NONLINEAR PROGRAMMING

Theory and Algorithms

Third Edition

MOKHTAR S. BAZARAA
Georgia Institute of Technology
School of Industrial and Systems Engineering
Atlanta, Georgia

HANIF D. SHERALI
Virginia Polytechnic Institute and State University
Grado Department of Industrial and Systems Engineering
Blacksburg, Virginia

C. M. SHETTY
Georgia Institute of Technology
School of Industrial and Systems Engineering
Atlanta, Georgia

A JOHN WILEY & SONS, INC., PUBLICATION

Contents

Chapter 3 Convex Functions and Generalizations

Convex and concave functions have many special and important properties. For example, any local minimum of a convex function over a convex set is also a global minimum. In this chapter we introduce the important topics of convex and concave functions and develop some of their properties. As we shall learn in this and later chapters, these properties can be utilized in developing suitable optimality conditions and computational schemes for optimization problems that involve convex and concave functions.

Following is an outline of the chapter.

Section 3.1: Definitions and Basic Properties We introduce convex and concave functions and develop some of their basic properties. Continuity of convex functions is proved, and the concept of a directional derivative is introduced.

Section 3.2: Subgradients of Convex Functions A convex function has a convex epigraph and hence has a supporting hyperplane. This leads to the important notion of a subgradient of a convex function.

Section 3.3: Differentiable Convex Functions In this section we give some characterizations of differentiable convex functions. These are helpful tools for checking convexity of simple differentiable functions.

Section 3.4: Minima and Maxima of Convex Functions This section is important, since it deals with the questions of minimizing and maximizing a convex function over a convex set. A necessary and sufficient condition for a minimum is developed, and we provide a characterization for the set of alternative optimal solutions. We also show that the maximum occurs at an extreme point. This fact is particularly important if the convex set is polyhedral.

Section 3.5: Generalizations of Convex Functions Various relaxations of convexity and concavity are possible. We present quasiconvex and pseudoconvex functions and develop some of their properties. We then discuss various types of convexity at a point. These types of convexity are sometimes sufficient for optimality, as shown in Chapter 4. (This section can be omitted by beginning readers, and later references to generalized convexity properties can largely be substituted simply by convexity.)

3.1 Definitions and Basic Properties

In this section we deal with some basic properties of convex and concave functions. In particular, we investigate their continuity and differentiability properties.

3.1.1 Definition

Let $f: S \to R$, where S is a nonempty convex set in R^n. The function f is said to be *convex* on S if

$$f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \le \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2)$$

for each $\mathbf{x}_1, \mathbf{x}_2 \in S$ and for each $\lambda \in (0, 1)$. The function f is called *strictly convex* on S if the above inequality is true as a strict inequality for each distinct $\mathbf{x}_1$ and $\mathbf{x}_2$ in S and for each $\lambda \in (0, 1)$. The function $f: S \to R$ is called *concave* (*strictly concave*) on S if $-f$ is convex (strictly convex) on S.

Now let us consider the geometric interpretation of convex and concave functions. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two distinct points in the domain of f, and consider the point $\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$, with $\lambda \in (0, 1)$. Note that $\lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2)$ gives the weighted average of $f(\mathbf{x}_1)$ and $f(\mathbf{x}_2)$, while $f[\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2]$ gives the value of f at the point $\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$. So for a convex function f, the value of f at points on the line segment $\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ is less than or equal to the height of the chord joining the points $[\mathbf{x}_1, f(\mathbf{x}_1)]$ and $[\mathbf{x}_2, f(\mathbf{x}_2)]$. For a concave function, the chord is (on or) below the function itself. Hence, a function is both convex and concave if and only if it is *affine*. Figure 3.1 shows some examples of convex and concave functions.

The following are some examples of convex functions. By taking the negatives of these functions, we get some examples of concave functions.

1. $f(x) = 3x + 4$.
2. $f(x) = |x|$.
3. $f(x) = x^2 - 2x$.
4. $f(x) = -x^{1/2}$ if $x \ge 0$.
5. $f(x_1, x_2) = 2x_1^2 + x_2^2 - 2x_1 x_2$.
6. $f(x_1, x_2, x_3) = x_1^4 + 2x_2^2 + 3x_3^2 - 4x_1 - 4x_2 x_3$.

Note that in each of the above examples, except for Example 4, the function f is convex over R^n. In Example 4 the function is not defined for $x < 0$. One can readily construct examples of functions that are convex over a region but not over R^n. For instance, $f(x) = x^3$ is not convex over R but is convex over $S = \{x : x \ge 0\}$.

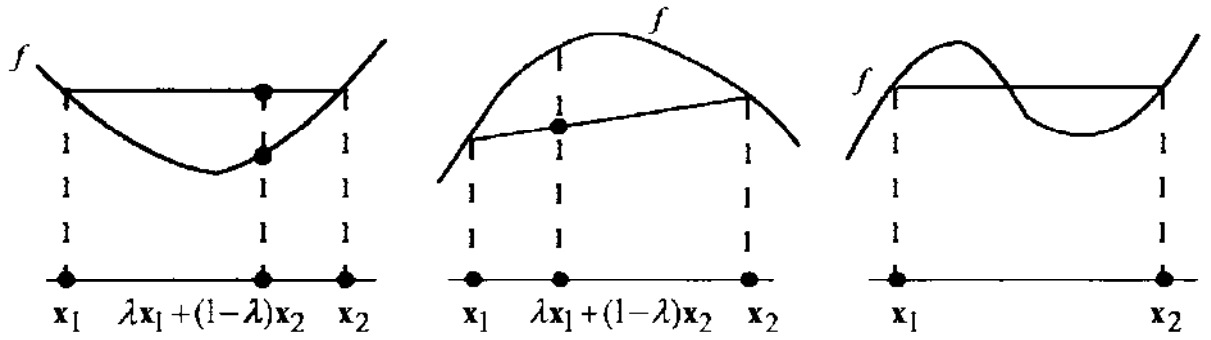


Figure 3.1 Convex and concave functions.

The examples above cite some arbitrary illustrative instances of convex functions. In contrast, we give below some particularly important instances of convex functions that arise very often in practice and that are useful to remember.

1 Let $f_1, f_2, \ldots, f_k: R^n \to R$ be convex functions. Then:

(a) $f(\mathbf{x}) = \sum_{j=1}^{k} \alpha_j f_j(\mathbf{x})$, where $\alpha_j > 0$ for $j = 1, 2, \ldots, k$ is a convex function (see Exercise 3.8).

(b) $f(\mathbf{x}) = \max\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_k(\mathbf{x})\}$ is a convex function (see Exercise 3.9).

2. Suppose that g: $R^n \to R$ is a concave function. Let $S = \{\mathbf{x} : g(\mathbf{x}) > 0\}$, and define f: $S \to R$ as $f(\mathbf{x}) = 1/g(\mathbf{x})$. Then f is convex over S (see Exercise 3.11).

3. Let g: $R \to R$ be a nondecreasing, univariate, convex function, and let h: $R^n \to R$ be a convex function. Then the composite function f: $R^n \to R$ defined as $f(\mathbf{x}) = g[h(\mathbf{x})]$ is a convex function (see Exercise 3.10).

4. Let g: $R^m \to R$ be a convex function, and let $\mathbf{h}$: $R^n \to R^m$ be an affine function of the form $\mathbf{h}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$, where $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{b}$ is an $m \times 1$ vector. Then the composite function f: $R^n \to R$ defined as $f(\mathbf{x}) = g[\mathbf{h}(\mathbf{x})]$ is a convex function (see Exercise 3.16).

From now on, we concentrate on convex functions. Results for concave functions can be obtained easily by noting that f is concave if and only if $-f$ is convex.

Associated with a convex function f is the set $S_\alpha = \{\mathbf{x} \in S : f(\mathbf{x}) \leq \alpha\}$, $\alpha \in R$, usually referred to as a *level set*. Sometimes this set is called a *lower-level set*, to differentiate it from the *upper-level set* $\{\mathbf{x} \in S : f(\mathbf{x}) \geq \alpha\}$, which has

properties similar to these for concave functions. Lemma 3.1.2 shows that S_α is convex for each real number α. Hence, if g_i: $R^n \to R$ is convex for $i = 1,\ldots, m$, the set $\{\mathbf{x} : g_i(\mathbf{x}) \le 0, i = 1,\ldots, m\}$ is a convex set.

3.1.2 Lemma

Let S be a nonempty convex set in R^n, and let f: $S \to R$ be a convex function. Then the level set $S_\alpha = \{\mathbf{x} \in S : f(\mathbf{x}) \le \alpha\}$, where α is a real number, is a convex set.

Proof

Let $\mathbf{x}_1, \mathbf{x}_2 \in S_\alpha$. Thus, $\mathbf{x}_1, \mathbf{x}_2 \in S$ and $f(\mathbf{x}_1) \le \alpha$ and $f(\mathbf{x}_2) \le \alpha$. Now let $\lambda \in (0, 1)$ and $\mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$. By the convexity of S, we have that $\mathbf{x} \in S$. Furthermore, by the convexity of f,

$$f(\mathbf{x}) \le \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2) \le \lambda\alpha + (1-\lambda)\alpha = \alpha.$$

Hence, $\mathbf{x} \in S_\alpha$, and therefore, S_α is convex.

Continuity of Convex Functions

An important property of convex and concave functions is that they are continuous on the interior of their domain. This fact is proved below.

3.1.3 Theorem

Let S be a nonempty convex set in R^n, and let f: $S \to R$ be convex. Then f is continuous on the interior of S.

Proof

Let $\bar{\mathbf{x}} \in \text{int } S$. To prove continuity of f at $\bar{\mathbf{x}}$, we need to show that given $\varepsilon > 0$, there exists a $\delta > 0$ such that $\|\mathbf{x} - \bar{\mathbf{x}}\| \le \delta$ implies that $|f(\mathbf{x}) - f(\bar{\mathbf{x}})| \le \varepsilon$. Since $\bar{\mathbf{x}} \in \text{int } S$, there exists a $\delta' > 0$ such that $\|\mathbf{x} - \bar{\mathbf{x}}\| \le \delta'$ implies that $\mathbf{x} \in S$. Construct θ as follows.

$$\theta = \max_{1 \le i \le n} \{\max[f(\bar{\mathbf{x}} + \delta'\mathbf{e}_i) - f(\bar{\mathbf{x}}), f(\bar{\mathbf{x}} - \delta'\mathbf{e}_i) - f(\bar{\mathbf{x}})]\}, \tag{3.1}$$

where $\mathbf{e}_i$ is a vector of zeros except for a 1 at the ith position. Note that $0 \le \theta < \infty$. Let

$$\delta = \min\left(\frac{\delta'}{n}, \frac{\varepsilon\delta'}{n\theta}\right). \tag{3.2}$$

Choose an $\mathbf{x}$ with $\|\mathbf{x}-\bar{\mathbf{x}}\| \le \delta$. If $x_i - \bar{x}_i \ge 0$, let $\mathbf{z}_i = \delta'\mathbf{e}_i$; otherwise, let $\mathbf{z}_i = -\delta'\mathbf{e}_i$. Then $\mathbf{x}-\bar{\mathbf{x}} = \sum_{i=1}^{n}\alpha_i\mathbf{z}_i$, where $\alpha_i \ge 0$ for $i = 1,\ldots, n$. Furthermore,

$$\|\mathbf{x}-\bar{\mathbf{x}}\| = \delta'\left(\sum_{i=1}^{n}\alpha_i^2\right)^{1/2}. \tag{3.3}$$

From (3.2), and since $\|\mathbf{x}-\bar{\mathbf{x}}\| \le \delta$, it follows that $\alpha_i \le 1/n$ for $i = 1,\ldots, n$. Hence, by the convexity of f, and since $0 \le n\alpha_i \le 1$, we get

$$\begin{aligned} f(\mathbf{x}) = f\left(\bar{\mathbf{x}} + \sum_{i=1}^{n}\alpha_i\mathbf{z}_i\right) &= f\left[\frac{1}{n}\sum_{i=1}^{n}(\bar{\mathbf{x}} + n\alpha_i\mathbf{z}_i)\right] \\ &\le \frac{1}{n}\sum_{i=1}^{n} f(\bar{\mathbf{x}} + n\alpha_i\mathbf{z}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} f[(1-n\alpha_i)\bar{\mathbf{x}} + n\alpha_i(\bar{\mathbf{x}} + \mathbf{z}_i)] \\ &\le \frac{1}{n}\sum_{i=1}^{n}[(1-n\alpha_i)f(\bar{\mathbf{x}}) + n\alpha_i f(\bar{\mathbf{x}} + \mathbf{z}_i)]. \end{aligned}$$

Therefore, $f(\mathbf{x}) - f(\bar{\mathbf{x}}) \le \sum_{i=1}^{n}\alpha_i[f(\bar{\mathbf{x}} + \mathbf{z}_i) - f(\bar{\mathbf{x}})]$. From (3.1) it is obvious that $f(\bar{\mathbf{x}} + \mathbf{z}_i) - f(\bar{\mathbf{x}}) \le \theta$ for each i; and since $\alpha_i \ge 0$, it follows that

$$f(\mathbf{x}) - f(\bar{\mathbf{x}}) \le \theta\sum_{i=1}^{n}\alpha_i. \tag{3.4}$$

Noting (3.3) and (3.2), it follows that $\alpha_i \le \varepsilon/n\theta$, and (3.4) implies that $f(\mathbf{x}) - f(\bar{\mathbf{x}}) \le \varepsilon$. So far, we have shown that $\|\mathbf{x}-\bar{\mathbf{x}}\| \le \delta$ implies that $f(\mathbf{x}) - f(\bar{\mathbf{x}}) \le \varepsilon$. By definition, this establishes the *upper semicontinuity* of f at $\bar{\mathbf{x}}$. To complete the proof, we need to establish the *lower semicontinuity* of f at $\bar{\mathbf{x}}$ as well, that is, to show that $f(\bar{\mathbf{x}}) - f(\mathbf{x}) \le \varepsilon$. Let $\mathbf{y} = 2\bar{\mathbf{x}} - \mathbf{x}$ and note that $\|\mathbf{y}-\bar{\mathbf{x}}\| \le \delta$. Therefore, as above,

$$f(\mathbf{y}) - f(\bar{\mathbf{x}}) \le \varepsilon. \tag{3.5}$$

But $\bar{\mathbf{x}} = (1/2)\mathbf{y} + (1/2)\mathbf{x}$, and by the convexity of f, we have

$$f(\bar{\mathbf{x}}) \le (1/2)f(\mathbf{y}) + (1/2)f(\mathbf{x}). \tag{3.6}$$

Combining (3.5) and (3.6) above, it follows that $f(\bar{\mathbf{x}}) - f(\mathbf{x}) \le \varepsilon$, and the proof is complete.

Note that convex and concave functions may not be continuous everywhere. However, by Theorem 3.1.3, points of discontinuity only are allowed at the boundary of S, as illustrated by the following convex function defined on $S = \{x : -1 \le x \le 1\}$:

$$f(x) = \begin{cases} x^2 & \text{for } |x| < 1 \\ 2 & \text{for } |x| = 1. \end{cases}$$

Directional Derivative of Convex Functions

The concept of directional derivatives is particularly useful in the motivation and development of some optimality criteria and computational procedures in nonlinear programming, where one is interested in finding a direction along which the function decreases or increases.

3.1.4 Definition

Let S be a nonempty set in R^n, and let $f: S \to R$. Let $\bar{\mathbf{x}} \in S$ and $\mathbf{d}$ be a nonzero vector such that $\bar{\mathbf{x}} + \lambda \mathbf{d} \in S$ for $\lambda > 0$ and sufficiently small. The *directional derivative* of f at $\bar{\mathbf{x}}$ along the vector $\mathbf{d}$, denoted by $f'(\bar{\mathbf{x}}; \mathbf{d})$, is given by the following limit if it exists:

$$f'(\bar{\mathbf{x}}; \mathbf{d}) = \lim_{\lambda \to 0^+} \frac{f(\bar{\mathbf{x}} + \lambda \mathbf{d}) - f(\bar{\mathbf{x}})}{\lambda}.$$

In particular, the limit in Definition 3.1.4 exists for globally defined convex and concave functions as shown below. As evident from the proof of the following lemma, if $f: S \to R$ is convex on S, the limit exists if $\bar{\mathbf{x}} \in \text{int } S$, but might be $-\infty$ if $\bar{\mathbf{x}} \in \partial S$, even if f is continuous at $\bar{\mathbf{x}}$, as seen in Figure 3.2.

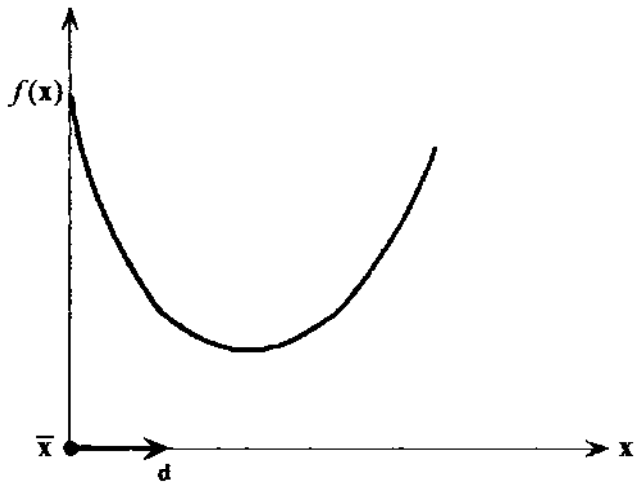


Figure 3.2 Nonexistence of the directional derivative of f at $\bar{\mathbf{x}}$ in the direction $\mathbf{d}$.

3.1.5 Lemma

Let f: $R^n \to R$ be a convex function. Consider any point $\bar{\mathbf{x}} \in R^n$ and a nonzero direction $\mathbf{d} \in R^n$. Then the directional derivative $f'(\bar{\mathbf{x}};\mathbf{d})$, of f at $\bar{\mathbf{x}}$ in the direction $\mathbf{d}$, exists.

Proof

Let $\lambda_2 > \lambda_1 > 0$. Noting the convexity of f, we have

$$f(\bar{\mathbf{x}}+\lambda_1\mathbf{d}) = f\left[\frac{\lambda_1}{\lambda_2}(\bar{\mathbf{x}}+\lambda_2\mathbf{d})+\left(1-\frac{\lambda_1}{\lambda_2}\right)\bar{\mathbf{x}}\right]$$
$$\leq \frac{\lambda_1}{\lambda_2} f(\bar{\mathbf{x}}+\lambda_2\mathbf{d})+\left(1-\frac{\lambda_1}{\lambda_2}\right) f(\bar{\mathbf{x}}).$$

This inequality implies that

$$\frac{f(\bar{\mathbf{x}}+\lambda_1\mathbf{d})-f(\bar{\mathbf{x}})}{\lambda_1} \leq \frac{f(\bar{\mathbf{x}}+\lambda_2\mathbf{d})-f(\bar{\mathbf{x}})}{\lambda_2}.$$

Thus, the difference quotient $[f(\bar{\mathbf{x}}+\lambda\mathbf{d})-f(\bar{\mathbf{x}})]/\lambda$ is monotone decreasing (nonincreasing) as $\lambda \to 0^+$.

Now, given any $\lambda \geq 0$, we also have, by the convexity of f, that

$$f(\bar{\mathbf{x}}) = f\left[\frac{\lambda}{1+\lambda}(\bar{\mathbf{x}}-\mathbf{d})+\frac{1}{1+\lambda}(\bar{\mathbf{x}}+\lambda\mathbf{d})\right]$$
$$\leq \frac{\lambda}{1+\lambda} f(\bar{\mathbf{x}}-\mathbf{d})+\frac{1}{1+\lambda} f(\bar{\mathbf{x}}+\lambda\mathbf{d}).$$

So

$$\frac{f(\bar{\mathbf{x}}+\lambda\mathbf{d})-f(\bar{\mathbf{x}})}{\lambda} \geq f(\bar{\mathbf{x}})-f(\bar{\mathbf{x}}-\mathbf{d}).$$

Hence, the monotone decreasing sequence of values $[f(\bar{\mathbf{x}}+\lambda\mathbf{d})-f(\bar{\mathbf{x}})]/\lambda$, as $\lambda \to 0^+$, is bounded from below by the constant $f(\bar{\mathbf{x}})-f(\bar{\mathbf{x}}-\mathbf{d})$. Hence, the limit in the theorem exists and is given by

$$\lim_{\lambda\to 0^+} \frac{f(\bar{\mathbf{x}}+\lambda\mathbf{d})-f(\bar{\mathbf{x}})}{\lambda} = \inf_{\lambda>0} \frac{f(\bar{\mathbf{x}}+\lambda\mathbf{d})-f(\bar{\mathbf{x}})}{\lambda}.$$

3.2 Subgradients of Convex Functions

In this section, we introduce the important concept of subgradients of convex and concave functions via supporting hyperplanes to the epigraphs of convex functions and to the hypographs of concave functions.

Epigraph and Hypograph of a Function

A function f on S can be fully described by the set $\{[\mathbf{x}, f(\mathbf{x})] : \mathbf{x} \in S\} \subset R^{n+1}$, which is referred to as the *graph* of the function. One can construct two sets that are related to the graph of f: the epigraph, which consists of points above the graph of f, and the hypograph, which consists of points below the graph of f. These notions are clarified in Definition 3.2.1.

3.2.1 Definition

Let S be a nonempty set in R^n, and let $f: S \to R$. The *epigraph* of f, denoted by epi f, is a subset of R^{n+1} defined by

$$\{(\mathbf{x}, y) : \mathbf{x} \in S,\ y \in R,\ y \geq f(\mathbf{x})\}.$$

The *hypograph* of f, denoted by hyp f, is a subset of R^{n+1} defined by

$$\{(\mathbf{x}, y) : \mathbf{x} \in S,\ y \in R,\ y \leq f(\mathbf{x})\}.$$

Figure 3.3 illustrates the epigraphs and hypographs of several functions. In Figure 3.3*a*, neither the epigraph nor the hypograph of f is a convex set. But in Figure 3.3*b* and *c*, respectively, the epigraph and hypograph of f are convex sets. It turns out that a function is convex if and only if its epigraph is a convex set and, equivalently, that a function is concave if and only if its hypograph is a convex set.

3.2.2 Theorem

Let S be a nonempty convex set in R^n, and let $f: S \to R$. Then f is convex if and only if epi f is a convex set.

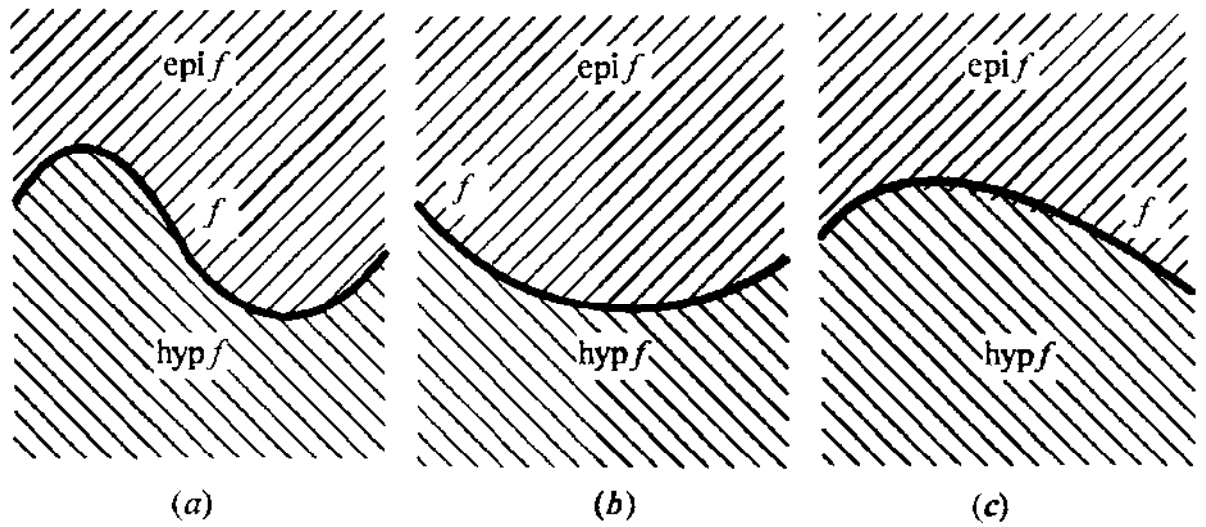


Figure 3.3 Epigraphs and hypographs.

Proof

Assume that f is convex, and let $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_1, y_2) \in \text{epi } f$; that is, $\mathbf{x}_1$, $\mathbf{x}_2 \in S$, $y_1 \geq f(\mathbf{x}_1)$, and $y_2 \geq f(\mathbf{x}_2)$. Let $\lambda \in (0, 1)$. Then

$$\lambda y_1 + (1-\lambda)y_2 \geq \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2) \geq f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2),$$

where the last inequality follows by the convexity of f. Note that $\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in S$. Thus, $[\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \lambda y_1 + (1-\lambda)y_2] \in \text{epi } f$, and hence epi f is convex. Conversely, assume that epi f is convex, and let $\mathbf{x}_1$, $\mathbf{x}_2 \in S$. Then $[\mathbf{x}_1, f(\mathbf{x}_1)]$ and $[\mathbf{x}_2, f(\mathbf{x}_2)]$ belong to epi f, and by the convexity of epi f, we must have

$$[\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2)] \in \text{epi } f \qquad \text{for } \lambda \in (0, 1).$$

In other words, $\lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2) \geq f[\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2]$ for each $\lambda \in (0, 1)$; that is, f is convex. This completes the proof.

Theorem 3.2.2 can be used to verify the convexity or concavity of a given function f. Making use of this result, it is clear that the functions illustrated in Figure 3.3 are (*a*) neither convex nor concave, (*b*) convex, and (*c*) concave.

Since the epigraph of a convex function and the hypograph of a concave function are convex sets, they have supporting hyperplanes at points of their boundary. These supporting hyperplanes lead to the notion of subgradients, which is defined below.

3.2.3 Definition

Let S be a nonempty convex set in R^n, and let $f: S \to R$ be convex. Then $\boldsymbol{\xi}$ is called a *subgradient of* f at $\bar{\mathbf{x}} \in S$ if

$$f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^t(\mathbf{x} - \bar{\mathbf{x}}) \qquad \text{for all } \mathbf{x} \in S.$$

Similarly, let $f: S \to R$ be concave. Then $\boldsymbol{\xi}$ is called a *subgradient* of f at $\bar{\mathbf{x}} \in S$ if

$$f(\mathbf{x}) \leq f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^t(\mathbf{x} - \bar{\mathbf{x}}) \qquad \text{for all } \mathbf{x} \in S.$$

From Definition 3.2.3 it follows immediately that the collection of subgradients of f at $\bar{\mathbf{x}}$ (known as the *subdifferential* of f at $\bar{\mathbf{x}}$) is a convex set. Figure 3.4 shows examples of subgradients of convex and concave functions. From the figure we see that the function $f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^t(\mathbf{x} - \bar{\mathbf{x}})$ corresponds to a supporting hyperplane of the epigraph or the hypograph of the function f. The subgradient vector $\boldsymbol{\xi}$ corresponds to the slope of the supporting hyperplane.

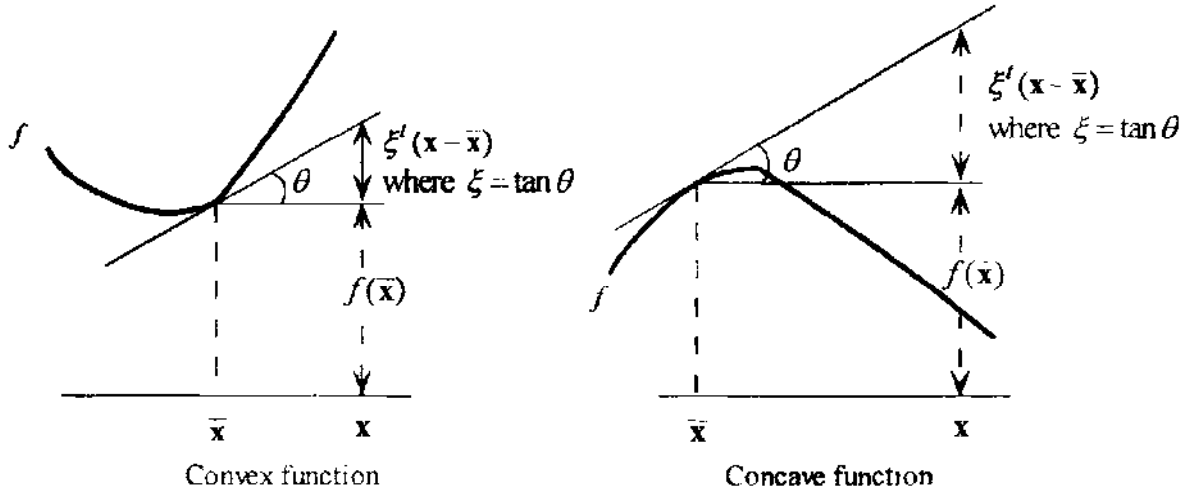


Figure 3.4 Geometric interpretation of subgradients.

3.2.4 Example

Let $f(x) = \min\{f_1(x), f_2(x)\}$, where f_1 and f_2 are as defined below:

$$f_1(x) = 4 - |x|, \qquad x \in R$$

$$f_2(x) = 4 - (x-2)^2, \qquad x \in R.$$

Since $f_2(x) \geq f_1(x)$ for $1 \leq x \leq 4$, f can be represented as follows:

$$f(x) = \begin{cases} 4 - x, & 1 \leq x \leq 4 \\ 4 - (x-2)^2, & \text{otherwise.} \end{cases}$$

In Figure 3.5 the concave function f is shown in dark lines. Note that $\xi = -1$ is the slope and hence the subgradient of f at any point x in the open interval (1, 4). If $x < 1$ or $x > 4$, $\xi = -2(x-2)$ is the unique subgradient of f. At the points $x = 1$ and $x = 4$, the subgradients are not unique because many supporting hyperplanes exist. At $x = 1$, the family of subgradients is characterized by $\lambda \nabla f_1(1) + (1-\lambda)\nabla f_2(1) = \lambda(-1) + (1-\lambda)(2) = 2 - 3\lambda$ for $\lambda \in [0, 1]$. In other words, any ξ in the interval $[-1, 2]$ is a subgradient of f at $x = 1$, and this corresponds to the slopes of the family of supporting hyperplanes of f at $x = 1$. At $x = 4$, the family of subgradients is characterized by $\lambda \nabla f_1(4) + (1-\lambda)\nabla f_2(4) = \lambda(-1) + (1-\lambda)(-4) = -4 + 3\lambda$ for $\lambda \in [0, 1]$. In other words, any ξ in the interval $[-4, -1]$ is a subgradient of f at $x = 4$. Exercise 3.27 addresses the general characterization of subgradients of functions of the form $f(x) = \min\{f_1(x), f_2(x)\}$.

The following theorem shows that every convex or concave function has at least one subgradient at points in the interior of its domain. The proof relies on the fact that a convex set has a supporting hyperplane at points of the boundary.

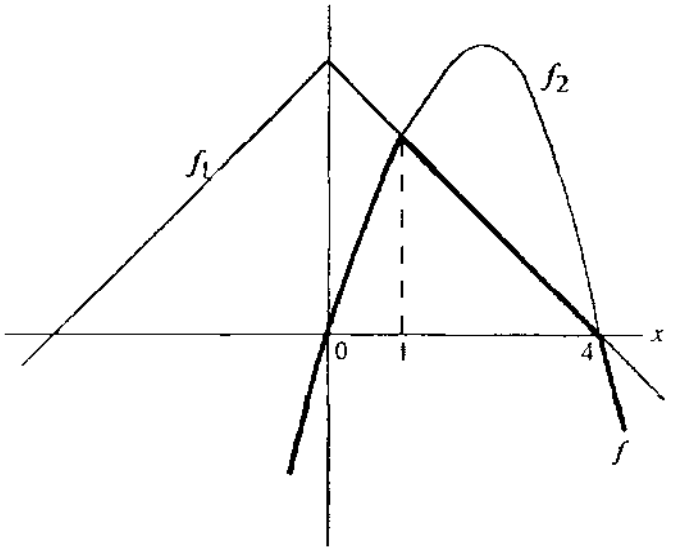


Figure 3.5 Setup for Example 3.2.4.

3.2.5 Theorem

Let S be a nonempty convex set in R^n, and let $f: S \to R$ be convex. Then for $\bar{\mathbf{x}} \in \text{int } S$, there exists a vector ξ such that the hyperplane

$$H = \{(\mathbf{x}, y) : y = f(\bar{\mathbf{x}}) + \xi^t(\mathbf{x} - \bar{\mathbf{x}})\}$$

supports epi f at $[\bar{\mathbf{x}}, f(\bar{\mathbf{x}})]$. In particular,

$$f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \xi^t(\mathbf{x} - \bar{\mathbf{x}}) \qquad \text{for each } \mathbf{x} \in S;$$

that is, ξ is a subgradient of f at $\bar{\mathbf{x}}$.

Proof

By Theorem 3.2.2, epi f is convex. Noting that $[\bar{\mathbf{x}}, f(\bar{\mathbf{x}})]$ belongs to the boundary of epi f, by Theorem 2.4.7 there exists a nonzero vector $(\xi_0, \mu) \in R^n \times R$ such that

$$\xi_0^t(\mathbf{x} - \bar{\mathbf{x}}) + \mu[y - f(\bar{\mathbf{x}})] \leq 0 \qquad \text{for all } (\mathbf{x}, y) \in \text{epi } f. \tag{3.7}$$

Note that μ is not positive, because otherwise, inequality (3.7) will be contradicted by choosing y sufficiently large. We now show that $\mu < 0$. By contradiction, suppose that $\mu = 0$. Then $\xi_0^t(\mathbf{x} - \bar{\mathbf{x}}) \leq 0$ for all $\mathbf{x} \in S$. Since $\bar{\mathbf{x}} \in \text{int } S$, there exists a $\lambda > 0$ such that $\bar{\mathbf{x}} + \lambda\xi_0 \in S$ and hence $\lambda\xi_0^t\xi_0 \leq 0$. This implies that $\xi_0 = \mathbf{0}$ and $(\xi_0, \mu) = (\mathbf{0}, 0)$, contradicting the fact that (ξ_0, μ) is a nonzero vector. Therefore, $\mu < 0$. Denoting $\xi_0 / |\mu|$ by ξ and dividing the inequality in (3.7) by $|\mu|$, we get

$$\xi^t(\mathbf{x} - \bar{\mathbf{x}}) - y + f(\bar{\mathbf{x}}) \leq 0 \qquad \text{for all } (\mathbf{x}, y) \in \text{epi } f. \tag{3.8}$$

In particular, the hyperplane $H = \{(\mathbf{x}, y) : y = f(\bar{\mathbf{x}}) + \xi^t(\mathbf{x} - \bar{\mathbf{x}})\}$ supports epi f at $[\bar{\mathbf{x}}, f(\bar{\mathbf{x}})]$. By letting $y = f(\bar{\mathbf{x}})$ in (3.8), we get $f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \xi^t(\mathbf{x} - \bar{\mathbf{x}})$ for all $\mathbf{x} \in S$, and the proof is complete.

Corollary

Let S be a nonempty convex set in R^n, and let $f: S \to R$ be strictly convex. Then for $\bar{\mathbf{x}} \in \text{int } S$ there exists a vector ξ such that

$$f(\mathbf{x}) > f(\bar{\mathbf{x}}) + \xi^t(\mathbf{x} - \bar{\mathbf{x}}) \qquad \text{for all } \mathbf{x} \in S,\ \mathbf{x} \neq \bar{\mathbf{x}}.$$

Proof

By Theorem 3.2.5 there exists a vector ξ such that

$$f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \xi^t(\mathbf{x} - \bar{\mathbf{x}}) \qquad \text{for all } \mathbf{x} \in S. \tag{3.9}$$

By contradiction, suppose that there is an $\hat{\mathbf{x}} \neq \bar{\mathbf{x}}$ such that $f(\hat{\mathbf{x}}) = f(\bar{\mathbf{x}}) + \xi^t(\hat{\mathbf{x}} - \bar{\mathbf{x}})$. Then, by the strict convexity of f for $\lambda \in (0, 1)$, we get

$$f[\lambda\bar{\mathbf{x}} + (1-\lambda)\hat{\mathbf{x}}] < \lambda f(\bar{\mathbf{x}}) + (1-\lambda) f(\hat{\mathbf{x}}) = f(\bar{\mathbf{x}}) + (1-\lambda)\xi^t(\hat{\mathbf{x}} - \bar{\mathbf{x}}). \tag{3.10}$$

But letting $\mathbf{x} = \lambda\bar{\mathbf{x}} + (1-\lambda)\hat{\mathbf{x}}$ in (3.9), we must have

$$f[\lambda\bar{\mathbf{x}} + (1-\lambda)\hat{\mathbf{x}}] \geq f(\bar{\mathbf{x}}) + (1-\lambda)\xi^t(\hat{\mathbf{x}} - \bar{\mathbf{x}}),$$

contradicting (3.10). This proves the corollary.

The converse of Theorem 3.2.5 is not true in general. In other words, if corresponding to each point $\bar{\mathbf{x}} \in \text{int } S$ there is a subgradient of f, then f is not necessarily a convex function. To illustrate, consider the following example, where f is defined on $S = \{(x_1, x_2) : 0 \leq x_1, x_2 \leq 1\}$:

$$f(x_1, x_2) = \begin{cases} 0, & 0 \leq x_1 \leq 1, \quad 0 < x_2 \leq 1 \\ \dfrac{1}{4} - \left(x_1 - \dfrac{1}{2}\right)^2, & 0 \leq x_1 \leq 1, \quad x_2 = 0. \end{cases}$$

For each point in the interior of the domain, the zero vector is a subgradient of f. However, f is not convex on S since epi f is clearly not a convex set. However, as the following theorem shows, f is indeed convex on int S.

3.2.6 Theorem

Let S be a nonempty convex set in R^n, and let $f: S \to R$. Suppose that for each point $\bar{\mathbf{x}} \in \text{int } S$ there exists a subgradient vector ξ such that

$$f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \xi^t(\mathbf{x} - \bar{\mathbf{x}}) \qquad \text{for each } \mathbf{x} \in S.$$

Then, f is convex on int S.

Proof

Let $\mathbf{x}_1, \mathbf{x}_2 \in \text{int } S$, and let $\lambda \in (0, 1)$. By Corollary 1 to Theorem 2.2.2, int S is convex, and we must have $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in \text{int } S$. By assumption, there exists a subgradient ξ of f at $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$. In particular, the following two inequalities hold true:

$$f(\mathbf{x}_1) \geq f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] + (1-\lambda)\xi^t(\mathbf{x}_1 - \mathbf{x}_2)$$

$$f(\mathbf{x}_2) \geq f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] + \lambda\xi^t(\mathbf{x}_2 - \mathbf{x}_1).$$

Multiplying the above two inequalities by λ and $(1 - \lambda)$, respectively, and adding, we obtain

$$\lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2) \geq f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2],$$

and the result follows.

3.3 Differentiable Convex Functions

We now focus on differentiable convex and concave functions. First, consider the following definition of differentiability.

3.3.1 Definition

Let S be a nonempty set in R^n, and let f: $S \to R$. Then f is said to be *differentiable* at $\bar{\mathbf{x}} \in \text{int } S$ if there exist a vector $\nabla f(\bar{\mathbf{x}})$, called the *gradient vector*, and a function α: $R^n \to R$ such that

$$f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) + \|\mathbf{x} - \bar{\mathbf{x}}\| \alpha(\bar{\mathbf{x}}; \mathbf{x} - \bar{\mathbf{x}}) \qquad \text{for each } \mathbf{x} \in S,$$

where $\lim_{\mathbf{x} \to \bar{\mathbf{x}}} \alpha(\bar{\mathbf{x}}; \mathbf{x} - \bar{\mathbf{x}}) = 0$. The function f is said to be differentiable on the open set $S' \subseteq S$ if it is differentiable at each point in S'. The representation of f above is called a *first-order (Taylor series) expansion* of f at (or about) the point $\bar{\mathbf{x}}$; and without the implicitly defined *remainder term* involving the function α, the resulting representation is called a *first-order (Taylor series) approximation* of f at (or about) the point $\bar{\mathbf{x}}$.

Note that if f is differentiable at $\bar{\mathbf{x}}$, there could only be one gradient vector, and this vector is given by

$$\nabla f(\bar{\mathbf{x}}) = \left(\frac{\partial f(\bar{\mathbf{x}})}{\partial x_1}, \ldots, \frac{\partial f(\bar{\mathbf{x}})}{\partial x_n} \right)^t \equiv (f_1(\bar{\mathbf{x}}), \ldots, f_n(\bar{\mathbf{x}}))^t,$$

where $f_i(\bar{\mathbf{x}}) \equiv \partial f(\bar{\mathbf{x}})/\partial x_i$ is the partial derivative of f with respect to x_i at $\bar{\mathbf{x}}$ (see Exercise 3.36, and review Appendix A.4).

