multivariate normal mixtures. J Multivariate Anal 100:1367–1383

[45] Chen SS, Donoho DL, Saunders MA (1998) Atomic decomposition by basis pursuit. SIAM J Sci Comput 20:33–61

[46] Cheney W (2001) Analysis for applied mathematics. Springer, New York

[47] Choi SC, Wette R (1969) Maximum likelihood estimation of the parameters of the gamma distribution and their bias. Technometrics 11:683–690

[48] Ciarlet PG (1989) Introduction to numerical linear algebra and optimization. Cambridge University Press, Cambridge

[49] Claerbout J, Muir F (1973) Robust modeling with erratic data. Geophysics 38:826–844

[50] Clarke CA, Price Evans DA, McConnell RB, Sheppard PM (1959) Secretion of blood group antigens and peptic ulcers. Br Med J 1:603–607

[51] Conn AR, Gould NIM, Toint PL (1991) Convergence of quasi-Newton matrices generated by the symmetric rank one update. Math Program 50:177–195

[52] Conte SD, deBoor C (1972) Elementary numerical analysis. McGraw-Hill, New York

[53] Cox DR (1970) Analysis of binary data. Methuen, London

[54] Danskin JM (1966) The theory of max-min, with applications. SIAM J Appl Math 14:641–664

[55] Daubechies I, Defrise M, De Mol C (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Comm Pure Appl Math 57:1413–1457

[56] Davidon WC (1959) Variable metric methods for minimization. AEC Research and Development Report ANL–5990, Argonne National Laboratory, Argonne

[57] Davis JA, Smith TW (2008) General social surveys, 1972–2008 [machine-readable data le]. Roper Center for Public Opinion Research, University of Connecticut, Storrs

[58] Debreu G (1952) Definite and semidefinite quadratic forms. Econometrica 20:295–300

[59] de Leeuw J (1994) Block relaxation algorithms in statistics. In: Bock HH, Lenski W, Richter MM (eds) Information systems and data analysis. Springer, New York, pp 308–325

[60] de Leeuw J (2006) Some majorization techniques. Preprint series, UCLA Department of Statistics.

[61] de Leeuw J, Heiser WJ (1980) Multidimensional scaling with restrictions on the configuration. In: Krishnaiah PR (ed) Multivariate analysis, vol V. North-Holland, Amsterdam, pp 501–522

[62] de Leeuw J, Lange K (2009) Sharp quadratic majorization in one dimension. Comput Stat Data Anal 53:2471–2484

[63] Delfour MC (2012) Introduction to optimization and semidifferential calculus. SIAM, Philadelphia

[64] Demmel J (1997) Applied numerical linear algebra. SIAM, Philadelphia

[65] Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J Roy Stat Soc B 39:1–38

[66] Dennis JE Jr, Schnabel RB (1996) Numerical methods for unconstrained optimization and nonlinear equations. SIAM, Philadelphia

[67] De Pierro AR (1993) On the relation between the ISRA and EM algorithm for positron emission tomography. IEEE Trans Med Imag 12:328–333

[68] DePree JD, Swartz CW (1988) Introduction to real analysis. Wiley, Hoboken

[69] de Souza PN, Silva J-N (2001) Berkeley problems in mathematics, 2nd edn. Springer, New York

[70] Deutsch F (2001) Best approximation in inner product spaces. Springer, New York

[71] Devijver PA (1985) Baum's forward-backward algorithm revisited. Pattern Recogn Lett 3:369–373

[72] Ding C, Li T, Jordan MI (2010) Convex and semi-nonnegative matrix factorizations. IEEE Trans Pattern Anal Mach Intell 32:45–55

[73] Dobson AJ (1990) An introduction to generalized linear models. Chapman & Hall, London

[74] Donoho DL (2006) Compressed sensing. IEEE Trans Inform Theor 52:1289–1306

[75] Donoho D, Johnstone I (1994) Ideal spatial adaptation by wavelet shrinkage. Biometrika 81:425–455

[76] Duan J-C, Simonato J-G (1993) Multiplicity of solutions in maximum likelihood factor analysis. J Stat Comput Simul 47:37–47

[77] Duchi J, Shalev-Shwartz S, Singer Y, Chandra T (2008) Efficient projections onto the l_1-ball for learning in high dimensions. In: Proceedings of the 25th international conference on machine learning (ICML 2008). ACM, New York, pp 272–279