The following lemma shows that a differentiable convex function has only one subgradient, the gradient vector. Hence, the results of the preceding section can easily be specialized to the differentiable case, in which the gradient vector replaces subgradients.

3.3.2 Lemma

Let S be a nonempty convex set in R^n, and let $f: S \to R$ be convex. Suppose that f is differentiable at $\bar{\mathbf{x}} \in \text{int } S$. Then the collection of subgradients of f at $\bar{\mathbf{x}}$ is the singleton set $\{\nabla f(\bar{\mathbf{x}})\}$.

Proof

By Theorem 3.2.5, the set of subgradients of f at $\bar{\mathbf{x}}$ is not empty. Now, let $\boldsymbol{\xi}$ be a subgradient of f at $\bar{\mathbf{x}}$. As a result of Theorem 3.2.5 and the differentiability of f at $\bar{\mathbf{x}}$, for any vector $\mathbf{d}$ and for λ sufficiently small, we get

$$f(\bar{\mathbf{x}} + \lambda\mathbf{d}) \geq f(\bar{\mathbf{x}}) + \lambda\boldsymbol{\xi}^t\mathbf{d}$$

$$f(\bar{\mathbf{x}} + \lambda\mathbf{d}) = f(\bar{\mathbf{x}}) + \lambda\nabla f(\bar{\mathbf{x}})^t\mathbf{d} + \lambda\|\mathbf{d}\|\alpha(\bar{\mathbf{x}}; \lambda\mathbf{d}).$$

Subtracting the equation from the inequality, we obtain

$$0 \geq \lambda[\boldsymbol{\xi} - \nabla f(\bar{\mathbf{x}})]^t\mathbf{d} - \lambda\|\mathbf{d}\|\alpha(\bar{\mathbf{x}}; \lambda\mathbf{d}).$$

If we divide by $\lambda > 0$ and let $\lambda \to 0^+$, it follows that $[\boldsymbol{\xi} - \nabla f(\bar{\mathbf{x}})]^t\mathbf{d} \leq 0$. Choosing $\mathbf{d} = \boldsymbol{\xi} - \nabla f(\bar{\mathbf{x}})$, the last inequality implies that $\boldsymbol{\xi} = \nabla f(\bar{\mathbf{x}})$. This completes the proof.

In the light of Lemma 3.3.2, we give the following important characterization of differentiable convex functions. The proof is immediate from Theorems 3.2.5 and 3.2.6 and Lemma 3.3.2.

3.3.3 Theorem

Let S be a nonempty open convex set in R^n, and let $f: S \to R$ be differentiable on S. Then f is convex if and only if for any $\bar{\mathbf{x}} \in S$, we have

$$f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) \qquad \text{for each } \mathbf{x} \in S.$$

Similarly, f is strictly convex if and only if for each $\bar{\mathbf{x}} \in S$, we have

$$f(\mathbf{x}) > f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) \qquad \text{for each } \mathbf{x} \neq \bar{\mathbf{x}} \text{ in } S.$$

There are two evident implications of the above result that find use in various contexts. The first is that if we have an optimization problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in X$, where f is a convex function, then given any point $\bar{\mathbf{x}}$, the affine

function $f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}})$ bounds f from below. Hence, the minimum of $f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}})$ over X (or over a relaxation of X) yields a lower bound on the optimum value of the given optimization problem, which can prove to be useful in an algorithmic approach. A second point in the same spirit is that this affine bounding function can be used to derive polyhedral outer approximations. For example, consider the set $X = \{\mathbf{x} : g_i(\mathbf{x}) \le 0,\ i = 1,\ldots,\ m\}$, where g_i is a convex function for each $i = 1,\ldots,\ m$. Given any point $\bar{\mathbf{x}}$, construct the polyhedral set $\bar{X} = \{\mathbf{x} : g_i(\bar{\mathbf{x}}) + \nabla g_i(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) \le 0,\ i = 1,\ldots,\ m\}$. Note that the polyhedral set $\bar{X}$ contains X and, hence, affords an *outer linearization* of this set, since for any $\mathbf{x} \in X$, we have $0 \ge g_i(\mathbf{x}) \ge g_i(\bar{\mathbf{x}}) + \nabla g_i(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}})$ for $i = 1,\ldots,\ m$ by Theorem 3.3.3. Such representations play a central role in many successive approximation algorithms for various nonlinear optimization problems.

The following theorem gives another necessary and sufficient characterization of differentiable convex functions. For a function of one variable, the characterization reduces to the slope being nondecreasing.

3.3.4 Theorem

Let S be a nonempty open convex set in R^n and let $f: S \to R$ be differentiable on S. Then f is convex if and only if for each $\mathbf{x}_1,\ \mathbf{x}_2 \in S$ we have

$$[\nabla f(\mathbf{x}_2) - \nabla f(\mathbf{x}_1)]^t(\mathbf{x}_2 - \mathbf{x}_1) \ge 0.$$

Similarly, f is strictly convex if and only if, for each distinct $\mathbf{x}_1,\ \mathbf{x}_2 \in S$, we have

$$[\nabla f(\mathbf{x}_2) - \nabla f(\mathbf{x}_1)]^t(\mathbf{x}_2 - \mathbf{x}_1) > 0.$$

Proof

Assume that f is convex, and let $\mathbf{x}_1,\ \mathbf{x}_2 \in S$. By Theorem 3.3.3 we have

$$f(\mathbf{x}_1) \ge f(\mathbf{x}_2) + \nabla f(\mathbf{x}_2)^t(\mathbf{x}_1 - \mathbf{x}_2)$$

$$f(\mathbf{x}_2) \ge f(\mathbf{x}_1) + \nabla f(\mathbf{x}_1)^t(\mathbf{x}_2 - \mathbf{x}_1).$$

Adding the two inequalities, we get $[\nabla f(\mathbf{x}_2) - \nabla f(\mathbf{x}_1)]^t(\mathbf{x}_2 - \mathbf{x}_1) \ge 0$. To show the converse, let $\mathbf{x}_1,\ \mathbf{x}_2 \in S$. By the mean value theorem,

$$f(\mathbf{x}_2) - f(\mathbf{x}_1) = \nabla f(\mathbf{x})^t(\mathbf{x}_2 - \mathbf{x}_1), \tag{3.11}$$

where $\mathbf{x} = \lambda\mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2$ for some $\lambda \in (0,\ 1)$. By assumption, $[\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}_1)]^t(\mathbf{x} - \mathbf{x}_1) \ge 0$; that is, $(1 - \lambda)[\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}_1)]^t(\mathbf{x}_2 - \mathbf{x}_1) \ge 0$. This implies that

$\nabla f(\mathbf{x})^t (\mathbf{x}_2 - \mathbf{x}_1) \geq \nabla f(\mathbf{x}_1)^t (\mathbf{x}_2 - \mathbf{x}_1)$. By (3.11) we get $f(\mathbf{x}_2) \geq f(\mathbf{x}_1) + \nabla f(\mathbf{x}_1)^t (\mathbf{x}_2 - \mathbf{x}_1)$, so by Theorem 3.3.3, f is convex. The strict case is similar and the proof is complete.

Even though Theorems 3.3.3 and 3.3.4 provide necessary and sufficient characterizations of convex functions, checking these conditions is difficult from a computational standpoint. A simple and more manageable characterization, at least for quadratic functions, can be obtained, provided that the function is twice differentiable.

Twice Differentiable Convex and Concave Functions

A function f that is differentiable at $\bar{\mathbf{x}}$ is said to be twice differentiable at $\bar{\mathbf{x}}$ if the *second-order (Taylor series) expansion* representation of Definition 3.3.5 exists.

3.3.5 Definition

Let S be a nonempty set in R^n, and let f: $S \to R$. Then f is said to be *twice differentiable* at $\bar{\mathbf{x}} \in \text{int } S$ if there exist a vector $\nabla f(\bar{\mathbf{x}})$, and an $n \times n$ symmetric matrix $\mathbf{H}(\bar{\mathbf{x}})$, called the *Hessian matrix*, and a function α: $R^n \to R$ such that

$$f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla f(\mathbf{x})^t (\mathbf{x} - \bar{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^t \mathbf{H}(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) + \|\mathbf{x} - \bar{\mathbf{x}}\|^2 \alpha(\bar{\mathbf{x}}; \mathbf{x} - \bar{\mathbf{x}})$$

for each $\mathbf{x} \in S$, where $\lim_{\mathbf{x} \to \bar{\mathbf{x}}} \alpha(\bar{\mathbf{x}}; \mathbf{x} - \bar{\mathbf{x}}) = 0$. The function f is said to be twice differentiable on the open set $S' \subseteq S$ if it is twice differentiable at each point in S'.

It may be noted that for twice differentiable functions, the Hessian matrix $\mathbf{H}(\bar{\mathbf{x}})$ is comprised of the second-order partial derivatives $f_{ij}(\bar{\mathbf{x}}) \equiv \partial^2 f(\bar{\mathbf{x}}) / \partial x_i \, \partial x_j$ for $i = 1, \ldots, n, j = 1, \ldots, n$, and is given as follows:

$$\mathbf{H}(\bar{\mathbf{x}}) = \begin{bmatrix} f_{11}(\bar{\mathbf{x}}) & f_{12}(\bar{\mathbf{x}}) & \cdots & f_{1n}(\bar{\mathbf{x}}) \\ f_{21}(\bar{\mathbf{x}}) & f_{22}(\bar{\mathbf{x}}) & \cdots & f_{2n}(\bar{\mathbf{x}}) \\ \cdots & \cdots & & \cdots \\ \cdots & \cdots & & \cdots \\ f_{n1}(\bar{\mathbf{x}}) & f_{n2}(\bar{\mathbf{x}}) & \cdots & f_{nn}(\bar{\mathbf{x}}) \end{bmatrix}.$$

In expanded form, the foregoing representation can be written as

$$f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \sum_{j=1}^{n} f_j(\bar{\mathbf{x}})(x_j - \bar{x}_j) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \bar{x}_i)(x_j - \bar{x}_j) f_{ij}(\bar{\mathbf{x}})$$

$$+ \|\mathbf{x} - \bar{\mathbf{x}}\|^2 \alpha(\bar{\mathbf{x}}; \mathbf{x} - \bar{\mathbf{x}}).$$

Again, without the remainder term associated with the function α, this representation is known as a *second-order (Taylor series) approximation* at (or about) the point $\bar{\mathbf{x}}$.

3.3.6 Examples

Example 1. Let $f(x_1, x_2) = 2x_1 + 6x_2 - 2x_1^2 - 3x_2^2 + 4x_1x_2$. Then we have

$$\nabla f(\bar{\mathbf{x}}) = \begin{bmatrix} 2 - 4\bar{x}_1 + 4\bar{x}_2 \\ 6 - 6\bar{x}_2 + 4\bar{x}_1 \end{bmatrix} \quad \text{and} \quad \mathbf{H}(\bar{\mathbf{x}}) = \begin{bmatrix} -4 & 4 \\ 4 & -6 \end{bmatrix}.$$

For example, taking $\bar{\mathbf{x}} = (0,0)^t$, the second-order expansion of this function is given by

$$f(x_1, x_2) = (2,6)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{1}{2}(x_1, x_2)\begin{bmatrix} -4 & 4 \\ 4 & -6 \end{bmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Note that there is no remainder term here since the given function is quadratic, so the above representation is exact.

Example 2. Let $f(x_1, x_2) = e^{2x_1 + 3x_2}$. Then we get

$$\nabla f(\bar{\mathbf{x}}) = \begin{bmatrix} 2e^{2\bar{x}_1 + 3\bar{x}_2} \\ 3e^{2\bar{x}_1 + 3\bar{x}_2} \end{bmatrix} \quad \text{and} \quad \mathbf{H}(\bar{\mathbf{x}}) = \begin{bmatrix} 4e^{2\bar{x}_1 + 3\bar{x}_2} & 6e^{2\bar{x}_1 + 3\bar{x}_2} \\ 6e^{2\bar{x}_1 + 3\bar{x}_2} & 9e^{2\bar{x}_1 + 3\bar{x}_2} \end{bmatrix}.$$

Hence, the second-order expansion of this function about the point $\bar{\mathbf{x}} = (2,1)^t$ is given by

$$f(\bar{\mathbf{x}}) = e^7 + (2e^7, 3e^7)\begin{pmatrix} x_1 - 2 \\ x_2 - 1 \end{pmatrix} + \frac{1}{2}(x_1 - 2, x_2 - 1)\begin{bmatrix} 4e^7 & 6e^7 \\ 6e^7 & 9e^7 \end{bmatrix}\begin{pmatrix} x_1 - 2 \\ x_2 - 1 \end{pmatrix}$$

$$+ \|\mathbf{x} - \bar{\mathbf{x}}\|^2 \alpha(\bar{\mathbf{x}}; \mathbf{x} - \bar{\mathbf{x}}).$$

Theorem 3.3.7 shows that f is convex on S if and only if its Hessian matrix is *positive semidefinite* (PSD) everywhere in S; that is, for any $\bar{\mathbf{x}}$ in S, we have $\mathbf{x}^t\mathbf{H}(\bar{\mathbf{x}})\mathbf{x} \geq 0$ for all $\mathbf{x} \in R^n$. Symmetrically, a function f is concave on S if and only if its Hessian matrix is *negative semidefinite* (NSD) everywhere in S,

that is, for any $\bar{\mathbf{x}} \in S$, we have $\mathbf{x}^t\mathbf{H}(\bar{\mathbf{x}})\mathbf{x} \le 0$ for all $\mathbf{x} \in R^n$. A matrix that is neither positive nor negative semidefinite is called *indefinite* (ID).

3.3.7 Theorem

Let S be a nonempty open convex set in R^n, and let f: $S \to R$ be twice differentiable on S. Then f is convex if and only if the Hessian matrix is positive semidefinite at each point in S.

Proof

Suppose that f is convex, and let $\bar{\mathbf{x}} \in S$. We need to show that $\mathbf{x}^t\mathbf{H}(\bar{\mathbf{x}})\mathbf{x} \ge 0$ for each $\mathbf{x} \in R^n$. Since S is open, then for any given $\mathbf{x} \in R^n$, $\bar{\mathbf{x}} + \lambda\mathbf{x} \in S$ for $|\lambda| \ne 0$ and sufficiently small. By Theorem 3.3.3 and by the twice differentiability of f, we get the following two expressions:

$$f(\bar{\mathbf{x}} + \lambda\mathbf{x}) \ge f(\bar{\mathbf{x}}) + \lambda\nabla f(\bar{\mathbf{x}})^t\mathbf{x} \tag{3.12}$$

$$f(\bar{\mathbf{x}} + \lambda\mathbf{x}) = f(\bar{\mathbf{x}}) + \lambda\nabla f(\bar{\mathbf{x}})^t\mathbf{x} + \frac{1}{2}\lambda^2\mathbf{x}^t\mathbf{H}(\bar{\mathbf{x}})\mathbf{x} + \lambda^2\|\mathbf{x}\|^2\alpha(\bar{\mathbf{x}};\lambda\mathbf{x}). \tag{3.13}$$

Subtracting (3.13) from (3.12), we get

$$\frac{1}{2}\lambda^2\mathbf{x}^t\mathbf{H}(\bar{\mathbf{x}})\mathbf{x} + \lambda^2\|\mathbf{x}\|^2\alpha(\bar{\mathbf{x}};\lambda\mathbf{x}) \ge 0.$$

Dividing by $\lambda^2 > 0$ and letting $\lambda \to 0$, it follows that $\mathbf{x}^t\mathbf{H}(\bar{\mathbf{x}})\mathbf{x} \ge 0$. Conversely, suppose that the Hessian matrix is positive semidefinite at each point in S. Consider $\mathbf{x}$ and $\bar{\mathbf{x}}$ in S. Then, by the mean value theorem, we have

$$f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^t\mathbf{H}(\hat{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}), \tag{3.14}$$

where $\hat{\mathbf{x}} = \lambda\bar{\mathbf{x}} + (1 - \lambda)\mathbf{x}$ for some $\lambda \in (0, 1)$. Note that $\hat{\mathbf{x}} \in S$ and hence, by assumption, $\mathbf{H}(\hat{\mathbf{x}})$ is positive semidefinite. Therefore, $(\mathbf{x} - \bar{\mathbf{x}})^t\mathbf{H}(\hat{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) \ge 0$, and from (3.14), we conclude that

$$f(\mathbf{x}) \ge f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}).$$

Since the above inequality is true for each $\mathbf{x}$, $\bar{\mathbf{x}}$ in S, f is convex by Theorem 3.3.3. This completes the proof.

Theorem 3.3.7 is useful in checking the convexity or concavity of a twice differentiable function. In particular, if the function is quadratic, the Hessian matrix is independent of the point under consideration. Hence, checking its convexity reduces to checking the positive semidefiniteness of a constant matrix.

Results analogous to Theorem 3.3.7 can be obtained for the strict convex and concave cases. It turns out that if the Hessian matrix is positive definite at each point in S, the function is strictly convex. In other words, if for any given point $\bar{\mathbf{x}}$ in S, we have $\mathbf{x}'\mathbf{H}(\bar{\mathbf{x}})\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$ in R^n, then f is strictly convex. This follows readily from the proof of Theorem 3.3.7. However, if f is strictly convex, its Hessian matrix is positive semidefinite, but not necessarily positive definite everywhere in S, unless, for example, if f is quadratic. The latter is seen by writing (3.12) as a strict inequality for $\lambda \mathbf{x} \neq \mathbf{0}$ and noting that the remainder term in (3.13) is then absent. To illustrate, consider the strictly convex function defined by $f(x) = x^4$. The Hessian matrix $\mathbf{H}(x) = 12x^2$ is positive definite for all nonzero x but is positive semidefinite, and not positive definite, at $x = 0$. The following theorem records this fact.

3.3.8 Theorem

Let S be a nonempty open convex set in R^n, and let $f: S \to R$ be twice differentiable on S. If the Hessian matrix is positive definite at each point in S, f is strictly convex. Conversely, if f is strictly convex, the Hessian matrix is positive semidefinite at each point in S. However, if f is strictly convex and quadratic, its Hessian is positive definite.

The foregoing result can be strengthened somewhat while providing some additional insights into the second-order characterization of convexity. Consider, for example, the univariate function $f(x) = x^4$ addressed above, and let us show how we can argue that this function is strictly convex despite the fact that $f''(0) = 0$. Since $f''(x) \geq 0$ for all $x \in R$, we have by Theorem 3.3.7 that f is convex. Hence, by Theorem 3.3.3, all that we need to show is that for any point $\bar{x}$, the supporting hyperplane $y = f(\bar{x}) + f'(\bar{x})(x - \bar{x})$ to the epigraph of the function touches this epigraph only at the given point $(x, y) = (\bar{x}, f(\bar{x}))$. On the contrary, if this supporting hyperplane also touches the epigraph at some other point $(\hat{x}, f(\hat{x}))$, we have $f(\hat{x}) = f(\bar{x}) + f'(\bar{x})(\hat{x} - \bar{x})$. But this means that for any $x_\lambda = \lambda\bar{x} + (1 - \lambda)\hat{x}$, $0 \leq \lambda \leq 1$, we have, upon using Theorem 3.3.3 and the convexity of f,

$$\lambda f(\bar{x}) + (1 - \lambda) f(\hat{x}) = f(\bar{x}) + f'(\bar{x})(x_\lambda - \bar{x}) \leq f(x_\lambda) \leq \lambda f(\bar{x}) + (1 - \lambda) f(\hat{x}).$$

Hence, equality holds true throughout, and the supporting hyperplane touches the graph of the function at all convex combinations $(x_\lambda, f(x_\lambda))$ as well. In fact, we obtain $f(x_\lambda) = \lambda f(\bar{x}) + (1 - \lambda) f(\hat{x})$ for all $0 \leq \lambda \leq 1$, so $f''(x_\lambda) = 0$ at the uncountably infinite number of points x_λ for all $0 < \lambda < 1$. This contradicts the fact that $f''(x) = 0$ only at $x = 0$ from the above example, and therefore, the function is strictly convex. As a result, if we lose positive definiteness of a

univariate convex function at only a finite (or countably infinite) number of points, we can still claim that this function is strictly convex.

Staying with univariate functions for the time being, if the function is infinitely differentiable, we can derive a necessary and sufficient condition for the function to be strictly convex. [By an *infinitely differentiable function f*: $R^n \to R$, we mean one for which for any $\bar{\mathbf{x}}$ in R^n, derivatives of all orders exist and so are continuous; are uniformly bounded in values; and for which the infinite Taylor series expansion of $f(\mathbf{x})$ about $f(\bar{\mathbf{x}})$ gives an infinite series representation of the value of f. Of course, this infinite series can possibly have only a finite number of terms, as, for example, when derivatives of order exceeding some value all vanish.]

3.3.9 Theorem

Let S be a nonempty open convex set in R, and let f: $S \to R$ be infinitely differentiable. Then f is strictly convex on S if and only if for each $\bar{x} \in S$, there exists an even n such that $f^{(n)}(\bar{x}) > 0$, while $f^{(j)}(\bar{x}) = 0$ for any $1 < j < n$, where $f^{(j)}$ denotes the jth-order derivative of f.

Proof

Let $\bar{x}$ be any point in S, and consider the infinite Taylor series expansion of f about $\bar{x}$ for a perturbation $h \neq 0$ and small enough:

$$f(\bar{x}+h) = f(\bar{x}) + hf'(\bar{x}) + \frac{h^2}{2!}f''(\bar{x}) + \frac{h^3}{3!}f'''(\bar{x}) + \cdots$$

If f is strictly convex, then by Theorem 3.3.3 we have that $f(\bar{x}+h) > f(\bar{x}) + hf'(\bar{x})$ for $h \neq 0$. Using this above, we get that for all $h \neq 0$ and sufficiently small,

$$\frac{h^2}{2!}f''(\bar{x}) + \frac{h^3}{3!}f'''(\bar{x}) + \frac{h^4}{4!}f^{(4)}(\bar{x}) + \cdots > 0.$$

Hence, not all derivatives of order greater than or equal to 2 at $\bar{x}$ can be zero. Moreover, since by making h sufficiently small, we can make the first nonzero term above dominate the rest of the expansion, and since h can be of either sign, it follows that this first nonzero derivative must be of an even order and positive for the inequality to hold true.

Conversely, suppose that given any $\bar{x} \in S$, there exists an even n such that $f^{(n)}(\bar{x}) > 0$, while $f^{(j)}(\bar{x}) = 0$ for $1 < j < n$. Then, as above, we have $(\bar{x}+h) \in S$ and $f(\bar{x}+h) > f(\bar{x}) + hf'(\bar{x})$ for all $-\delta < h < \delta$, for some $\delta > 0$ and sufficiently small. Now the hypothesis given also asserts that $f''(\bar{x}) \geq 0$ for all $\bar{x} \in S$, so by Theorem 3.3.7 we know that f is convex. Consequently, for any $\bar{h} \neq 0$, with $(\bar{x}+\bar{h}) \in S$, we get $f(\bar{x}+\bar{h}) \geq f(\bar{x}) + \bar{h}f'(\bar{x})$ by Theorem 3.3.3. To

complete the proof, we must show that this inequality is indeed strict. On the contrary, if $f(\bar{x}+\bar{h}) = f(\bar{x}) + \bar{h}f'(\bar{x})$, we get

$$\begin{aligned}\lambda f(\bar{x}+\bar{h}) + (1-\lambda)f(\bar{x}) &= f(\bar{x}) + \lambda\bar{h}f'(\bar{x}) \le f(\bar{x}+\lambda\bar{h}) \\ &= f[\lambda(\bar{x}+\bar{h}) + (1-\lambda)\bar{x}] \le \lambda f(\bar{x}+\bar{h}) + (1-\lambda)f(\bar{x})\end{aligned}$$

for all $0 \le \lambda \le 1$. But this means that equality holds throughout and that $f(\bar{x} + \lambda\bar{h}) = f(\bar{x}) + \lambda\bar{h}f'(\bar{x})$ for all $0 \le \lambda \le 1$. By taking λ close enough to zero, we can contradict the statement that $f(\bar{x}+h) > f(\bar{x}) + hf'(\bar{x})$ for all $-\delta < h < \delta$, and this completes the proof.

To illustrate, when $f(x) = x^4$, we have $f'(x) = 4x^3$ and $f''(x) = 12x^2$. Hence, for $\bar{x} \neq 0$, the first nonzero derivative as in Theorem 3.3.9 is of order 2 and is positive. Furthermore, for $\bar{x} = 0$, we have $f''(\bar{x}) = f'''(\bar{x}) = 0$ and $f^{(4)}(\bar{x}) = 24 > 0$; so by Theorem 3.3.9, we can conclude that f is strictly convex.

Now let us turn to the multivariate case. The following result provides an insightful connection between the univariate and multivariate cases and permits us to derive results for the latter case from those for the former case. For notational simplicity, we have stated this result for f: $R^n \to R$, although one can readily restate it for f: $S \to R$, where S is some nonempty convex subset of R^n.

3.3.10 Theorem

Consider a function f: $R^n \to R$, and for any point $\bar{\mathbf{x}} \in R^n$ and a nonzero direction $\mathbf{d} \in R^n$, define $F_{(\bar{\mathbf{x}},\mathbf{d})}(\lambda) = f(\bar{\mathbf{x}} + \lambda\mathbf{d})$ as a function of $\lambda \in R$. Then f is (strictly) convex if and only if $F_{(\bar{\mathbf{x}};\mathbf{d})}$ is (strictly) convex for all $\bar{\mathbf{x}}$ and $\mathbf{d} \neq \mathbf{0}$ in R^n.

Proof

Given any $\bar{\mathbf{x}}$ and $\mathbf{d} \neq \mathbf{0}$ in R^n, let us write $F_{(\bar{\mathbf{x}};\mathbf{d})}(\lambda)$ simply as $F(\lambda)$ for convenience. If f is convex, then for any λ_1 and λ_2 in R and for any $0 \le \alpha \le 1$, we have

$$\begin{aligned}F(\alpha\lambda_1 + (1-\alpha)\lambda_2) &= f(\alpha[\bar{\mathbf{x}} + \lambda_1\mathbf{d}] + (1-\alpha)[\bar{\mathbf{x}} + \lambda_2\mathbf{d}]) \\ &\le \alpha f(\bar{\mathbf{x}} + \lambda_1\mathbf{d}) + (1-\alpha)f(\bar{\mathbf{x}} + \lambda_2\mathbf{d}) = \alpha F(\lambda_1) + (1-\alpha)F(\lambda_2).\end{aligned}$$

Hence, F is convex. Conversely, suppose that $F_{(\bar{\mathbf{x}};\mathbf{d})}(\lambda)$, $\lambda \in R$, is convex for all $\bar{\mathbf{x}}$ and $\mathbf{d} \neq \mathbf{0}$ in R^n. Then, for any $\mathbf{x}_1$ and $\mathbf{x}_2$ in R^n and $0 \le \lambda \le 1$, we have

$$\begin{aligned}\lambda f(\mathbf{x}_1)+(1-\lambda)f(\mathbf{x}_2) &= \lambda f[\mathbf{x}_1+0(\mathbf{x}_2-\mathbf{x}_1)]+(1-\lambda)f[\mathbf{x}_1+1(\mathbf{x}_2-\mathbf{x}_1)]\\ &= \lambda F_{[\mathbf{x}_1;(\mathbf{x}_2-\mathbf{x}_1)]}(0)+(1-\lambda)F_{[\mathbf{x}_1;(\mathbf{x}_2-\mathbf{x}_1)]}(1)\\ &\geq F_{[\mathbf{x}_1;(\mathbf{x}_2-\mathbf{x}_1)]}(1-\lambda)\\ &= f[\mathbf{x}_1+(1-\lambda)(\mathbf{x}_2-\mathbf{x}_1)] = f[\lambda\mathbf{x}_1+(1-\lambda)\mathbf{x}_2],\end{aligned}$$

so f is convex. The argument for the strictly convex case is similar, and this completes the proof.

This insight of examining f: $R^n \to R$ via its univariate cross sections $F_{(\bar{\mathbf{x}};\mathbf{d})}$ can be very useful both as a conceptual tool for viewing f and as an analytical tool for deriving various results. For example, writing $F(\lambda) = F_{(\bar{\mathbf{x}};\mathbf{d})}(\lambda) = f(\bar{\mathbf{x}}+\lambda\mathbf{d})$, for any given $\bar{\mathbf{x}}$ and $\mathbf{d} \neq \mathbf{0}$ in R^n, we have from the univariate Taylor series expansion (assuming infinite differentiability) that

$$F(\lambda) = F(0)+\lambda F'(0)+\frac{\lambda^2}{2!}F''(0)+\frac{\lambda^3}{3!}F'''(0)+\cdots.$$

By using the chain rule for differentiation, we obtain

$$F'(\lambda) = \nabla f(\bar{\mathbf{x}}+\lambda\mathbf{d})^t\,\mathrm{d} = \sum_i f_i(\bar{\mathbf{x}}+\lambda\mathrm{d})d_i$$

$$F''(\lambda) = \mathbf{d}^t\mathbf{H}(\bar{\mathbf{x}}+\lambda\mathbf{d})\mathbf{d} = \sum_i\sum_j f_{ij}(\bar{\mathbf{x}}+\lambda\mathbf{d})d_i d_j$$

$$F'''(\lambda) = \sum_i\sum_j\sum_k f_{ijk}(\bar{\mathbf{x}}+\lambda\mathbf{d})d_i d_j d_k, \text{ etc.}$$

Substituting above, this gives the corresponding multivariate Taylor series expansion as

$$f(\bar{\mathbf{x}}+\lambda\mathbf{d}) = f(\bar{\mathbf{x}})+\lambda\nabla f(\bar{\mathbf{x}})^t\mathbf{d}+\frac{\lambda^2}{2!}\mathbf{d}^t\mathbf{H}(\bar{\mathbf{x}})\mathbf{d}+\frac{\lambda^3}{3!}\sum_i\sum_j\sum_k f_{ijk}(\bar{\mathbf{x}})d_i d_j d_k+\cdots.$$

As another example, using the second-order derivative result for characterizing the convexity of a univariate function along with Theorem 3.3.10, we can derive that f: $R^n \to R$ is convex if and only if $F''_{(\bar{\mathbf{x}};\mathbf{d})}(\lambda) \geq 0$ for all $\lambda \in R$, $\bar{\mathbf{x}} \in R^n$, and $\mathbf{d} \in R^n$. But since $\bar{\mathbf{x}}$ and $\mathbf{d}$ can be chosen arbitrarily, this is equivalent to requiring that $F''_{(\bar{\mathbf{x}};\mathbf{d})}(0) \geq 0$ for all $\bar{\mathbf{x}}$ and $\mathbf{d}$ in R^n. From above, this translates to the statement that $\mathbf{d}^t\mathbf{H}(\bar{\mathbf{x}})\mathbf{d} \geq 0$ for all $\mathbf{d} \in R^n$, for each $\bar{\mathbf{x}} \in R^n$, or that $\mathbf{H}(\bar{\mathbf{x}})$ is positive semidefinite for all $\bar{\mathbf{x}} \in R^n$, as in Theorem 3.3.7. In a similar manner, or by using the multivariate Taylor series expansion directly as in the proof of

Theorem 3.3.9, we can assert that an infinitely differentiable function f: $R^n \to R$ is strictly convex if and only if for each $\bar{\mathbf{x}}$ and $\mathbf{d} \neq \mathbf{0}$ in R^n, the first nonzero derivative term $[F^{(j)}(0)]$ of order greater than or equal to 2 in the Taylor series expansion above exists, is of even order, and is positive. We leave the details of exploring this result to the reader in Exercise 3.38.

We present below an efficient (polynomial-time) algorithm for checking the definiteness of a (symmetric) Hessian matrix $\mathbf{H}(\bar{\mathbf{x}})$ using elementary Gauss–Jordan operations. Appendix A cites a characterization of definiteness in terms of eigenvalues which finds use in some analytical proofs but is not an algorithmically convenient alternative. Moreover, if one needs to check for the definiteness of a matrix $\mathbf{H}(\mathbf{x})$ that is a function of $\mathbf{x}$, this eigenvalue method is very cumbersome, if not virtually impossible, to use. Although the method presented below can also get messy in such instances, it is overall a more simple and efficient approach.

We begin by considering a 2×2 Hessian matrix $\mathbf{H}$ in Lemma 3.3.11, where the argument $\bar{\mathbf{x}}$ has been suppressed for convenience. This is then generalized in an inductive fashion to an $n \times n$ matrix in Theorem 3.3.12.

3.3.11 Lemma

Consider a symmetric matrix $\mathbf{H} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$. Then $\mathbf{H}$ is positive semidefinite if and only if $a \geq 0$, $c \geq 0$, and $ac - b^2 \geq 0$, and is positive definite if and only if the foregoing inequalities are all strict.

Proof

By definition, $\mathbf{H}$ is positive semidefinite if and only if $\mathbf{d}^t\mathbf{H}\mathbf{d} = ad_1^2 + 2bd_1d_2 + cd_2^2 \geq 0$ for all $(d_1, d_2)^t \in R^2$. Hence, if $\mathbf{H}$ is positive semidefinite, we must clearly have $a \geq 0$ and $c \geq 0$. Moreover, if $a = 0$, we must have $b = 0$, so $ac - b^2 = 0$; or else, by taking $d_2 = 1$ and $d_1 = -Mb$ for $M > 0$ and large enough, we would obtain $\mathbf{d}^t\mathbf{H}\mathbf{d} < 0$, a contradiction. On the other hand, if $a > 0$, then completing the squares, we get

$$\mathbf{d}^t\mathbf{H}\mathbf{d} = a\left(d_1^2 + \frac{2bd_1d_2}{a} + \frac{b^2}{a^2}d_2^2\right) + d_2^2\left(c - \frac{b^2}{a}\right) = a\left(d_1 + \frac{b}{a}d_2\right)^2 + d_2^2\left(\frac{ac - b^2}{a}\right).$$

Hence, we must again have $(ac - b^2) \geq 0$, since otherwise, by taking $d_2 = 1$ and $d_1 = -b/a$, we would get $\mathbf{d}^t\mathbf{H}\mathbf{d} = (ac - b^2)/a < 0$, a contradiction. Hence, the condition of the theorem holds true. Conversely, suppose that $a \geq 0$, $c \geq 0$, and

$ac - b^2 \geq 0$. If $a = 0$, this gives $b = 0$, so $\mathbf{d}^t\mathbf{H}\mathbf{d} = cd_2^2 \geq 0$. On the other hand, if $a > 0$, by completing the squares as above we get

$$\mathbf{d}^t\mathbf{H}\mathbf{d} = a\left(d_1 + \frac{b}{a}d_2^2\right) + d_2^2\left(\frac{ac - b^2}{a}\right) \geq 0.$$

Hence, **H** is positive semidefinite. The proof of positive definiteness is similar, and this completes the proof.

We remark here that since a matrix **H** is negative semidefinite (negative definite) if and only if **–H** is positive semidefinite (positive definite), we get from Lemma 3.3.11 that **H** is negative semidefinite if and only if $a \leq 0$, $c \leq 0$, and $ac - b^2 \geq 0$, and that **H** is negative definite if and only if these inequalities are all strict. Theorem 3.3.12 is stated for checking positive semidefiniteness or positive definiteness of **H**. By replacing **H** by **–H**, we could test symmetrically for negative semidefiniteness or negative definiteness. If the matrix turns out to be neither positive semidefinite nor negative semidefinite, it is indefinite. Also, we assume below that **H** is symmetric, being the Hessian of a twice differentiable function for our purposes. In general, if **H** is not symmetric, then since $\mathbf{d}^t\mathbf{H}\mathbf{d} = \mathbf{d}^t\mathbf{H}^t\mathbf{d} = \mathbf{d}^t[(\mathbf{H} + \mathbf{H}^t)/2]\mathbf{d}$, we can check for the definiteness of **H** by using the symmetric matrix $(\mathbf{H} + \mathbf{H}^t)/2$ below.

3.3.12 Theorem (Checking for PSD/PD)

Let **H** be a symmetric $n \times n$ matrix with elements h_{ij}.

(a) If $h_{ii} \leq 0$ for any $i \in \{1,...,n\}$, **H** is not positive definite; and if $h_{ii} < 0$ for any $i \in \{1,...,n\}$, **H** is not positive semidefinite.

(b) If $h_{ii} = 0$ for any $i \in \{1,...,n\}$, we must have $h_{ij} = h_{ji} = 0$ for all $j = 1,..., n$ as well, or else **H** is not positive semidefinite.

(c) If $n = 1$, **H** is positive semidefinite (positive definite) if and only if $h_{11} \geq 0$ (> 0). Otherwise, if $n \geq 2$, let

$$\mathbf{H} = \begin{bmatrix} h_{11} & \mathbf{q}^t \\ \mathbf{q} & \mathbf{G} \end{bmatrix}$$

in partitioned form, where $\mathbf{q} = \mathbf{0}$ if $h_{11} = 0$, and otherwise, $h_{11} > 0$. Perform elementary Gauss–Jordan operations using the first row of **H** to reduce it to the following matrix in either case:

$$\mathbf{H} = \begin{bmatrix} h_{11} & \mathbf{q}^t \\ \mathbf{0} & \mathbf{G}_{\text{new}} \end{bmatrix}.$$

Then $\mathbf{G}_{new}$ is a symmetric $(n-1)\times(n-1)$ matrix, and $\mathbf{H}$ is positive semidefinite if and only if $\mathbf{G}_{new}$ is positive semidefinite. Moreover, if $h_{11}>0$, $\mathbf{H}$ is positive definite if and only if $\mathbf{G}_{new}$ is positive definite.

Proof

(a) Since $\mathbf{d}^t\mathbf{H}\mathbf{d} = d_i^2 h_{ii}$ whenever $d_j = 0$ for all $j \neq i$, Part (a) of the theorem is obviously true.

(b) Suppose that for some $i \neq j$, we have $h_{ii} = 0$ and $h_{ij} \neq 0$. Then, by taking $d_k = 0$ for all $k \neq i$ or j, we get $\mathbf{d}^t\mathbf{H}\mathbf{d} = 2h_{ij}d_i d_j + d_j^2 h_{jj}$, which can be made negative as in the proof of Lemma 3.3.11 by taking $d_j = 1$ and $d_i = -h_{ij}M$ for $M > 0$ and sufficiently large. This establishes Part (b).

(c) Finally, suppose that $\mathbf{H}$ is given in partitioned form as in Part (c). If $n = 1$, the result is trivial. Otherwise, for $n \geq 2$, let $\mathbf{d}^t = (d_1, \delta^t)$. If $h_{11} = 0$, by assumption we also have $\mathbf{q} = \mathbf{0}$, and then $\mathbf{G}_{new} = \mathbf{G}$. Moreover, in this case, $\mathbf{d}^t\mathbf{H}\mathbf{d} = \delta^t\mathbf{G}_{new}\delta$, so $\mathbf{H}$ is positive semidefinite if and only if $\mathbf{G}_{new}$ is positive semidefinite. On the other hand, if $h_{11} > 0$, we get

$$\mathbf{d}^t\mathbf{H}\mathbf{d} = (d_1, \delta^t)\begin{bmatrix} h_{11} & \mathbf{q}^t \\ \mathbf{q} & \mathbf{G} \end{bmatrix}\begin{pmatrix} d_1 \\ \delta \end{pmatrix} = d_1^2 h_{11} + 2d_1(\mathbf{q}^t\delta) + \delta^t\mathbf{G}\delta.$$

But by the Gauss–Jordan reduction process, we have

$$\mathbf{G}_{new} = \mathbf{G} - \frac{1}{h_{11}}\begin{bmatrix} q_1\mathbf{q}^t \\ q_2\mathbf{q}^t \\ \vdots \\ q_n\mathbf{q}^t \end{bmatrix} = \mathbf{G} - \frac{1}{h_{11}}\mathbf{q}\mathbf{q}^t,$$

which is a symmetric matrix. By substituting this above, we get

$$\mathbf{d}^t\mathbf{H}\mathbf{d} = d_1^2 h_{11} + 2d_1(\mathbf{q}^t\delta) + \delta^t\left(\mathbf{G}_{new} + \frac{1}{h_{11}}\mathbf{q}\mathbf{q}^t\right)\delta = \delta^t\mathbf{G}_{new}\delta + h_{11}\left(d_1 + \frac{\mathbf{q}^t\delta}{h_{11}}\right)^2.$$

Hence, it can readily be verified that $\mathbf{d}^t\mathbf{H}\mathbf{d} \geq 0$ for all $\mathbf{d} \in R^n$ if and only if $\delta^t\mathbf{G}_{new}\delta \geq 0$ for all $\delta \in R^{n-1}$, because $h_{11}(d_1 + \mathbf{q}^t\delta/h_{11})^2 \geq 0$, and the latter term can be made zero by selecting $d_1 = -\mathbf{q}^t\delta/h_{11}$, if necessary. By the same argu-

ment, $\mathbf{d}^t\mathbf{H}\mathbf{d} > 0$ for all $\mathbf{d} \neq \mathbf{0}$ in R^n if and only if $\boldsymbol{\delta}^t\mathbf{G}_{\text{new}}\boldsymbol{\delta} > 0$ for all $\boldsymbol{\delta} \neq \mathbf{0}$ in R^{n-1}, and this completes the proof.