[78] Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge

[79] Dykstra RL (1983) An algorithm for restricted least squares estimation. J Am Stat Assoc 78:837–842

[80] Edgeworth FY (1887) On observations relating to several quantities. Hermathena 6:279–285

[81] Edgeworth FY (1888) On a new method of reducing observations relating to several quantities. Phil Mag 25:184–191

[82] Edwards CH Jr (1973) Advanced calculus of several variables. Academic, New York

[83] Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32:407–499

[84] Ekeland I (1974) On the variational principle. J Math Anal Appl 47:324–353

[85] Elsner L, Koltracht L, Neumann M (1992) Convergence of sequential and asynchronous nonlinear paracontractions. Numer Math 62:305–319

[86] Everitt BS (1977) The analysis of contingency tables. Chapman & Hall, London

[87] Fang S-C, Puthenpura S (1993) Linear optimization and extensions: theory and algorithms. Prentice-Hall, Englewood Cliffs

[88] Fazel M, Hindi M, Boyd S (2003) Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices. Proc Am Contr Conf 3:2156–2162

[89] Feller W (1971) An introduction to probability theory and its applications, vol 2, 2nd edn. Wiley, Hoboken

[90] Fessler JA, Clinthorne NH, Rogers WL (1993) On complete-data spaces for PET reconstruction algorithms. IEEE Trans Nucl Sci 40:1055–1061

[91] Fiacco AV, McCormick GP (1968) Nonlinear programming: sequential unconstrained minimization techniques. Wiley, Hoboken

[92] Fletcher R (2000) Practical methods of optimization, 2nd edn. Wiley, Hoboken

[93] Fletcher R, Powell MJD (1963) A rapidly convergent descent method for minimization. Comput J 6:163–168

[94] Fletcher R, Reeves CM (1964) Function minimization by conjugate gradients. Comput J 7:149–154

[95] Flury B, Zoppè A (2000) Exercises in EM. Am Stat 54:207–209

[96] Forsgren A, Gill PE, Wright MH (2002) Interior point methods for nonlinear optimization. SIAM Rev 44:523–597

[97] Franklin J (1983) Mathematical methods of economics. Am Math Mon 90:229–244

[98] Friedman J, Hastie T, Tibshirani R (2007) Pathwise coordinate optimization. Ann Appl Stat 1:302–332

[99] Friedman J, Hastie T, Tibshirani R (2009) Regularized paths for generalized linear models via coordinate descent. Technical Report, Department of Statistics, Stanford University

[100] Fu WJ (1998) Penalized regressions: the bridge versus the lasso. J Comput Graph Stat 7:397–416

[101] Gabriel KR, Zamir S (1979) Lower rank approximation of matrices by least squares with any choice of weights. Technometrics 21:489–498

[102] Gelfand IM, Fomin SV (1963) Calculus of variations. Prentice-Hall, Englewood Cliffs

[103] Geman S, McClure D (1985) Bayesian image analysis: an application to single photon emission tomography. In: Proceedings of the statistical computing section. American Statistical Association, Washington, DC, pp 12–18

[104] Gifi A (1990) Nonlinear multivariate analysis. Wiley, Hoboken

[105] Gill PE, Murray W, Wright MH (1991) Numerical linear algebra and optimization, vol 1. Addison-Wesley, Redwood City

[106] Goldstein T, Osher S (2009) The split Bregman method for ℓ_1-regularized problems. SIAM J Imag Sci 2:323–343

[107] Golub GH, Van Loan CF (1996) Matrix computations, 3rd edn. Johns Hopkins University Press, Baltimore

[108] Gordon RA (1998) The use of tagged partitions in elementary real analysis. Am Math Mon 105:107–117

[109] Gould NIM (2008) How good are projection methods for convex feasibility problems? Comput Optim Appl 40:1–12

[110] Green PJ (1984) Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives (with discussion). J Roy Stat Soc B 46:149–192

[111] Green PJ (1990) Bayesian reconstruction for emission tomography data using a modified EM algorithm. IEEE Trans Med Imag 9:84–94

[112] Green PJ (1990) On use of the EM algorithm for penalized likelihood estimation. J Roy Stat Soc B 52:443–452

[113] Grimmett GR, Stirzaker DR (1992) Probability and random processes, 2nd edn. Oxford University Press, Oxford