Observe that Theorem 3.3.12 prompts a polynomial-time algorithm for checking the PSD/PD of a symmetric $n \times n$ matrix $\mathbf{H}$. We first scan the diagonal elements to see if either condition (a) or (b) leads to the conclusion that the matrix is not PSD/PD. If this does not terminate the process, we perform the Gauss–Jordan reduction as in Part (c) and arrive at a matrix $\mathbf{G}_{\text{new}}$ of one lesser dimension for which we may now perform the same test as on $\mathbf{H}$. When $\mathbf{G}_{\text{new}}$ is finally a 2×2 matrix, we can use Lemma 3.3.11, or we can continue to reduce it to a 1×1 matrix and hence determine the PSD/PD of $\mathbf{H}$. Since each pass through the inductive step of the algorithm is of *complexity* $O(n^2)$ (read as "*of order* n^2" and meaning that the number of elementary arithmetic operations, comparison, etc., involved are bounded above by Kn^2 for some constant K) and the number of inductive steps is of $O(n)$, the overall process is of polynomial complexity $O(n^3)$. Because the algorithm basically works toward reducing the matrix to an upper triangular matrix, it is sometimes called a *superdiagonalization algorithm.* This algorithm affords a proof for the following useful result, which can alternatively be proved using the eigenvalue characterization of definiteness (see Exercise 3.42).

Corollary

Let $\mathbf{H}$ be an $n \times n$ symmetric matrix. Then $\mathbf{H}$ is positive definite if and only if it is positive semidefinite and nonsingular.

Proof

If $\mathbf{H}$ is positive definite, it is positive semidefinite; and since the superdiagonalization algorithm reduces the matrix $\mathbf{H}$ to an upper triangular matrix with positive diagonal elements via elementary row operations, $\mathbf{H}$ is nonsingular. Conversely, if $\mathbf{H}$ is positive semidefinite and nonsingular, the superdiagonalization algorithm must always encounter nonzero elements along the diagonal because $\mathbf{H}$ is nonsingular, and these must be positive because $\mathbf{H}$ is positive semidefinite. Hence, $\mathbf{H}$ is positive definite.

3.3.13 Examples

Example 1. Consider Example 1 of Section 3.3.6. Here we have

$$\mathbf{H}(\mathbf{x}) = \begin{bmatrix} -4 & 4 \\ 4 & -6 \end{bmatrix}$$

so

$$-\mathbf{H}(\mathbf{x}) = \begin{bmatrix} 4 & -4 \\ -4 & 6 \end{bmatrix}.$$

By Lemma 3.3.11 we conclude that $-\mathbf{H}(\mathbf{x})$ is positive definite, so $\mathbf{H}(\mathbf{x})$ is negative definite and the function f is strictly concave.

Example 2. Consider the function $f(x_1, x_2) = x_1^3 + 2x_2^2$. Here we have

$$\nabla f(\mathbf{x}) = \begin{bmatrix} 3x_1^2 \\ 4x_2 \end{bmatrix} \quad \text{and} \quad \mathbf{H}(\mathbf{x}) = \begin{bmatrix} 6x_1 & 0 \\ 0 & 4 \end{bmatrix}.$$

By Lemma 3.3.11, whenever $x_1 < 0$, $\mathbf{H}(\mathbf{x})$ is indefinite. However, $\mathbf{H}(\mathbf{x})$ is positive definite for $x_1 > 0$, so f is strictly convex over $\{\mathbf{x} : x_1 > 0\}$.

Example 3. Consider the matrix

$$\mathbf{H} = \begin{bmatrix} 2 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 4 \end{bmatrix}.$$

Note that the matrix is not negative semidefinite. To check PSD/PD, apply the superdiagonalization algorithm and reduce H to

$$\begin{bmatrix} 2 & 1 & 2 \\ 0 & 3/2 & 2 \\ 0 & 2 & 2 \end{bmatrix} \quad \text{which gives} \quad \mathbf{G}_{\text{new}} = \begin{bmatrix} 3/2 & 2 \\ 2 & 2 \end{bmatrix}.$$

Now the diagonals of $\mathbf{G}_{\text{new}}$ are positive, but $\det(\mathbf{G}_{\text{new}}) = -1$. Hence, H is not positive semidefinite. Alternatively, we could have verified this by continuing to reduce $\mathbf{G}_{\text{new}}$ to obtain the matrix

$$\begin{bmatrix} 3/2 & 2 \\ 0 & -2/3 \end{bmatrix}.$$

Since the resulting second diagonal element (i.e., the reduced $\mathbf{G}_{\text{new}}$) is negative, H is not positive semidefinite. Since H is not negative semidefinite either, it is indefinite.

3.4 Minima and Maxima of Convex Functions

In this section we consider the problems of minimizing and maximizing a convex function over a convex set and develop necessary and/or sufficient conditions for optimality.

Minimizing a Convex Function

The case of maximizing a concave function is similar to that of minimizing a convex function. We develop the latter in detail and ask the reader to draw the analogous results for the concave case.

3.4.1 Definition

Let f: $R^n \rightarrow R$ and consider the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$. A point $\mathbf{x} \in S$ is called a *feasible solution* to the problem. If $\bar{\mathbf{x}} \in S$ and $f(\mathbf{x}) \geq f(\bar{\mathbf{x}})$ for each $\mathbf{x} \in S$, $\bar{\mathbf{x}}$ is called an *optimal solution*, a *global optimal solution*, or simply a *solution* to the problem. The collection of optimal solutions are called *alternative optimal solutions*. If $\bar{\mathbf{x}} \in S$ and if there exists an ε-neighborhood $N_\varepsilon(\bar{\mathbf{x}})$ around $\bar{\mathbf{x}}$ such that $f(\mathbf{x}) \geq f(\bar{\mathbf{x}})$ for each $\mathbf{x} \in S \cap N_\varepsilon(\bar{\mathbf{x}})$, $\bar{\mathbf{x}}$ is called a *local optimal solution*. Similarly, if $\bar{\mathbf{x}} \in S$ and if $f(\mathbf{x}) > f(\bar{\mathbf{x}})$ for all $\mathbf{x} \in S \cap N_\varepsilon(\bar{\mathbf{x}})$, $\mathbf{x} \neq \bar{\mathbf{x}}$, for some $\varepsilon > 0$, $\bar{\mathbf{x}}$ is called a *strict local optimal solution*. On the other hand, if $\bar{\mathbf{x}} \in S$ is the *only* local minimum in $S \cap N_\varepsilon(\bar{\mathbf{x}})$, for some ε-neighborhood $N_\varepsilon(\bar{\mathbf{x}})$ around $\bar{\mathbf{x}}$, $\bar{\mathbf{x}}$ is called a *strong* or *isolated local optimal solution*. All these types of local optima or minima are sometimes also referred to as *relative minima*. Figure 3.6 illustrates instances of local and global minima for the problem of minimizing $f(\mathbf{x})$ subject to $\mathbf{x} \in S$, where f and S are shown in the figure.

The points in S corresponding to A, B, and C are also both strict and strong local minima, whereas those corresponding to the flat segment of the graph between D and E are local minima that are neither strict nor strong. Note that if $\bar{\mathbf{x}}$ is a strong or isolated local minimum, it is also a strict minimum. To see this, consider the ε-neighborhood $N_\varepsilon(\bar{\mathbf{x}})$ characterizing the strong local minimum nature of $\bar{\mathbf{x}}$. Then we must also have $f(\mathbf{x}) > f(\bar{\mathbf{x}})$ for all $\mathbf{x} \in S \cap$

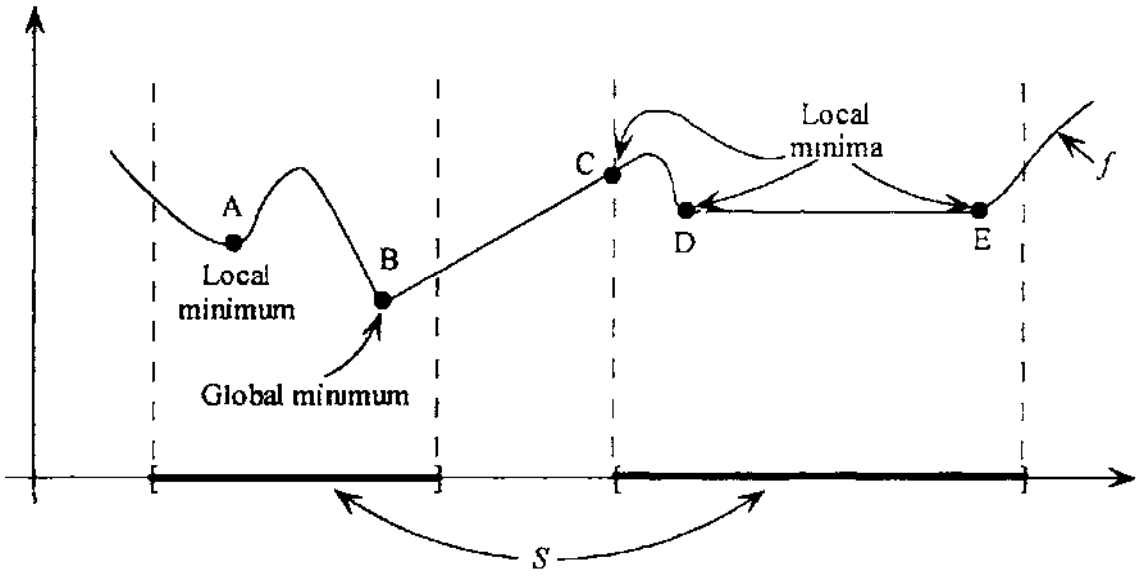


Figure 3.6 Local and global minima.

$N_\varepsilon(\bar{\mathbf{x}})$, because otherwise, suppose that there exists an $\hat{\mathbf{x}} \in S \cap N_\varepsilon(\bar{\mathbf{x}})$ such that $f(\hat{\mathbf{x}}) = f(\bar{\mathbf{x}})$. Note that $\hat{\mathbf{x}}$ is an alternative optimal solution within $S \cap N_\varepsilon(\bar{\mathbf{x}})$, so there exists some $0 < \varepsilon' < \varepsilon$ such that $f(\mathbf{x}) \geq f(\hat{\mathbf{x}})$ for all $\mathbf{x} \in S \cap N_{\varepsilon'}(\hat{\mathbf{x}})$. But this contradicts the isolated local minimum status of $\bar{\mathbf{x}}$, and hence $\bar{\mathbf{x}}$ must also be a strict local minimum. On the other hand, a strict local minimum need not be an isolated local minimum. Figure 3.7 illustrates two such instances. In Figure 3.7*a*, $S = R$ and $f(x) = 1$ for $x = 1$ and is equal to 2 otherwise. Note that the point of discontinuity $\bar{x} = 1$ of f is a strict local minimum but is not isolated, since any ε-neighborhood about $\bar{x}$ contains points other than $\bar{x} = 1$, all of which are also local minima. Figure 3.7*b* illustrates another case in which $f(x) = x^2$, a strictly convex function; but $S = \{1/2^k, k = 0,1,2,...\} \cup \{0\}$ is a nonconvex set. Here, the point $\bar{\mathbf{x}} = 1/2^k$ for any integer $k \geq 0$ is an isolated and therefore a strict local minimum because it can be captured as the unique feasible solution in $S \cap N_\varepsilon(\bar{\mathbf{x}})$ for some sufficiently small $\varepsilon > 0$. However, although $\bar{x} = 0$ is clearly a strict local minimum (it is, in fact, the unique global minimum), it is not isolated because any ε-neighborhood about $\bar{x} = 0$ contains other local minima of the foregoing type.

Nonetheless, for optimization problems, $\min\{f(\mathbf{x}): \mathbf{x} \in S\}$, where f is a convex function and S is a convex set, which are known as *convex programming problems* and that are of interest to us in this section, a strict local minimum is also a strong local minimum, as shown in Theorem 3.4.2 (see Exercise 3.47 for a weaker sufficient condition). The principal result here is that each local minimum of a convex program is also a global minimum. This fact is quite useful in the optimization process, since it enables us to stop with a global optimal solution if the search in the vicinity of a feasible point does not lead to an improving feasible solution.

3.4.2 Theorem

Let S be a nonempty convex set in R^n, and let f: $S \to R$ be convex on S. Consider the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$. Suppose that $\bar{\mathbf{x}} \in S$ is a local optimal solution to the problem.

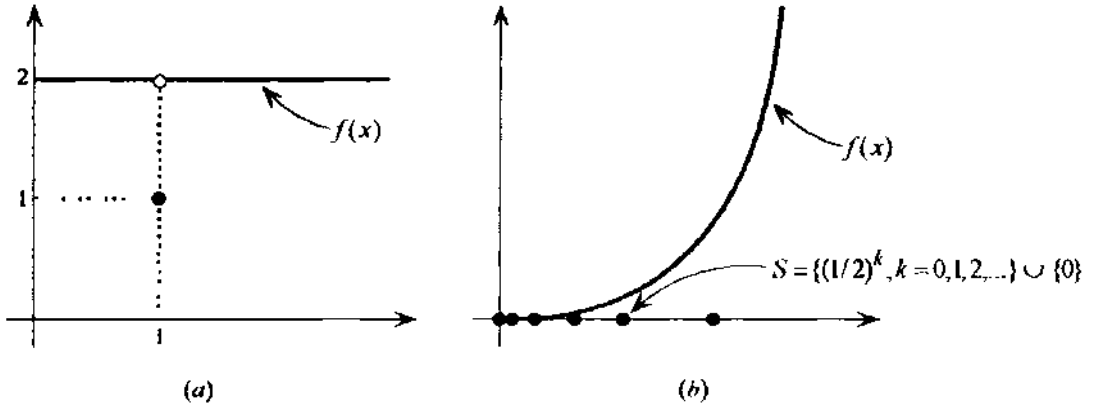


Figure 3.7 Strict local minima are not necessarily strong local minima.

1. Then $\bar{\mathbf{x}}$ is a global optimal solution.
2. If either $\bar{\mathbf{x}}$ is a strict local minimum or f is strictly convex, $\bar{\mathbf{x}}$ is the unique global optimal solution and is also a strong local minimum.

Proof

Since $\bar{\mathbf{x}}$ is a local optimal solution, there exists an ε-neighborhood $N_\varepsilon(\bar{\mathbf{x}})$ around $\bar{\mathbf{x}}$ such that

$$f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) \qquad \text{for each } \mathbf{x} \in S \cap N_\varepsilon(\bar{\mathbf{x}}). \tag{3.15}$$

By contradiction, suppose that $\bar{\mathbf{x}}$ is not a global optimal solution so that $f(\hat{\mathbf{x}}) < f(\bar{\mathbf{x}})$ for some $\hat{\mathbf{x}} \in S$. By the convexity of f, the following is true for each $0 \leq \lambda \leq 1$:

$$f(\lambda\hat{\mathbf{x}} + (1-\lambda)\bar{\mathbf{x}}) \leq \lambda f(\hat{\mathbf{x}}) + (1-\lambda) f(\bar{\mathbf{x}}) < \lambda f(\bar{\mathbf{x}}) + (1-\lambda) f(\bar{\mathbf{x}}) = f(\bar{\mathbf{x}}).$$

But for $\lambda > 0$ and sufficiently small, $\lambda\hat{\mathbf{x}} + (1-\lambda)\bar{\mathbf{x}} \in S \cap N_\varepsilon(\bar{\mathbf{x}})$. Hence, the above inequality contradicts (3.15), and Part 1 is proved.

Next, suppose that $\bar{\mathbf{x}}$ is a strict local minimum. By Part 1 it is a global minimum. Now, on the contrary , if there exists an $\hat{\mathbf{x}} \in S$ such that $f(\hat{\mathbf{x}}) = f(\bar{\mathbf{x}})$, then defining $\mathbf{x}_\lambda = \lambda\hat{\mathbf{x}} + (1-\lambda)\bar{\mathbf{x}}$ for $0 \leq \lambda \leq 1$, we have, by the convexity of f and S that $f(\mathbf{x}_\lambda) \leq \lambda f(\hat{\mathbf{x}}) + (1-\lambda) f(\bar{\mathbf{x}}) = f(\bar{\mathbf{x}})$, and $\mathbf{x}_\lambda \in S$ for all $0 \leq \lambda \leq 1$. By taking $\lambda \to 0^+$, since we can make $\mathbf{x}_\lambda \in N_\varepsilon(\bar{\mathbf{x}}) \cap S$ for any $\varepsilon > 0$, this contradicts the strict local optimality of $\bar{\mathbf{x}}$. Hence, $\bar{\mathbf{x}}$ is the unique global minimum. Therefore, it must also be an isolated local minimum, since any other local minimum in $N_\varepsilon(\bar{\mathbf{x}}) \cap S$ for any $\varepsilon > 0$ would also be a global minimum, which is a contradiction.

Finally, suppose that $\bar{\mathbf{x}}$ is a local optimal solution and that f is strictly convex. Since strict convexity implies convexity, then by Part 1, $\bar{\mathbf{x}}$ is a global optimal solution. By contradiction, suppose that $\bar{\mathbf{x}}$ is not the unique global optimal solution, so that there exists an $\mathbf{x} \in S$, $\mathbf{x} \neq \bar{\mathbf{x}}$ such that $f(\mathbf{x}) = f(\bar{\mathbf{x}})$. By strict convexity,

$$f\left(\frac{1}{2}\mathbf{x} + \frac{1}{2}\bar{\mathbf{x}}\right) < \frac{1}{2} f(\mathbf{x}) + \frac{1}{2} f(\bar{\mathbf{x}}) = f(\bar{\mathbf{x}}).$$

By the convexity of S, $(1/2)\mathbf{x} + (1/2)\bar{\mathbf{x}} \in S$, and the above inequality violates the global optimality of $\bar{\mathbf{x}}$. Hence, $\bar{\mathbf{x}}$ is the unique global minimum and, as above, is also a strong local minimum. This completes the proof.

We now develop a necessary and sufficient condition for the existence of a global solution. If such an optimal solution does not exist, then $\inf\{f(\mathbf{x}) : \mathbf{x} \in S\}$ is finite but is not achieved at any point in S, or it is equal to $-\infty$.

3.4.3 Theorem

Let f: $R^n \to R$ be a convex function, and let S be a nonempty convex set in R^n. Consider the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$. The point $\bar{\mathbf{x}} \in S$ is an optimal solution to this problem if and only if f has a subgradient ξ at $\bar{\mathbf{x}}$ such that $\xi^t(\mathbf{x} - \bar{\mathbf{x}}) \geq 0$ for all $\mathbf{x} \in S$.

Proof

Suppose that $\xi^t(\mathbf{x} - \bar{\mathbf{x}}) \geq 0$ for all $\mathbf{x} \in S$, where ξ is a subgradient of f at $\bar{\mathbf{x}}$. By the convexity of f, we have

$$f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \xi^t(\mathbf{x} - \bar{\mathbf{x}}) \geq f(\bar{\mathbf{x}}) \qquad \text{for all } \mathbf{x} \in S,$$

and hence $\bar{\mathbf{x}}$ is an optimal solution to the given problem.

To show the converse, suppose that $\bar{\mathbf{x}}$ is an optimal solution to the problem and construct the following two sets in R^{n+1}:

$$\Lambda_1 = \{(\mathbf{x} - \bar{\mathbf{x}}, y) : \mathbf{x} \in R^n,\ y > f(\mathbf{x}) - f(\bar{\mathbf{x}})\}$$

$$\Lambda_2 = \{(\mathbf{x} - \bar{\mathbf{x}}, y) : \mathbf{x} \in S,\ y \leq 0\}.$$

The reader may easily verify that both Λ_1 and Λ_2 are convex sets. Also, $\Lambda_1 \cap \Lambda_2 = \varnothing$ because otherwise there would exist a point $(\mathbf{x}, y)$ such that

$$\mathbf{x} \in S, \qquad 0 \geq y > f(\mathbf{x}) - f(\bar{\mathbf{x}}),$$

contradicting the assumption that $\bar{\mathbf{x}}$ is an optimal solution to the problem. By Theorem 2.4.8 there is a hyperplane that separates Λ_1 and Λ_2; that is, there exist a nonzero vector (ξ_0, μ) and a scalar α such that

$$\xi_0^t(\mathbf{x} - \bar{\mathbf{x}}) + \mu y \leq \alpha, \quad \forall \mathbf{x} \in R^n,\ y > f(\mathbf{x}) - f(\bar{\mathbf{x}}) \tag{3.16}$$

$$\xi_0^t(\mathbf{x} - \bar{\mathbf{x}}) + \mu y \geq \alpha, \quad \forall \mathbf{x} \in S,\ y \leq 0. \tag{3.17}$$

If we let $\mathbf{x} = \bar{\mathbf{x}}$ and $y = 0$ in (3.17), it follows that $\alpha \leq 0$. Next, letting $\mathbf{x} = \bar{\mathbf{x}}$ and $y = \varepsilon > 0$ in (3.16), it follows that $\mu\varepsilon \leq \alpha$. Since this is true for every $\varepsilon > 0$, $\mu \leq 0$ and $\alpha \geq 0$. To summarize, we have shown that $\mu \leq 0$ and $\alpha = 0$. If $\mu = 0$, from (3.16), $\xi_0^t(\mathbf{x} - \bar{\mathbf{x}}) \leq 0$ for each $\mathbf{x} \in R^n$. If we let $\mathbf{x} = \bar{\mathbf{x}} + \xi_0$, it follows that

$$0 \geq \xi_0^t(\mathbf{x} - \bar{\mathbf{x}}) = \|\xi_0\|^2$$

and hence $\xi_0 = 0$. Since $(\xi_0, \mu) \neq (0, 0)$, we must have $\mu < 0$. Dividing (3.16) and (3.17) by $-\mu$ and denoting $-\xi_0/\mu$ by ξ, we get the following inequalities:

$$y \geq \xi^t(\mathbf{x} - \bar{\mathbf{x}}), \quad \forall \mathbf{x} \in R^n, \quad y > f(\mathbf{x}) - f(\bar{\mathbf{x}}) \tag{3.18}$$

$$\xi^t(\mathbf{x} - \bar{\mathbf{x}}) - y \geq 0, \quad \forall \mathbf{x} \in S, \quad y \leq 0. \tag{3.19}$$

By letting $y = 0$ in (3.19), we get $\xi^t(\mathbf{x} - \bar{\mathbf{x}}) \geq 0$ for all $\mathbf{x} \in S$. From (3.18) it is obvious that

$$f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \xi^t(\mathbf{x} - \bar{\mathbf{x}}) \qquad \text{for all } \mathbf{x} \in R^n.$$

Therefore, ξ is a subgradient of f at $\bar{\mathbf{x}}$ with the property that $\xi^t(\mathbf{x} - \bar{\mathbf{x}}) \geq 0$ for all $\mathbf{x} \in S$, and the proof is complete.

Corollary 1

Under the assumptions of Theorem 3.4.3, if S is open, $\bar{\mathbf{x}}$ is an optimal solution to the problem if and only if there exists a zero subgradient of f at $\bar{\mathbf{x}}$. In particular, if $S = R^n$, $\bar{\mathbf{x}}$ is a global minimum if and only if there exists a zero subgradient of f at $\bar{\mathbf{x}}$.

Proof

By the theorem, $\bar{\mathbf{x}}$ is an optimal solution if and only if $\xi^t(\mathbf{x} - \bar{\mathbf{x}}) \geq 0$ for each $\mathbf{x} \in S$, where ξ is a subgradient of f at $\bar{\mathbf{x}}$. Since S is open, $\mathbf{x} = \bar{\mathbf{x}} - \lambda\xi \in S$ for some positive λ. Therefore, $-\lambda\|\xi\|^2 \geq 0$; that is, $\xi = \mathbf{0}$.

Corollary 2

In addition to the assumptions of the theorem, suppose that f is differentiable. Then $\bar{\mathbf{x}}$ is an optimal solution if and only if $\nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) \geq 0$ for all $\mathbf{x} \in S$. Furthermore, if S is open, $\bar{\mathbf{x}}$ is an optimal solution if and only if $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$.

Note the important implications of Theorem 3.4.3. First, the theorem gives a necessary and sufficient characterization of optimal solutions. This characterization reduces to the well-known condition of vanishing derivatives if f is differentiable and S is open. Another important implication is that if we reach a nonoptimal point $\bar{\mathbf{x}}$, where $\nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) < 0$ for some $\mathbf{x} \in S$, there is an obvious way to proceed to an improving solution. This can be achieved by moving from $\bar{\mathbf{x}}$ in the direction $\mathbf{d} = \mathbf{x} - \bar{\mathbf{x}}$. The actual size of the step can be determined by solving a *line search problem*, which is a one-dimensional minimization subproblem of the following form: Minimize $f[\bar{\mathbf{x}} + \lambda \mathbf{d}]$ subject to $\lambda \geq 0$ and $\bar{\mathbf{x}} + \lambda\mathbf{d} \in S$. This procedure is called the *method of feasible directions* and is discussed in more detail in Chapter 10.

To provide additional insights, let us dwell for awhile on Corollary 2, which addresses the differentiable case for Theorem 3.4.3. Figure 3.8 illustrates

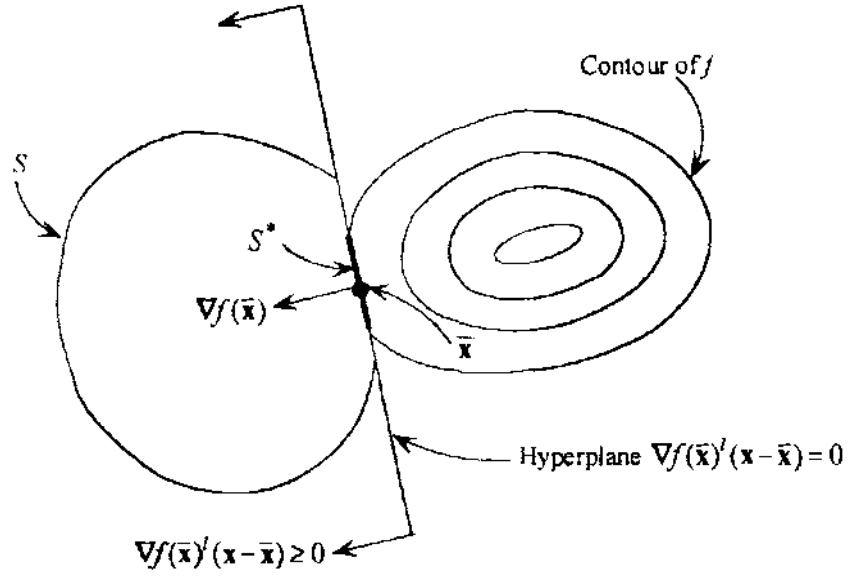


Figure 3.8 Geometry for Theorems 3.4.3 and 3.4.4.

the geometry of the result. Now suppose that for the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$, we have f differentiable and convex, but S is an arbitrary set. Suppose further that it turns out that the directional derivative $f'(\bar{\mathbf{x}};\mathbf{x}-\bar{\mathbf{x}}) = \nabla f(\bar{\mathbf{x}})^t(\mathbf{x}-\bar{\mathbf{x}}) \geq 0$ for all $\mathbf{x} \in S$. The proof of the theorem actually shows that $\bar{\mathbf{x}}$ is a global minimum regardless of S, since for any solution $\hat{\mathbf{x}}$ that improves over $\bar{\mathbf{x}}$, we have, by the convexity of f, that $f(\bar{\mathbf{x}}) > f(\hat{\mathbf{x}}) \geq f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t(\hat{\mathbf{x}}-\bar{\mathbf{x}})$, which implies that $\nabla f(\bar{\mathbf{x}})^t(\hat{\mathbf{x}}-\bar{\mathbf{x}}) < 0$, whereas $\nabla f(\bar{\mathbf{x}})^t(\mathbf{x}-\bar{\mathbf{x}}) \geq 0$ for all $\mathbf{x} \in S$. Hence, the hyperplane $\nabla f(\bar{\mathbf{x}})^t(\mathbf{x}-\bar{\mathbf{x}}) = 0$ separates S from solutions that improve over $\bar{\mathbf{x}}$. [For the nondifferentiable case, the hyperplane $\boldsymbol{\xi}^t(\mathbf{x}-\bar{\mathbf{x}}) = 0$ plays a similar role.] However, if f is not convex, the directional derivative $\nabla f(\bar{\mathbf{x}})^t(\mathbf{x}-\bar{\mathbf{x}})$ being nonnegative for all $\mathbf{x} \in S$ does not even necessarily imply that $\bar{\mathbf{x}}$ is a local minimum. For example, for the problem to minimize $f(x) = x^3$ subject to $-1 \leq x \leq 1$, we have the condition $f'(\bar{x})(x-\bar{x}) \geq 0$ for all $x \in S$ being satisfied at $\bar{x} = 0$, since $f'(0) = 0$, but $\bar{x} = 0$ is not even a local minimum for this problem.

Conversely, suppose that f is differentiable but arbitrary otherwise and that S is a convex set. Then, if $\bar{\mathbf{x}}$ is a global minimum, we must have $f'(\bar{\mathbf{x}};\mathbf{x}-\bar{\mathbf{x}}) = \nabla f(\bar{\mathbf{x}})^t(\mathbf{x}-\bar{\mathbf{x}}) \geq 0$. This follows because, otherwise, if $\nabla f(\bar{\mathbf{x}})^t(\mathbf{x}-\bar{\mathbf{x}}) < 0$, we could move along the direction $\mathbf{d} = \mathbf{x} - \bar{\mathbf{x}}$ and, as above, the objective value would fall for sufficiently small step lengths, whereas $\bar{\mathbf{x}} + \lambda\mathbf{d}$ would remain feasible for $0 \leq \lambda \leq 1$ by the convexity of S. Note that this explains a more general concept: if f is differentiable but f and S are otherwise arbitrary, and if $\bar{\mathbf{x}}$ is a local minimum of f over S, then for any direction $\mathbf{d}$ for which $\bar{\mathbf{x}} + \lambda\mathbf{d}$ remains feasible for $0 < \lambda \leq \delta$ for some $\delta > 0$, we must have a nonnegative

directional derivative of f at $\bar{\mathbf{x}}$ in the direction d; that is, we must have $f'(\bar{\mathbf{x}};\mathbf{d}) = \nabla f(\bar{\mathbf{x}})^t \mathbf{d} \geq 0$.

Now let us turn our attention back to convex programming problems. The following result and its corollaries characterize the set of alternative optimal solutions and show, in part, that the gradient of the objective function (assuming twice differentiability) is a constant over the optimal solution set, and that for a quadratic objective function, the optimal solution set is in fact polyhedral. (See Figure 3.8 to identify the set of alternative optimal solutions S^* defined by the theorem in light of Theorem 3.4.3.)

3.4.4 Theorem

Consider the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$, where f is a convex and twice differentiable function and S is a convex set, and suppose that there exists an optimal solution $\bar{\mathbf{x}}$. Then the set of alternative optimal solutions is characterized by the set

$$S^* = \{\mathbf{x} \in S : \nabla f(\bar{\mathbf{x}})^t (\mathbf{x} - \bar{\mathbf{x}}) \leq 0 \text{ and } \nabla f(\mathbf{x}) = \nabla f(\bar{\mathbf{x}})\}.$$

Proof

Denote the set of alternative optimal solutions as $\bar{S}$, say, and note that $\bar{\mathbf{x}} \in \bar{S} \neq \varnothing$. Consider any $\hat{\mathbf{x}} \in S^*$. By the convexity of f and the definition of S^*, we have $\hat{\mathbf{x}} \in S$ and

$$f(\bar{\mathbf{x}}) \geq f(\hat{\mathbf{x}}) + \nabla f(\hat{\mathbf{x}})^t (\bar{\mathbf{x}} - \hat{\mathbf{x}}) = f(\hat{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t (\bar{\mathbf{x}} - \hat{\mathbf{x}}) \geq f(\hat{\mathbf{x}}),$$

so we must have $\hat{\mathbf{x}} \in \bar{S}$ by the optimality of $\bar{\mathbf{x}}$. Hence, $S^* \subseteq \bar{S}$.

Conversely, suppose that $\hat{\mathbf{x}} \in \bar{S}$, so that $\hat{\mathbf{x}} \in S$ and $f(\hat{\mathbf{x}}) = f(\bar{\mathbf{x}})$. This means that $f(\bar{\mathbf{x}}) = f(\hat{\mathbf{x}}) \geq f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t (\hat{\mathbf{x}} - \bar{\mathbf{x}})$ or that $\nabla f(\bar{\mathbf{x}})^t (\hat{\mathbf{x}} - \bar{\mathbf{x}}) \leq 0$. But by Corollary 2 to Theorem 3.4.3, we have $\nabla f(\bar{\mathbf{x}})^t (\hat{\mathbf{x}} - \bar{\mathbf{x}}) \geq 0$. Hence, $\nabla f(\bar{\mathbf{x}})^t (\hat{\mathbf{x}} - \bar{\mathbf{x}}) = 0$. By interchanging the roles of $\bar{\mathbf{x}}$ and $\hat{\mathbf{x}}$, we obtain $\nabla f(\hat{\mathbf{x}})^t (\bar{\mathbf{x}} - \hat{\mathbf{x}}) = 0$ symmetrically. Therefore,

$$[\nabla f(\bar{\mathbf{x}}) - \nabla f(\hat{\mathbf{x}})]^t (\bar{\mathbf{x}} - \hat{\mathbf{x}}) = 0. \tag{3.20}$$

Now we have

$$\begin{aligned} [\nabla f(\bar{\mathbf{x}}) - \nabla f(\hat{\mathbf{x}})] &= \nabla f[\hat{\mathbf{x}} + \lambda(\bar{\mathbf{x}} - \hat{\mathbf{x}})]_{\lambda=0}^{\lambda=1} \\ &= \int_{\lambda=0}^{\lambda=1} \mathbf{H}[\hat{\mathbf{x}} + \lambda(\bar{\mathbf{x}} - \hat{\mathbf{x}})](\bar{\mathbf{x}} - \hat{\mathbf{x}})\, d\lambda = \mathbf{G}(\bar{\mathbf{x}} - \hat{\mathbf{x}}), \end{aligned} \tag{3.21}$$

where $\mathbf{G} = \int_0^1 \mathbf{H}[\hat{\mathbf{x}} + \lambda(\bar{\mathbf{x}} - \hat{\mathbf{x}})]\, d\lambda$ and where the integral of the matrix is performed componentwise. But note that $\mathbf{G}$ is positive semidefinite because $\mathbf{d}^t\mathbf{G}\mathbf{d} = \int_0^1 \mathbf{d}^t \mathbf{H}[\hat{\mathbf{x}} + \lambda(\bar{\mathbf{x}} - \hat{\mathbf{x}})]\mathbf{d}\ \ d\lambda \geq 0$ for all $\mathbf{d} \in R^n$, since $\mathbf{d}^t\mathbf{H}[\hat{\mathbf{x}} + \lambda(\bar{\mathbf{x}} - \hat{\mathbf{x}})]\mathbf{d}$ is a nonnegative function of λ by the convexity of f. Hence, by (3.20) and (3.21), we get $0 = (\bar{\mathbf{x}} - \hat{\mathbf{x}})^t[\nabla f(\bar{\mathbf{x}}) - \nabla f(\hat{\mathbf{x}})] = (\bar{\mathbf{x}} - \hat{\mathbf{x}})^t \mathbf{G}(\bar{\mathbf{x}} - \hat{\mathbf{x}})$. But the positive semidefiniteness of $\mathbf{G}$ implies that $\mathbf{G}(\bar{\mathbf{x}} - \hat{\mathbf{x}}) = 0$ by a standard result (see Exercise 3.41). Therefore, by (3.21), we have $\nabla f(\bar{\mathbf{x}}) = \nabla f(\hat{\mathbf{x}})$. We have hence shown that $\hat{\mathbf{x}} \in S$, $\nabla f(\bar{\mathbf{x}})^t\ (\hat{\mathbf{x}} - \bar{\mathbf{x}}) \leq 0$, and $\nabla f(\hat{\mathbf{x}}) = \nabla f(\bar{\mathbf{x}})$. This means that $\hat{\mathbf{x}} \in S^*$, and thus $\bar{S} \subseteq S^*$. This, together with $S^* \subseteq \bar{S}$, completes the proof.

Corollary 1

The set S^* of alternative optimal solutions can equivalently be defined as

$$S^* = \{\mathbf{x} \in S : \nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) = 0 \text{ and } \nabla f(\mathbf{x}) = \nabla f(\bar{\mathbf{x}})\}.$$

Proof

The proof follows from the definition of S^* in Theorem 3.4.4 and the fact that $\nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) \geq 0$ for all $\mathbf{x} \in S$ by Corollary 2 to Theorem 3.4.3.

Corollary 2

Suppose that f is a quadratic function given by $f(\mathbf{x}) = \mathbf{c}^t\mathbf{x} + (1/2)\mathbf{x}^t\mathbf{H}\mathbf{x}$ and that S is polyhedral. Then S^* is a polyhedral set given by

$$S^* = \{\mathbf{x} \in S : \mathbf{c}^t(\mathbf{x} - \bar{\mathbf{x}}) \leq 0,\ \mathbf{H}(\mathbf{x} - \bar{\mathbf{x}}) = 0\} = \{\mathbf{x} \in S : \mathbf{c}^t(\mathbf{x} - \bar{\mathbf{x}}) = 0,$$

$$\mathbf{H}(\mathbf{x} - \bar{\mathbf{x}}) = 0\}.$$

Proof

The proof follows by direct substitution in Theorem 3.4.4 and Corollary 1, noting that $\nabla f(\mathbf{x}) = \mathbf{c} + \mathbf{H}\mathbf{x}$.

3.4.5 Example

$$\begin{aligned} \text{Minimize } & \left(x_1 - \frac{3}{2}\right)^2 + (x_2 - 5)^2 \\ \text{subject to } & -x_1 + x_2 \leq 2 \\ & 2x_1 + 3x_2 \leq 11 \\ & -x_1 \leq 0 \\ & -x_2 \leq 0. \end{aligned}$$

Clearly, $f(x_1, x_2) = (x_1 - 3/2)^2 + (x_2 - 5)^2$ is a convex function, which gives the square of the distance from the point (3/2, 5). The convex polyhedral set S is represented by the above four inequalities. The problem is depicted in Figure 3.9. From the figure, clearly the optimal point is (1, 3). The gradient vector of f at the point (1, 3) is $\nabla f(1,3) = (-1,-4)^t$. We see geometrically that the vector (–1, –4) makes an angle of $< 90°$ with each vector of the form $(x_1 - 1, x_2 - 3)$, where $(x_1, x_2) \in S$. Thus, the optimality condition of Theorem 3.4.3 is verified and, by Theorem 3.4.4, (1, 3) is the unique optimum.

To illustrate further, suppose that it is claimed that $\hat{\mathbf{x}} = (0,0)^t$ is an optimal point. By Theorem 3.4.4, this cannot be true since we have $\nabla f(\bar{\mathbf{x}})^t(\hat{\mathbf{x}} - \bar{\mathbf{x}}) = 13 > 0$ when $\bar{\mathbf{x}} = (1,3)^t$. Similarly, by Theorem 3.4.3, we can easily verify that $\hat{\mathbf{x}}$ is not optimal. Note that $\nabla f(0,0) = (-3,-10)^t$ and actually, for each nonzero $\mathbf{x} \in S$, we have $-3x_1 - 10x_2 < 0$. Hence, the origin could not be an optimal point. Moreover, we can improve f by moving from $\mathbf{0}$ in the direction $\mathbf{x} - 0$ for any $\mathbf{x} \in S$. In this case, the best local direction is $-\nabla f(0,0)$, that is, the direction (3, 10). In Chapter 10 we discuss methods for finding a particular direction among many alternatives.

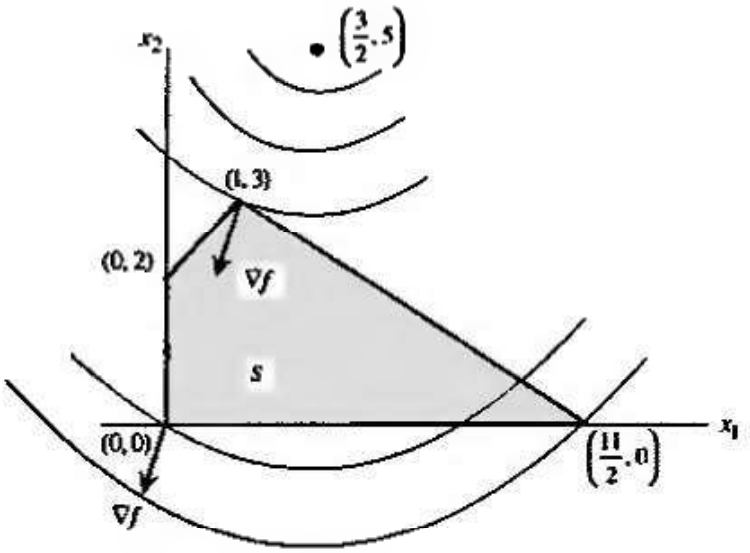


Figure 3.9 Setup for Example 3.4.5.

Maximizing a Convex Function

We now develop a necessary condition for a maximum of a convex function over a convex set. Unfortunately, this condition is not sufficient. Therefore, it is possible, and actually not unlikely, that several local maxima satisfying the condition of Theorem 3.4.6 exist. Unlike the minimization case, there exists no local information at such solutions that could lead us to better points. Hence, maximizing a convex function is usually a much harder task than minimizing a convex function. Again, minimizing a concave function is similar to maximizing a convex function, and hence the development for the concave case is left to the reader.

3.4.6 Theorem

Let f: $R^n \to R$ be a convex function, and let S be a nonempty convex set in R^n. Consider the problem to maximize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$. If $\bar{\mathbf{x}} \in S$ is a local optimal solution, $\xi^t(\mathbf{x} - \bar{\mathbf{x}}) \le 0$ for each $\mathbf{x} \in S$, where ξ is any subgradient of f at $\bar{\mathbf{x}}$.

Proof

Suppose that $\bar{\mathbf{x}} \in S$ is a local optimal solution. Then an ε-neighborhood $N_\varepsilon(\bar{\mathbf{x}})$ exists such that $f(\mathbf{x}) \le f(\bar{\mathbf{x}})$ for each $\mathbf{x} \in S \cap N_\varepsilon(\bar{\mathbf{x}})$. Let $\mathbf{x} \in S$, and note that $\bar{\mathbf{x}} + \lambda(\mathbf{x} - \bar{\mathbf{x}}) \in S \cap N_\varepsilon(\bar{\mathbf{x}})$ for $\lambda > 0$ and sufficiently small. Hence,

$$f[\bar{\mathbf{x}} + \lambda(\mathbf{x} - \bar{\mathbf{x}})] \le f(\bar{\mathbf{x}}). \tag{3.22}$$

Let ξ be a subgradient of f at $\bar{\mathbf{x}}$. By the convexity of f, we have

$$f[\bar{\mathbf{x}} + \lambda(\mathbf{x} - \bar{\mathbf{x}})] - f(\bar{\mathbf{x}}) \ge \lambda\xi^t(\mathbf{x} - \bar{\mathbf{x}}).$$

The above inequality, together with (3.20), implies that $\lambda\xi^t(\mathbf{x} - \bar{\mathbf{x}}) \le 0$, and dividing by $\lambda > 0$, the result follows.

Corollary

In addition to the assumptions of the theorem, suppose that f is differentiable. If $\bar{\mathbf{x}} \in S$ is a local optimal solution, $\nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) \le 0$ for all $\mathbf{x} \in S$.

Note that the above result is, in general, necessary but not sufficient for optimality. To illustrate, let $f(x) = x^2$ and $S = \{x : -1 \le x \le 2\}$. The maximum of f over S is equal to 4 and is achieved at $x = 2$. However, at $\bar{x} = 0$, we have $\nabla f(\bar{x}) = 0$ and hence $\nabla f(\bar{x})^t(x - \bar{x}) = 0$ for each $x \in S$. Clearly, the point $\bar{x} = 0$ is not even a local maximum. Referring to Example 3.4.5, discussed earlier, we

have two local maxima, (0, 0) and (11/2, 0). Both points satisfy the necessary condition of Theorem 3.4.6. If we are currently at the local optimal point (0, 0), unfortunately no local information exists that will lead us toward the global maximum point (11/2, 0). Also, if we are at the global maximum point (11/2, 0), there is no convenient local criterion that tells us that we are at the optimal point.

Theorem 3.4.7 shows that a convex function achieves a maximum over a compact polyhedral set at an extreme point. This result has been utilized by several computational schemes for solving such problems. We ask the reader to think for a moment about the case when the objective function is linear and, hence, both convex and concave. Theorem 3.4.7 could be extended to the case where the convex feasible region is not polyhedral.

3.4.7 Theorem

Let $f: R^n \to R$ be a convex function, and let S be a nonempty compact polyhedral set in R^n. Consider the problem to maximize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$. An optimal solution $\bar{\mathbf{x}}$ to the problem then exists, where $\bar{\mathbf{x}}$ is an extreme point of S.

Proof

By Theorem 3.1.3, note that f is continuous. Since S is compact, f assumes a maximum at $\mathbf{x}' \in S$. If $\mathbf{x}'$ is an extreme point of S, the result is at hand. Otherwise, by Theorem 2.6.7, $\mathbf{x}' = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j$, where $\sum_{j=1}^{k} \lambda_j = 1$, $\lambda_j > 0$, and $\mathbf{x}_j$ is an extreme point of S for $j = 1,..., k$. By the convexity of f, we have

$$f(\mathbf{x}') = f\left(\sum_{j=1}^{k} \lambda_j \mathbf{x}_j\right) \leq \sum_{j=1}^{k} \lambda_j f(\mathbf{x}_j).$$

But since $f(\mathbf{x}') \geq f(\mathbf{x}_j)$ for $j = 1,..., k$, the above inequality implies that $f(\mathbf{x}') = f(\mathbf{x}_j)$ for $j = 1,..., k$. Thus, the extreme points $\mathbf{x}_1,..., \mathbf{x}_k$ are optimal solutions to the problem, and the proof is complete.

3.5 Generalizations of a Convex Functions

In this section we present various types of functions that are similar to convex and concave functions but that share only some of their desirable properties. As we shall learn, many of the results presented later in the book do not require the restrictive assumption of convexity, but rather, the less restrictive assumptions of quasiconvexity, pseudoconvexity, and convexity at a point.

Quasiconvex Functions

Definition 3.5.1 introduces quasiconvex functions. From the definition it is apparent that every convex function is also quasiconvex.

3.5.1 Definition

Let f: $S \to R$, where S is a nonempty convex set in R^n. The function f is said to be *quasiconvex* if for each $\mathbf{x}_1$ and $\mathbf{x}_2 \in S$, the following inequality is true:

$$f[\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] \leq \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\} \quad \text{for each } \lambda \in (0, 1).$$

The function f is said to be *quasiconcave* if $-f$ is quasiconvex.

From Definition 3.5.1, a function f is quasiconvex if whenever $f(\mathbf{x}_2) \geq f(\mathbf{x}_1)$, $f(\mathbf{x}_2)$ is greater than or equal to f at all convex combinations of $\mathbf{x}_1$ and $\mathbf{x}_2$. Hence, if f increases from its value at a point along any direction, it must remain nondecreasing in that direction. Therefore, its univariate cross section is either monotone or unimodal (see Exercise 3.57). A function f is quasiconcave if whenever $f(\mathbf{x}_2) \geq f(\mathbf{x}_1)$, f at all convex combinations of $\mathbf{x}_1$ and $\mathbf{x}_2$ is greater than or equal to $f(\mathbf{x}_1)$. Figure 3.10 shows some examples of quasiconvex and quasiconcave functions. We shall concentrate on quasiconvex functions; the reader is advised to draw all the parallel results for quasiconcave functions. A function that is both quasiconvex and quasiconcave is called *quasimonotone* (see Figure 3.10d).

We have learned in Section 3.2 that a convex function can be characterized by the convexity of its epigraph. We now learn that a quasiconvex function can be characterized by the convexity of its level sets. This result is given in Theorem 3.5.2.

3.5.2 Theorem

Let f: $S \to R$ where S is a nonempty convex set in R^n. The function f is quasiconvex if and only if $S_\alpha = \{\mathbf{x} \in S : f(\mathbf{x}) \leq \alpha\}$ is convex for each real number α.

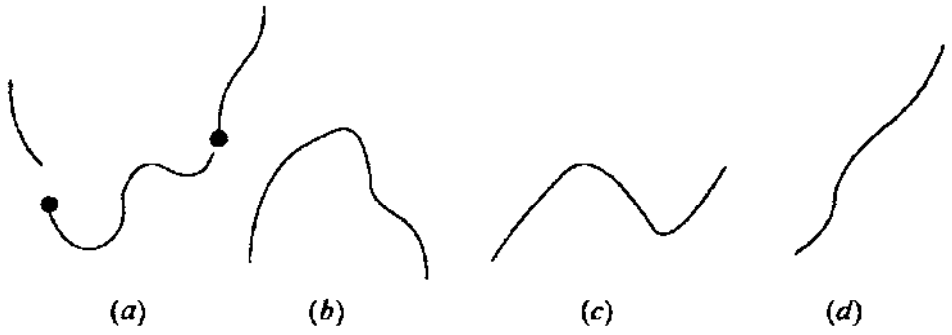


Figure 3.10 Quasiconvex and quasiconcave functions: (*a*) quasiconvex, (*b*) quasiconcave, (*c*) neither quasiconvex nor quasiconcave, (*d*) quasimonotone.

Proof

Suppose that f is quasiconvex, and let $\mathbf{x}_1, \mathbf{x}_2 \in S_\alpha$. Therefore, $\mathbf{x}_1, \mathbf{x}_2 \in S$ and $\max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\} \leq \alpha$. Let $\lambda \in (0,1)$, and let $\mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$. By the convexity of S, $\mathbf{x} \in S$. Furthermore, by the quasiconvexity of f, $f(\mathbf{x}) \leq \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\} \leq \alpha$. Hence, $\mathbf{x} \in S_\alpha$ and thus S_α is convex. Conversely, suppose that S_α is convex for each real number α. Let $\mathbf{x}_1, \mathbf{x}_2 \in S$. Furthermore, let $\lambda \in (0,1)$ and $\mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$. Note that $\mathbf{x}_1, \mathbf{x}_2 \in S_\alpha$ for $\alpha = \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\}$. By assumption, S_α is convex, so that $\mathbf{x} \in S_\alpha$. Therefore, $f(\mathbf{x}) \leq \alpha = \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\}$. Hence, f is quasiconvex, and the proof is complete.

The level set S_α defined in Theorem 3.5.2 is sometimes referred to as a *lower-level set*, to differentiate it from the *upper-level set* $\{\mathbf{x} \in S : f(\mathbf{x}) \geq \alpha\}$, which is convex for all $\alpha \in R$ if and only if f is quasiconcave. Also, it can be shown (see Exercise 3.59) that f is quasimonotone if and only if the *level surface* $\{\mathbf{x} \in S : f(\mathbf{x}) = \alpha\}$ is convex for all $\alpha \in R$.

We now give a result analogous to Theorem 3.4.7. Theorem 3.5.3 shows that the maximum of a continuous quasiconvex function over a compact polyhedral set occurs at an extreme point.

3.5.3 Theorem

Let S be a nonempty compact polyhedral set in R^n, and let $f: R^n \to R$ be quasiconvex and continuous on S. Consider the problem to maximize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$. Then an optimal solution $\bar{\mathbf{x}}$ to the problem exists, where $\bar{\mathbf{x}}$ is an extreme point of S.

Proof

Note that f is continuous on S and hence attains a maximum, say, at $\mathbf{x}' \in S$. If there is an extreme point whose objective is equal to $f(\mathbf{x}')$, the result is at hand. Otherwise, let $\mathbf{x}_1, \ldots, \mathbf{x}_k$ be the extreme points of S, and assume that $f(\mathbf{x}') > f(\mathbf{x}_j)$ for $j = 1, \ldots, k$. By Theorem 2.6.7, $\mathbf{x}'$ can be represented as

$$\mathbf{x}' = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j$$

$$\sum_{j=1}^{k} \lambda_j = 1$$

$$\lambda_j \geq 0, \qquad j = 1, \ldots, k.$$

Since $f(\mathbf{x}') > f(\mathbf{x}_j)$ for each j, then

$$f(\mathbf{x}') > \max_{1 \le j \le k} f(\mathbf{x}_j) = \alpha. \tag{3.23}$$

Now, consider the set $S_\alpha = \{\mathbf{x} : f(\mathbf{x}) \le \alpha\}$. Note that $\mathbf{x}_j \in S_\alpha$ for $j = 1,\ldots, k$, and by the quasiconvexity of f, S_α is convex. Hence, $\mathbf{x}' = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j$ belongs to S_α. This implies that $f(\mathbf{x}') \le \alpha$, which contradicts (3.23). This contradiction shows that $f(\mathbf{x}') = f(\mathbf{x}_j)$ for some extreme point $\mathbf{x}_j$, and the proof is complete.

Differentiable Quasiconvex Functions

The following theorem gives a necessary and sufficient characterization of a differentiable quasiconvex function. (See Appendix B for a second-order characterization in terms of *bordered Hessian determinants*.)

3.5.4 Theorem

Let S be a nonempty open convex set in R^n, and let $f: S \to R$ be differentiable on S. Then f is quasiconvex if and only if either one of the following equivalent statements holds true:

1. If $\mathbf{x}_1, \mathbf{x}_2 \in S$ and $f(\mathbf{x}_1) \le f(\mathbf{x}_2)$, $\nabla f(\mathbf{x}_2)^t(\mathbf{x}_1 - \mathbf{x}_2) \le 0$.
2. If $\mathbf{x}_1, \mathbf{x}_2 \in S$ and $\nabla f(\mathbf{x}_2)^t(\mathbf{x}_1 - \mathbf{x}_2) > 0$, $f(\mathbf{x}_1) > f(\mathbf{x}_2)$.

Proof

Obviously, statements 1 and 2 are equivalent. We shall prove Part 1. Let f be quasiconvex, and let $\mathbf{x}_1, \mathbf{x}_2 \in S$ be such that $f(\mathbf{x}_1) \le f(\mathbf{x}_2)$. By the differentiability of f at $\mathbf{x}_2$, for $\lambda \in (0, 1)$, we have

$$f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] - f(\mathbf{x}_2) = \lambda \nabla f(\mathbf{x}_2)^t(\mathbf{x}_1 - \mathbf{x}_2) + \lambda \|\mathbf{x}_1 - \mathbf{x}_2\| \alpha[\mathbf{x}_2; \lambda(\mathbf{x}_1 - \mathbf{x}_2)],$$

where $\alpha[\mathbf{x}_2; \lambda(\mathbf{x}_1 - \mathbf{x}_2)] \to 0$ as $\lambda \to 0$. By the quasiconvexity of f, we have $f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] \le f(\mathbf{x}_2)$, and hence the above equation implies that

$$\lambda \nabla f(\mathbf{x}_2)^t(\mathbf{x}_1 - \mathbf{x}_2) + \lambda \|\mathbf{x}_1 - \mathbf{x}_2\| \alpha[\mathbf{x}_2; \lambda(\mathbf{x}_1 - \mathbf{x}_2)] \le 0.$$

Dividing by λ and letting $\lambda \to 0$, we get $\nabla f(\mathbf{x}_2)^t(\mathbf{x}_1 - \mathbf{x}_2) \le 0$.

Conversely, suppose that $\mathbf{x}_1, \mathbf{x}_2 \in S$ and that $f(\mathbf{x}_1) \le f(\mathbf{x}_2)$. We need to show that given Part 1, we have $f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] \le f(\mathbf{x}_2)$ for each $\lambda \in (0, 1)$. We do this by showing that the set

$$L = \{\mathbf{x} : \mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \lambda \in (0,1), f(\mathbf{x}) > f(\mathbf{x}_2)\}$$

is empty. By contradiction, suppose that there exists an $\mathbf{x}' \in L$. Therefore, $\mathbf{x}' = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ for some $\lambda \in (0, 1)$ and $f(\mathbf{x}') > f(\mathbf{x}_2)$. Since f is differentiable, it is continuous, and there must exist a $\delta \in (0, 1)$ such that

$$f[\mu\mathbf{x}' + (1-\mu)\mathbf{x}_2] > f(\mathbf{x}_2) \qquad \text{for each } \mu \in [\delta, 1] \tag{3.24}$$

and $f(\mathbf{x}') > f[\delta\mathbf{x}' + (1-\delta)\mathbf{x}_2]$. By this inequality and the mean value theorem, we must have

$$0 < f(\mathbf{x}') - f[\delta\mathbf{x}' + (1-\delta)\mathbf{x}_2] = (1-\delta)\nabla f(\hat{\mathbf{x}})^t(\mathbf{x}' - \mathbf{x}_2), \tag{3.25}$$

where $\hat{\mathbf{x}} = \hat{\mu}\mathbf{x}' + (1-\hat{\mu})\mathbf{x}_2$ for some $\hat{\mu} \in (\delta, 1)$. From (3.24) it is clear that $f(\hat{\mathbf{x}}) > f(\mathbf{x}_2)$. Dividing (3.25) by $1 - \delta > 0$, it follows that $\nabla f(\hat{\mathbf{x}})^t(\mathbf{x}' - \mathbf{x}_2) > 0$, which in turn implies that

$$\nabla f(\hat{\mathbf{x}})^t(\mathbf{x}_1 - \mathbf{x}_2) > 0. \tag{3.26}$$

But on the other hand, $f(\hat{\mathbf{x}}) > f(\mathbf{x}_2) \geq f(\mathbf{x}_1)$, and $\hat{\mathbf{x}}$ is a convex combination of $\mathbf{x}_1$ and $\mathbf{x}_2$, say $\hat{\mathbf{x}} = \hat{\lambda}\mathbf{x}_1 + (1-\hat{\lambda})\mathbf{x}_2$, where $\hat{\lambda} \in (0,1)$. By the assumption of the theorem, $\nabla f(\hat{\mathbf{x}})^t(\mathbf{x}_1 - \hat{\mathbf{x}}) \leq 0$, and thus we must have

$$0 \geq \nabla f(\hat{\mathbf{x}})^t(\mathbf{x}_1 - \hat{\mathbf{x}}) = (1-\hat{\lambda})\nabla f(\hat{\mathbf{x}})^t(\mathbf{x}_1 - \mathbf{x}_2).$$

The above inequality is not compatible with (3.26). Therefore, L is empty, and the proof is complete.

To illustrate Theorem 3.5.4, let $f(x) = x^3$. To check its quasiconvexity, suppose that $f(x_1) \leq f(x_2)$, that is, $x_1^3 \leq x_2^3$. This is true only if $x_1 \leq x_2$. Now consider $\nabla f(x_2)(x_1 - x_2) = 3(x_1 - x_2)x_2^2$. Since $x_1 \leq x_2$, $3(x_1 - x_2)x_2^2 \leq 0$. Therefore, $f(x_1) \leq f(x_2)$ implies that $\nabla f(x_2)(x_1 - x_2) \leq 0$, and by the theorem we have that f is quasiconvex. As another illustration, let $f(x_1, x_2) = x_1^3 + x_2^3$. Let $\mathbf{x}_1 = (2, -2)^t$ and $\mathbf{x}_2 = (1,0)^t$. Note that $f(\mathbf{x}_1) = 0$ and $f(\mathbf{x}_2) = 1$, so that $f(\mathbf{x}_1) < f(\mathbf{x}_2)$. But on the other hand, $\nabla f(\mathbf{x}_2)^t(\mathbf{x}_1 - \mathbf{x}_2) = (3,0)(1,-2)^t = 3$. By the necessary part of the theorem, f is not quasiconvex. This also shows that the sum of two quasiconvex functions is not necessarily quasiconvex.

Strictly Quasiconvex Functions

Strictly quasiconvex and strictly quasiconcave functions are especially important in nonlinear programming because they ensure that a local minimum and a local

maximum over a convex set are, respectively, a global minimum and a global maximum.

3.5.5 Definition

Let f: $S \to R$, where S is a nonempty convex set in R^n. The function f is said to be *strictly quasiconvex* if for each $\mathbf{x}_1, \mathbf{x}_2 \in S$ with $f(\mathbf{x}_1) \neq f(\mathbf{x}_2)$, we have

$$f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] < \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\} \qquad \text{for each } \lambda \in (0, 1).$$

The function f is called *strictly quasiconcave* if $-f$ is strictly quasiconvex. Strictly quasiconvex functions are also sometimes referred to as *semi-strictly quasiconvex*, *functionally convex*, or *explicitly quasiconvex*.

Note from Definition 3.5.5 that every convex function is strictly quasiconvex. Figure 3.11 gives examples of strictly quasiconvex and strictly quasiconcave functions. Also, the definition precludes any "flat spots" from occurring anywhere except at extremizing points. This is formalized by the following theorem, which shows that a local minimum of a strictly quasiconvex function over a convex set is also a global minimum. This property is not enjoyed by quasiconvex functions, as seen in Figure 3.10*a*.

3.5.6 Theorem

Let f: $R^n \to R$ be strictly quasiconvex. Consider the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$, where S is a nonempty convex set in R^n. If $\bar{\mathbf{x}}$ is a local optimal solution, $\bar{\mathbf{x}}$ is also a global optimal solution.

Proof

Assume, on the contrary, that there exists an $\hat{\mathbf{x}} \in S$ with $f(\hat{\mathbf{x}}) < f(\bar{\mathbf{x}})$. By the convexity of S, $\lambda\hat{\mathbf{x}} + (1-\lambda)\bar{\mathbf{x}} \in S$ for each $\lambda \in (0, 1)$. Since $\bar{\mathbf{x}}$ is a local minimum by assumption, $f(\bar{\mathbf{x}}) \leq f[\lambda\hat{\mathbf{x}} + (1-\lambda)\bar{\mathbf{x}}]$ for all $\lambda \in (0, \delta)$ and for some $\delta \in$

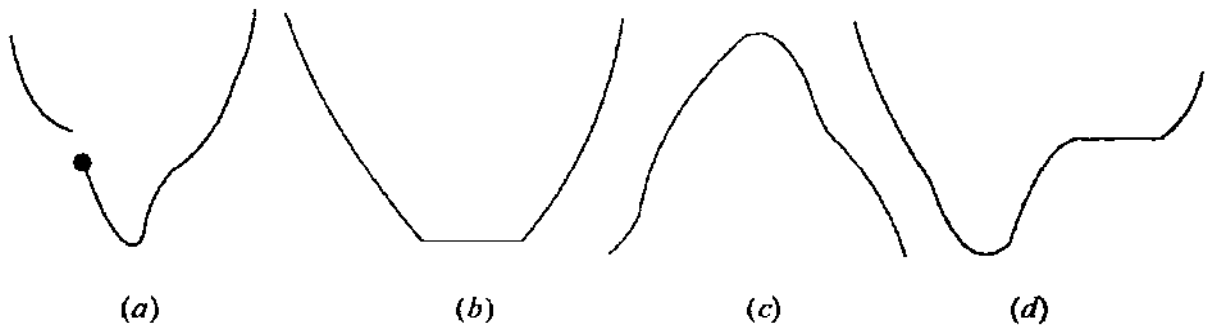


Figure 3.11 Strictly quasiconvex and strictly quasiconcave functions: (*a*) strictly quasiconvex, (*b*) strictly quasiconvex, (*c*) strictly quasiconcave, (*d*) neither strictly quasiconvex nor quasiconcave.

(0, 1). But because f is strictly quasiconvex and $f(\hat{\mathbf{x}}) < f(\bar{\mathbf{x}})$, we have that $f[\lambda\hat{\mathbf{x}} + (1-\lambda)\bar{\mathbf{x}}] < f(\bar{\mathbf{x}})$ for each $\lambda \in (0, 1)$. This contradicts the local optimality of $\bar{\mathbf{x}}$, and the proof is complete.

As seen from Definition 3.1.1, every strictly convex function is indeed a convex function. But every strictly quasiconvex function is not quasiconvex. To illustrate, consider the following function given by Karamardian [1967]:

$$f(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x \neq 0. \end{cases}$$

By Definition 3.5.5, f is strictly quasiconvex. However, f is not quasiconvex, since for $x_1 = 1$ and $x_2 = -1$, $f(x_1) = f(x_2) = 0$, but $f[(1/2)x_1 + (1/2)x_2] = f(0) = 1 > f(x_2)$. If f is lower semicontinuous, however, then as shown below, strict quasiconvexity implies quasiconvexity, as one would usually expect from the word *strict*. (For a definition of lower semicontinuity, refer to Appendix A.)

3.5.7 Lemma

Let S be a nonempty convex set in R^n and let $f: S \to R$ be strictly quasiconvex and lower semicontinuous. Then f is quasiconvex.

Proof

Let $\mathbf{x}_1$ and $\mathbf{x}_2 \in S$. If $f(\mathbf{x}_1) \neq f(\mathbf{x}_2)$, then by the strict quasiconvexity of f, we must have $f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] < \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\}$ for each $\lambda \in (0, 1)$. Now, suppose that $f(\mathbf{x}_1) = f(\mathbf{x}_2)$. To show that f is quasiconvex, we need to show that $f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] \leq f(\mathbf{x}_1)$ for each $\lambda \in (0, 1)$. By contradiction, suppose that $f[\mu\mathbf{x}_1 + (1-\mu)\mathbf{x}_2] > f(\mathbf{x}_1)$ for some $\mu \in (0, 1)$. Denote $\mu\mathbf{x}_1 + (1-\mu)\mathbf{x}_2$ by $\mathbf{x}$. Since f is lower semicontinuous, there exists a $\lambda \in (0, 1)$ such that

$$f(\mathbf{x}) > f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}] > f(\mathbf{x}_1) = f(\mathbf{x}_2). \tag{3.27}$$

Note that $\mathbf{x}$ can be represented as a convex combination of $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}$ and $\mathbf{x}_2$. Hence, by the strict quasiconvexity of f and since $f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}] > f(\mathbf{x}_2)$, we have $f(\mathbf{x}) < f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}]$, contradicting (3.27). This completes the proof.

Strongly Quasiconvex Functions

From Theorem 3.5.6 it followed that a local minimum of a strictly quasiconvex function over a convex set is also a global optimal solution. However, strict quasiconvexity does not assert uniqueness of the global optimal solution. We shall define here another version of quasiconvexity, called *strong quasiconvexity*, which assures uniqueness of the global minimum when it exists.

3.5.8 Definition

Let S be a nonempty convex set in R^n, and let f: $S \to R$. The function f is said to be *strongly quasiconvex* if for each $\mathbf{x}_1, \mathbf{x}_2 \in S$, with $\mathbf{x}_1 \neq \mathbf{x}_2$, we have

$$f[\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] < \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\}$$

for each $\lambda \in (0, 1)$. The function f is said to be *strongly quasiconcave* if $-f$ is strongly quasiconvex. (We caution the reader that such a function is sometimes referred to in the literature as being *strictly quasiconvex*, whereas a function satisfying Definition 3.5.5 is called *semi-strictly quasiconvex*. This is done because of Karamardian's example given above and Property 3 below.)

From Definition 3.5.8 and from Definitions 3.1.1, 3.5.1, and 3.5.5, the following statements hold true:

1. Every strictly convex function is strongly quasiconvex.
2. Every strongly quasiconvex function is strictly quasiconvex.
3. Every strongly quasiconvex function is quasiconvex even in the absence of any semicontinuity assumption.

Figure 3.11*a* illustrates a case where the function is both strongly quasiconvex and strictly quasiconvex, whereas the function represented in Figure 3.11*b* is strictly quasiconvex but not strongly quasiconvex. The key to strong quasiconvexity is that it enforces strict unimodality (see Exercise 3.58). This leads to the following property.

3.5.9 Theorem

Let f: $R^n \to R$ be strongly quasiconvex. Consider the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$, where S is a nonempty convex set in R^n. If $\bar{\mathbf{x}}$ is a local optimal solution, $\bar{\mathbf{x}}$ is the unique global optimal solution.

Proof

Since $\bar{\mathbf{x}}$ is a local optimal solution, there exists an ε-neighborhood $N_\varepsilon(\bar{\mathbf{x}})$ around $\bar{\mathbf{x}}$ such that $f(\bar{\mathbf{x}}) \leq f(\mathbf{x})$ for all $\mathbf{x} \in S \cap N_\varepsilon(\bar{\mathbf{x}})$. Suppose, by contradiction to the conclusion of the theorem, that there exists a point $\hat{\mathbf{x}} \in S$ such that $\hat{\mathbf{x}} \neq \bar{\mathbf{x}}$ and $f(\hat{\mathbf{x}}) \leq f(\bar{\mathbf{x}})$. By strong quasiconvexity it follows that

$$f[\lambda \hat{\mathbf{x}} + (1-\lambda)\bar{\mathbf{x}}] < \max\{f(\hat{\mathbf{x}}), f(\bar{\mathbf{x}})\} = f(\bar{\mathbf{x}})$$

for all $\lambda \in (0, 1)$. But for λ small enough, $\lambda \hat{\mathbf{x}} + (1-\lambda)\bar{\mathbf{x}} \in S \cap N_\varepsilon(\bar{\mathbf{x}})$, so that the above inequality violates the local optimality of $\bar{\mathbf{x}}$. This completes the proof.

Pseudoconvex Functions

The astute reader might already have observed that differentiable strongly (or strictly) quasiconvex functions do not share the particular property of convex

functions, which says that if $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$ at some point $\bar{\mathbf{x}}$, $\bar{\mathbf{x}}$ is a global minimum of f. Figure 3.12*c* illustrates this fact. This motivates the definition of pseudoconvex functions that share this important property with convex functions, and leads to a generalization of various derivative-based optimality conditions.

3.5.10 Definition

Let S be a nonempty open set in R^n, and let $f: S \to R$ be differentiable on S. The function f is said to be *pseudoconvex* if for each $\mathbf{x}_1, \mathbf{x}_2 \in S$ with $\nabla f(\mathbf{x}_1)^t (\mathbf{x}_2 - \mathbf{x}_1) \geq 0$, we have $f(\mathbf{x}_2) \geq f(\mathbf{x}_1)$; or equivalently, if $f(\mathbf{x}_2) < f(\mathbf{x}_1)$, $\nabla f(\mathbf{x}_1)^t (\mathbf{x}_2 - \mathbf{x}_1) < 0$. The function f is said to be *pseudoconcave* if $-f$ is pseudoconvex.

The function f is said to be *strictly pseudoconvex* if for each distinct $\mathbf{x}_1, \mathbf{x}_2 \in S$ satisfying $\nabla f(\mathbf{x}_1)^t (\mathbf{x}_2 - \mathbf{x}_1) \geq 0$, we have $f(\mathbf{x}_2) \geq f(\mathbf{x}_1)$; or equivalently, if for each distinct $\mathbf{x}_1, \mathbf{x}_2 \in S$, $f(\mathbf{x}_2) \leq f(\mathbf{x}_1)$ implies that $\nabla f(\mathbf{x}_1)^t (\mathbf{x}_2 - \mathbf{x}_1) < 0$. The function f is said to be *strictly pseudoconcave* if $-f$ is strictly pseudoconvex.

Figure 3.12*a* illustrates a pseudoconvex function. From the definition of pseudoconvexity it is clear that if $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$ at any point $\bar{\mathbf{x}}$, $f(\mathbf{x}) \geq f(\bar{\mathbf{x}})$ for all $\bar{\mathbf{x}}$; so $\bar{\mathbf{x}}$ is a global minimum for f. Hence, the function in Figure 3.12*c* is neither pseudoconvex nor pseudoconcave. In fact, the definition asserts that if the directional derivative of f at any point $\mathbf{x}_1$ in the direction $(\mathbf{x}_2 - \mathbf{x}_1)$ is nonnegative, the function values are nondecreasing in that direction (see Exercise 3.69). Furthermore, observe that the pseudoconvex functions shown in Figure 3.12 are also strictly quasiconvex, which is true in general, as shown by Theorem 3.5.11. The reader may note that the function in Figure 3.8*c* is not pseudoconvex, yet it is strictly quasiconvex.

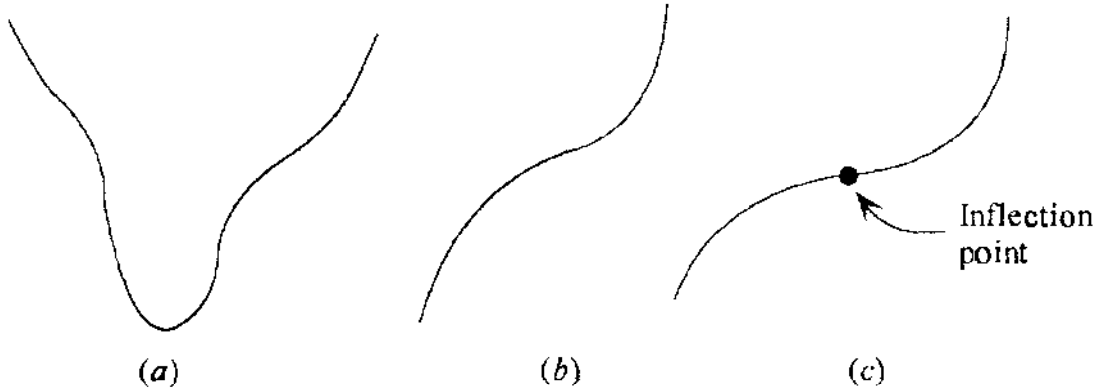


Figure 3.12 Pseudoconvex and pseudoconcave functions: (*a*) pseudoconvex, (*b*) both pseudoconvex and pseudoconcave, (*c*) neither pseudoconvex nor pseudoconcave.

3.5.11 Theorem

Let S be a nonempty open convex set in R^n, and let f: $S \to R$ be a differentiable pseudoconvex function on S. Then f is both strictly quasiconvex and quasiconvex.

Proof

We first show that f is strictly quasiconvex. By contradiction, suppose that there exist $\mathbf{x}_1, \mathbf{x}_2 \in S$ such that $f(\mathbf{x}_1) \neq f(\mathbf{x}_2)$ and $f(\mathbf{x}') \geq \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\}$, where $\mathbf{x}' = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ for some $\lambda \in (0, 1)$. Without loss of generality, assume that $f(\mathbf{x}_1) < f(\mathbf{x}_2)$, so that

$$f(\mathbf{x}') \geq f(\mathbf{x}_2) > f(\mathbf{x}_1). \tag{3.28}$$

Note, by the pseudoconvexity of f, that $\nabla f(\mathbf{x}')^t(\mathbf{x}_1 - \mathbf{x}') < 0$. Now since $\nabla f(\mathbf{x}')^t(\mathbf{x}_1 - \mathbf{x}') < 0$ and $\mathbf{x}_1 - \mathbf{x}' = -(1-\lambda)(\mathbf{x}_2 - \mathbf{x}')/\lambda$, $\nabla f(\mathbf{x}')^t(\mathbf{x}_2 - \mathbf{x}') > 0$; and hence, by the pseudoconvexity of f, we must have $f(\mathbf{x}_2) \geq f(\mathbf{x}')$. Therefore, by (3.28), we get $f(\mathbf{x}_2) = f(\mathbf{x}')$. Also, since $\nabla f(\mathbf{x}')^t(\mathbf{x}_2 - \mathbf{x}') > 0$, there exists a point $\hat{\mathbf{x}} = \mu\mathbf{x}' + (1-\mu)\mathbf{x}_2$ with $\mu \in (0, 1)$ such that

$$f(\hat{\mathbf{x}}) > f(\mathbf{x}') = f(\mathbf{x}_2).$$

Again, by the pseudoconvexity of f, we have $\nabla f(\hat{\mathbf{x}})^t(\mathbf{x}_2 - \hat{\mathbf{x}}) < 0$. Similarly, $\nabla f(\hat{\mathbf{x}})^t(\mathbf{x}' - \hat{\mathbf{x}}) < 0$. Summarizing, we must have

$$\nabla f(\hat{\mathbf{x}})^t(\mathbf{x}_2 - \hat{\mathbf{x}}) < 0$$

$$\nabla f(\hat{\mathbf{x}})^t(\mathbf{x}' - \hat{\mathbf{x}}) < 0.$$

Note that $\mathbf{x}_2 - \hat{\mathbf{x}} = \mu(\hat{\mathbf{x}} - \mathbf{x}')/(1-\mu)$, and hence the above two inequalities are not compatible. This contradiction shows that f is strictly quasiconvex. By Lemma 3.5.7, then f is also quasiconvex, and the proof is complete.

In Theorem 3.5.12 we see that every strictly pseudoconvex function is strongly quasiconvex.

3.5.12 Theorem

Let S be a nonempty open convex set in R^n, and let f: $S \to R$ be a differentiable strictly pseudoconvex function. Then f is strongly quasiconvex.

Proof

By contradiction, suppose that there exist distinct $\mathbf{x}_1, \mathbf{x}_2 \in S$ and $\lambda \in (0, 1)$ such that $f(\mathbf{x}) \geq \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\}$, where $\mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$. Since $f(\mathbf{x}_1)$

$\leq f(\mathbf{x})$, we have, by the strict pseudoconvexity of f, that $\nabla f(\mathbf{x})^t(\mathbf{x}_1 - \mathbf{x}) < 0$ and hence

$$\nabla f(\mathbf{x})^t(\mathbf{x}_1 - \mathbf{x}_2) < 0. \tag{3.29}$$

Similarly, since $f(\mathbf{x}_2) \leq f(\mathbf{x})$, we have

$$\nabla f(\mathbf{x})^t(\mathbf{x}_2 - \mathbf{x}_1) < 0. \tag{3.30}$$

The two inequalities (3.29) and (3.30) are not compatible, and hence f is strongly quasiconvex. This completes the proof.

We remark here in connection with Theorems 3.5.11 and 3.5.12, for the special case in which f is quadratic, that f is pseudoconvex if and only if f is strictly quasiconvex, which holds true if and only if f is quasiconvex. Moreover, we also have that f is strictly pseudoconvex if and only if f is strongly quasiconvex. Hence, all these properties become equivalent to each other for quadratic functions (see Exercise 3.55). Also, Appendix B provides a bordered Hessian determinant characterization for checking the pseudoconvexity and the strict pseudoconvexity of quadratic functions.

Thus far we have discussed various types of convexity and concavity. Figure 3.13 summarizes the implications among these types of convexity. These implications either follow from the definitions or from the various results proved in this section. A similar figure can be constructed for the concave case.

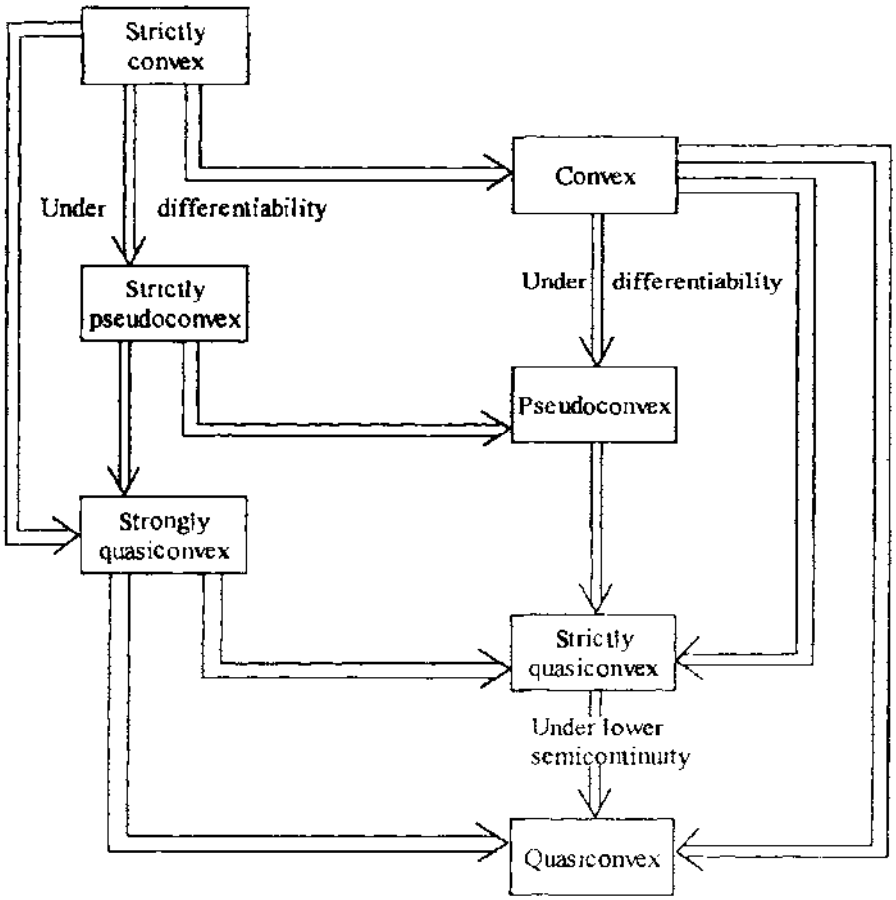


Figure 3.13 Relationship among various types of convexity.

Convexity at a Point

Another useful concept in optimization is the notion of convexity or concavity at a point. In some cases the requirement of a convex or concave function may be too strong and really not essential. Instead, convexity or concavity at a point may be all that is needed.

3.5.13 Definition

Let S be a nonempty convex set in R^n, and let $f: S \to R$. The following are relaxations of various forms of convexity presented in this chapter:

Convexity at $\bar{\mathbf{x}}$. The function f is said to be convex at $\bar{\mathbf{x}} \in S$ if

$$f[\lambda\bar{\mathbf{x}} + (1-\lambda)\mathbf{x}] \leq \lambda f(\bar{\mathbf{x}}) + (1-\lambda) f(\mathbf{x})$$

for each $\lambda \in (0, 1)$ and each $\mathbf{x} \in S$.