[114] Groenen PJF, Nalbantov G, Bioch JC (2007) Nonlinear support vector machines through iterative majorization and I-splines. In: Lenz HJ, Decker R (eds) Studies in classification, data analysis, and knowledge organization. Springer, Heidelberg, pp 149–161

[115] Guillemin V, Pollack A (1974) Differential topology. Prentice-Hall, Englewood Cliffs

[116] Güler O (2010) Foundations of optimization. Springer, New York

[117] Hämmerlin G, Hoffmann K-H (1991) Numerical mathematics. Springer, New York

[118] Hardy GH, Littlewood JE, Pólya G (1952) Inequalities, 2nd edn. Cambridge University Press, Cambridge

[119] Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York

[120] He L, Marquina A, Osher S (2005) Blind deconvolution using TV regularization and Bregman iteration. Int J Imag Syst Technol 15, 74–83

[121] Heiser WJ (1987) Correspondence analysis with least absolute residuals. Comput Stat Data Anal 5:337–356

[122] Heiser WJ (1995) Convergent computing by iterative majorization: theory and applications in multidimensional data analysis. In: Krzanowski WJ (ed) Recent advances in descriptive multivariate analysis. Clarendon, Oxford, pp 157–189

[123] Henrici P (1982) Essentials of numerical analysis with pocket calculator demonstrations. Wiley, Hoboken

[124] Herman GT (1980) Image reconstruction from projections: the fundamentals of computerized tomography. Springer, New York

[125] Hestenes MR (1981) Optimization theory: the finite dimensional case. Robert E Krieger Publishing, Huntington

[126] Hestenes MR, Karush WE (1951) A method of gradients for the calculation of the characteristic roots and vectors of a real symmetric matrix. J Res Natl Bur Stand 47:471–478

[127] Hestenes MR, Stiefel E (1952) Methods of conjugate gradients for solving linear systems. J Res Natl Bur Stand 29:409–439

[128] Higham NJ (2008) Functions of matrices: theory and computation. SIAM, Philadelphia

[129] Hille E (1959) Analytic function theory, vol 1. Blaisdell, New York

[130] Hiriart-Urruty J-B (1986) When is a point x satisfying $\nabla f(x) = \mathbf{0}$ a global minimum of $f(x)$? Am Math Mon 93:556–558

[131] Hiriart-Urruty J-B, Claude Lemaréchal C (2001) Fundamentals of convex analysis. Springer, New York

[132] Hochstadt H (1986) The functions of mathematical physics. Dover, New York

[133] Hoel PG, Port SC, Stone CJ (1971) Introduction to probability theory. Houghton Mifflin, Boston

[134] Hoffman K (1975) Analysis in Euclidean space. Prentice-Hall, Englewood Cliffs

[135] Hoffman K, Kunze R (1971) Linear algebra, 2nd edn. Prentice-Hall, Englewood Cliffs

[136] Horn RA, Johnson CR (1985) Matrix analysis. Cambridge University Press, Cambridge

[137] Horn RA, Johnson CR (1991) Topics in matrix analysis. Cambridge University Press, Cambridge

[138] Householder AS (1975) The theory of matrices in numerical analysis. Dover, New York

[139] Hrusa W, Troutman JL (1981) Elementary characterization of classical minima. Am Math Mon 88:321–327

[140] Hunter DR (2004) MM algorithms for generalized Bradley-Terry models. Ann Stat 32:386–408

[141] Hunter DR, Lange K (2000) Quantile regression via an MM algorithm. J Comput Graph Stat 9:60–77

[142] Hunter DR, Lange K (2004) A tutorial on MM algorithms. Am Stat 58:30–37

[143] Hunter DR, Li R (2005) Variable selection using MM algorithms. Ann Stat 33:1617–1642

[144] Jamshidian M, Jennrich RI (1995) Acceleration of the EM algorithm by using quasi-Newton methods. J Roy Stat Soc B 59:569–587

[145] Jamshidian M, Jennrich RI (1997) Quasi-Newton acceleration of the EM algorithm. J Roy Stat Soc B 59:569–587

[146] Jennrich RI, Moore RH (1975) Maximum likelihood estimation by means of nonlinear least squares. In: Proceedings of the statistical computing section. American Statistical Association, Washington, DC, pp 57–65

[147] Jia R-Q, Zhao H, Zhao W (2009) Convergence analysis of the Bregman method for the variational model of image denoising. Appl Comput Harmon Anal 27:367–379