Strict convexity at $\bar{\mathbf{x}}$. The function f is said to be strictly convex at $\bar{\mathbf{x}} \in S$ if

$$f[\lambda\bar{\mathbf{x}} + (1-\lambda)\mathbf{x}] < \lambda f(\bar{\mathbf{x}}) + (1-\lambda) f(\mathbf{x})$$

for each $\lambda \in (0, 1)$ and for each $\mathbf{x} \in S$, $\mathbf{x} \neq \bar{\mathbf{x}}$.

Quasiconvexity at $\bar{\mathbf{x}}$. The function f is said to be quasiconvex at $\bar{\mathbf{x}} \in S$ if

$$f[\lambda\bar{\mathbf{x}} + (1-\lambda)\mathbf{x}] \leq \max\{f(\mathbf{x}), f(\bar{\mathbf{x}})\}$$

for each $\lambda \in (0, 1)$ and each $\mathbf{x} \in S$.

Strict quasiconvexity at $\bar{\mathbf{x}}$. The function is said to be strictly quasiconvex at $\bar{\mathbf{x}} \in S$ if

$$f[\lambda\bar{\mathbf{x}} + (1-\lambda)\mathbf{x}] < \max\{f(\mathbf{x}), f(\bar{\mathbf{x}})\}$$

for each $\lambda \in (0, 1)$ and each $\mathbf{x} \in S$ such that $f(\mathbf{x}) \neq f(\bar{\mathbf{x}})$.

Strong quasiconvexity at $\bar{\mathbf{x}}$. The function f is said to be strongly quasiconvex at $\bar{\mathbf{x}} \in S$ if

$$f[\lambda\bar{\mathbf{x}} + (1-\lambda)\mathbf{x}] < \max\{f(\mathbf{x}), f(\bar{\mathbf{x}})\}$$

for each $\lambda \in (0,1)$ and each $\mathbf{x} \in S$, $\mathbf{x} \neq \bar{\mathbf{x}}$.

Pseudoconvexity at $\bar{\mathbf{x}}$. The function f is said to be pseudoconvex at $\bar{\mathbf{x}} \in S$ if $\nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) \geq 0$ for $\mathbf{x} \in S$ implies that $f(\mathbf{x}) \geq f(\bar{\mathbf{x}})$.

Strict pseudoconvexity at $\bar{\mathbf{x}}$. The function f is said to be strictly pseudoconvex at $\bar{\mathbf{x}} \in S$ if $\nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) \geq 0$ for $\mathbf{x} \in S$, $\mathbf{x} \neq \bar{\mathbf{x}}$, implies that $f(\mathbf{x}) > f(\bar{\mathbf{x}})$.

Various types of concavity at a point can be stated in a similar fashion. Figure 3.14 shows some types of convexity at a point. As the figure suggests, these types of convexity at a point represent a significant relaxation of the concept of convexity.

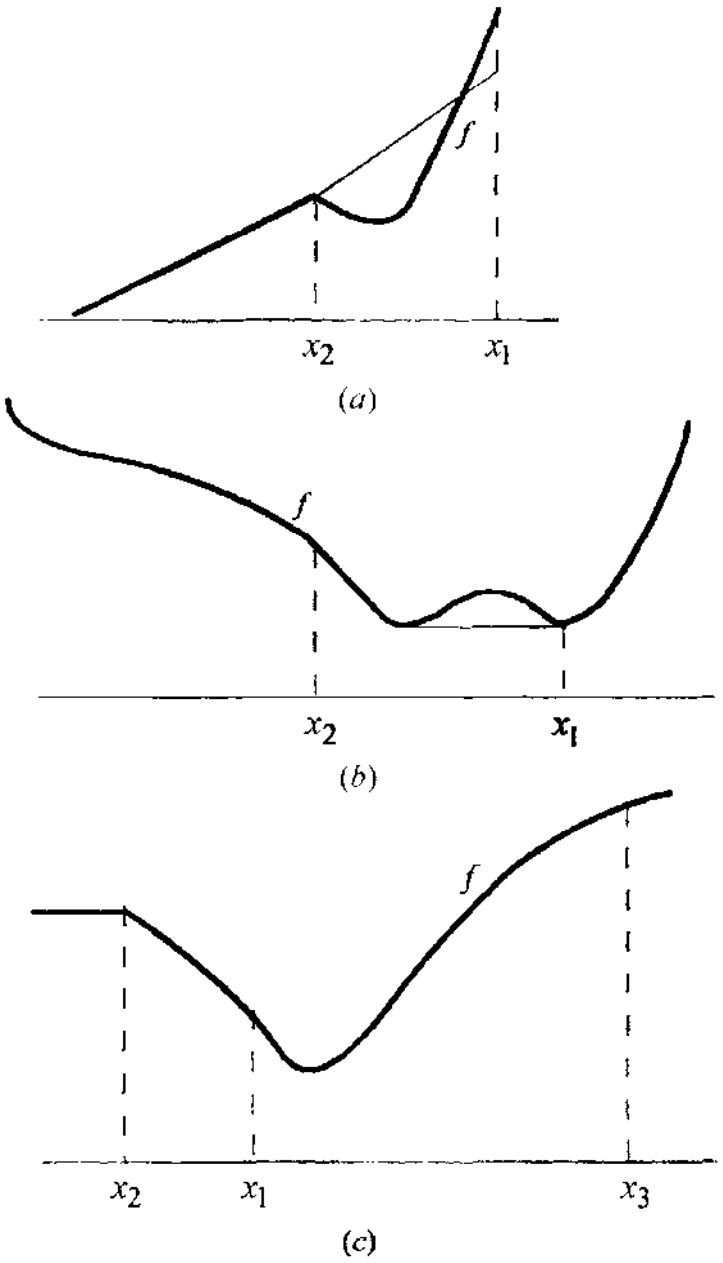


Figure 3.14 Various types of convexity at a point. (*a*) Convexity and strict convexity: f is convex but not strictly convex at x_1; f is both convex and strictly convex at x_2. (*b*) Pseudoconvexity and strict pseudoconvexity: f is pseudoconvex but not strictly pseudoconvex at x_1; f is both pseudoconvex and strictly pseudoconvex at x_2. (*c*) Quasiconvexity, strict quasiconvexity, and strong quasiconvexity: f is quasiconvex but neither strictly quasiconvex nor strongly quasiconvex at x_1; f is both quasiconvex and strictly quasiconvex at x_2 but not strongly quasiconvex at x_2; f is quasiconvex, strictly quasiconvex, and strongly quasiconvex at x_3.

We specify below some important results related to convexity of a function f at a point, where f: $S \to R$ and S is a nonempty convex set in R^n. Of course, not all the results developed throughout this chapter hold true. However, several of these results hold true and are summarized below. The proofs are similar to the corresponding theorems in this chapter.

1. Let f be both convex and differentiable at $\bar{\mathbf{x}}$. Then $f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}})$ for each $\mathbf{x} \in S$. If f is strictly convex, strict inequality holds for $\mathbf{x} \neq \bar{\mathbf{x}}$.
2. Let f be both convex and twice differentiable at $\bar{\mathbf{x}}$. Then the Hessian matrix $\mathbf{H}(\bar{\mathbf{x}})$ is positive semidefinite.
3. Let f be convex at $\bar{\mathbf{x}} \in S$, and let $\bar{\mathbf{x}}$ be an optimal solution to the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$. Then $\bar{\mathbf{x}}$ is a global optimal solution.
4. Let f be convex and differentiable at $\bar{\mathbf{x}} \in S$. Then $\bar{\mathbf{x}}$ is an optimal solution to the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$ if and only if $\nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) \geq 0$ for each $\mathbf{x} \in S$. In particular, if $\bar{\mathbf{x}} \in$ int S, $\bar{\mathbf{x}}$ is an optimal solution if and only if $\nabla f(\bar{\mathbf{x}}) = 0$.
5. Let f be convex and differentiable at $\bar{\mathbf{x}} \in S$. Suppose that $\bar{\mathbf{x}}$ is an optimal solution to the problem to maximize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$. Then $\nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) \leq 0$ for each $\mathbf{x} \in S$.
6. Let f be both quasiconvex and differentiable at $\bar{\mathbf{x}}$, and let $\mathbf{x} \in S$ be such that $f(\mathbf{x}) \leq f(\bar{\mathbf{x}})$. Then $\nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) \leq 0$.
7. Suppose that $\bar{\mathbf{x}}$ is a local optimal solution to the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$. If f is strictly quasiconvex at $\bar{\mathbf{x}}$, $\bar{\mathbf{x}}$ is a global optimal solution. If f is strongly quasiconvex at $\bar{\mathbf{x}}$, $\bar{\mathbf{x}}$ is the unique global optimal solution.
8. Consider the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$, and let $\bar{\mathbf{x}} \in S$ be such that $\nabla f(\bar{\mathbf{x}}) = 0$. If f is pseudoconvex at $\bar{\mathbf{x}}$, $\bar{\mathbf{x}}$ is a global optimal solution; and if f is strictly pseudoconvex at $\bar{\mathbf{x}}$, $\bar{\mathbf{x}}$ is the unique global optimal solution.

Exercises

[3.1] Which of the following functions is convex, concave, or neither? Why?

a. $f(x_1, x_2) = 2x_1^2 - 4x_1x_2 - 8x_1 + 3x_2$

b. $f(x_1, x_2) = x_1 e^{-(x_1 + 3x_2)}$

c. $f(x_1, x_2) = -x_1^2 - 3x_2^2 + 4x_1x_2 + 10x_1 - 10x_2$

d. $f(x_1, x_2, x_3) = 2x_1x_2 + 2x_1^2 + x_2^2 + 2x_3^2 - 5x_1x_3$

e. $f(x_1, x_2, x_3) = -2x_1^2 - 3x_2^2 - 2x_3^2 + 8x_1x_2 + 3x_1x_3 + 4x_2x_3$

[3.2] Over what subset of $\{x : x > 0\}$ is the univariate function $f(x) = e^{-ax^b}$ convex, where $a > 0$ and $b \geq 1$?

[3.3] Prove or disprove concavity of the following function defined over $S = \{(x_1, x_2) : -1 \leq x_1 \leq 1, -1 \leq x_2 \leq 1\}$:

$$f(x_1, x_2) = 10 - 3(x_2 - x_1^2)^2.$$

Repeat for a convex set $S \subseteq \{(x_1, x_2) : x_1^2 \geq x_2\}$.

[3.4] Over what domain is the function $f(x) = x^2(x^2 - 1)$ convex? Is it strictly convex over the region(s) specified? Justify your answer.

[3.5] Show that a function f: $R^n \to R$ is affine if and only if f is both convex and concave. [A function f is *affine* if it is of the form $f(\mathbf{x}) = \alpha + \mathbf{c}^t\mathbf{x}$, where a is a scalar and $\mathbf{c}$ is an n-vector.]

[3.6] Let S be a nonempty convex set in R^n, and let f: $S \to R$. Show that f is convex if and only if for any integer $k \geq 2$, the following holds true: $\mathbf{x}_1, \ldots, \mathbf{x}_k \in S$ implies that $f(\sum_{j=1}^k \lambda_j \mathbf{x}_j) \leq \sum_{j=1}^k \lambda_j f(\mathbf{x}_j)$, where $\sum_{j=1}^k \lambda_j = 1$, $\lambda_j \geq 0$ for $j = 1, \ldots, k$.

[3.7] Let S be a nonempty convex set in R^n, and let f: $S \to R$. Show that f is concave if and only if hyp f is convex.

[3.8] Let $f_1, f_2, \ldots, f_k$: $R^n \to R$ be convex functions. Consider the function f defined by $f(\mathbf{x}) = \sum_{j=1}^k \alpha_j f_j(\mathbf{x})$, where $\alpha_j > 0$ for $j = 1, 2, \ldots, k$. Show that f is convex. State and prove a similar result for concave functions.

[3.9] Let $f_1, f_2, \ldots, f_k$: $R^n \to R$ be convex functions. Consider the function f defined by $f(\mathbf{x}) = \max\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_k(\mathbf{x})\}$. Show that f is convex. State and prove a similar result for concave functions.

[3.10] Let h: $R^n \to R$ be a convex function, and let g: $R \to R$ be a nondecreasing convex function. Consider the composite function f: $R^n \to R$ defined by $f(\mathbf{x}) = g[h(\mathbf{x})]$. Show that f is convex.

[3.11] Let g: $R^n \to R$ be a concave function, and let f be defined by $f(\mathbf{x}) = 1/g(\mathbf{x})$. Show that f is convex over $S = \{\mathbf{x} : g(\mathbf{x}) > 0\}$. State a symmetric result interchanging the convex and concave functions.

[3.12] Let S be a nonempty convex set in R^n, and let f: $R^n \to R$ be defined as follows:

$$f(\mathbf{y}) = \inf\{\|\mathbf{y} - \mathbf{x}\| : \mathbf{x} \in S\}.$$

Note that $f(\mathbf{y})$ gives the distance from $\mathbf{y}$ to the set S and is called the *distance function*. Prove that f is convex.

[3.13] Let $S = \{(x_1, x_2) : x_1^2 + x_2^2 \leq 4\}$. Let f be the distance function defined in Exercise 3.12. Find the function f explicitly.

[3.14] Let S be a nonempty, bounded convex set in R^n, and let f: $R^n \to R$ be defined as follows:

$$f(\mathbf{y}) = \sup\{\mathbf{y}^t \mathbf{x} : \mathbf{x} \in S\}.$$

The function f is called the *support function* of S. Prove that f is convex. Also, show that if $f(\mathbf{y}) = \mathbf{y}^t \bar{\mathbf{x}}$, where $\bar{\mathbf{x}} \in S$, $\bar{\mathbf{x}}$ is a subgradient of f at $\mathbf{y}$.

[3.15] Let $S = A \cup B$, where

$$A = \{(x_1, x_2) : x_1 < 0, x_1^2 + x_2^2 \leq 4\}$$

$$B = \{(x_1, x_2) : x_1 \geq 0, -2 \leq x_2 \leq 2\}.$$

Find the support function defined in Exercise 3.14 explicitly.

[3.16] Let g: $R^m \to R$ be a convex function, and let $\mathbf{h} : R^n \to R^m$ be an affine function of the form $\mathbf{h}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$, where $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{b}$ is an $m \times 1$ vector. Then show that the composite function f: $R^n \to R$ defined as $f(\mathbf{x}) = g[\mathbf{h}(\mathbf{x})]$ is a convex function. Also, assuming twice differentiability of g, derive an expression for the Hessian of f.

[3.17] Let F be a *cumulative distribution function* for a random variable b, that is, $F(y) = \text{Prob}(b \leq y)$. Show that $\phi(z) = \int_{-\infty}^{z} F(y)\, dy$ is a convex function. Is ϕ convex for any nondecreasing function F?

[3.18] A function f: $R^n \to R$ is called a *gauge function* if it satisfies the following equality:

$$f(\lambda \mathbf{x}) = \lambda f(\mathbf{x}) \qquad \text{for all } \mathbf{x} \in R^n \text{ and all } \lambda \geq 0.$$

Further, a gauge function is said to be *subadditive* if it satisfies the following inequality:

$$f(\mathbf{x}) + f(\mathbf{y}) \geq f(\mathbf{x} + \mathbf{y}) \qquad \text{for all } \mathbf{x}, \mathbf{y} \in R^n.$$

Prove that subadditivity is equivalent to convexity of gauge functions.

[3.19] Let f: $S \to R$ be defined as

$$f(\mathbf{x}) = \frac{(\boldsymbol{\alpha}^t \mathbf{x})^2}{\boldsymbol{\beta}^t \mathbf{x}},$$

where S is a convex subset of R^n, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors in R^n, and where $\boldsymbol{\beta}^t \mathbf{x} > 0$ for all $\mathbf{x} \in S$. Derive an explicit expression for the Hessian of f, and hence verify that f is convex over S.

[3.20] Consider a quadratic function f: $R^n \to R$ and suppose that f is convex on S, where S is a nonempty convex set in R^n. Show that:

a. The function f is convex on $M(S)$, where $M(S)$ is the *affine manifold* containing S defined by $M(S) = \{\mathbf{y} : \mathbf{y} = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j,\ \sum_{j=1}^{k} \lambda_j = 1,\ \mathbf{x}_j \in S \text{ for all } j, \text{ for } k \geq 1\}$.

b. The function f is convex on $L(S)$, the *linear subspace* parallel to $M(S)$, defined by $L(S) = \{\mathbf{y} - \mathbf{x} : \mathbf{y} \in M(S) \text{ and } \mathbf{x} \in S\}$. (This result is credited to Cottle [1967].)

[3.21] Let h: $R^n \to R$ be convex, and let $\mathbf{A}$ be an $m \times n$ matrix. Consider the function h: $R^m \to R$ defined as follows:

$$h(\mathbf{y}) = \inf\{f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{y}\}.$$

Show that h is convex.

[3.22] Let S be a nonempty convex set in R^n, and let f: $R^n \to R$ and $\mathbf{g}$: $R^n \to R^m$ be convex. Consider the *perturbation function* $\phi : R^m \to R$ defined below:

$$\phi(\mathbf{y}) = \inf\{f(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \leq \mathbf{y}, \mathbf{x} \in S\}.$$

a. Prove that ϕ is convex.
b. Show that if $\mathbf{y}_1 \leq \mathbf{y}_2$, $\phi(\mathbf{y}_1) \geq \phi(\mathbf{y}_2)$.

[3.23] Let f: $R^n \to R$ be lower semicontinuous. Show that the level set $S_\alpha = \{\mathbf{x} : f(\mathbf{x}) \leq \alpha\}$ is closed for all $\alpha \in R$.

[3.24] Let f be a convex function on R^n. Prove that the set of subgradients of f at a given point forms a closed convex set.

[3.25] Let f: $R^n \to R$ be convex. Show that $\boldsymbol{\xi}$ is a subgradient of f at $\bar{\mathbf{x}}$ if and only if the hyperplane $\{(\mathbf{x}, y) : y = f(\bar{\mathbf{x}}) + \boldsymbol{\xi}^t(\mathbf{x} - \bar{\mathbf{x}})\}$ supports epi f at $[\bar{\mathbf{x}}, f(\bar{\mathbf{x}})]$. State and prove a similar result for concave functions.

[3.26] Let f: $R^n \to R$ be defined by $f(\mathbf{x}) = \|\mathbf{x}\|$. Prove that subgradients of f are characterized as follows: If $\mathbf{x} = \mathbf{0}$, $\boldsymbol{\xi}$ is a subgradient of f at $\mathbf{x}$ if and only if

$\|\xi\| \leq 1$. On the other hand, if $\mathbf{x} \neq \mathbf{0}$, ξ is a subgradient of f at $\mathbf{x}$ if and only if $\|\xi\| = 1$ and $\xi^t \mathbf{x} = \|\mathbf{x}\|$. Use this result to show that f is differentiable at each $\mathbf{x} \neq \mathbf{0}$, and characterize the gradient vector.

[3.27] Let f_1, f_2: $R^n \to R$ be differentiable convex functions. Consider the function f defined by $f(\mathbf{x}) = \max\{f_1(\mathbf{x}), f_2(\mathbf{x})\}$. Let $\bar{\mathbf{x}}$ be such that $f(\bar{\mathbf{x}}) = f_1(\bar{\mathbf{x}}) = f_2(\bar{\mathbf{x}})$. Show that ξ is a subgradient of f at $\bar{\mathbf{x}}$ if and only if

$$\xi = \lambda \nabla f_1(\bar{\mathbf{x}}) + (1-\lambda)\nabla f_2(\bar{\mathbf{x}}), \qquad \text{where } \lambda \in [0, 1].$$

Generalize the result to several convex functions and state a similar result for concave functions.

[3.28] Consider the function θ defined by the following optimization problem for any $\mathbf{u} \geq \mathbf{0}$, where X is a compact polyhedral set.

$$\begin{aligned} \theta(\mathbf{u}) = \text{Minimize } & \mathbf{c}^t \mathbf{x} + \mathbf{u}^t (\mathbf{A}\mathbf{x} - \mathbf{b}) \\ \text{subject to } & \mathbf{x} \in X. \end{aligned}$$

a. Show that θ is concave.
b. Characterize the subgradients of θ at any given $\mathbf{u}$.

[3.29] In reference to Exercise 3.28, find the function θ explicitly and describe the set of subgradients at each point $\mathbf{u} \geq \mathbf{0}$ if

$$\mathbf{A} = \begin{bmatrix} 3 & 2 \\ -1 & 2 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 6 \\ 4 \end{bmatrix}, \qquad \mathbf{c} = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$$

$$X = \{(x_1, x_2) : 0 \leq x_1 \leq 3/2,\ 0 \leq x_2 \leq 3/2\}.$$

[3.30] Consider the function θ defined by the following optimization problem:

$$\begin{aligned} \theta(u_1, u_2) = \text{Minimize } & x_1(2-u_1) + x_2(3-u_2) \\ \text{subject to } & x_1^2 + x_2^2 \leq 4. \end{aligned}$$

a. Show that θ is concave.
b. Evaluate θ at the point (2, 3).
c. Find the collection of subgradients of θ at (2, 3).

[3.31] Let f: $S \to R$, where $S \subseteq R^n$ is a nonempty convex set. Then the *convex envelope* of f over S, denoted $f_S(\mathbf{x})$, $\mathbf{x} \in S$, is a convex function such that $f_S(\mathbf{x}) \leq f(\mathbf{x})$ for all $\mathbf{x} \in S$; and if g is any other convex function for which $g(\mathbf{x}) \leq f(\mathbf{x})$ for all $\mathbf{x} \in S$, $f_S(\mathbf{x}) \geq g(\mathbf{x})$ for all $\mathbf{x} \in S$. Hence f_S is the pointwise supremum over all convex underestimators of f over S. Show that $\min\{f(\mathbf{x}) : \mathbf{x} \in S\} = \min\{f_S(\mathbf{x}) : \mathbf{x} \in S\}$, assuming that the minima exist, and that

$$\{\mathbf{x}^* \in S : f(\mathbf{x}^*) \le f(\mathbf{x}) \text{ for all } \mathbf{x} \in S\}$$

$$\subseteq \{\mathbf{x}^* \in S : f_S(\mathbf{x}^*) \le f_S(\mathbf{x}) \text{ for all } \mathbf{x} \in S\}.$$

[3.32] Let f: $S \to R$ be a concave function, where $S \subseteq R^n$ is a nonempty polytope with vertices $\mathbf{x}_1, ..., \mathbf{x}_E$. Show that the convex envelope (see Exercise 3.31) of f over S is given by

$$f_S(\mathbf{x}) = \min\left\{\sum_{i=1}^{E} \lambda_i f(\mathbf{x}_i) : \sum_{i=1}^{E} \lambda_i \mathbf{x}_i = \mathbf{x}, \sum_{i=1}^{E} \lambda_i = 1, \lambda_i \ge 0 \text{ for } i = 1, ..., E\right\}.$$

Hence, show that if S is a simplex in R^n, f_S is an affine function that attains the same values as f over all the vertices of S. (This result is due to Falk and Hoffman [1976].)

[3.33] Let f: $S \to R$ and f_S : $S \to R$ be as defined in Exercise 3.31. Show that if f is continuous, the epigraph $\{(\mathbf{x}, y) : y \ge f_S(\mathbf{x}),\ \mathbf{x} \in S,\ y \in R\}$ of f_S over S is the closure of the convex hull of the epigraph $\{(\mathbf{x}, y) : y \ge f(\mathbf{x}),\ \mathbf{x} \in S,\ y \in R\}$ of f over S. Give an example to show that the epigraph of the latter set is not necessarily closed.

[3.34] Let $f(x, y) = xy$ be a bivariate bilinear function, and let S be a polytope in R^2 having no edge with a finite, positive slope. Define $\Lambda = \{(\alpha, \beta, \gamma) \in R^3 : \alpha x_k + \beta y_k + \gamma \le x_k y_k \text{ for } k = 1, ..., K\}$, where (x_k, y_k), $k = 1, ..., K$, are the vertices of S. Referring to Exercise 3.31, show that if S is two-dimensional, the set of extreme points $(\alpha_e, \beta_e, \gamma_e)$, $e = 1, ..., E$, of Λ is nonempty and that $f_S(x, y) = \max\{\alpha_e x + \beta_e y + \gamma_e, e = 1, ..., E\}$. On the other hand, if S is one-dimensional and given by the convex hull of (x_1, y_1) and (x_2, y_2), show that there exists a solution $(\alpha_1, \beta_1, \gamma_1)$ to the system $\alpha x_k + \beta y_k + \gamma = x_k y_k$ for $k = 1, 2$, and in this case, $f_S(x, y) = \alpha_1 x + \beta_1 y + \gamma_1$. Specialize this result to verify that if $S = \{(x, y) : a \le x \le b,\ c \le y \le d\}$, where $a < b$ and $c < d$, then $f_S(x, y) = \max\{dx + by - bd,\ cx + ay - ac\}$. (This result is due to Sherali and Alameddine [1990].)

[3.35] Consider a triangle S having vertices (0, 1), (2, 0), and (1, 2) and let $f(x, y) = xy$ be a bivariate, bilinear function. Show that the convex envelope f_S of f over S (see Exercise 3.31) is given by

$$f_S(x, y) = \begin{cases} -y + \dfrac{3y^2}{2 - x + y} & \text{for } (x, y) \ne (2, 0) \\ 0 & \text{for } (x, y) = (2, 0) \end{cases} \qquad \text{for } (x, y) \in S.$$

Can you generalize your approach to finding the convex envelope of f over a triangle having a single edge that has a finite, positive slope? (This result is due to Sherali and Alameddine [1990].)

[3.36] Let f: $R^n \to R$ be a differentiable function. Show that the gradient vector is given by

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_n} \right)^t.$$

[3.37] Let f: $R^n \to R$, be a differentiable function. The *linear approximation* of f at a given point $\bar{\mathbf{x}}$ is given by

$$f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t (\mathbf{x} - \bar{\mathbf{x}}).$$

If f is twice differentiable at $\bar{\mathbf{x}}$, the *quadratic approximation* of f at $\bar{\mathbf{x}}$ is given by

$$f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t (\mathbf{x} - \bar{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^t \mathbf{H}(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}).$$

Let $f(x_1, x_2) = e^{2x_1^2 - x_2^2} - 3x_1 + 5x_2$. Give the linear and quadratic approximations of f at (1, 1). Are these approximations convex, concave, or neither? Why?

[3.38] Consider the function f: $R^n \to R$, and suppose that f is infinitely differentiable. Then show that f is strictly convex if and only if for each $\bar{\mathbf{x}}$ and $\mathbf{d}$ in R^n, the first nonzero derivative term of order greater than or equal to 2 in the Taylor series expansion exists, is of even order, and is positive.

[3.39] Consider the function f: $R^3 \to R$, given by $f(\mathbf{x}) = \mathbf{x}^t \mathbf{A} \mathbf{x}$, where

$$\mathbf{A} = \begin{bmatrix} 2 & 2 & 3 \\ 1 & 3 & 1 \\ 1 & 2 & \theta \end{bmatrix}.$$

What is the Hessian of f? For what values of θ is f strictly convex?

[3.40] Consider the function $f(x) = x^3$, defined over the set $S = \{x \in R : x \geq 0\}$. Show that f is strictly convex over S. Noting that $f''(0) = 0$ and $f'''(0) = 6$, comment on the application of Theorem 3.3.9.

[3.41] Let $\mathbf{H}$ be an $n \times n$ symmetric, positive semidefinite matrix, and suppose that $\mathbf{x}^t \mathbf{H} \mathbf{x} = 0$ for some $\mathbf{x} \in R^n$. Then show that $\mathbf{H}\mathbf{x} = \mathbf{0}$. (*Hint*: Consider the diagonal of the quadratic form $\mathbf{x}^t \mathbf{H} \mathbf{x}$ via the transformation $\mathbf{x} = \mathbf{Q}\mathbf{y}$, where the columns of $\mathbf{Q}$ are the normalized eigenvectors of $\mathbf{H}$.)

[3.42] Let $\mathbf{H}$ be an $n \times n$ symmetric matrix. Using the eigenvalue characterization of definiteness, verify that $\mathbf{H}$ is positive definite if and only if it is positive semidefinite and nonsingular.

[3.43] Suppose that $\mathbf{H}$ is an $n \times n$ symmetric matrix. Show how Theorem 3.3.12 demonstrates that $\mathbf{H}$ is positive definite if and only if it can be premultiplied by a series of n lower triangular Gauss–Jordan reduction matrices $\mathbf{L}_1, \ldots, \mathbf{L}_n$ to yield an upper triangular matrix $\mathbf{U}$ with positive diagonal elements. (Letting $\mathbf{L}^{-1} = \mathbf{L}_n \cdots \mathbf{L}_1$, we obtain $\mathbf{H} = \mathbf{LU}$, where $\mathbf{L}$ is lower triangular. This is known as the *LU-decomposition* of $\mathbf{H}$; see Appendix A.2.) Furthermore, show that $\mathbf{H}$ is positive definite if and only if there exists a lower triangular matrix $\mathbf{L}$ with positive diagonal elements such that $\mathbf{H} = \mathbf{LL}^t$. (This is known as the *Cholesky factorization* of $\mathbf{H}$; see Appendix A.2.)

[3.44] Suppose that $S \neq \varnothing$ is closed and convex. Let f: $S \to R$ be differentiable on int S. State if the following are true or false, justifying your answer:

a. If f is convex on S, $f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}})$ for all $\mathbf{x} \in S$, $\bar{\mathbf{x}} \in \text{int } S$.

b. If $f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}})$ for all $\mathbf{x} \in S$ and $\bar{\mathbf{x}} \in \text{int } S$, f is convex on S.

[3.45] Consider the following problem:

$$\begin{aligned} &\text{Minimize } (x_1 - 4)^2 + (x_2 - 6)^2 \\ &\text{subject to } x_2 \geq x_1^2 \\ &\qquad\qquad x_2 \leq 4. \end{aligned}$$

Write a necessary condition for optimality and verify that it is satisfied by the point (2, 4). Is this the optimal point? Why?

[3.46] Use Theorem 3.4.3 to prove that every local minimum of a convex function over a convex set is also a global minimum.

[3.47] Consider the problem to minimize $\{f(\mathbf{x}) : \mathbf{x} \in S\}$ and suppose that there exists an $\varepsilon > 0$ such that $N_\varepsilon(\bar{\mathbf{x}}) \cap S$ is a convex set and that $f(\bar{\mathbf{x}}) \leq f(\mathbf{x})$ for all $\mathbf{x} \in N_\varepsilon(\bar{\mathbf{x}}) \cap S$.

a. Show that if $\mathbf{H}(\bar{\mathbf{x}})$ is positive definite, $\bar{\mathbf{x}}$ is both a strict and a strong local minimum.

b. Show that if $\bar{\mathbf{x}}$ is a strict local minimum and f is pseudoconvex on $N_\varepsilon(\bar{\mathbf{x}}) \cap S$, $\bar{\mathbf{x}}$ is also a strong local minimum.

[3.48] Let f: $R^n \to R$ be a convex function, and suppose that $f(\mathbf{x} + \lambda\mathbf{d}) \geq f(\mathbf{x})$ for all $\lambda \in (0, \delta)$, where $\delta > 0$. Show that $f(\mathbf{x} + \lambda\mathbf{d})$ is a nondecreasing function of λ. In particular, show that $f(\mathbf{x} + \lambda\mathbf{d})$ is a strictly increasing function of λ if f is strictly convex.

[3.49] Consider the following problem:

$$\begin{aligned} \text{Maximize } & \mathbf{c}^t\mathbf{x} + \frac{1}{2}\mathbf{x}^t\mathbf{H}\mathbf{x} \\ \text{subject to } & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0}, \end{aligned}$$

where $\mathbf{H}$ is a symmetric negative definite matrix, $\mathbf{A}$ is an $m \times n$ matrix, $\mathbf{c}$ is an n-vector, and $\mathbf{b}$ is an m-vector. Write the necessary and sufficient condition for optimality of Theorem 3.4.3, and simplify it using the special structure of this problem.

[3.50] Consider the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$, where f: $R^n \to R$ is a differentiable convex function and S is a nonempty convex set in R^n. Prove that $\bar{\mathbf{x}}$ is an optimal solution if and only if $\nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) \geq 0$ for each $\mathbf{x} \in S$. State and prove a similar result for the maximization of a concave function. (This result was proved in the text as Corollary 2 to Theorem 3.4.3. In this exercise the reader is asked to give a direct proof without resorting to subgradients.)

[3.51] A vector $\mathbf{d}$ is called a *direction of descent* of f at $\bar{\mathbf{x}}$ if there exists a $\delta > 0$ such that $f(\bar{\mathbf{x}} + \lambda\mathbf{d}) < f(\bar{\mathbf{x}})$ for each $\lambda \in (0, \delta)$. Suppose that f is convex. Show that $\mathbf{d}$ is a direction of descent if and only if $f'(\bar{\mathbf{x}}; \mathbf{d}) < 0$. Does the result hold true without the convexity of f?

[3.52] Consider the following problem:

$$\begin{aligned} \text{Maximize } & f(\mathbf{x}) \\ \text{subject to } & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0}, \end{aligned}$$

where $\mathbf{A}$ is an $m \times n$ matrix with rank m and f is a differentiable convex function. Consider the extreme point $(\mathbf{x}_B^t, \mathbf{x}_N^t) = (\bar{\mathbf{b}}^t, \mathbf{0}^t)$, where $\bar{\mathbf{b}} = \mathbf{B}^{-1}\mathbf{b} \geq \mathbf{0}$ and $\mathbf{A} = [\mathbf{B}, \mathbf{N}]$. Decompose $\nabla f(\mathbf{x})$ accordingly into $\nabla_B f(\mathbf{x})$ and $\nabla_N f(\mathbf{x})$. Show that the necessary condition of Theorem 3.4.6 holds true if $\nabla_N f(\mathbf{x})^t - \nabla_B f(\mathbf{x})^t \mathbf{B}^{-1}\mathbf{N} \leq \mathbf{0}$. If this condition holds, is it necessarily true that $\mathbf{x}$ is a local maximum? Prove or give a counterexample.

If $\nabla_N f(\mathbf{x})^t - \nabla_B f(\mathbf{x})^t \mathbf{B}^{-1}\mathbf{N} \leq \mathbf{0}$, choose a positive component j and increase its corresponding nonbasic variable x_j until a new extreme point is reached. Show that this process results in a new extreme point having a larger objective value. Does this method guarantee convergence to a global optimal solution? Prove or give a counterexample.

[3.53] Apply the procedure of Exercise 3.52 to the following problem starting with the extreme point (1/2, 3, 0, 0):

$$\begin{aligned} \text{Maximize } & (x_1 - 2)^2 + (x_2 - 5)^2 \\ \text{subject to } & -2x_1 + x_2 + x_3 = 2 \\ & 2x_1 + 3x_2 + x_4 = 10 \\ & x_1, \quad x_2, \quad x_3, \quad x_4 \geq 0. \end{aligned}$$

[3.54] Consider the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$, where f: $R^n \to R$ is convex and S is a nonempty convex set in R^n. The cone of feasible directions of S at $\mathbf{x} \in S$ is defined by

$$D = \{\mathbf{d} : \text{there exists a } \delta > 0 \text{ such that } \bar{\mathbf{x}} + \lambda \mathbf{d} \in S \text{ for } \lambda \in (0, \delta)\}.$$

Show that $\bar{\mathbf{x}}$ is an optimal solution to the problem if and only if $f'(\bar{\mathbf{x}}; \mathbf{d}) \geq 0$ for each $\mathbf{d} \in D$. Compare this result with the necessary and sufficient condition of Theorem 3.4.3. Specialize the result to the case where $S = R^n$.

[3.55] Let f: $R^n \to R$ be a quadratic function. Show that f is quasiconvex if and only if it is strictly quasiconvex, which holds true if and only if it is pseudoconvex. Furthermore, show that f is strongly quasiconvex if and only if it is strictly pseudoconvex.

[3.56] Let h: $R^n \to R$ be a quasiconvex function, and let g: $R \to R$ be a nondecreasing function. Then show that the composite function f: $R^n \to R$ defined as $f(\mathbf{x}) = g[h(\mathbf{x})]$ is quasiconvex.

[3.57] Let f: $S \subseteq R \to R$ be a univariate function, where S is some interval on the real line. Define f as *unimodal* on S if there exists an $x^* \in S$ at which f attains a minimum and f is nondecreasing on the interval $\{x \in S : x \geq x^*\}$, whereas it is nonincreasing on the interval $\{x \in S : x \leq x^*\}$. Assuming that f attains a minimum on S, show that f is quasiconvex if and only if it is unimodal on S.

[3.58] Let f: $S \to R$ be a continuous function, where S is a convex subset of R^n. Show that f is quasimonotone if and only if the level surface $\{\mathbf{x} \in S : f(\mathbf{x}) = \alpha\}$ is a convex set for all $\alpha \in R$.

[3.59] Let f: $S \to R$ be a differentiable function, where S is an open, convex subset of R^n. Show that f is quasimonotone if and only if for every $\mathbf{x}_1$ and $\mathbf{x}_2$ in S, $f(\mathbf{x}_1) \geq f(\mathbf{x}_2)$ implies that $\nabla f(\mathbf{x}_2)^t(\mathbf{x}_1 - \mathbf{x}_2) \geq 0$ and $f(\mathbf{x}_1) \leq f(\mathbf{x}_2)$ implies that $\nabla f(\mathbf{x}_2)^t(\mathbf{x}_1 - \mathbf{x}_2) \leq 0$. Hence, show that f is quasimonotone if and only if $f(\mathbf{x}_1) \geq f(\mathbf{x}_2)$ implies that $\nabla f(\mathbf{x}_\lambda)^t(\mathbf{x}_1 - \mathbf{x}_2) \geq 0$ for all $\mathbf{x}_1$ and $\mathbf{x}_2$ in S and for all $\mathbf{x}_\lambda = \lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2$, where $0 \leq \lambda \leq 1$.

[3.60] Let f: $S \to R$, where f is lower semicontinuous and where S is a convex subset of R^n. Define f as being *strongly unimodal* on S if for each $\mathbf{x}_1$ and $\mathbf{x}_2$ in

S for which the function $F(\lambda) = f[\mathbf{x}_1 + \lambda(\mathbf{x}_2 - \mathbf{x}_1)]$, $0 \le \lambda \le 1$, attains a minimum at a point $\lambda^* > 0$, we have that $F(0) > F(\lambda) > F(\lambda^*)$ for all $0 < \lambda < \lambda^*$. Show that f is strongly quasiconvex on S if and only if it is strongly unimodal on S (see Exercise 8.10).

[3.61] Let g: $S \to R$ and h: $S \to R$, where S is a nonempty convex set in R^n. Consider the function f: $S \to R$ defined by $f(\mathbf{x}) = g(\mathbf{x})/h(\mathbf{x})$. Show that f is quasiconvex if the following two conditions hold true:

a. g is convex on S, and $g(\mathbf{x}) \ge 0$ for each $\mathbf{x} \in S$.

b. h is concave on S, and $h(\mathbf{x}) > 0$ for each $\mathbf{x} \in S$.

(*Hint*: Use Theorem 3.5.2.)

[3.62] Show that the function f defined in Exercise 3.61 is quasiconvex if the following two conditions hold true:

a. g is convex on S, and $g(\mathbf{x}) \le 0$ for each $\mathbf{x} \in S$.

b. h is convex on S, and $h(\mathbf{x}) > 0$ for each $\mathbf{x} \in S$.

[3.63] Let g: $S \to R$ and h: $S \to R$, where S is a nonempty convex set in R^n. Consider the function f: $S \to R$ defined by $f(\mathbf{x}) = g(\mathbf{x})h(\mathbf{x})$. Show that f is quasiconvex if the following two conditions hold true:

a. g is convex, and $g(\mathbf{x}) \le 0$ for each $\mathbf{x} \in S$.

b. h is concave, and $h(\mathbf{x}) > 0$ for each $\mathbf{x} \in S$.

[3.64] In each of Exercises 3.61, 3.62, and 3.63, show that f is pseudoconvex provided that S is open and that g and h are differentiable.

[3.65] Let $\mathbf{c}_1$, $\mathbf{c}_2$ be nonzero vectors in R^n, and α_1, α_2 be scalars. Let $S = \{\mathbf{x} : \mathbf{c}_2^t\mathbf{x} + \alpha_2 > 0\}$. Consider the function $f : S \to R$ defined as follows:

$$f(\mathbf{x}) = \frac{\mathbf{c}_1^t\mathbf{x} + \alpha_1}{\mathbf{c}_2^t\mathbf{x} + \alpha_2}.$$

Show that f is both pseudoconvex and pseudoconcave. (Functions that are both pseudoconvex and pseudoconcave are called *pseudolinear*.)

[3.66] Consider the quadratic function f: $R^n \to R$ defined by $f(\mathbf{x}) = \mathbf{x}^t\mathbf{H}\mathbf{x}$. The function f is said to be *positive subdefinite* if $\mathbf{x}^t\mathbf{H}\mathbf{x} < 0$ implies that $\mathbf{H}\mathbf{x} \ge \mathbf{0}$ or $\mathbf{H}\mathbf{x} \le \mathbf{0}$ for each $\mathbf{x} \in R^n$. Prove that f is quasiconvex on the *nonnegative orthant*, $R_+^n = \{\mathbf{x} \in R^n : \mathbf{x} \ge \mathbf{0}\}$ if and only if it is positive subdefinite. (This result is credited to Martos [1969].)