[148] Karlin S, Taylor HM (1975) A first course in stochastic processes, 2nd edn. Academic, New York

[149] Karush W (1939) Minima of functions of several variables with inequalities as side conditions. Master's Thesis, Department of Mathematics, University of Chicago, Chicago

[150] Keener JP (1993) The Perron-Frobenius theorem and the ranking of football teams. SIAM Rev 35:80–93

[151] Kelley CT (1999) Iterative methods for optimization. SIAM, Philadelphia

[152] Khalfan HF, Byrd RH, Schnabel RB (1993) A theoretical and experimental study of the symmetric rank-one update. SIAM J Optim 3:1–24

[153] Kiers HAL (1997) Weighted least squares fitting using ordinary least squares algorithms. Psychometrika 62:251–266

[154] Kingman JFC (1993) Poisson processes. Oxford University Press, Oxford

[155] Komiya H (1988) Elementary proof for Sion's minimax theorem. Kodai Math J 11:5–7

[156] Kosowsky JJ, Yuille AL (1994) The invisible hand algorithm: solving the assignment problem with statistical physics. Neural Network 7:477–490

[157] Kruskal JB (1964) Nonmetric multidimensional scaling: a numerical method. Psychometrika 29:115–129

[158] Kruskal JB (1965) Analysis of factorial experiments by estimating monotone transformations of the data. J Roy Stat Soc B 27:251–263

[159] Ku HH, Kullback S (1974) Log-linear models in contingency table analysis. Biometrics 10:452–458

[160] Kuhn S (1991) The derivative á la Carathéodory. Am Math Mon 98:40–44

[161] Kuhn HW, Tucker AW (1951) Nonlinear programming. In: Proceedings of the second Berkeley symposium on mathematical statistics and probability. University of California Press, Berkeley

[162] Lange K (1994) An adaptive barrier method for convex programming. Meth Appl Anal 1:392–402

[163] Lange K (1995) A gradient algorithm locally equivalent to the EM algorithm. J Roy Stat Soc B 57:425–437

[164] Lange K (1995) A quasi-Newton acceleration of the EM algorithm. Stat Sin 5:1–18

[165] Lange K (2002) Mathematical and statistical methods for genetic analysis, 2nd edn. Springer, New York

[166] Lange K (2010) Numerical analysis for statisticians, 2nd edn. Springer, New York

[167] Lange K, Carson R (1984) EM reconstruction algorithms for emission and transmission tomography. J Comput Assist Tomogr 8:306–316

[168] Lange K, Fessler JA (1995) Globally convergent algorithms for maximum a posteriori transmission tomography. IEEE Trans Image Process 4:1430–1438

[169] Lange K, Wu T (2008) An MM algorithm for multicategory vertex discriminant analysis. J Comput Graph Stat 17:527–544

[170] Lange K, Zhou H (2012) MM algorithms for geometric and signomial programming. Math Program, Series A, DOI 10.1007/s10107-012-0612-1

[171] Lange K, Hunter D, Yang I (2000) Optimization transfer using surrogate objective functions (with discussion). J Comput Graph Stat 9:1–59

[172] Lax PD (2007) Linear algebra and its applications, 2nd edn. Wiley, Hoboken

[173] Ledoita O, Wolf M (2004) A well-conditioned estimator for large-dimensional covariance matrices. J Multivar Anal 88:365–411

[174] Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791

[175] Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. Adv Neural Inf Process Syst 13:556–562

[176] Lehmann EL (1986) Testing statistical hypotheses, 2nd edn. Wiley, Hoboken

[177] Levina E, Rothman A, Zhu J (2008) Sparse estimation of large covariance matrices via a nested lasso penalty. Ann Appl Stat 2:245–263

[178] Li Y, Arce GR (2004) A maximum likelihood approach to least absolute deviation regression. EURASIP J Appl Signal Process 2004:1762–1769

[179] Little RJA, Rubin DB (1987) Statistical analysis with missing data. Wiley, Hoboken

[180] Louis TA (1982) Finding the observed information matrix when using the EM algorithm. J Roy Stat Soc B 44:226–233

[181] Luce RD (1959) Individual choice behavior: a theoretical analysis. Wiley, Hoboken

[182] Luce RD (1977) The choice axiom after twenty years. J Math Psychol 15:215–233

[183] Luenberger DG (1984) Linear and nonlinear programming, 2nd edn. Addison-Wesley, Reading