[3.67] The function f defined in Exercise 3.66 is said to be *strictly positive subdefinite* if $\mathbf{x}^t\mathbf{H}\mathbf{x} < 0$ implies that $\mathbf{H}\mathbf{x} > \mathbf{0}$ or $\mathbf{H}\mathbf{x} < \mathbf{0}$ for each $\mathbf{x} \in R^n$. Prove that f is pseudoconvex on the nonnegative orthant excluding $\mathbf{x} = \mathbf{0}$ if and only if it is strictly positive subdefinite. (This result is credited to Martos [1969].)

[3.68] Let f: $S \to R$ be a continuously differentiable convex function, where S is some open interval in R. Then show that f is (strictly) pseudoconvex if and only if whenever $f'(\bar{x}) = 0$ for any $\bar{x} \in S$, this implies that $\bar{x}$ is a (strict) local minimum of f on S. Generalize this result to the multivariate case.

[3.69] Let f: $S \to R$ be pseudoconvex, and suppose that for some $\mathbf{x}_1$ and $\mathbf{x}_2$ in R^n, we have $\nabla f(\mathbf{x}_1)'(\mathbf{x}_2 - \mathbf{x}_1) \geq 0$. Show that the function $F(\lambda) = f[\mathbf{x}_1 + \lambda(\mathbf{x}_2 - \mathbf{x}_1)]$ is nondecreasing for $\lambda \geq 0$.

[3.70] Let f: $S \to R$ be a twice differentiable univariate function, where S is some open interval in R. Then show that f is (strictly) pseudoconvex if and only if whenever $f'(\bar{x}) = 0$ for any $\bar{x} \in S$, we have that either $f''(\bar{x}) > 0$ or that $f''(\bar{x}) = 0$ and $\bar{x}$ is a (strict) local minimum of f over S. Generalize this result to the multivariate case.

[3.71] Let **f**: $R^n \to R^m$ and **g**: $R^n \to R^k$ be differentiable and convex. Let ϕ: $R^{m+k} \to R$ satisfy the following: If $\mathbf{a}_2 \geq \mathbf{a}_1$ and $\mathbf{b}_2 \geq \mathbf{b}_1$, $\phi(\mathbf{a}_2, \mathbf{b}_2) \geq \phi(\mathbf{a}_1, \mathbf{b}_1)$. Consider the function h: $R^n \to R$ defined by $h(\mathbf{x}) = \phi(\mathbf{f}(\mathbf{x}), \mathbf{g}(\mathbf{x}))$. Show the following:

a. If ϕ is convex, h is convex.
b. If ϕ is pseudoconvex, h is pseudoconvex.
c. If ϕ is quasiconvex, h is quasiconvex.

[3.72] Let g_1, g_2: $R^n \to R$, and let $\alpha \in [0,1]$. Consider the function G_α: $R^n \to R$ defined as

$$G_\alpha(\mathbf{x}) = \frac{1}{2}\left[g_1(\mathbf{x}) + g_2(\mathbf{x}) - \sqrt{g_1^2(\mathbf{x}) + g_2^2(\mathbf{x}) - 2\alpha g_1(\mathbf{x}) g_2(\mathbf{x})} \right],$$

where $\sqrt{\ }$ denotes the positive square root.

a. Show that $G_\alpha(\mathbf{x}) \geq 0$ if and only if $g_1(\mathbf{x}) \geq 0$ and $g_2(\mathbf{x}) \geq 0$, that is, minimum $\{g_1(\mathbf{x}), g_2(\mathbf{x})\} \geq 0$.
b. If g_1 and g_2 are differentiable, show that G_α is differentiable at $\mathbf{x}$ for each $\alpha \in [0,1)$ provided that $g_1(\mathbf{x})$, $g_2(\mathbf{x}) \neq 0$.
c. Now suppose that g_1 and g_2 are concave. Show that G_α is concave for α in the interval [0, 1]. Does this result hold true for $\alpha \in (-1,0)$?
d. Suppose that g_1 and g_2 are quasiconcave. Show that G_α is quasiconcave for $\alpha = 1$.
e. Let $g_1(\mathbf{x}) = -x_1^2 - x_2^2 + 4$ and $g_2(\mathbf{x}) = 2x_1 + x_2 - 1$. Obtain an explicit expression for G_α, and verify parts a, b, and c.

This exercise describes a general method for combining two constraints of the form $g_1(\mathbf{x}) \geq 0$ and $g_2(\mathbf{x}) \geq 0$ into an equivalent single constraint of the

form $G_\alpha(\mathbf{x}) \geq 0$. This procedure could be applied successively to reduce a problem with several constraints into an equivalent single constrained problem. The procedure is due to Rvačev [1963].

[3.73] Let g_1, g_2: $R^n \to R$, and let $\alpha \in [0,1]$. Consider the function G_α: $R^n \to R$ defined by

$$G_\alpha(\mathbf{x}) = \frac{1}{2}\left[g_1(\mathbf{x}) + g_2(\mathbf{x}) + \sqrt{g_1^2(\mathbf{x}) + g_2^2(\mathbf{x}) - 2\alpha g_1(\mathbf{x}) g_2(\mathbf{x})} \right],$$

where $\sqrt{\ }$ denotes the positive square root.

a. Show that $G_\alpha(\mathbf{x}) \geq 0$ if and only if maximum $\{g_1(\mathbf{x}), g_2(\mathbf{x})\} \geq 0$.

b. If g_1 and g_2 are differentiable, show that G_α is differentiable at x for each $\alpha \in [0,1)$, provided that $g_1(\mathbf{x})$, $g_2(\mathbf{x}) \neq 0$.

c. Now suppose that g_1 and g_2 are convex. Show that G_α is convex for $\alpha \in [0,1]$. Does the result hold true for $\alpha \in (-1,0)$?

d. Suppose that g_1 and g_2 are quasiconvex. Show that G_α is quasiconvex for $\alpha = 1$.

e. In some optimization problems, the restriction that the variable $x = 0$ or 1 arises. Show that this restriction is equivalent to maximum $\{g_1(x), g_2(x)\} \geq 0$, where $g_1(x) = -x^2$ and $g_2(x) = -(x-1)^2$. Find the function G_α explicitly, and verify statements a, b, and c.

This exercise describes a general method for combining the *either-or constraints* of the form $g_1(\mathbf{x}) \geq 0$ or $g_2(\mathbf{x}) \geq 0$ into a single constraint of the form $G_\alpha(\mathbf{x}) \geq 0$, and is due to Rvačev [1963].

Notes and References

In this chapter we deal with the important topic of convex and concave functions. The recognition of these functions is generally traced to Jensen [1905, 1906]. For earlier related works on the subject, see Hadamard [1893] and Hölder [1889].

In Section 3.1, several results related to continuity and directional derivatives of a convex function are presented. In particular, we show that a convex function is continuous on the interior of the domain. See, for example, Rockafellar [1970]. Rockafellar also discusses the *convex extension* to R^n of a convex function f: $S \subset R^n \to R$, which takes on finite values over a convex subset S of R^n, by letting $f(\mathbf{x}) = \infty$ for $\mathbf{x} \notin S$. Accordingly, a set of arithmetic operations involving ∞ also needs to be defined. In this case, S is referred to as the *effective domain* of f. Also, a *proper convex function* is then defined as a convex function for which $f(\mathbf{x}) < \infty$ for at least one point x and for which $f(\mathbf{x}) > -\infty$ for all x.

In Section 3.2 we discuss subgradients of convex functions. Many of the properties of differentiable convex functions are retained by replacing the gradient vector by a subgradient. For this reason, subgradients have been used frequently in the optimization of nondifferentiable functions. See, for example, Bertsekas [1975], Demyanov and Pallaschke [1985], Demyanov and Vasilev [1985], Held and Karp [1970], Held et al. [1974], Kiwiel [1985], Sherali et al. [2000], Shor [1985], and Wolfe [1976]. (See also, Chapter 8.)

In Section 3.3 we give some properties of differentiable convex functions. For further study of these topics as well as other properties of convex functions, refer to Eggleston [1958], Fenchel [1953], Roberts and Varberg [1973], and Rockafellar [1970]. The superdiagonalization algorithm derived from Theorem 3.3.12 provides an efficient polynomial-time algorithm for checking definiteness properties of matrices. This method is intimately related with LU and Cholesky factorization techniques (see Exercise 3.43, and refer to Section A.2, Fletcher [1985], Luenberger [1973a], and Murty [1983] for further details).

Section 3.4 treats the subject of minima and maxima of convex functions over convex sets. Robinson [1987] discusses the distinction between strict and strong local minima. For general functions, the study of minima and maxima is quite complicated. As shown in Section 3.4, however, every local minimum of a convex function over a convex set is also a global minimum, and the maximum of a convex function over a convex set occurs at an extreme point. For an excellent study of optimization of convex functions, see Rockafellar [1970]. The characterization of the optimal solution set for convex programs is due to Mangasarian [1988]. This paper also extends the results given in Section 3.4 to subdifferentiable convex functions.

In Section 3.5 we examine other classes of functions that are related to convex functions; namely, quasiconvex and pseudoconvex functions. The class of quasiconvex functions was first studied by De Finetti [1949]. Arrow and Enthoven [1961] derived necessary and sufficient conditions for quasiconvexity on the nonnegative orthant assuming twice differentiability. Their results were extended by Ferland [1972]. Note that a local minimum of a quasiconvex function over a convex set is not necessarily a global minimum. This result holds true, however, for a strictly quasiconvex function. Ponstein [1967] introduced the concept of strongly quasiconvex functions, which ensures that the global minimum is unique, a property that is not enjoyed by strictly quasiconvex functions. The notion of pseudoconvexity was introduced by Mangasarian [1965]. The significance of the class of pseudoconvex functions stems from the fact that every point with a zero gradient is a global minimum. Matrix theoretic characterizations (see, e.g., Exercises 3.66 and 3.67) of quadratic pseudoconvex and quasiconvex functions have been presented by Cottle and Ferland [1972] and by Martos [1965, 1967b, 1969, 1975]. For further reading on this topic, refer to Avriel et al. [1988], Fenchel [1953], Greenberg and Pierskalla [1971], Karamardian [1967], Mangasarian [1969a], Ponstein [1967], Schaible [1981a,b], and Schaible and Ziemba [1981]. The last four references give excellent surveys on this topic, and the results of Exercises 3.55 to 3.60 and 3.68 to 3.70 are discussed in detail by Avriel et al. [1988] and Schaible [1981a,b]. Karamardian

and Schaible [1990] also present various tests for checking generalized properties for differentiable functions. See also Section B.2.

Exercises 3.31 to 3.34 deal with *convex envelopes* of nonconvex functions. This construct plays an important role in global optimization techniques for nonconvex programming problems. For additional information on this subject, we refer the reader to Al-Khayyal and Falk [1983], Falk [1976], Grotzinger [1985], Horst and Tuy [1990], Pardalos and Rosen [1987], Sherali [1997], and Sherali and Alameddine [1990].

This page is intentionally left blank

Part 2

Optimality Conditions and Duality

This page is intentionally left blank

Chapter 4 The Fritz John and Karush–Kuhn–Tucker Optimality Conditions

In Chapter 3 we derived an optimality condition for a problem of the following form: Minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$, where f is a convex function and S is a convex set. The necessary and sufficient condition for $\bar{\mathbf{x}}$ to solve the problem was shown to be

$$\nabla f(\bar{\mathbf{x}})^t (\mathbf{x} - \bar{\mathbf{x}}) \geq 0 \qquad \text{for all } \mathbf{x} \in S.$$

In this chapter the nature of the set S will be specified more explicitly in terms of inequality and/or equality constraints. A set of first-order necessary conditions are derived without any convexity assumptions that are sharper than the above in the sense that they explicitly consider the constraint functions and are more easily verifiable, since they deal with a system of equations. Under suitable convexity assumptions, these necessary conditions are also sufficient for optimality. These optimality conditions lead to *classical* or *direct optimization techniques* for solving unconstrained and constrained problems that construct these conditions and then attempt to find a solution to them. In contrast, we discuss several *indirect methods* in Chapters 8 through 11, which iteratively improve the current solution, converging to a point that can be shown to satisfy these optimality conditions. A discussion of second-order necessary and/or sufficient conditions for unconstrained as well as for constrained problems is also provided.

Readers who are unfamiliar with generalized convexity concepts from Section 3.5 may substitute any references to such properties by related convexity assumptions for ease in reading.

Following is an outline of the chapter.

Section 4.1: Unconstrained Problems We consider briefly optimality conditions for unconstrained problems. First- and second-order conditions are discussed.

Section 4.2: Problems Having Inequality Constraints Both the Fritz John (FJ) and the Karush–Kuhn–Tucker (KKT) conditions for problems having

inequality constraints are derived. The nature and value of solutions satisfying these conditions are emphasized.

Section 4.3: Problems Having Inequality and Equality Constraints This section extends the results of the preceding section to problems having both inequality and equality constraints.

Section 4.4: Second-Order Necessary and Sufficient Optimality Conditions for Constrained Problems Similar to the unconstrained case discussed in Section 4.1, we develop second-order necessary and sufficient optimality conditions as an extension to the first-order conditions developed in Sections 4.2 and 4.3 for inequality and equality constrained problems. Many results and algorithms in nonlinear programming assume the existence of a local optimal solution that satisfies the second-order sufficiency conditions.

4.1 Unconstrained Problems

An unconstrained problem is a problem of the form to minimize $f(\mathbf{x})$ without any constraints on the vector $\mathbf{x}$. Unconstrained problems seldom arise in practical applications. However, we consider such problems here because optimality conditions for constrained problems become a logical extension of the conditions for unconstrained problems. Furthermore, as shown in Chapter 9, one strategy for solving a constrained problem is to solve a sequence of unconstrained problems.

We recall below the definitions of local and global minima for unconstrained problems as a special case of Definition 3.4.1, where the set S is replaced by R^n.

4.1.1 Definition

Consider the problem of minimizing $f(\mathbf{x})$ over R^n, and let $\bar{\mathbf{x}} \in R^n$. If $f(\bar{\mathbf{x}}) \leq f(\mathbf{x})$ for all $\mathbf{x} \in R^n$, $\bar{\mathbf{x}}$ is called a *global minimum*. If there exists an ε-neighborhood $N_\varepsilon(\bar{\mathbf{x}})$ around $\bar{\mathbf{x}}$ such that $f(\bar{\mathbf{x}}) \leq f(\mathbf{x})$ for each $\mathbf{x} \in N_\varepsilon(\bar{\mathbf{x}})$, $\bar{\mathbf{x}}$ is called a *local minimum*, while if $f(\bar{\mathbf{x}}) < f(\mathbf{x})$ for all $\mathbf{x} \in N_\varepsilon(\bar{\mathbf{x}})$, $\mathbf{x} \neq \bar{\mathbf{x}}$, for some $\varepsilon > 0$, $\bar{\mathbf{x}}$ is called a *strict local minimum*. Clearly, a global minimum is also a local minimum.

Necessary Optimality Conditions

Given a point $\bar{\mathbf{x}}$ in R^n, we wish to determine, if possible, whether or not the point is a local or a global minimum of a function f. For this purpose we need to characterize a minimizing solution. Fortunately, the differentiability assumption of f provides a means for obtaining this characterization. The corollary to Theorem 4.1.2 gives a first-order necessary condition for $\bar{\mathbf{x}}$ to be a local optimum. Theorem 4.1.3 gives a second-order necessary condition using the Hessian matrix.

4.1.2 Theorem

Suppose that f: $R^n \to R$ is differentiable at $\bar{\mathbf{x}}$. If there is a vector $\mathbf{d}$ such that $\nabla f(\bar{\mathbf{x}})^t \mathbf{d} < 0$, there exists a $\delta > 0$ such that $f(\bar{\mathbf{x}} + \lambda \mathbf{d}) < f(\bar{\mathbf{x}})$ for each $\delta \in (0, \delta)$, so that $\mathbf{d}$ is a *descent direction* of f at $\bar{\mathbf{x}}$.

Proof

By the differentiability of f at $\bar{\mathbf{x}}$, we must have

$$f(\bar{\mathbf{x}} + \lambda \mathbf{d}) = f(\bar{\mathbf{x}}) + \lambda \nabla f(\bar{\mathbf{x}})^t \mathbf{d} + \lambda \|\mathbf{d}\| \alpha(\bar{\mathbf{x}}; \lambda \mathbf{d}),$$

where $\alpha(\bar{\mathbf{x}}; \lambda \mathbf{d}) \to 0$ as $\lambda \to 0$. Rearranging the terms and dividing by λ, $\lambda \neq 0$, we get

$$\frac{f(\bar{\mathbf{x}} + \lambda \mathbf{d}) - f(\bar{\mathbf{x}})}{\lambda} = \nabla f(\bar{\mathbf{x}})^t \mathbf{d} + \|\mathbf{d}\| \alpha(\bar{\mathbf{x}}; \lambda \mathbf{d}).$$

Since $\nabla f(\bar{\mathbf{x}})^t \mathbf{d} < 0$ and $\alpha(\bar{\mathbf{x}}; \lambda \mathbf{d}) \to 0$ as $\lambda \to 0$, there exists a $\delta > 0$ such that $\nabla f(\bar{\mathbf{x}})^t \mathbf{d} + \|\mathbf{d}\| \alpha(\bar{\mathbf{x}}; \lambda \mathbf{d}) < 0$ for all $\lambda \in (0, \delta)$. The result then follows.

Corollary

Suppose that f: $R^n \to R$ is differentiable at $\bar{\mathbf{x}}$. If $\bar{\mathbf{x}}$ is a local minimum, $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$.

Proof

Suppose that $\nabla f(\bar{\mathbf{x}}) \neq \mathbf{0}$. Then, letting $\mathbf{d} = -\nabla f(\bar{\mathbf{x}})$, we get $\nabla f(\bar{\mathbf{x}})^t \mathbf{d} = -\|\nabla f(\bar{\mathbf{x}})\|^2 < 0$; and by Theorem 4.1.2, there is a $\delta > 0$ such that $f(\bar{\mathbf{x}} + \lambda \mathbf{d}) < f(\bar{\mathbf{x}})$ for $\lambda \in (0, \delta)$, contradicting the assumption that $\bar{\mathbf{x}}$ is a local minimum. Hence, $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$.

The condition above uses the gradient vector whose components are the first partials of f. Hence, it is called a *first-order condition*. Necessary conditions can also be stated in terms of the Hessian matrix $\mathbf{H}$, whose elements are the second partials of f, and are then called *second-order conditions*. One such condition is given below.

4.1.3 Theorem

Suppose that f: $R^n \to R$ is twice differentiable at $\bar{\mathbf{x}}$. If $\bar{\mathbf{x}}$ is a local minimum, $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$ and $\mathbf{H}(\bar{\mathbf{x}})$ is positive semidefinite.

Proof

Consider an arbitrary direction $\mathbf{d}$. Then from the differentiability of f at $\bar{\mathbf{x}}$, we have

$$f(\bar{\mathbf{x}}+\lambda\mathbf{d}) = f(\bar{\mathbf{x}}) + \lambda\nabla f(\bar{\mathbf{x}})^t\mathbf{d} + \frac{1}{2}\lambda^2\mathbf{d}^t\mathbf{H}(\bar{\mathbf{x}})\mathbf{d} + \lambda^2\|\mathbf{d}\|^2\alpha(\bar{\mathbf{x}};\lambda\mathbf{d}), \quad (4.1)$$

where $\alpha(\bar{\mathbf{x}};\lambda\mathbf{d}) \to 0$ as $\lambda \to 0$. Since $\bar{\mathbf{x}}$ is a local minimum, from the corollary to Theorem 4.1.2, we have $\nabla f(\bar{\mathbf{x}}) = 0$. Rearranging the terms in (4.1) and dividing by $\lambda^2 > 0$, we get

$$\frac{f(\bar{\mathbf{x}}+\lambda\mathbf{d}) - f(\bar{\mathbf{x}})}{\lambda^2} = \frac{1}{2}\mathbf{d}^t\mathbf{H}(\bar{\mathbf{x}})\mathbf{d} + \|\mathbf{d}\|^2\alpha(\bar{\mathbf{x}};\lambda\mathbf{d}). \quad (4.2)$$

Since $\bar{\mathbf{x}}$ is a local minimum, $f(\bar{\mathbf{x}}+\lambda\mathbf{d}) \geq f(\bar{\mathbf{x}})$ for λ sufficiently small. From (4.2) it is thus clear that $(1/2)\mathbf{d}^t\mathbf{H}(\bar{\mathbf{x}})\mathbf{d} + \|\mathbf{d}\|^2\alpha(\bar{\mathbf{x}};\lambda\mathbf{d}) \geq 0$ for λ sufficiently small. By taking the limit as $\lambda \to 0$, it follows that $\mathbf{d}^t\mathbf{H}(\bar{\mathbf{x}})\mathbf{d} \geq 0$; and hence, since $\mathbf{d}$ was arbitrary, $\mathbf{H}(\bar{\mathbf{x}})$ is positive semidefinite.

Sufficient Optimality Conditions

The conditions discussed thus far are necessary conditions; that is, they must be true for every local optimal solution. On the other hand, a point satisfying these conditions need not be a local minimum. Theorem 4.1.4 gives a sufficient condition for a local minimum.

4.1.4 Theorem

Suppose that f: $R^n \to R$ is twice differentiable at $\bar{\mathbf{x}}$. If $\nabla f(\bar{\mathbf{x}}) = 0$ and $\mathbf{H}(\bar{\mathbf{x}})$ is positive definite, $\bar{\mathbf{x}}$ is a strict local minimum.

Proof

Since f is twice differentiable at $\bar{\mathbf{x}}$, we must have, for each $\bar{\mathbf{x}} \in R^n$,

$$f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t(\mathbf{x}-\bar{\mathbf{x}}) + \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})^t\mathbf{H}(\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}}) + \|\mathbf{x}-\bar{\mathbf{x}}\|^2\alpha(\bar{\mathbf{x}};\mathbf{x}-\bar{\mathbf{x}}), \quad (4.3)$$

where $\alpha(\bar{\mathbf{x}};\mathbf{x}-\bar{\mathbf{x}}) \to 0$ as $\mathbf{x} \to \bar{\mathbf{x}}$. Suppose, by contradiction, that $\bar{\mathbf{x}}$ is not a strict local minimum; that is, suppose that there exists a sequence $\{\mathbf{x}_k\}$ converging to $\bar{\mathbf{x}}$ such that $f(\mathbf{x}_k) \leq f(\bar{\mathbf{x}})$, $\mathbf{x}_k \neq \bar{\mathbf{x}}$, for each k. Considering this sequence, noting that $\nabla f(\bar{\mathbf{x}}) = 0$ and that $f(\mathbf{x}_k) \leq f(\bar{\mathbf{x}})$, and denoting $(\mathbf{x}_k - \bar{\mathbf{x}})/\|\mathbf{x}_k - \bar{\mathbf{x}}\|$ by $\mathbf{d}_k$, (4.3) then implies that

$$\frac{1}{2}\mathbf{d}_k^t \mathbf{H}(\bar{\mathbf{x}})\mathbf{d}_k + \alpha(\bar{\mathbf{x}}; \mathbf{x}_k - \bar{\mathbf{x}}) \leq 0 \qquad \text{for each } k. \tag{4.4}$$

But $\|\mathbf{d}_k\| = 1$ for each k; and hence there exists an index set $\mathcal{K}$ such that $\{\mathbf{d}_k\}_{\mathcal{K}}$ converges to $\mathbf{d}$, where $\|\mathbf{d}\| = 1$. Considering this subsequence and the fact that $\alpha(\bar{\mathbf{x}}; \mathbf{x}_k - \bar{\mathbf{x}}) \to 0$ as $k \in \mathcal{K}$ approaches ∞, (4.4) implies that $\mathbf{d}^t \mathbf{H}(\bar{\mathbf{x}})\mathbf{d} \leq 0$. This contradicts the assumption that $\mathbf{H}(\bar{\mathbf{x}})$ is positive definite since $\|\mathbf{d}\| = 1$. Therefore, $\bar{\mathbf{x}}$ is indeed a strict local minimum.

Essentially, note that assuming f to be twice continuously differentiable, since $\mathrm{H}(\bar{\mathbf{x}})$ is positive definite, we have that $\mathbf{H}(\mathbf{x})$ is positive definite in an ε-neighborhood of $\bar{\mathbf{x}}$, so f is strictly convex in an ε-neighborhood of $\bar{\mathbf{x}}$. Therefore, as follows from Theorem 3.4.2, $\bar{\mathbf{x}}$ is a strict local minimum, that is, it is the unique global minimum over $N_\varepsilon(\bar{\mathbf{x}})$ for some $\varepsilon > 0$. In fact, noting the second part of Theorem 3.4.2, we can conclude that $\bar{\mathbf{x}}$ is also a strong or isolated local minimum in this case.

In Theorem 4.1.5, we show that the necessary condition $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$ is also sufficient for $\bar{\mathbf{x}}$ to be a global minimum if f is pseudoconvex at $\bar{\mathbf{x}}$. In particular, if $\nabla f(\bar{\mathbf{x}}) = 0$ and if $\mathbf{H}(\mathbf{x})$ is positive semidefinite for all $\mathbf{x}$, f is convex, and therefore also pseudoconvex. Consequently, $\bar{\mathbf{x}}$ is a global minimum. This is also evident from Theorem 3.3.3 or from Corollary 2 to Theorem 3.4.3.

4.1.5 Theorem

Let f: $R^n \to R$ be pseudoconvex at $\bar{\mathbf{x}}$. Then $\bar{\mathbf{x}}$ is a global minimum if and only if $\nabla f(\bar{\mathbf{x}}) = 0$.

Proof

By the corollary to Theorem 4.1.2, if $\bar{\mathbf{x}}$ is a global minimum, $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$. Now suppose that $\nabla f(\bar{\mathbf{x}}) = 0$, so that $\nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) = 0$ for each $\mathbf{x} \in R^n$. By the pseudoconvexity of f at $\bar{\mathbf{x}}$, it then follows that $f(\mathbf{x}) \geq f(\bar{\mathbf{x}})$ for each $\mathbf{x} \in R^n$, and the proof is complete.

Theorem 4.1.5 provides a necessary *and* sufficient optimality condition in terms of the first-order derivative alone when f is pseudoconvex. In a similar manner, we can derive necessary *and* sufficient conditions for local optimality in terms of higher-order derivatives when f is infinitely differentiable, as an extension to the foregoing results. Toward this end, consider the following result for the ***univariate*** case.

4.1.6 Theorem

Let f: $R \to R$ be an infinitely differentiable univariate function. Then $\bar{x} \in R$ is a local minimum if and only if either $f^{(j)}(\bar{x}) = 0$ for all $j = 1, 2, \ldots$, or else there exists an even $n \geq 2$ such that $f^{(n)}(\bar{x}) > 0$ while $f^{(j)}(\bar{x}) = 0$ for all $1 \leq j < n$, where $f^{(j)}$ denotes the jth-order derivative of f.

Proof

We know that $\bar{x}$ is a local minimum of f if and only if $f(\bar{x} + h) - f(\bar{x}) \geq 0$ for all sufficiently small values of $|h|$. Using the infinite Taylor series representation of $f(\bar{x} + h)$, this holds true if and only if

$$hf^{(1)}(\bar{x}) + \frac{h^2}{2!} f^{(2)}(\bar{x}) + \frac{h^3}{3!} f^{(3)}(\bar{x}) + \frac{h^4}{4!} f^{(4)}(\bar{x}) + \cdots \geq 0$$

for all $|h|$ small enough. Similar to the proof of Theorem 3.3.9, it is readily verified that the foregoing inequality holds true if and only if the condition of the theorem is satisfied, and this completes the proof.

Before proceeding, we remark here that for a local maximum, the condition of Theorem 4.1.6 remains the same, except that we require $f^{(n)}(\bar{x}) < 0$ in lieu of $f^{(n)}(\bar{x}) > 0$. Observe also, noting Theorem 3.3.9, that the above result essentially asserts that for the case under discussion, $\bar{x}$ is a local minimum if and only if f is locally convex about $\bar{x}$. This result can be partially extended, at least in theory, to the case of multivariate functions. Toward this end, suppose that $\bar{\mathbf{x}} \in R^n$ is a local minimum for f: $R^n \to R$. Then this holds true if and only if $f(\bar{\mathbf{x}} + \lambda \mathbf{d}) \geq f(\bar{\mathbf{x}})$ for all $\mathbf{d} \in R^n$ and for all sufficiently small values of $|\lambda|$. Assuming f to be infinitely differentiable, this asserts that for all $\mathbf{d} \in R^n$, $\|\mathbf{d}\| = 1$, we must equivalently have

$$f(\bar{\mathbf{x}} + \lambda \mathbf{d}) - f(\bar{\mathbf{x}}) = \lambda \nabla f(\bar{\mathbf{x}})^t \mathbf{d} + \frac{\lambda^2}{2!} \mathbf{d}^t \mathbf{H}(\bar{\mathbf{x}}) \mathbf{d}$$

$$+ \frac{\lambda^3}{3!} \sum_i \sum_j \sum_k f_{ijk}(\bar{\mathbf{x}}) d_i d_j d_k + \cdots \geq 0$$

for all $-\delta \leq \lambda \leq \delta$, for some $\delta > 0$. Consequently, the first nonzero derivative term, if it exists, must correspond to an even power of λ and must be positive in value.

Note that the foregoing concluding statement is *not* sufficient to claim local optimality of $\bar{\mathbf{x}}$. The difficulty is that it might be the case that this

statement holds true, implying that for any $\mathbf{d} \in R^n$, $\|\mathbf{d}\| = 1$, we have $f(\bar{\mathbf{x}} + \lambda\mathbf{d}) \geq f(\bar{\mathbf{x}})$ for all $-\delta_{\mathbf{d}} \leq \lambda \leq \delta_{\mathbf{d}}$ for some $\delta_{\mathbf{d}} > 0$, which depends on $\mathbf{d}$, but then, $\delta_{\mathbf{d}}$ might get vanishingly small as $\mathbf{d}$ varies, so that we cannot assert the existence of a $\delta > 0$ such that $f(\bar{\mathbf{x}} + \lambda\mathbf{d}) \geq f(\bar{\mathbf{x}})$ for all $-\delta \leq \lambda \leq \delta$. In this case, by moving along curves instead of along straight lines, improving values of f might be accessible in the immediate neighborhood of $\bar{\mathbf{x}}$. On the other hand, a valid sufficient condition by Theorem 4.1.5 is that $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$ and that f is convex (or pseudoconvex) over an ε-neighborhood about $\bar{\mathbf{x}}$, for some $\varepsilon > 0$. However, this might not be easy to check, and we might need to assess the situation numerically by examining values of the function at perturbations about the point $\bar{\mathbf{x}}$ (refer also to Exercise 4.19).

To illustrate the above point, consider the following example due to the Italian mathematician Peano. Let $f(x_1, x_2) = (x_2^2 - x_1)(x_2^2 - 2x_1) = 2x_1^2 - 3x_1x_2^2 + x_2^4$. Then we have, at $\bar{\mathbf{x}} = (0,0)^t$,

$$\nabla f(\mathbf{0}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \ \mathbf{H}(\mathbf{0}) = \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix}, \ f_{122}(\mathbf{0}) = f_{212}(\mathbf{0}) = f_{221}(\mathbf{0}) = -6, \ f_{2222}(\mathbf{0}) = 24,$$

and all other partial derivatives of f of order 3 or higher are zeros. Hence, we obtain by the Taylor series expansion

$$\begin{aligned} f(\bar{\mathbf{x}} + \lambda\mathbf{d}) - f(\bar{\mathbf{x}}) &= \frac{\lambda^2}{2}(4d_1^2) + \frac{\lambda^3}{6}(-18d_1d_2^2) + \frac{\lambda^4}{24}(24d_2^4) \\ &= 2\lambda^2\left(d_1 - \frac{3\lambda}{4}d_2^2\right)^2 - \frac{1}{8}\lambda^4 d_2^4. \end{aligned}$$

Note that for any $\mathbf{d} = (d_1, d_2)^t$, $\|\mathbf{d}\| = 1$, if $d_1 \neq 0$, the given necessary condition holds true because the second-order term is positive. On the other hand, if $d_1 = 0$, we must have $d_2 \neq 0$, and the condition holds true again because the first nonzero term is of order 4 and is positive. However, $\bar{\mathbf{x}} = (0,0)^t$ is not a local minimum, as evident from Figure 4.1. We have $f(0, 0) = 0$, while there exist negative values of f in any ε-neighborhood about the point (0, 0). In fact, taking $\mathbf{d} = (\sin\theta, \cos\theta)^t$, we have $f(\bar{\mathbf{x}} + \lambda\mathbf{d}) - f(\bar{\mathbf{x}}) = 2\sin^2\theta\lambda^2 - 3\sin\theta\cos^2\theta\lambda^3 + \cos^4\theta\lambda^4$; and for this to be nonnegative for all $-\delta_\theta \leq \lambda \leq \delta_\theta$, $\delta_\theta > 0$, we observe that as $\theta \to 0^+$, we get $\delta_\theta \to 0^+$ as well (see Exercise 4.11), although at $\theta = 0$ we get $\delta_\theta = \infty$. Hence, we cannot derive a $\delta > 0$ such that $f(\bar{\mathbf{x}} + \lambda\mathbf{d}) - f(\bar{\mathbf{x}}) \geq 0$, for all $\mathbf{d} \in R^n$ and $-\delta \leq \lambda \leq \delta$, so $\bar{\mathbf{x}}$ is not a local minimum.

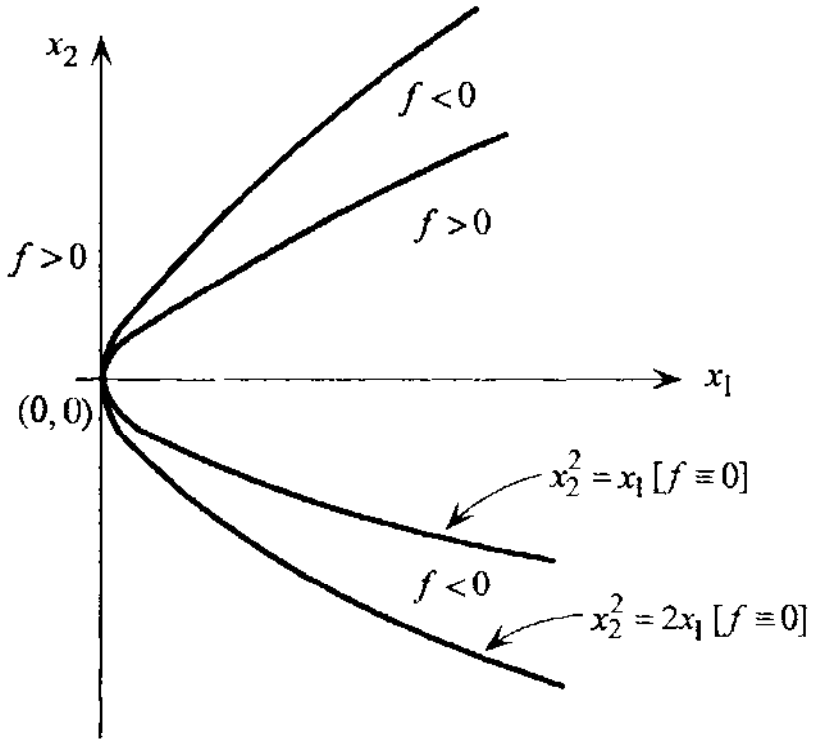


Figure 4.1 Regions of zero, positive, and negative values of $f(x_1, x_2) = (x_2^2 - x_1)\,(x_2^2 - 2x_1)$.

To afford further insight into the multivariate case, let us examine a situation in which f: $R^n \to R$ is twice continuously differentiable, and at a given point $\bar{\mathbf{x}} \in R^n$, we have that $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$ but $\mathbf{H}(\bar{\mathbf{x}})$ is indefinite. Hence, there exist directions $\mathbf{d}_1$ and $\mathbf{d}_2$ in R^n such that $\mathbf{d}_1^t \mathbf{H}(\bar{\mathbf{x}})\mathbf{d}_1 > 0$ and $\mathbf{d}_2^t \mathbf{H}(\bar{\mathbf{x}})\mathbf{d}_2 < 0$. Defining $F_{(\bar{\mathbf{x}};\mathbf{d}_j)}(\lambda) = f(\bar{\mathbf{x}} + \lambda \mathbf{d}_j) \equiv F_{\mathbf{d}_j}(\lambda)$, say, for $j = 1, 2$, and denoting derivatives by primes, we get

$$F'_{\mathbf{d}_j}(\lambda) = \nabla f(\bar{\mathbf{x}} + \lambda \mathbf{d}_j)^t \mathbf{d}_j \quad \text{and} \quad F''_{\mathbf{d}_j}(\lambda) = \mathbf{d}_j^t \mathbf{H}(\bar{\mathbf{x}} + \lambda \mathbf{d}_j)\mathbf{d}_j \quad \text{for } j = 1, 2.$$

Hence, for $j = 1$, we have $F'_{\mathbf{d}_1}(0) = 0$, $F''_{\mathbf{d}_1}(0) > 0$; and moreover, by continuity of the second derivative, $F''_{\mathbf{d}_1}(\lambda) > 0$, for $|\lambda|$ sufficiently small. Hence, $F_{\mathbf{d}_1}(\lambda)$ is strictly convex in some ε-neighborhood of $\lambda = 0$, achieving a strict local minimum at $\lambda = 0$. Similarly, for $j = 2$, noting that $F'_{\mathbf{d}_2}(0) = 0$ and $F''_{\mathbf{d}_2}(0) < 0$, we conclude that $F_{\mathbf{d}_2}(\lambda)$ is strictly concave in some ε-neighborhood of $\lambda = 0$, achieving a strict local maximum at $\lambda = 0$. Hence, as foretold by Theorem 4.1.3, $\bar{\mathbf{x}} = \mathbf{0}$ is neither a local minimum nor a local maximum. Such a point $\bar{\mathbf{x}}$ is called a *saddle point* (or an *inflection point*). Figure 4.2 illustrates the situation. Observe the convex and concave cross sections of the function in the respective directions $\mathbf{d}_1$ and $\mathbf{d}_2$ about the point $\bar{\mathbf{x}}$ at which $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$, which gives the function the appearance of a saddle in the vicinity of $\bar{\mathbf{x}}$.

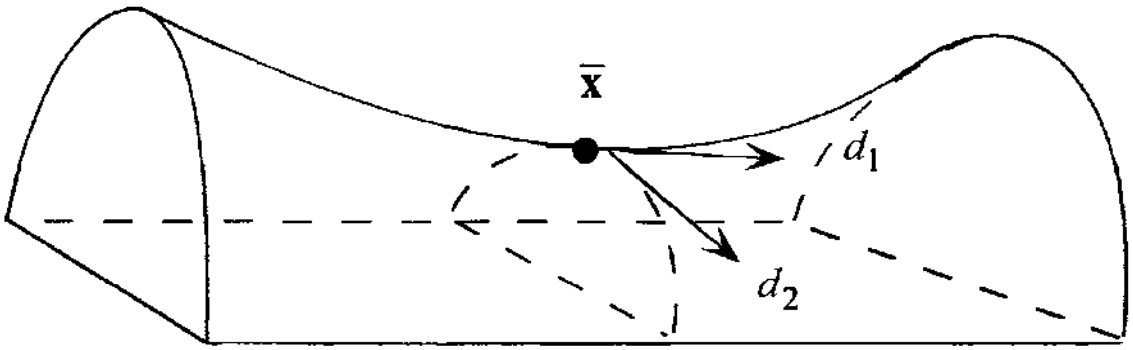


Figure 4.2 Saddle point at $\bar{\mathbf{x}}$.

4.1.7 Examples

Example 1: Univariate Function To illustrate the necessary and sufficient conditions of this section, consider the problem to minimize $f(x) = (x^2 - 1)^3$. First, let us determine the candidate points for optimality satisfying the first-order necessary condition that $\nabla f(x) = 0$. Note that $\nabla f(x) \equiv f'(x) = 6x(x^2 - 1)^2 = 0$ when $x = 0$, 1, or -1. Hence, our candidate points for local optimality are $\bar{x} = 0$, 1, or -1. Now let us examine the second-order derivatives. We have $H(x) = f''(x) = 24x^2(x^2 - 1) + 6(x^2 - 1)^2$, and hence $H(1) = H(-1) = 0$ and $H(0) = 6$. Since H is positive definite at $\bar{x} = 0$, we have by Theorem 4.1.4 that $\bar{x} = 0$ is a strict local minimum. However, at $x = +1$ or -1, H is both positive and negative semidefinite; and although it satisfies the second-order necessary condition of Theorem 4.1.3, this is not sufficient for us to conclude anything about the behavior of f at these points. Hence, we continue and examine the third-order derivative $f'''(x) = 48x(x^2 - 1) + 48x^3 + 24x(x^2 - 1)$. Evaluating this at the two candidate points $\bar{x} = \pm 1$ in question, we obtain $f'''(1) = 48 > 0$ and $f'''(-1) = -48 < 0$. By Theorem 4.1.6 it follows that we have neither a local minimum nor a local maximum at these points, and these points are merely inflection points.