[184] Magnus JR, Neudecker H (1988) Matrix differential calculus with applications in statistics and econometrics. Wiley, Hoboken

[185] Maher MJ (1982) Modelling association football scores. Stat Neerl 36:109–118

[186] Mangasarian OL, Fromovitz S (1967) The Fritz John necessary optimality conditions in the presence of equality and inequality constraints. J Math Anal Appl 17:37–47

[187] Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis. Academic, New York

[188] Marsden JE, Hoffman MJ (1993) Elementary classical analysis, 2nd edn. W H Freeman & Co, New York

[189] Mazumder R, Hastie T, Tibshirani R (2010) Spectral regularization algorithms for learning large incomplete matrices. J Mach Learn Res 11:2287–2322

[190] McLachlan GJ, Do K-A, Ambroise C (2004) Analyzing microarray gene expression data. Wiley, Hoboken

[191] McLachlan GJ, Krishnan T (2008) The EM algorithm and extensions, 2nd edn. Wiley, Hoboken

[192] McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, Hoboken

[193] McLeod RM (1980) The generalized Riemann integral. Mathematical Association of America, Washington, DC

[194] McShane EJ (1973) The Lagrange multiplier rule. Am Math Mon 80:922–925

[195] Meyer RR (1976) Sufficient conditions for the convergence of monotonic mathematical programming algorithms. J Comput Syst Sci 12:108–121

[196] Michelot C (1986) A finite algorithm for finding the projection of a point onto the canonical simplex in R^n. J Optim Theor Appl 50:195–200

[197] Miller KS (1987) Some eclectic matrix theory. Robert E Krieger Publishing, Malabar

[198] Moré JJ, Sorensen DC (1983) Computing a trust region step. SIAM J Sci Stat Comput 4:553–572

[199] Narayanan A (1991) Algorithm AS 266: maximum likelihood estimation of the parameters of the Dirichlet distribution. Appl Stat 40:365–374

[200] Nazareth L (1979) A relationship between the BFGS and conjugate gradient algorithms and its implications for new algorithms. SIAM J Numer Anal 16:794–800

[201] Nedelman J, Wallenuis T (1986) Bernoulli trials, Poisson trials, surprising variances, and Jensen's inequality. Am Stat 40:286–289

[202] Nelder JA, Wedderburn RWM (1972) Generalized linear models. J Roy Stat Soc A 135:370–384

[203] Nemirovski AS, Todd MJ (2008) Interior-point methods for optimization. Acta Numerica 17:191–234

[204] Nocedal J (1991) Theory of algorithms for unconstrained optimization. Acta Numerica 1991:199–242

[205] Nocedal J, Wright S (2006) Numerical optimization, 2nd edn. Springer, New York

[206] Orchard T, Woodbury MA (1972) A missing information principle: theory and applications. In: Proceedings of the 6th Berkeley symposium on mathematical statistics and probability. University of California Press, Berkeley, pp 697–715

[207] Ortega JM (1990) Numerical analysis: a second course. Society for Industrial and Applied Mathematics, Philadelphia

[208] Osher S, Burger M, Goldfarb D, Xu J, Yin W (2005) An iterative regularization method for total variation based image restoration. Multiscale Model Simul 4:460–489

[209] Osher S, Mao T, Dong B, Yin W (2011) Fast linearized Bregman iteration for compressive sensing and sparse denoising. Comm Math Sci 8:93–111

[210] Park MY, Hastie T (2008) Penalized logistic regression for detecting gene interactions. Biostatistics 9:30–50

[211] Pauca VP, Piper J, Plemmons RJ (2006) Nonnegative matrix factorization for spectral data analysis. Linear Algebra Appl 416:29–47

[212] Peressini AL, Sullivan FE, Uhl JJ Jr (1988) The mathematics of nonlinear programming. Springer, New York

[213] Polya G (1954) Induction and analogy in mathematics. Volume I of mathematics and plausible reasoning. Princeton University Press, Princeton

[214] Portnoy S, Koenker R (1997) The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. Stat Sci 12:279–300

[215] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in Fortran: the art of scientific computing, 2nd edn. Cambridge University Press, Cambridge

[216] Rabiner L (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–285

[217] Ranola JM, Ahn S, Sehl ME, Smith DJ, Lange K (2010) A Poisson model for random multigraphs. Bioinformatics 26:2004–2011

[218] Rao CR (1973) Linear statistical inference and its applications, 2nd edn. Wiley, Hoboken

[219] Robertson T, Wright FT, Dykstra RL (1988) Order restricted statistical inference. Wiley, Hoboken

[220] Romano G (1995) New results in subdifferential calculus with applications to convex analysis. Appl Math Optim 32:213–234

[221] Rockafellar RT (1996) Convex analysis. Princeton University Press, Princeton

[222] Royden HL (1988) Real analysis, 3rd edn. Macmillan, London

[223] Rudin W (1979) Principles of mathematical analysis, 3rd edn. McGraw-Hill, New York

[224] Rudin LI, Osher S, Fatemi E (1992) Nonlinear total variation based noise removal algorithms. Physica D 60:259–268

[225] Rustagi JS (1976) Variational methods in statistics. Academic, New York

[226] Ruszczyński A (2006) Nonlinear optimization. Princeton University Press, Princeton

[227] Sabatti C, Lange K (2002) Genomewide motif identification using a dictionary model. Proc IEEE 90:1803–1810

[228] Sagan H (1969) Introduction to the calculus of variations. McGraw-Hill, New York

[229] Santosa F, Symes WW (1986) Linear inversion of band-limited reflection seimograms. SIAM J Sci Stat Comput 7:1307–1330

[230] Schmidt M, van den Berg E, Friedlander MP, Murphy K (2009) Optimizing costly functions with simple constraints: a limited-memory projected quasi-Newton algorithm. In: van Dyk D, Welling M (eds) Proceedings of The twelfth international conference on artificial intelligence and statistics (AISTATS), vol 5, pp 456–463

[231] Schölkopf B, Smola AJ (2002) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT, Cambridge

[232] Seber GAF, Lee AJ (2003) Linear regression analysis, 2nd edn. Wiley, Hoboken

[233] Segel LA (1977) Mathematics applied to continuum mechanics. Macmillan, New York

[234] Seneta E (1973) Non-negative matrices: an introduction to theory and applications. Wiley, Hoboken

[235] Sha F, Saul LK, Lee DD (2003) Multiplicative updates for nonnegative quadratic programming in support vector machines. In: Becker S, Thrun S, Obermayer K (eds) Advances in neural information processing systems 15. MIT, Cambridge, pp 1065–1073

[236] Silvapulle MJ, Sen PK (2005) Constrained statistical inference. Wiley, Hoboken

[237] Sinkhorn R (1967) Diagonal equivalence to matrices with prescribed row and column sums. Am Math Mon 74:402–405

[238] Sion M (1958) On general minimax theorems. Pac J Math 8:171–176

[239] Smith CAB (1957) Counting methods in genetical statistics. Ann Hum Genet 21:254–276

[240] Smith DR (1974) Variational methods in optimization. Dover, Mineola

[241] Sorensen DC (1997) Minimization of a large-scale quadratic function subject to spherical constraints. SIAM J Optim 7:141–161

[242] Srebro N, Jaakkola T (2003) Weighted low-rank approximations. In: Machine learning international workshop conference 2003. AAAI Press, 20:720–727

[243] Steele JM (2004) The Cauchy-Schwarz master class: an introduction to the art of inequalities. Cambridge University Press and the Mathematical Association of America, Cambridge

[244] Stein EM, Shakarchi R (2003) Complex analysis. Princeton University Press, Princeton

[245] Stern RJ, Wolkowicz H (1995) Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations. SIAM J Optim 5:286–313

[246] Stoer J, Bulirsch R (2002) Introduction to numerical analysis, 3rd edn. Springer, New York

[247] Strang G (1986) Introduction to applied mathematics. Wellesley-Cambridge, Wellesley

[248] Strang G (1986) The fundamental theorem of linear algebra. Am Math Mon 100:848–855

[249] Strang G (2003) Introduction to linear algebra, 3rd edn. Wellesley-Cambridge, Wellesley

[250] Swartz C, Thomson BS (1988) More on the fundamental theorem of calculus. Am Math Mon 95:644–648

[251] Tanner MA (1993) Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions, 2nd edn. Springer, New York

[252] Taylor H, Banks SC, McCoy JF (1979) Deconvolution with the ℓ_1 norm. Geophysics 44:39–52

[253] Teboulle M (1992) Entropic proximal mappings with applications to nonlinear programming. Math Oper Res 17:670–690