Example 2: Multivariate Function Consider the bivariate function $f(x_1, x_2) = x_1^3 + x_2^3$. Evaluating the gradient and the Hessian of f, we obtain

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 3x_1^2 \\ 3x_2^2 \end{pmatrix} \quad \text{and} \quad \mathbf{H}(\mathbf{x}) = \begin{bmatrix} 6x_1 & 0 \\ 0 & 6x_2 \end{bmatrix}.$$

The first-order necessary condition $\nabla f(\bar{\mathbf{x}}) = 0$ yields $\bar{\mathbf{x}} = (0, 0)^t$ as the single candidate point. However, $\mathbf{H}(\bar{\mathbf{x}})$ is the zero matrix; and although it satisfies the second-order necessary condition of Theorem 4.1.3, we need to examine higher-order derivatives to make any conclusive statement about the point $\bar{\mathbf{x}}$. Defining $F_{(\bar{\mathbf{x}};\mathbf{d})}(\lambda) = f(\bar{\mathbf{x}} + \lambda\mathbf{d}) \equiv F_{\mathbf{d}}(\lambda)$, say, we have $F'_{\mathbf{d}}(\lambda) = \nabla f(\bar{\mathbf{x}} + \lambda\mathbf{d})^t\mathbf{d}$, $F''_{\mathbf{d}}(\lambda) =$

$\mathbf{d}^t\mathbf{H}(\bar{\mathbf{x}}+\lambda\mathbf{d})\mathbf{d}$, and $F_{\mathbf{d}}'''(\lambda) = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} d_i d_j d_k f_{ijk}(\bar{\mathbf{x}}+\lambda\mathbf{d})$. Noting that $f_{111}(\mathbf{x}) = 6$, $f_{222}(\mathbf{x}) = 6$, and $f_{ijk}(\mathbf{x}) = 0$ otherwise, we obtain $F_{\mathbf{d}}'''(0) = 6d_1^3 + 6d_2^3$. Since there exist directions $\mathbf{d}$ for which the first nonzero derivative term at $\lambda = 0$ is $F_{\mathbf{d}}'''(0)$, which is of odd order, $\bar{\mathbf{x}} = (0,0)^t$ is an inflection point and is therefore neither a local minimum nor a local maximum. In fact, note that $F_{\mathbf{d}}''(\lambda) = 6\lambda(d_1^3 + d_2^3)$ can be made to take on opposite signs about $\lambda = 0$ along any direction $\mathbf{d}$ for which $d_1^3 + d_2^3 \neq 0$; so the function switches from a convex to a concave function, or vice versa, about the point $\mathbf{0}$ along any direction $\mathbf{d}$. Observe also that $\mathbf{H}$ is positive semidefinite over $\{\mathbf{x} : x_1 \geq 0,\ x_2 \geq 0\}$; and hence, over this region, the function is convex, yielding $\bar{\mathbf{x}} = (0,0)^t$ as a global minimum. Similarly, $\bar{\mathbf{x}} = (0,0)^t$ is a global maximum over the region $\{\mathbf{x} : x_1 \leq 0,\ x_2 \leq 0\}$.

4.2 Problems Having Inequality Constraints

In this section we first develop a necessary optimality condition for the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$ for a general set S. Later, we let S be more specifically defined as the feasible region of a nonlinear programming problem of the form to minimize $f(\mathbf{x})$ subject to $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$ and $\mathbf{x} \in X$.

Geometric Optimality Conditions

In Theorem 4.2.2 we develop a necessary optimality condition for the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$, using the cone of feasible directions defined below.

4.2.1 Definition

Let S be a nonempty set in R^n, and let $\bar{\mathbf{x}} \in \text{cl } S$. The *cone of feasible directions* of S at $\bar{\mathbf{x}}$, denoted by D, is given by

$$D = \{\mathbf{d} : \mathbf{d} \neq \mathbf{0}, \text{ and } \bar{\mathbf{x}} + \lambda\mathbf{d} \in S \text{ for all } \lambda \in (0,\delta) \text{ for some } \delta > 0\}.$$

Each nonzero vector $\mathbf{d} \in D$ is called a *feasible direction.* Moreover, given a function f: $R^n \to R$, the *cone of improving directions* at $\bar{\mathbf{x}}$, denoted by F, is given by

$$F = \{\mathbf{d} : f(\bar{\mathbf{x}} + \lambda\mathbf{d}) < f(\bar{\mathbf{x}}) \text{ for all } \lambda \in (0,\delta) \text{ for some } \delta > 0\}.$$

Each direction $\mathbf{d} \in F$ is called an *improving direction*, or a *descent direction*, of f at $\bar{\mathbf{x}}$.

From the above definitions, it is clear that a small movement from $\bar{\mathbf{x}}$ along a vector $\mathbf{d} \in D$ leads to feasible points, whereas a similar movement along a $\mathbf{d} \in F$ vector leads to solutions of improving objective value. Furthermore,

Chapter 8 Unconstrained Optimization

Unconstrained optimization deals with the problem of minimizing or maximizing a function in the absence of any restrictions. In this chapter we discuss both the minimization of a function of one variable and a function of several variables. Even though most practical optimization problems have side restrictions that must be satisfied, the study of techniques for unconstrained optimization is important for several reasons. Many algorithms solve a constrained problem by converting it into a sequence of unconstrained problems via Lagrangian multipliers, as illustrated in Chapter 6, or via penalty and barrier functions, as discussed in Chapter 9. Furthermore, most methods proceed by finding a direction and then minimizing along this direction. This line search is equivalent to minimizing a function of one variable without constraints or with simple constraints, such as lower and upper bounds on the variables. Finally, several unconstrained optimization techniques can be extended in a natural way to provide and motivate solution procedures for constrained problems.

Following is an outline of the chapter.

Section 8.1: Line Search Without Using Derivatives We discuss several procedures for minimizing strictly quasiconvex functions of one variable without using derivatives. Uniform search, dichotomous search, the golden section method, and the Fibonacci method are covered.

Section 8.2: Line Search Using Derivatives Differentiability is assumed, and the bisection search method and Newton's method are discussed.

Section 8.3: Some Practical Line Search Methods We describe the popular quadratic-fit line search method and present the Armijo rule for performing acceptable, inexact line searches.

Section 8.4: Closedness of the Line Search Algorithmic Map We show that the line search algorithmic map is closed, a property that is essential in convergence analyses. Readers who are not interested in convergence analyses may skip this section.

Section 8.5: Multidimensional Search Without Using Derivatives The cyclic coordinate method, the method of Hooke and Jeeves, and Rosenbrock's method are discussed. Convergence of these methods is also established.

Section 8.6: Multidimensional Search Using Derivatives We develop the steepest descent method and the method of Newton and analyze their convergence properties.

Section 8.7: Modification of Newton's Method: Levenberg–Marquardt and Trust Region Methods We describe different variants of Newton's method based on the Levenberg–Marquardt and trust region methods,

which ensure the global convergence of Newton's method. We also discuss some insightful connections between these methods.

Section 8.8: Methods Using Conjugate Directions: Quasi-Newton and Conjugate Gradient Methods The important concept of conjugacy is introduced. If the objective function is quadratic, then methods using conjugate directions are shown to converge in a finite number of steps. Various quasi-Newton/variable metric and conjugate gradient methods are covered based on the concept of conjugate directions, and their computational performance and convergence properties are discussed.

Section 8.9: Subgradient Optimization Methods We introduce the extension of the steepest descent algorithm to that of minimizing convex, nondifferentiable functions via subgradient-based directions. Variants of this technique that are related to conjugate gradient and variable metric methods are mentioned, and the crucial step of selecting appropriate step sizes in practice is discussed.

8.1 Line Search Without Using Derivatives

One-dimensional search is the backbone of many algorithms for solving a nonlinear programming problem. Many nonlinear programming algorithms proceed as follows. Given a point $\mathbf{x}_k$, find a direction vector $\mathbf{d}_k$ and then a suitable step size λ_k, yielding a new point $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$; the process is then repeated. Finding the step size λ_k involves solving the subproblem to minimize $f(\mathbf{x}_k + \lambda \mathbf{d}_k)$, which is a one-dimensional search problem in the variable λ. The minimization may be over all real λ, nonnegative λ, or λ such that $\mathbf{x}_k + \lambda \mathbf{d}_k$ is feasible.

Consider a function θ of one variable λ to be minimized. One approach to minimizing θ is to set the derivative θ' equal to 0 and then solve for λ. Note, however, that θ is usually defined implicitly in terms of a function f of several variables. In particular, given the vectors x and d, $\theta(\lambda) = f(\mathbf{x} + \lambda \mathbf{d})$. If f is not differentiable, then θ will not be differentiable. If f is differentiable, then $\theta'(\lambda) = \mathbf{d}^t \nabla f(\mathbf{x} + \lambda \mathbf{d})$. Therefore, to find a point λ with $\theta'(\lambda) = 0$, we have to solve the equation $\mathbf{d}^t \nabla f(\mathbf{x} + \lambda \mathbf{d}) = 0$, which is usually nonlinear in λ. Furthermore, λ satisfying $\theta'(\lambda) = 0$ is not necessarily a minimum; it may be a local minimum, a local maximum, or even a saddle point. For these reasons, and except for some special cases, we avoid minimizing θ by letting its derivative be equal to zero. Instead, we resort to some numerical techniques for minimizing the function θ.

In this section we discuss several methods that do not use derivatives for minimizing a function θ of one variable over a closed bounded interval. These methods fall under the categories of simultaneous line search and sequential line search problems. In the former case, the candidate points are determined *a priori*,

whereas in the sequential search, the values of the function at the previous iterations are used to determine the succeeding points.

Interval of Uncertainty

Consider the line search problem to minimize $\theta(\lambda)$ subject to $a \leq \lambda \leq b$. Since the exact location of the minimum of θ over $[a, b]$ is not known, this interval is called the *interval of uncertainty*. During the search procedure if we can exclude portions of this interval that do not contain the minimum, then the interval of uncertainty is reduced. In general, $[a, b]$ is called the *interval of uncertainty* if a minimum point $\bar{\lambda}$ lies in $[a, b]$, although its exact value is not known.

Theorem 8.1.1 shows that if the function θ is strictly quasiconvex, then the interval of uncertainty can be reduced by evaluating θ at two points within the interval.

8.1.1 Theorem

Let $\theta: R \rightarrow R$ be strictly quasiconvex over the interval $[a, b]$. Let $\lambda, \mu \in [a, b]$ be such that $\lambda < \mu$. If $\theta(\lambda) > \theta(\mu)$, then $\theta(z) \geq \theta(\mu)$ for all $z \in [a, \lambda)$. If $\theta(\lambda) \leq \theta(\mu)$, then $\theta(z) \geq \theta(\lambda)$ for all $z \in (\mu, b]$.

Proof

Suppose that $\theta(\lambda) > \theta(\mu)$, and let $z \in [a, \lambda)$. By contradiction, suppose that $\theta(z) < \theta(\mu)$. Since λ can be written as a convex combination of z and μ, and by the strict quasiconvexity of θ, we have

$$\theta(\lambda) < \max\{\theta(z), \ \theta(\mu)\} = \theta(\mu),$$

contradicting $\theta(\lambda) > \theta(\mu)$. Hence, $\theta(z) \geq \theta(\mu)$. The second part of the theorem can be proved similarly.

From Theorem 8.1.1, under strict quasiconvexity if $\theta(\lambda) > \theta(\mu)$, the new interval of uncertainty is $[\lambda, b]$. On the other hand, if $\theta(\lambda) \leq \theta(\mu)$, the new interval of uncertainty is $[a, \mu]$. These two cases are illustrated in Figure 8.1.

Literature on nonlinear programming frequently uses the concept of *strict unimodality* of θ to reduce the interval of uncertainty (see Exercise 3.60). In this book we are using the equivalent concept of strict quasiconvexity. (See Exercises 3.57, 3.60, and 8.10 for definitions of various forms of unimodality and their relationships with different forms of quasiconvexity.)

We now present several procedures for minimizing a strictly quasiconvex function over a closed bounded interval by iteratively reducing the interval of uncertainty.

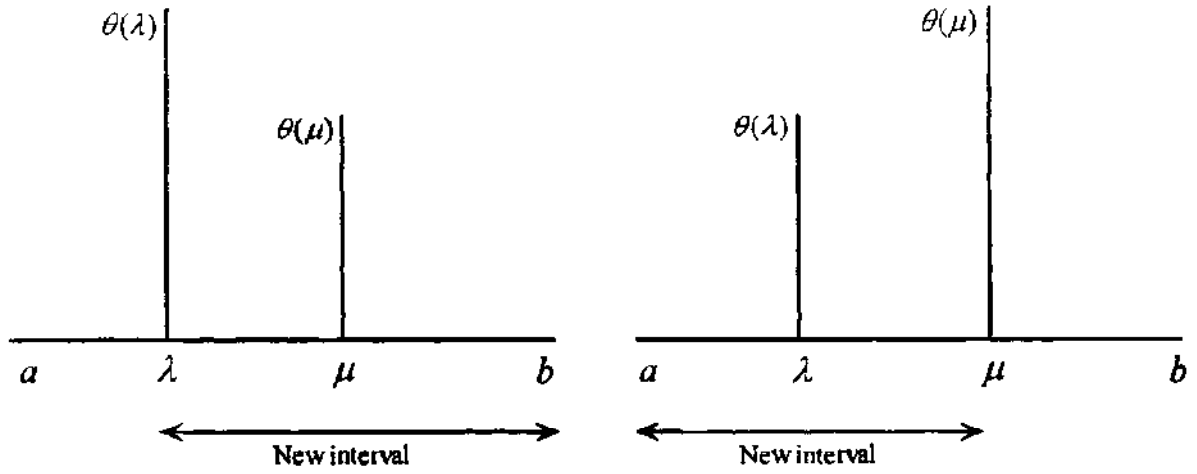


Figure 8.1 Reducing the interval of uncertainty.

Example of a Simultaneous Search: Uniform Search

Uniform search is an example of *simultaneous search*, where we decide beforehand the points at which the functional evaluations are to be made. The interval of uncertainty $[a_1, b_1]$ is divided into smaller subintervals via the *grid points* $a_1 + k\delta$ for $k = 1,..., n$, where $b_1 = a_1 + (n+1)\delta$, as illustrated in Figure 8.2. The function θ is evaluated at each of the n grid points. Let $\hat{\lambda}$ be a grid point having the smallest value of θ. If θ is strictly quasiconvex, it follows that a minimum of θ lies in the interval $[\hat{\lambda} - \delta, \hat{\lambda} + \delta]$.

Choice of the Grid Length δ

We see that the interval of uncertainty $[a_1, b_1]$ is reduced, after n functional evaluations, to an interval of length 2δ. Noting that $n = [(b_1 - a_1)/\delta] - 1$, if we desire a small final interval of uncertainty, then a large number n of function evaluations must be made. One technique that is often used to reduce the computational effort is to utilize a large grid size first and then switch to a finer grid size.

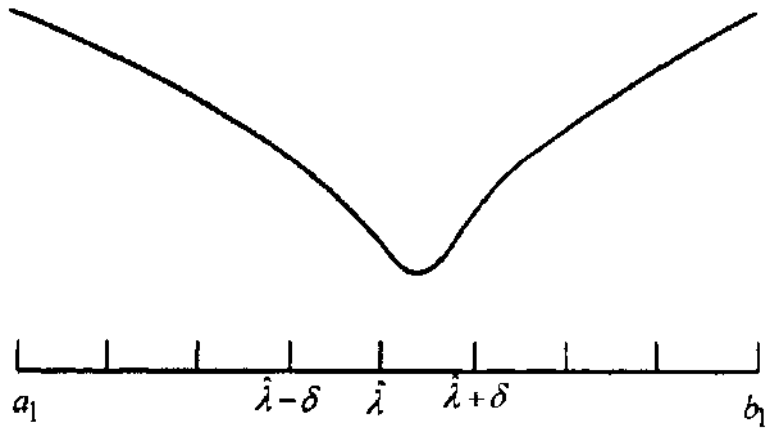


Figure 8.2 Uniform search

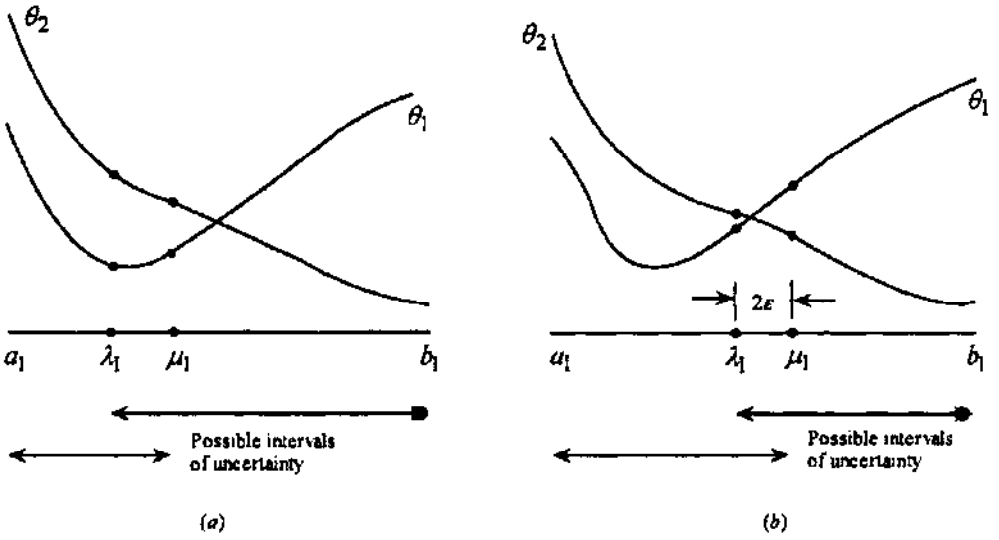


Figure 8.3 Possible intervals of uncertainty.

Sequential Search

As may be expected, more efficient procedures that utilize the information generated at the previous iterations in placing the subsequent iterate can be devised. Here, we discuss the following *sequential search procedures*: dichotomous search, the golden section method, and the Fibonacci method.

Dichotomous Search

Consider $\theta: R \to R$ to be minimized over the interval $[a_1, b_1]$. Suppose that θ is strictly quasiconvex. Obviously, the smallest number of functional evaluations that is needed to reduce the interval of uncertainty is two. In Figure 8.3 we consider the location of the two points λ_1 and μ_1. In Figure 8.3*a*, for $\theta = \theta_1$, note that $\theta(\lambda_1) < \theta(\mu_1)$; and hence, by Theorem 8.1.1, the new interval of uncertainty is $[a_1, \mu_1]$. However, for $\theta = \theta_2$, note that $\theta(\lambda_1) > \theta(\mu_1)$; hence, by Theorem 8.1.1 the new interval of uncertainty is $[\lambda_1, b_1]$. Thus, depending on the function θ, the length of the new interval of uncertainty is equal to $\mu_1 - a_1$ or $b_1 - \lambda_1$.

Note, however, that we do not know, *a priori*, whether $\theta(\lambda_1) < \theta(\mu_1)$ or $\theta(\lambda_1) > \theta(\mu_1)$.* Thus, the *optimal strategy* is to place λ_1 and μ_1 in such a way as to guard against the worst possible outcome, that is, to minimize the maximum of $\mu_1 - a_1$ and $b_1 - \lambda_1$. This can be accomplished by placing λ_1 and μ_1 at the midpoint of the interval $[a_1, b_1]$. If we do this, however, we would have only one trial point and would not be able to reduce the interval of

* If the equality $\theta(\lambda_1) = \theta(\mu_1)$ is true, then the interval of uncertainty can be reduced further to $[\lambda_1, \mu_1]$. It may be noted, however, that exact equality is quite unlikely to occur in practice.

uncertainty. Therefore, as shown in Figure 8.3*b*, λ_1 and μ_1 are placed symmetrically, each at a distance $\varepsilon > 0$ from the midpoint. Here, $\varepsilon > 0$ is a scalar that is sufficiently small so that the new length of uncertainty, $\varepsilon + (b_1 - a_1)/2$, is close enough to the theoretical optimal value of $(b_1 - a_1)/2$ and, in the meantime, would make the functional evaluations $\theta(\lambda_1)$ and $\theta(\mu_1)$ distinguishable.

In dichotomous search, we place each of the first two observations, λ_1 and μ_1, symmetrically at a distance ε from the midpoint $(a_1 + b_1)/2$. Depending on the values of θ at λ_1 and μ_1, a new interval of uncertainty is obtained. The process is then repeated by placing two new observations.

Summary of the Dichotomous Search Method

Following is a summary of the dichotomous method for minimizing a strictly quasiconvex function θ over the interval $[a_1, b_1]$.

Initialization Step Choose the distinguishability constant, $2\varepsilon > 0$, and the allowable final length of uncertainty, $\ell > 0$. Let $[a_1, b_1]$ be the initial interval of uncertainty, let $k = 1$, and go to the Main Step.

Main Step

1. If $b_k - a_k < \ell$, stop; the minimum point lies in the interval $[a_k, b_k]$. Otherwise, consider λ_k and μ_k defined below, and go to Step 2.

$$\lambda_k = \frac{a_k + b_k}{2} - \varepsilon, \qquad \mu_k = \frac{a_k + b_k}{2} + \varepsilon.$$

2. If $\theta(\lambda_k) < \theta(\mu_k)$, let $a_{k+1} = a_k$ and $b_{k+1} = \mu_k$. Otherwise, let $a_{k+1} = \lambda_k$ and $b_{k+1} = b_k$. Replace k by $k + 1$, and go to Step 1.

Note that the length of uncertainty at the beginning of iteration $k + 1$ is given by

$$(b_{k+1} - a_{k+1}) = \frac{1}{2^k}(b_1 - a_1) + 2\varepsilon\left(1 - \frac{1}{2^k}\right).$$

This formula can be used to determine the number of iterations needed to achieve the desired accuracy. Since each iteration requires two observations, the formula can also be used to determine the number of observations.

Golden Section Method

To compare the various line search procedures, the following reduction ratio will be of use:

$$\frac{\text{length of interval of uncertainty after } v \text{ observations are taken}}{\text{length of interval of uncertainty before taking the observations}}.$$

Obviously, more efficient schemes correspond to small ratios. In dichotomous search, the reduction ratio above is approximately $(0.5)^{v/2}$. We now describe the more efficient golden section method for minimizing a strictly quasiconvex function, whose reduction ratio is given by $(0.618)^{v-1}$.

At a general iteration k of the golden section method, let the interval of uncertainty be $[a_k, b_k]$. By Theorem 8.1.1, the new interval of uncertainty $[a_{k+1}, b_{k+1}]$ is given by $[\lambda_k, b_k]$ if $\theta(\lambda_k) > \theta(\mu_k)$ and by $[a_k, \mu_k]$ if $\theta(\lambda_k) \leq \theta(\mu_k)$. The points λ_k and μ_k are selected such that the following hold true.

1. The length of the new interval of uncertainty $b_{k+1} - a_{k+1}$ does not depend on the outcome of the kth iteration, that is, on whether $\theta(\lambda_k) > \theta(\mu_k)$ or $\theta(\lambda_k) \leq \theta(\mu_k)$. Therefore, we must have $b_k - \lambda_k = \mu_k - a_k$. Thus, if λ_k is of the form

$$\lambda_k = a_k + (1-\alpha)(b_k - a_k), \tag{8.1}$$

where $\alpha \in (0,1)$, μ_k must be of the form

$$\mu_k = a_k + \alpha(b_k - a_k) \tag{8.2}$$

so that

$$b_{k+1} - a_{k+1} = \alpha(b_k - a_k).$$

2. As λ_{k+1} and μ_{k+1} are selected for the purpose of a new iteration, either λ_{k+1} coincides with μ_k or μ_{k+1} coincides with λ_k. If this can be realized, then during iteration $k+1$, only one extra observation is needed. To illustrate, consider Figure 8.4 and the following two cases.

Case 1: $\theta(\lambda_k) > \theta(\mu_k)$. In this case, $a_{k+1} = \lambda_k$ and $b_{k+1} = b_k$. To satisfy $\lambda_{k+1} = \mu_k$, and applying (8.1) with k replaced by $k+1$, we get

$$\mu_k = \lambda_{k+1} = a_{k+1} + (1-\alpha)(b_{k+1} - a_{k+1}) = \lambda_k + (1-\alpha)(b_k - \lambda_k).$$

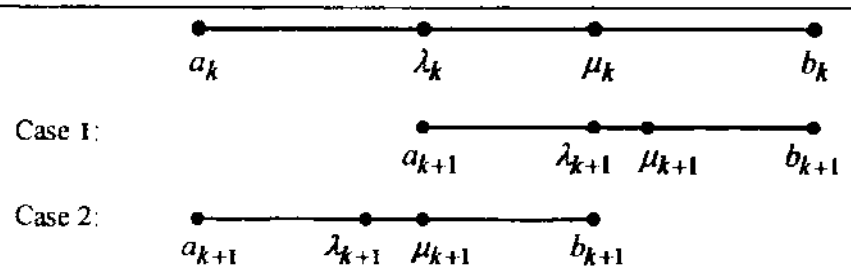


Figure 8.4 Golden section rule.

Substituting the expressions of λ_k and μ_k from (8.1) and (8.2) into the above equation, we get $\alpha^2 + \alpha - 1 = 0$.

Case 2: $\theta(\lambda_k) \leq \theta(\mu_k)$. In this case, $a_{k+1} = a_k$ and $b_{k+1} = \mu_k$. To satisfy $\mu_{k+1} = \lambda_k$, and applying (8.2) with k replaced by $k + 1$, we get

$$\lambda_k = \mu_{k+1} = a_{k+1} + \alpha(b_{k+1} - a_{k+1}) = a_k + \alpha(\mu_k - a_k).$$

Noting (8.1) and (8.2), the above equation gives $\alpha^2 + \alpha - 1 = 0$. The roots of the equation $\alpha^2 + \alpha - 1 = 0$ are $\alpha \cong 0.618$ and $\alpha \cong -1.618$. Since α must be in the interval (0, 1), then $\alpha \cong 0.618$. To summarize, if at iteration k, μ_k and λ_k are chosen according to (8.1) and (8.2), where $\alpha = 0.618$, then the interval of uncertainty is reduced by a factor of 0.618. At the first iteration, two observations are needed at λ_1 and μ_1, but at each subsequent iteration only one evaluation is needed, since either $\lambda_{k+1} = \mu_k$ or $\mu_{k+1} = \lambda_k$.

Summary of the Golden Section Method

Following is a summary of the golden section method for minimizing a strictly quasiconvex function over the interval $[a_1, b_1]$.

Initialization Step Choose an allowable final length of uncertainty $\ell > 0$. Let $[a_1, b_1]$ be the initial interval of uncertainty, and let $\lambda_1 = a_1 + (1 - \alpha)(b_1 - a_1)$ and $\mu_1 = a_1 + \alpha(b_1 - a_1)$, where $\alpha = 0.618$. Evaluate $\theta(\lambda_1)$ and $\theta(\mu_1)$, let $k = 1$, and go to the Main Step.

Main Step

1. If $b_k - a_k < \ell$, stop; the optimal solution lies in the interval $[a_k, b_k]$. Otherwise, if $\theta(\lambda_k) > \theta(\mu_k)$, go to Step 2; and if $\theta(\lambda_k) \leq \theta(\mu_k)$, go to Step 3.
2. Let $a_{k+1} = \lambda_k$ and $b_{k+1} = b_k$. Furthermore, let $\lambda_{k+1} = \mu_k$, and let $\mu_{k+1} = a_{k+1} + \alpha(b_{k+1} - a_{k+1})$. Evaluate $\theta(\mu_{k+1})$ and go to Step 4.
3. Let $a_{k+1} = a_k$ and $b_{k+1} = \mu_k$. Furthermore, let $\mu_{k+1} = \lambda_k$, and let $\lambda_{k+1} = a_{k+1} + (1 - \alpha)(b_{k+1} - a_{k+1})$. Evaluate $\theta(\lambda_{k+1})$ and go to Step 4.
4. Replace k by $k + 1$ and go to Step 1.

8.1.2 Example

Consider the following problem:

$$\begin{aligned} &\text{Minimize} && \lambda^2 + 2\lambda \\ &\text{subject to} && -3 \leq \lambda \leq 5. \end{aligned}$$

Clearly, the function θ to be minimized is strictly quasiconvex, and the initial interval of uncertainty is of length 8. We reduce this interval of uncertainty to one whose length is at most 0.2. The first two observations are located at

$$\lambda_1 = -3 + 0.382(8) = 0.056, \qquad \mu_1 = -3 + 0.618(8) = 1.944.$$

Note that $\theta(\lambda_1) < \theta(\mu_1)$. Hence, the new interval of uncertainty is [–3, 1.944]. The process is repeated, and the computations are summarized in Table 8.1. The values of θ that are computed at each iteration are indicated by an asterisk. After eight iterations involving nine observations, the interval of uncertainty is [–1.112, –0.936], so that the minimum can be estimated to be the midpoint –1.024. Note that the true minimum is in fact –1.0.

Fibonacci Search

The Fibonacci method is a line search procedure for minimizing a strictly quasiconvex function θ over a closed bounded interval. Similar to the golden section method, the Fibonacci search procedure makes two functional evaluations at the first iteration and then only one evaluation at each of the subsequent iterations. However, the procedure differs from the golden section method in that the reduction of the interval of uncertainty varies from one iteration to another.

The procedure is based on the Fibonacci sequence $\{F_v\}$, defined as follows:

$$\begin{aligned} F_{v+1} &= F_v + F_{v-1}, \quad v = 1, 2, \ldots \\ F_0 &= F_1 = 1. \end{aligned} \tag{8.3}$$

The sequence is therefore 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, ... At iteration k, suppose that the interval of uncertainty is $[a_k, b_k]$. Consider the two points λ_k and μ_k given below, where n is the total number of functional evaluations planned.

$$\lambda_k = a_k + \frac{F_{n-k-1}}{F_{n-k+1}}(b_k - a_k), \qquad k = 1, \ldots, n-1 \tag{8.4}$$

Table 8.1 Summary of Computations for the Golden Section Method

Iteration k	a_k	b_k	λ_k	μ_k	$\theta(\lambda_k)$	$\theta(\mu_k)$
1	–3.000	5.000	0.056	1.944	0.115*	7.667*
2	–3.000	1.944	–1.112	0.056	–0.987*	0.115
3	–3.000	0.056	–1.832	–1.112	–0.308*	–0.987
4	–1.832	0.056	–1.112	–0.664	–0.987	–0.887*
5	–1.832	–0.664	–1.384	–1.112	–0.853*	–0.987
6	–1.384	–0.664	–1.112	–0.936	–0.987	–0.996*
7	–1.112	–0.664	–0.936	–0.840	–0.996	–0.974*
8	–1.112	–0.840	–1.016	–0.936	–1.000*	–0.996
9	–1.112	–0.936				

$$\mu_k = a_k + \frac{F_{n-k}}{F_{n-k+1}}(b_k - a_k), \qquad k = 1,\ldots, n-1. \tag{8.5}$$

By Theorem 8.1.1, the new interval of uncertainty $[a_{k+1}, b_{k+1}]$ is given by $[\lambda_k, b_k]$ if $\theta(\lambda_k) > \theta(\mu_k)$ and is given by $[a_k, \mu_k]$ if $\theta(\lambda_k) \leq \theta(\mu_k)$. In the former case, noting (8.4) and letting $v = n - k$ in (8.3), we get

$$\begin{aligned} b_{k+1} - a_{k+1} &= b_k - \lambda_k \\ &= b_k - a_k - \frac{F_{n-k-1}}{F_{n-k+1}}(b_k - a_k) \\ &= \frac{F_{n-k}}{F_{n-k+1}}(b_k - a_k). \end{aligned} \tag{8.6}$$

In the latter case, noting (8.5), we get

$$b_{k+1} - a_{k+1} = \mu_k - a_k = \frac{F_{n-k}}{F_{n-k+1}}(b_k - a_k). \tag{8.7}$$

Thus, in either case, the interval of uncertainty is reduced by the factor F_{n-k}/F_{n-k+1}.

We now show that at iteration $k + 1$, either $\lambda_{k+1} = \mu_k$ or $\mu_{k+1} = \lambda_k$, so that only one functional evaluation is needed. Suppose that $\theta(\lambda_k) > \theta(\mu_k)$. Then, by Theorem 8.1.1, $a_{k+1} = \lambda_k$ and $b_{k+1} = b_k$. Thus, applying (8.4) with k replaced by $k + 1$, we get

$$\begin{aligned} \lambda_{k+1} &= a_{k+1} + \frac{F_{n-k-2}}{F_{n-k}}(b_{k+1} - a_{k+1}) \\ &= \lambda_k + \frac{F_{n-k-2}}{F_{n-k}}(b_k - \lambda_k). \end{aligned}$$

Substituting for λ_k from (8.4), we get

$$\lambda_{k+1} = a_k + \frac{F_{n-k-1}}{F_{n-k+1}}(b_k - a_k) + \frac{F_{n-k-2}}{F_{n-k}}\left(1 - \frac{F_{n-k-1}}{F_{n-k+1}}\right)(b_k - a_k).$$

Letting $v = n - k$ in (8.3), it follows that $1 - (F_{n-k-1}/F_{n-k+1}) = F_{n-k}/F_{n-k+1}$. Substituting in the above equation, we get

$$\lambda_{k+1} = a_k + \frac{F_{n-k-1} + F_{n-k-2}}{F_{n-k+1}}(b_k - a_k).$$

Now let $v = n - k - 1$ in (8.3), and noting (8.5) it follows that

$$\lambda_{k+1} = a_k + \frac{F_{n-k}}{F_{n-k+1}}(b_k - a_k) = \mu_k.$$

Similarly, if $\theta(\lambda_k) \leq \theta(\mu_k)$, the reader can easily verify that $\mu_{k+1} = \lambda_k$. Thus, in either case, only one observation is needed at iteration $k + 1$.

To summarize, at the first iteration, two observations are made, and at each subsequent iteration, only one observation is necessary. Thus, at the end of iteration $n - 2$, we have completed $n - 1$ functional evaluations. Furthermore, for $k = n - 1$, it follows from (8.4) and (8.5) that $\lambda_{n-1} = \mu_{n-1} = (1/2)(a_{n-1} + b_{n-1})$. Since either $\lambda_{n-1} = \mu_{n-2}$ or $\mu_{n-1} = \lambda_{n-2}$, theoretically no new observations are to be made at this stage. However, in order to reduce the interval of uncertainty further, the last observation is placed slightly to the right or to the left of the midpoint $\lambda_{n-1} = \mu_{n-1}$, so that $(1/2)(b_{n-1} - a_{n-1})$ is the length of the final interval of uncertainty $[a_n, b_n]$.

Choosing the Number of Observations

Unlike the dichotomous search method and the golden section procedure, the Fibonacci method requires that the total number of observations n be chosen beforehand. This is because placement of the observations is given by (8.4) and (8.5) and, hence, is dependent on n. From (8.6) and (8.7), the length of the interval of uncertainty is reduced at iteration k by the factor F_{n-k}/F_{n-k+1}. Hence, at the end of $n - 1$ iterations, where n total observations have been made, the length of the interval of uncertainty is reduced from $b_1 - a_1$ to $b_n - a_n = (b_1 - a_1)/F_n$. Therefore, n must be chosen such that $(b_1 - a_1)/F_n$ reflects the accuracy required.

Summary of the Fibonacci Search Method

The following is a summary of the Fibonacci search method for minimizing a strictly quasiconvex function over the interval $[a_1, b_1]$.

Initialization Step Choose an allowable final length of uncertainty $\ell > 0$ and a distinguishability constant $\varepsilon > 0$. Let $[a_1, b_1]$ be the initial interval of uncertainty, and choose the number of observations n to be taken such that $F_n > (b_1 - a_1)/\ell$. Let $\lambda_1 = a_1 + (F_{n-2}/F_n)(b_1 - a_1)$ and $\mu_1 = a_1 + (F_{n-1}/F_n)(b_1 - a_1)$. Evaluate $\theta(\lambda_1)$ and $\theta(\mu_1)$, let $k = 1$, and go to the Main Step.

Main Step

1. If $\theta(\lambda_k) > \theta(\mu_k)$, go to Step 2; and if $\theta(\lambda_k) \leq \theta(\mu_k)$, go to Step 3.
2. Let $a_{k+1} = \lambda_k$ and $b_{k+1} = b_k$. Furthermore, let $\lambda_{k+1} = \mu_k$, and let $\mu_{k+1} = a_{k+1} + (F_{n-k-1}/F_{n-k})(b_{k+1} - a_{k+1})$. If $k = n - 2$, go to Step 5; otherwise, evaluate $\theta(\mu_{k+1})$ and go to Step 4.

3. Let $a_{k+1} = a_k$ and $b_{k+1} = \mu_k$. Furthermore, let $\mu_{k+1} = \lambda_k$, and let $\lambda_{k+1} = a_{k+1} + (F_{n-k-2} / F_{n-k})(b_{k+1} - a_{k+1})$. If $k = n - 2$, go to Step 5; otherwise, evaluate $\theta(\lambda_{k+1})$ and go to Step 4.
4. Replace k by $k + 1$ and go to Step 1.
5. Let $\lambda_n = \lambda_{n-1}$ and $\mu_n = \lambda_{n-1} + \varepsilon$. If $\theta(\lambda_n) > \theta(\mu_n)$, let $a_n = \lambda_n$ and $b_n = b_{n-1}$. Otherwise, if $\theta(\lambda_n) \leq \theta(\mu_n)$, let $a_n = a_{n-1}$ and $b_n = \lambda_n$. Stop; the optimal solution lies in the interval $[a_n, b_n]$.

8.1.3 Example

Consider the following problem:

$$\begin{aligned} &\text{Minimize} && \lambda^2 + 2\lambda \\ &\text{subject to} && -3 \leq \lambda \leq 5. \end{aligned}$$

Note that the function is strictly quasiconvex on the interval and that the true minimum occurs at $\lambda = -1$. We reduce the interval of uncertainty to one whose length is, at most, 0.2. Hence, we must have $F_n > 8/0.2 = 40$, so that $n = 9$. We adopt the distinguishability constant $\varepsilon = 0.01$.

The first two observations are located at

$$\lambda_1 = -3 + \frac{F_7}{F_9}(8) = 0.054545, \qquad \mu_1 = -3 + \frac{F_8}{F_9}(8) = 1.945454.$$

Note that $\theta(\lambda_1) < \theta(\mu_1)$. Hence, the new interval of uncertainty is [–3.000000, 1.945454]. The process is repeated, and the computations are summarized in Table 8.2. The values of θ that are computed at each iteration are indicated by an asterisk. Note that at $k = 8$, $\lambda_k = \mu_k = \lambda_{k-1}$, so that no functional evaluations are needed at this stage. For $k = 9$, $\lambda_k = \lambda_{k-1} = -0.963636$ and $\mu_k = \lambda_k + \varepsilon = -0.953636$. Since $\theta(\mu_k) > \theta(\lambda_k)$, the final interval of uncertainty $[a_9, b_9]$ is [–1.109091, –0.963636], whose length ℓ is 0.145455. We approximate the minimum to be the midpoint –1.036364. Note from Example 8.1.2 that with the same number of observations $n = 9$, the golden section method gave a final interval of uncertainty whose length is 0.176.

Comparison of Derivative-Free Line Search Methods

Given a function θ that is strictly quasiconvex on the interval $[a_1, b_1]$, obviously each of the methods discussed in this section will yield a point λ in a finite number of steps such that $|\lambda - \bar{\lambda}| \leq \ell$, where ℓ is the length of the final interval of uncertainty and $\bar{\lambda}$ is the minimum point over the interval. In particular, given the length ℓ of the final interval of uncertainty, which reflects the desired degree

Table 8.2 Summary of Computations for the Fibonacci Search Method

Iteration k	a_k	b_k	λ_k	μ_k	$\theta(\lambda_k)$	$\theta(\mu_k)$
1	–3.000000	5.000000	0.054545	1.945454	0.112065*	7.675699*
2	–3.000000	1.945454	–1.109091	0.054545	–0.988099*	0.112065
3	–3.000000	0.054545	–1.836363	–1.109091	–0.300497*	–0.988099
4	–1.836363	0.054545	–1.109091	–0.672727	–0.988099	–0.892892*
5	–1.836363	–0.672727	–1.399999	–1.109091	–0.840001*	–0.988099
6	–1.399999	–0.672727	–1.109091	–0.963636	–0.988099	–0.998677*
7	–1.109091	–0.672727	–0.963636	–0.818182	–0.998677	–0.966942*
8	–1.109091	–0.818182	–0.963636	–0.963636	–0.998677	–0.998677
9	–1.109091	–0.963636	–0.963636	–0.953636	–0.998677	–0.997850*

of accuracy, the required number of observations n can be computed as the smallest positive integer satisfying the following relationships.