[254] Theobald CM (1975) An inequality for the trace of the product of two symmetric matrices. Math Proc Camb Phil Soc 77:265–267

[255] Thompson HB (1989) Taylor's theorem using the generalized Riemann integral. Am Math Mon 96:346–350

[256] Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc B 58:267–288

[257] Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused lasso. J Roy Stat Soc B 67:91–108

[258] Tikhomirov VM (1990) Stories about maxima and minima. American Mathematical Society, Providence

[259] Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley, Hoboken

[260] Trefethen LN, Bau D (1997) Numerical linear algebra. SIAM, Philadelphia

[261] Uherka DJ, Sergott AM (1977) On the continuous dependence of the roots of a polynomial on its coefficients. Am Math Mon 84:368–370

[262] Vandenberghe L, Boyd S, Wu S (1998) Determinant maximization with linear matrix inequality constraints. SIAM J Matrix Anal Appl 19:499–533

[263] Van Ruitenburg J (2005) Algorithms for parameter estimation in the Rasch model. Measurement and Research Department Reports 2005–4. CITO, Arnhem

[264] Vapnik V (1995) The nature of statistical learning theory. Springer, New York

[265] Vardi Y, Shepp LA, Kaufman L (1985) A statistical model for positron emission tomography. J Am Stat Assoc 80:8–37

[266] Von Neumann J (1928) Zur theorie der gesellschaftsspiele. Math Ann 100:295–320

[267] Wang L, Gordon MD, Zhu J (2006) Regularized least absolute deviations regression and an efficient algorithm for parameter tuning. In: Proceedings of the sixth international conference on data mining (ICDM'06). IEEE Computer Society, Washington, DC, pp 690–700

[268] Wang S, Yehya N, Schadt EE, Wang H, Drake TA, Lusis AJ (2006) Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. PLoS Genet 2:148–159

[269] Watson GA (1992) Characterization of the subdifferential of some matrix norms. Linear Algebra Appl 170:1039–1053

[270] Weeks DE, Lange K (1989) Trials, tribulations, and triumphs of the EM algorithm in pedigree analysis. IMA J Math Appl Med Biol 6:209–232

[271] Weiszfeld E (1937) On the point for which the sum of the distances to n given points is minimum. Ann Oper Res 167:741 (Translated from the French original [Tohoku Math J 43:335–386 (1937)] and annotated by Frank Plastria)

[272] Weston J, Elisseeff A, Schölkopf B, Tipping M (2003) Use of the zero-norm with linear models and kernel methods. J Mach Learn Res 3:1439–1461

[273] Whyte BM, Gold J, Dobson AJ, Cooper DA (1987) Epidemiology of acquired immunodeficiency syndrome in Australia. Med J Aust 147:65–69

[274] Wright MH (2005) The interior-point revolution in optimization: history, recent developments, and lasting consequences. Bull Am Math Soc 42:39–56

[275] Wu CF (1983) On the convergence properties of the EM algorithm. Ann Stat 11:95–103

[276] Wu TT, Lange K (2008) Coordinate descent algorithms for lasso penalized regression. Ann Appl Stat 2:224–244

[277] Wu TT, Lange K (2010) Multicategory vertex discriminant analysis for high-dimensional data. Ann Appl Stat 4:1698–1721

[278] Yee PL, Vyb́orný R (2000) The integral: an easy approach after Kurzweil and Henstock. Cambridge University Press, Cambridge

[279] Yin W, Osher S, Goldfarb D, Darbon J (2008) Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. SIAM J Imag Sci 1:143–168

[280] Zhang Z, Lange K, Ophoff R, Sabatti C (2010) Reconstructing DNA copy number by penalized estimation and imputation. Ann Appl Stat 4:1749–1773

[281] Zhou H, Lange K (2009) On the bumpy road to the dominant mode. Scand J Stat 37:612–631

[282] Zhou H, Lange K (2010) MM algorithms for some discrete multivariate distributions. J Comput Graph Stat 19:645–665

[283] Zhou H, Lange K (2012) A path algorithm for constrained estimation. J Comput Graph Stat DOI 10.1080/10618600.2012.681248

[284] Zhou H, Lange K (2012) Path following in the exact penalty method of convex programming (submitted)

Index

K. Lange, *Optimization*, Springer Texts in Statistics 95,
DOI 10.1007/978-1-4614-5838-8,