Uniform search method: $$n \geq \frac{b_1 - a_1}{\ell/2} - 1.$$

Dichotomous search method: $$(1/2)^{n/2} \leq \frac{\ell}{b_1 - a_1}.$$

Golden section method: $$(0.618)^{n-1} \leq \frac{\ell}{b_1 - a_1}.$$

Fibonacci search method: $$F_n \geq \frac{b_1 - a_1}{\ell}.$$

From the above expressions, we see that the number of observations needed is a function of the ratio $(b_1 - a_1)/\ell$. Hence, for a fixed ratio $(b_1 - a_1)/\ell$, the smaller the number of observations required, the more efficient is the algorithm. It should be evident that the most efficient algorithm is the Fibonacci method, followed by the golden section procedure, the dichotomous search method, and finally the uniform search method.

Also note that for n large enough, $1/F_n$ is asymptotic to $(0.618)^{n-1}$, so that the Fibonacci search method and the golden section are almost identical. It is worth mentioning that among the derivative-free methods that minimize strict quasiconvex functions over a closed bounded interval, the Fibonacci search method is the most efficient in that it requires the smallest number of observations for a given reduction in the length of the interval of uncertainty.

General Functions

The procedures discussed above all rely on the strict quasiconvexity assumption. In many problems this assumption does not hold true, and in any case, it cannot

be verified easily. One way to handle this difficulty, especially if the initial interval of uncertainty is large, is to divide it into smaller intervals, find the minimum over each subinterval and choose the smallest of the minima over the subintervals. (A more refined global optimization scheme could also be adopted; see the Notes and References section.) Alternatively, one can simply apply the method assuming strict quasiconvexity and allow the procedure to converge to some local minimum solution.

8.2 Line Search Using Derivatives

In the preceding section we discussed several line search procedures that use functional evaluations. In this section we discuss the bisection search method and Newton's method, both of which need derivative information.

Bisection Search Method

Suppose that we wish to minimize a function θ over a closed and bounded interval. Furthermore, suppose that θ is pseudoconvex and hence, differentiable. At iteration k, let the interval of uncertainty be $[a_k, b_k]$. Suppose that the derivative $\theta'(\lambda_k)$ is known, and consider the following three possible cases:

1. If $\theta'(\lambda_k) = 0$, then, by the pseudoconvexity of θ, λ_k is a minimizing point.
2. If $\theta'(\lambda_k) > 0$, then, for $\lambda > \lambda_k$ we have $\theta'(\lambda_k)(\lambda - \lambda_k) > 0$; and by the pseudoconvexity of θ it follows that $\theta(\lambda) \geq \theta(\lambda_k)$. In other words, the minimum occurs to the left of λ_k, so that the new interval of uncertainty $[a_{k+1}, b_{k+1}]$ is given by $[a_k, \lambda_k]$.
3. If $\theta'(\lambda_k) < 0$, then, for $\lambda < \lambda_k$, $\theta'(\lambda_k)(\lambda - \lambda_k) > 0$, so that $\theta(\lambda) \geq \theta(\lambda_k)$. Thus, the minimum occurs to the right of λ_k, so that the new interval of uncertainty $[a_{k+1}, b_{k+1}]$ is given by $[\lambda_k, b_k]$.

The position of λ_k in the interval $[a_k, b_k]$ must be chosen so that the maximum possible length of the new interval of uncertainty is minimized. That is, λ_k must be chosen so as to minimize the maximum of $\lambda_k - a_k$ and $b_k - \lambda_k$. Obviously, the optimal location of λ_k is the midpoint $(1/2)(a_k + b_k)$.

To summarize, at any iteration k, θ' is evaluated at the midpoint of the interval of uncertainty. Based on the value of θ', we either stop or construct a new interval of uncertainty whose length is half that at the previous iteration. Note that this procedure is very similar to the dichotomous search method except that at each iteration, only one derivative evaluation is required, as opposed to two functional evaluations for the dichotomous search method. However, the latter is akin to a finite difference derivative evaluation.

Convergence of the Bisection Search Method

Note that the length of the interval of uncertainty after n observations is equal to $(1/2)^n(b_1 - a_1)$, so that the method converges to a minimum point within any desired degree of accuracy. In particular, if the length of the final interval of uncertainty is fixed at ℓ, then n must be chosen to be the smallest integer such that $(1/2)^n \leq \ell/(b_1 - a_1)$.

Summary of the Bisection Search Method

We now summarize the bisection search procedure for minimizing a pseudoconvex function θ over a closed and bounded interval.

Initialization Step Let $[a_1, b_1]$ be the initial interval of uncertainty, and let ℓ be the allowable final interval of uncertainty. Let n be the smallest positive integer such that $(1/2)^n \leq \ell/(b_1 - a_1)$. Let k = 1 and go to the Main Step.

Main Step

1. Let $\lambda_k = (1/2)(a_k + b_k)$ and evaluate $\theta'(\lambda_k)$. If $\theta'(\lambda_k) = 0$, stop; λ_k is an optimal solution. Otherwise, go to Step 2 if $\theta'(\lambda_k) > 0$, and go to Step 3 if $\theta'(\lambda_k) < 0$.
2. Let $a_{k+1} = a_k$ and $b_{k+1} = \lambda_k$. Go to Step 4.
3. Let $a_{k+1} = \lambda_k$ and $b_{k+1} = b_k$. Go to Step 4.
4. If $k = n$, stop; the minimum lies in the interval $[a_{n+1}, b_{n+1}]$. Otherwise, replace k by $k + 1$ and repeat Step 1.

8.2.1 Example

Consider the following problem:

$$\begin{aligned} &\text{Minimize} \quad \lambda^2 + 2\lambda \\ &\text{subject to} \quad -3 \leq \lambda \leq 6. \end{aligned}$$

Suppose that we want to reduce the interval of uncertainty to an interval whose length ℓ is less than or equal to 0.2. Hence, the number of observations n satisfying $(1/2)^n \leq \ell/(b_1 - a_1) = 0.2/9 = 0.0222$ is given by $n = 6$. A summary of the computations using the bisection search method is given in Table 8.3. Note that the final interval of uncertainty is [–1.0313, –0.8907], so that the minimum could be taken as the midpoint, –0.961.

Table 8.3 Summary of Computations for the Bisection Search Method

Iteration k	a_k	b_k	λ_k	$\theta'(\lambda_k)$
1	–3.0000	6.0000	1.5000	5.0000
2	–3.0000	1.5000	–0.7500	0.5000
3	–3.0000	–0.7500	–1.8750	–1.7500
4	–1.8750	–0.7500	–1.3125	–0.6250
5	–1.3125	–0.7500	–1.0313	–0.0625
6	–1.0313	–0.7500	–0.8907	0.2186
7	–1.0313	–0.8907		

Newton's Method

Newton's method is based on exploiting the quadratic approximation of the function θ at a given point λ_k. This quadratic approximation q is given by

$$q(\lambda) = \theta(\lambda_k) + \theta'(\lambda_k)(\lambda - \lambda_k) + \frac{1}{2}\theta''(\lambda_k)(\lambda - \lambda_k)^2.$$

The point λ_{k+1} is taken to be the point where the derivative of q is equal to zero. This yields $\theta'(\lambda_k) + \theta''(\lambda_k)(\lambda_{k+1} - \lambda_k) = 0$, so that

$$\lambda_{k+1} = \lambda_k - \frac{\theta'(\lambda_k)}{\theta''(\lambda_k)}. \tag{8.8}$$

The procedure is terminated when $|\lambda_{k+1} - \lambda_k| < \varepsilon$, or when $|\theta'(\lambda_k)| < \varepsilon$, where ε is a prespecified termination scalar.

Note that the above procedure can only be applied for twice differentiable functions. Furthermore, the procedure is well defined only if $\theta''(\lambda_k) \neq 0$ for each k.

8.2.2 Example

Consider the function θ:

$$\theta(\lambda) = \begin{cases} 4\lambda^3 - 3\lambda^4 & \text{if } \lambda \geq 0 \\ 4\lambda^3 + 3\lambda^4 & \text{if } \lambda < 0. \end{cases}$$

Note that θ is twice differentiable everywhere. We apply Newton's method, starting from two different points. In the first case, $\lambda_1 = 0.40$; and as shown in Table 8.4, the procedure produces the point 0.002807 after six iterations. The reader can verify that the procedure indeed converges to the stationary point $\lambda = 0$. In the second case, $\lambda_1 = 0.60$, and the procedure oscillates between the points 0.60 and –0.60, as shown in Table 8.5.

Table 8.4 Summary of Computations for Newton's Method Starting from $\lambda_1 = 0.4$

Iteration k	λ_k	$\theta'(\lambda_k)$	$\theta''(\lambda_k)$	λ_{k+1}
1	0.400000	1.152000	3.840000	0.100000
2	0.100000	0.108000	2.040000	0.047059
3	0.047059	0.025324	1.049692	0.022934
4	0.022934	0.006167	0.531481	0.011331
5	0.11331	0.001523	0.267322	0.005634
6	0.005634	0.000379	0.134073	0.002807

Convergence of Newton's Method

The method of Newton, in general, does not converge to a stationary point starting with an arbitrary initial point. Observe that, in general, Theorem 7.2.3 cannot be applied as a result of the unavailability of a descent function. However, as shown in Theorem 8.2.3, if the starting point is sufficiently close to a stationary point, then a suitable descent function can be devised so that the method converges.

8.2.3 Theorem

Let $\theta: R \to R$ be continuously twice differentiable. Consider Newton's algorithm defined by the map $A(\lambda) = \lambda - \theta'(\lambda)/\theta''(\lambda)$. Let $\bar{\lambda}$ be such that $\theta'(\bar{\lambda}) = 0$ and $\theta''(\bar{\lambda}) \neq 0$. Let the starting point λ_1 be sufficiently close to $\bar{\lambda}$ so that there exist scalars k_1, $k_2 > 0$ with $k_1 k_2 < 1$ such that

1. $$\frac{1}{|\theta''(\lambda)|} \leq k_1$$

2. $$\frac{\left|\theta(\bar{\lambda}) - \theta'(\lambda) - \theta''(\lambda)(\bar{\lambda} - \lambda)\right|}{(\bar{\lambda} - \lambda)} \leq k_2$$

for each λ satisfying $|\lambda - \bar{\lambda}| \leq |\lambda_1 - \bar{\lambda}|$. Then the algorithm converges to $\bar{\lambda}$.

Table 8.5 Summary of Computations for Newton's Method Starting from $\lambda_1 = 0.6$

Iteration k	λ_k	$\theta'(\lambda_k)$	$\theta''(\lambda_k)$	λ_{k+1}
1	0.600	1.728	1.440	–0.600
2	–0.600	1.728	–1.440	0.600
3	0.600	1.728	1.440	–0.600
4	–0.600	1.728	–1.440	0.600

Proof

Let the solution set $\Omega = \{\bar{\lambda}\}$, and let $X = \{\lambda : |\lambda - \bar{\lambda}| \leq |\lambda_1 - \bar{\lambda}|\}$. We prove convergence by using Theorem 7.2.3. Note that X is compact and that the map **A** is closed on X. We now show that $\alpha(\lambda) = |\lambda - \bar{\lambda}|$ is indeed a descent function. Let $\lambda \in X$ and suppose that $\lambda \neq \bar{\lambda}$. Let $\hat{\lambda} \in \mathbf{A}(\lambda)$. Then, by the definition of **A** and since $\theta'(\bar{\lambda}) = 0$, we get

$$\hat{\lambda} - \bar{\lambda} = (\lambda - \bar{\lambda}) - \frac{1}{\theta''(\lambda)}[\theta'(\lambda) - \theta'(\bar{\lambda})]$$

$$= \frac{1}{\theta''(\lambda)}[\theta'(\bar{\lambda}) - \theta'(\lambda) - \theta''(\lambda)(\bar{\lambda} - \lambda)].$$

Noting the hypothesis of the theorem, it then follows that

$$|\hat{\lambda} - \bar{\lambda}| = \frac{1}{|\theta''(\lambda)|} \frac{|\theta'(\bar{\lambda}) - \theta'(\lambda) - \theta''(\lambda)(\bar{\lambda} - \lambda)|}{|\bar{\lambda} - \lambda|} |\lambda - \bar{\lambda}| \leq k_1 k_2 |\lambda - \bar{\lambda}| < |\lambda - \bar{\lambda}|.$$

Therefore, α is indeed a descent function, and the result follows immediately by the corollary to Theorem 7.2.3.

8.3 Some Practical Line Search Methods

In the preceding two sections we presented various line search methods that either use or do not use derivative-based information. Of these, the golden section method (which is a limiting form of Fibonacci's search method) and the bisection method are often applied in practice, sometimes in combination with other methods. However, these methods follow a restrictive pattern of placing subsequent observations and do not accelerate the process by adaptively exploiting information regarding the shape of the function. Although Newton's method tends to do this, it requires second-order derivative information and is not globally convergent. The quadratic-fit technique described in the discussion that follows adopts this philosophy, enjoys global convergence under appropriate assumptions such as pseudoconvexity, and is a very popular method.

We remark here that quite often in practice, whenever ill-conditioning effects are experienced with this method or if it fails to make sufficient progress during an iteration, a switchover to the bisection search procedure is typically made. Such a check for a possible switchover is referred to as a *safeguard technique*

Quadratic-Fit Line Search

Suppose that we are trying to minimize a continuous, strictly quasiconvex function $\theta(\lambda)$ over $\lambda \geq 0$, and assume that we have three points $0 \leq \lambda_1 < \lambda_2 < \lambda_3$ such that $\theta_1 \geq \theta_2$ and $\theta_2 \leq \theta_3$, where $\theta_j \equiv \theta(\lambda_j)$ for $j = 1, 2, 3$. Note that if $\theta_1 = \theta_2 = \theta_3$, then, by the nature of θ, it is easily verified that these must all be minimizing solutions (see Exercise 8.12). Hence, suppose that in addition, at least one of the inequalities $\theta_1 > \theta_2$ and $\theta_2 < \theta_3$ holds true. Let us refer to the conditions satisfied by these three points as the *three-point pattern* (TPP). To begin with, we can take $\lambda_1 = 0$ and examine a trial point $\hat{\lambda}$, which might be the step length of a line search at the previous iteration of an algorithm. Let $\hat{\theta} = \theta(\hat{\lambda})$. If $\hat{\theta} \geq \theta_1$, we can set $\lambda_3 = \hat{\lambda}$ and find λ_2 by repeatedly halving the interval $[\lambda_1, \lambda_3]$ until a TPP is obtained. On the other hand, if $\hat{\theta} < \theta_1$, we can set $\lambda_2 = \hat{\lambda}$ and find λ_3 by doubling the interval $[\lambda_1, \lambda_2]$ until a TPP is obtained.

Now, given the three points (λ_j, θ_j), $j = 1, 2, 3$, we can fit a quadratic curve passing through these points and find its minimizer $\bar{\lambda}$, which must lie in (λ_1, λ_3) by the TPP (see Exercise 8.11). There are three cases to consider. Denote $\bar{\theta} = \theta(\bar{\lambda})$ and let λ_{new} denote the revised set of three points $(\lambda_1, \lambda_2, \lambda_3)$ found as follows:

Case 1: $\bar{\lambda} > \lambda_2$ (see Figure 8.5). If $\bar{\theta} \geq \theta_2$, then we let $\lambda_{\text{new}} = (\lambda_1, \lambda_2, \bar{\lambda})$. On the other hand, if $\bar{\theta} \leq \theta_2$, we let $\lambda_{\text{new}} = (\lambda_2, \bar{\lambda}, \lambda_3)$. (Note that in case $\bar{\theta} = \theta_2$, either choice is permissible.)

Case 2: $\bar{\lambda} < \lambda_2$. Similar to Case 1, if $\bar{\theta} \geq \theta_2$, we let $\lambda_{\text{new}} = (\bar{\lambda}, \lambda_2, \lambda_3)$; and if $\bar{\theta} \leq \theta_2$, we let $\lambda_{\text{new}} = (\lambda_1, \bar{\lambda}, \lambda_2)$.

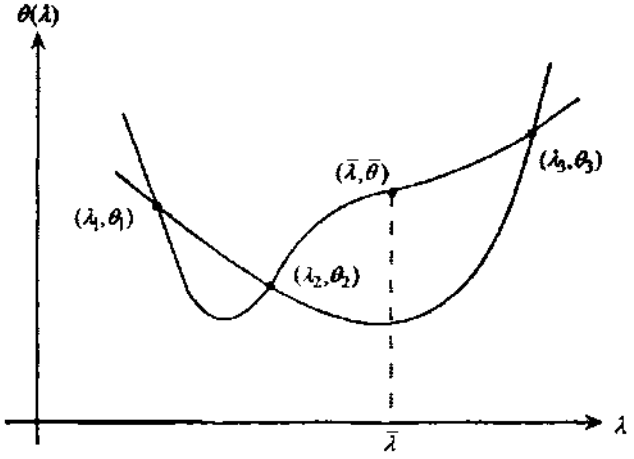


Figure 8.5 Quadratic-fit line search.

Case 3: $\bar{\lambda} = \lambda_2$. In this case, we do not have a distinct point to obtain a new TPP. If $\lambda_3 - \lambda_1 \leq \varepsilon$ for some convergence tolerance $\varepsilon > 0$, we stop with λ_2 as the prescribed step length. Otherwise, we place a new observation point $\bar{\lambda}$ at a distance $\varepsilon/2$ away from λ_2 toward λ_1 or λ_3, whichever is further. This yields the situation described by Case 1 or 2 above, and hence, a new set of points defining λ_{new} may be obtained accordingly.

Again, with respect to λ_{new} if $\theta_1 = \theta_2 = \theta_3$ or if $(\lambda_3 - \lambda_1) \leq \varepsilon$ [or if $\theta'(\lambda_2) = 0$ in the differentiable case, or if some other termination criterion such as an acceptable step length in an inexact line search as described next below holds true], then we terminate this process. Otherwise, λ_{new} satisfies the TPP, and the above procedure can be repeated using this new TPP.

Note that in Case 3 of the above procedure, when $\bar{\lambda} = \lambda_2$ the step of placing an observation in the vicinity of λ_2 is akin to evaluating $\theta'(\lambda_2)$ when θ is differentiable. In fact, if we assume that θ is pseudoconvex and continuously twice differentiable, and we apply a modified version of the foregoing procedure that uses derivatives to represent limiting cases of coincident observation values as described in Exercise 8.13, we can use Theorem 7.2.3 to demonstrate convergence to an optimal solution, given a starting solution $(\lambda_1, \lambda_2, \lambda_3)$ that satisfies the TPP.

Inexact Line Searches: Armijo's Rule

Very often in practice, we cannot afford the luxury of performing an exact line search because of the expense of excessive function evaluations, even if we terminate with some small accuracy tolerance $\varepsilon > 0$. On the other hand, if we sacrifice accuracy, we might impair the convergence of the overall algorithm that iteratively employs such a line search. However, if we adopt a line search that guarantees a sufficient degree of accuracy or descent in the function value in a well-defined sense, this might induce the overall algorithm to converge. Below we describe one popular definition of an acceptable step length known as Armijo's rule and refer the reader to the Notes and References section and Exercise 8.8 for other such exact line search criteria.

Armijo's rule is driven by two parameters, $0 < \varepsilon < 1$ and $\alpha > 1$, which manage the acceptable step length from being too large or too small, respectively. (Typical values are $\varepsilon = 0.2$ and $\alpha = 2$.) Suppose that we are minimizing some differentiable function f: $R^n \rightarrow R$ at the point $\bar{\mathrm{x}} \in R^n$ in the direction $\mathrm{d} \in R^n$, where $\nabla f(\bar{\mathrm{x}})^t \mathrm{d} < 0$. Hence, d is a descent direction. Define the line search function θ: $R \rightarrow R$ as $\theta(\lambda) = f(\bar{\mathrm{x}} + \lambda \mathrm{d})$ for $\lambda \geq 0$. Then the first-order approximation of θ at $\lambda = 0$ is given by $\theta(0) + \lambda\theta'(0)$ and is depicted in Figure 8.6. Now define

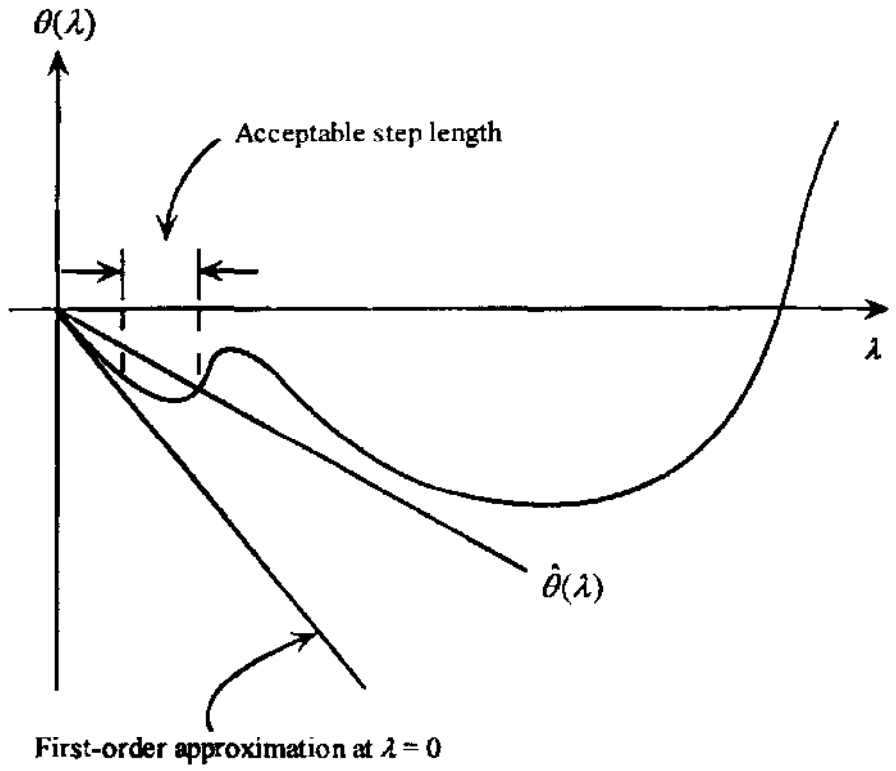


Figure 8.6 Armijo's rule.

$$\hat{\theta}(\lambda) = \theta(0) + \lambda\varepsilon\theta'(0) \qquad \text{for } \lambda \geq 0.$$

A step length $\bar{\lambda}$ is considered to be acceptable provided that $\theta(\bar{\lambda}) \leq \hat{\theta}(\bar{\lambda})$. However, to prevent $\bar{\lambda}$ from being too small, Armijo's rule also requires that $\theta(\alpha\bar{\lambda}) > \hat{\theta}(\alpha\bar{\lambda})$. This gives an acceptable range for $\bar{\lambda}$, as shown in Figure 8.6.

Frequently, Armijo's rule is adopted in the following manner. A fixed-step-length parameter $\bar{\lambda}$ is chosen. If $\theta(\bar{\lambda}) \leq \hat{\theta}(\bar{\lambda})$, then either $\bar{\lambda}$ is itself selected as the step size, or $\bar{\lambda}$ is sequentially doubled (assuming that $\alpha = 2$) to find the largest integer $t \geq 0$ for which $\theta(2^t\bar{\lambda}) \leq \hat{\theta}(2^t\bar{\lambda})$. On the other hand, if $\theta(\bar{\lambda}) > \hat{\theta}(\bar{\lambda})$, then $\bar{\lambda}$ is sequentially halved to find the smallest integer $t \geq 1$ for which $\theta(\bar{\lambda}/2^t) \leq \hat{\theta}(\bar{\lambda}/2^t)$. Later, in Section 8.6, we analyze the convergence of a steepest descent algorithm that employs such a line search criterion.

8.4 Closedness of the Line Search Algorithmic Map

In the preceding three sections we discussed several procedures for minimizing a function of one variable. Since the one-dimensional search is a component of most nonlinear programming algorithms, we show in this section that line search procedures define a closed map.

Consider the line search problem to minimize $\theta(\lambda)$ subject to $\lambda \in L$, where $\theta(\lambda) = f(\mathbf{x} + \lambda\mathbf{d})$ and L is a closed interval in R. This line search problem can be defined by the algorithmic map $\mathbf{M}$: $R^n \times R^n \to R^n$, defined by

$\mathbf{M}(\mathbf{x}, \mathrm{d}) = \{\mathbf{y} : \mathbf{y} = \mathbf{x} + \bar{\lambda}\mathbf{d}$ for some $\bar{\lambda} \in L$ and $f(\mathbf{y}) \leq f(\mathbf{x} + \lambda\mathbf{d})$ for each $\lambda \in L\}$.

Note that **M** is generally a point-to-set map because there can be more than one minimizing point **y**. Theorem 8.4.1 shows that the map **M** is closed. Thus, if the map **D** that determines the direction d is also closed, then, by Theorem 7.3.2 or its corollaries, if the additional conditions stated hold true, the overall algorithmic map **A** = **MD** is closed.

8.4.1 Theorem

Let f: $R^n \to R$, and let L be a closed interval in R. Consider the line search map **M**: $R^n \times R^n \to R^n$ defined by

$\mathbf{M}(\mathbf{x}, \mathrm{d}) = \{\mathbf{y} : \mathbf{y} = \mathbf{x} + \bar{\lambda}\mathbf{d}$ for some $\bar{\lambda} \in L$ and $f(\mathbf{y}) \leq f(\mathbf{x} + \lambda\mathbf{d})$ for each $\lambda \in L\}$.

If f is continuous at **x** and $\mathbf{d} \neq \mathbf{0}$, then **M** is closed at (**x**, **d**).

Proof

Suppose that $(\mathbf{x}_k, \mathrm{d}_k) \to (\mathbf{x}, \mathrm{d})$ and that $\mathbf{y}_k \to \mathbf{y}$, where $\mathbf{y}_k \in \mathbf{M}(\mathbf{x}_k, \mathrm{d}_k)$. We want to show that $\mathbf{y} \in \mathbf{M}(\mathbf{x}, \mathrm{d})$. First, note that $\mathbf{y}_k = \mathbf{x}_k + \lambda_k \mathbf{d}_k$, where $\lambda_k \in L$. Since $\mathbf{d} \neq \mathbf{0}$, $\mathbf{d}_k \neq \mathbf{0}$ for k large enough, and hence $\lambda_k = \|\mathbf{y}_k - \mathbf{x}_k\| / \|\mathrm{d}_k\|$. Taking the limit as $k \to \infty$, then $\lambda_k \to \bar{\lambda}$, where $\bar{\lambda} = \|\mathbf{y} - \mathbf{x}\| / \|\mathrm{d}\|$, and hence, $\mathbf{y} = \mathbf{x} + \bar{\lambda}\mathbf{d}$. Furthermore, since $\lambda_k \in L$ for each k, and since L is closed, $\bar{\lambda} \in L$. Now let $\lambda \in L$ and note that $f(\mathbf{y}_k) \leq f(\mathbf{x}_k + \lambda\mathbf{d}_k)$ for all k. Taking the limit as $k \to \infty$ and noting the continuity of f, we conclude that $f(\mathbf{y}) \leq f(\mathbf{x} + \lambda\mathbf{d})$. Thus, $\mathbf{y} \in \mathbf{M}(\mathbf{x}, \mathbf{d})$, and the proof is complete.

In nonlinear programming, line search is typically performed over one of the following intervals:

$$L = \{\lambda : \lambda \in R\}$$

$$L = \{\lambda : \lambda \geq 0\}$$

$$L = \{\lambda : a \leq \lambda \leq b\}.$$

In each of the above cases, L is closed and the theorem applies.

In Theorem 8.4.1 we required that the vector **d** be nonzero. Example 8.4.2 presents a case in which **M** is not closed if **d** = **0**. In most cases the direction vector d is nonzero over points outside the solution set Ω. Thus, **M** is closed at these points, and Theorem 7.2.3 can be applied to prove convergence.

8.4.2 Example

Consider the following problem:

Minimize $(x-2)^4$.

Here $f(x)=(x-2)^4$. Now consider the sequence $(x_k, d_k)=(1/k, 1/k)$. Clearly, x_k converges to $x = 0$ and d_k converges to $d = 0$. Consider the line search map **M** defined in Theorem 8.4.1, where $L=\{\lambda : \lambda \geq 0\}$. The point y_k is obtained by solving the problem to minimize $f(x_k + \lambda d_k)$ subject to $\lambda \geq 0$. The reader can verify that $y_k = 2$ for all k, so its limit y equals 2. Note, however, that **M**(0, 0) = {0}, so that $y \notin$ **M**(0, 0). This shows that **M** is not closed.

8.5 Multidimensional Search Without Using Derivatives

In this section we consider the problem of minimizing a function f of several variables without using derivatives. The methods described here proceed in the following manner. Given a vector **x**, a suitable direction **d** is first determined, and then f is minimized from **x** in the direction **d** by one of the techniques discussed earlier in this chapter.

Throughout the book we are required to solve a line search problem of the form to minimize $f(\mathbf{x}+\lambda \mathrm{d})$ subject to $\lambda \in L$, where L is typically of the form $L = R$, $L = \{\lambda : \lambda \geq 0\}$ or $L = \{\lambda : a \leq \lambda \leq b\}$. In the statements of the algorithms, for the purpose of simplicity we have assumed that a minimizing point $\bar{\lambda}$ exists. However, this may not be the case. Here, the optimal objective value of the line search problem may be unbounded, or else the optimal objective value may be finite but not achieved at any particular λ. In the first case, the original problem is unbounded and we may stop. In the latter case, λ could be chosen as $\bar{\lambda}$ such that $f(\mathbf{x}+\bar{\lambda}\mathrm{d})$ is sufficiently close to the value $\inf\{f(\mathbf{x} + \lambda \mathbf{d}) : \lambda \in L\}$.

Cyclic Coordinate Method

This method uses the coordinate axes as the search directions. More specifically, the method searches along the directions $\mathrm{d}_1,\ldots,\mathbf{d}_n$, where d_j is a vector of zeros except for a 1 at the jth position. Thus, along the search direction d_j, the variable x_j is changed while all other variables are kept fixed. The method is illustrated schematically in Figure 8.7 for the problem of Example 8.5.1.

Note that we are assuming here that the minimization is done in order over the dimensions 1,..., n at each iteration. In a variant known as the *Aitken double sweep method*, the search is conducted by minimizing over the dimensions 1,..., n and then back over the dimensions $n - 1$, $n - 2$,..., 1. This requires $n - 1$ line searches per iteration. Accordingly, if the function to be minimized is differentiable and its gradient is available, the *Gauss-Southwell* variant recommends that one select that coordinate direction for minimizing at each step that has the largest magnitude of the partial derivative component. These types of sequential one-dimensional minimizations are sometimes

referred to as *Gauss-Seidel iterations*, based on the *Gauss-Seidel method* for solving systems of equations.

Summary of the Cyclic Coordinate Method

We summarize below the cyclic coordinate method for minimizing a function of several variables without using any derivative information. As we show shortly, if the function is differentiable, the method converges to a stationary point.

As discussed in Section 7.2, several criteria could be used for terminating the algorithm. In the statement of the algorithm below, the termination criterion $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| < \varepsilon$ is used. Obviously, any of the other criteria could be used to stop the procedure.

Initialization Step Choose a scalar $\varepsilon > 0$ to be used for terminating the algorithm, and let $\mathbf{d}_1, ..., \mathbf{d}_n$ be the coordinate directions. Choose an initial point $\mathbf{x}_1$, let $\mathbf{y}_1 = \mathbf{x}_1$, let $k = j = 1$, and go to the Main Step.

Main Step

1. Let λ_j be an optimal solution to the problem to minimize $f(\mathbf{y}_j + \lambda \mathbf{d}_j)$ subject to $\lambda \in R$, and let $\mathbf{y}_{j+1} = \mathbf{y}_j + \lambda_j \mathbf{d}_j$. If $j < n$, replace j by $j + 1$, and repeat Step 1. Otherwise, if $j = n$, go to Step 2.
2. Let $\mathbf{x}_{k+1} = \mathbf{y}_{n+1}$. If $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| < \varepsilon$, then stop. Otherwise, let $\mathbf{y}_1 = \mathbf{x}_{k+1}$, let $j = 1$, replace k by $k + 1$, and go to Step 1.

8.5.1 Example

Consider the following problem:

$$\text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2.$$

Note that the optimal solution to this problem is (2, 1) with objective value equal to zero. Table 8.6 gives a summary of computations for the cyclic coordinate method starting from the initial point (0, 3). Note that at each iteration, the vectors $\mathbf{y}_2$ and $\mathbf{y}_3$ are obtained by performing a line search in the directions (1, 0) and (0, 1), respectively. Also note that significant progress is made during the first few iterations, whereas much slower progress is made during later iterations. After seven iterations, the point (2.22, 1.11), whose objective value is 0.0023, is reached.

In Figure 8.7 the contours of the objective function are given, and the points generated above by the cyclic coordinate method are shown. Note that at later iterations, slow progress is made because of the short orthogonal movements along the valley indicated by the dashed lines. Later, we analyze the convergence rate of steepest descent methods. The cyclic coordinate method tends to exhibit a performance characteristic over the n coordinate line searches similar to that of an iteration of the steepest descent method.

Table 8.6 Summary of Computations for the Cyclic Coordinate Method

Iteration k	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	j	$\mathbf{d}_j$	$\mathbf{y}_j$	λ_j	$\mathbf{y}_{j+1}$
1	(0.00, 3.00)	1	(1.0, 0.0)	(0.00,3.00)	3.13	(3.13, 3.00)
	52.00	2	(0.0, 1.0)	(3.13,3.00)	–1.44	(3.13, 1.56)
2	(3.13, 1.56)	1	(1.0, 0.0)	(3.13, 1.56)	–0.50	(2.63, 1.56)
	1.63	2	(0.0, 1.0)	(2.63, 1.56)	–0.25	(2.63, 1.31)
3	(2.63, 1.31)	1	(1.0, 0.0)	(2.63, 1.31)	–0.19	(2.44, 1.31)
	0.16	2	(0.0, 1.0)	(2.44, 1.31)	–0.09	(2.44, 1.22)
4	(2.44, 1.22)	1	(1.0, 0.0)	(2.44, 1.22)	–0.09	(2.35, 1.22)
	0.04	2	(0.0, 1.0)	(2.35, 1.22)	–0.05	(2.35, 1.17)
5	(2.35, 1.17)	1	(1.0, 0.0)	(2.35, 1.17)	–0.06	(2.29, 1.17)
	0.015	2	(0.0, 1.0)	(2.29, 1.17)	–0.03	(2.29, 1.14)
6	(2.29, 1.14)	1	(1.0, 0.0)	(2.29, 1.14)	–0.04	(2.25, 1.14)
	0.007	2	(0.0, 1.0)	(2.25, 1.14)	–0.02	(2.25, 1.12)
7	(2.25, 1.12)	1	(1.0, 0.0)	(2.25, 1.12)	–0.03	(2.22, 1.12)
	0.004	2	(0.0, 1.0)	(2.22,1.12)	–0.01	(2.22, 1.11)

Convergence of the Cyclic Coordinate Method

Convergence of the cyclic coordinate method to a stationary point follows immediately from Theorem 7.3.5 under the following assumptions:

1. The minimum of f along any line in R^n is unique.
2. The sequence of points generated by the algorithm is contained in a compact subset of R^n.

Note that the search directions used at each iteration are the coordinate vectors, so that the matrix of search directions $\mathbf{D} = \mathbf{I}$. Obviously, Assumption 1 of Theorem 7.3.5 holds true.

As an alternative approach, Theorem 7.2.3 could have been used to prove convergence after showing that the overall algorithmic map is closed at each $\mathbf{x}$ satisfying $\nabla f(\mathbf{x}) \neq \mathbf{0}$. In this case, the descent function α is taken as f itself, and the solution set is $\Omega = \{\mathbf{x} : \nabla f(\mathbf{x}) = \mathbf{0}\}$.

Acceleration Step

We learned from the foregoing analysis that when applied to a differentiable function, the cyclic coordinate method will converge to a point with zero gradient. In the absence of differentiability, however, the method can stall at a nonoptimal point. As shown in Figure 8.8*a*, searching along any of the

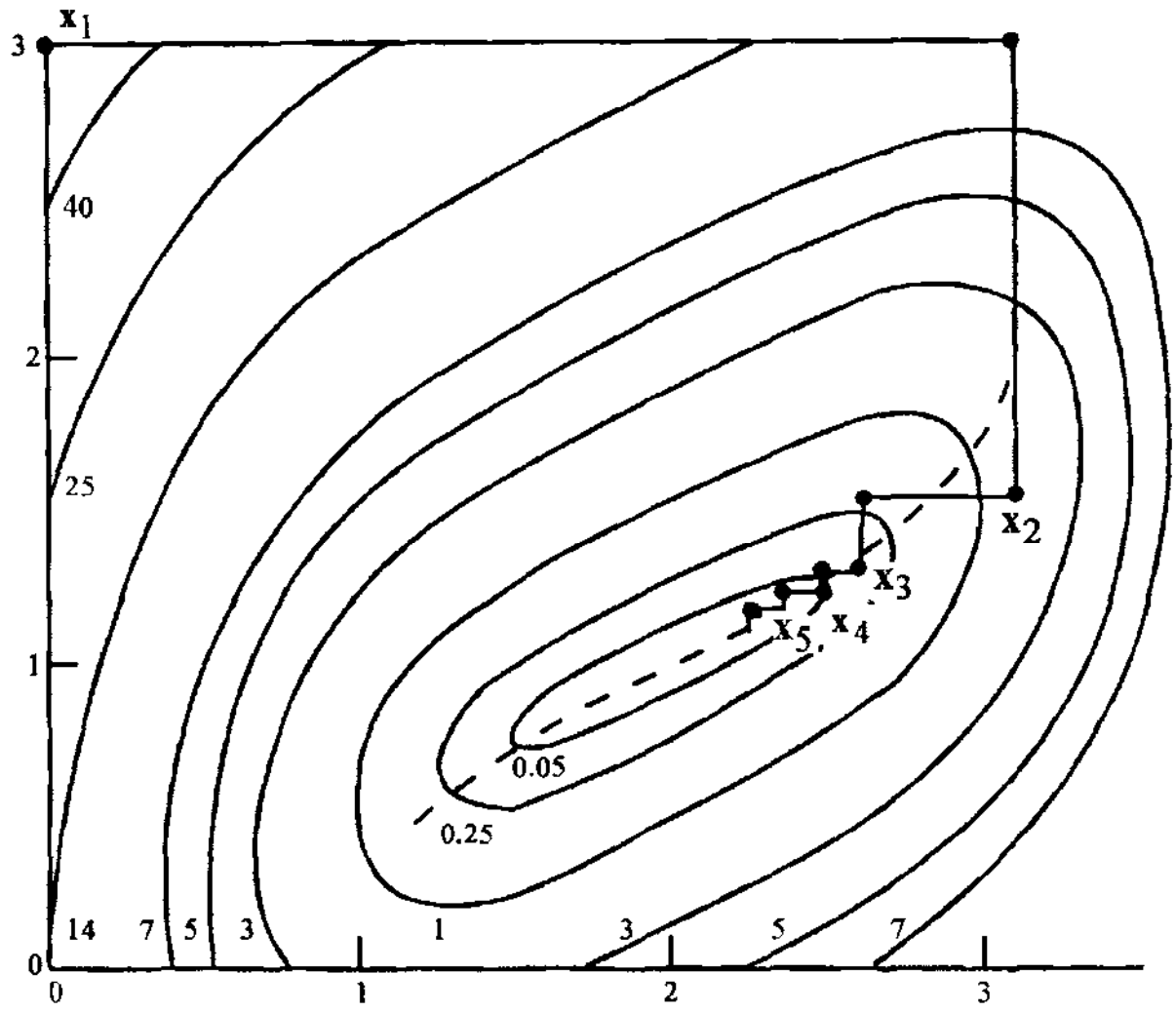


Figure 8.7 Cyclic coordinate method.

coordinate axes at the point $\mathbf{x}_2$ leads to no improvement of the objective function and results in premature termination. The reason for this premature termination is the presence of a sharp-edged valley caused by the nondifferentiability of f. As illustrated in Figure 8.8*b*, this difficulty could possibly be overcome by searching along the direction $\mathbf{x}_2 - \mathbf{x}_1$.

The search along a direction $\mathbf{x}_{k+1} - \mathbf{x}_k$ is frequently used in applying the cyclic coordinate method, even in the case where f is differentiable. The usual rule of thumb is to apply it at every pth iteration. This modification to the cyclic coordinate method frequently accelerates convergence, particularly when the sequence of points generated zigzags along a valley. Such a step is usually referred to as an *acceleration step* or a *pattern search step.*

Method of Hooke and Jeeves

The method of Hooke and Jeeves performs two types of search: exploratory search and pattern search. The first two iterations of the procedure are illustrated in Figure 8.9. Given $\mathbf{x}_1$, an exploratory search along the coordinate directions produces the point $\mathbf{x}_2$. Now a pattern search along the direction $\mathbf{x}_2 - \mathbf{x}_1$ leads to

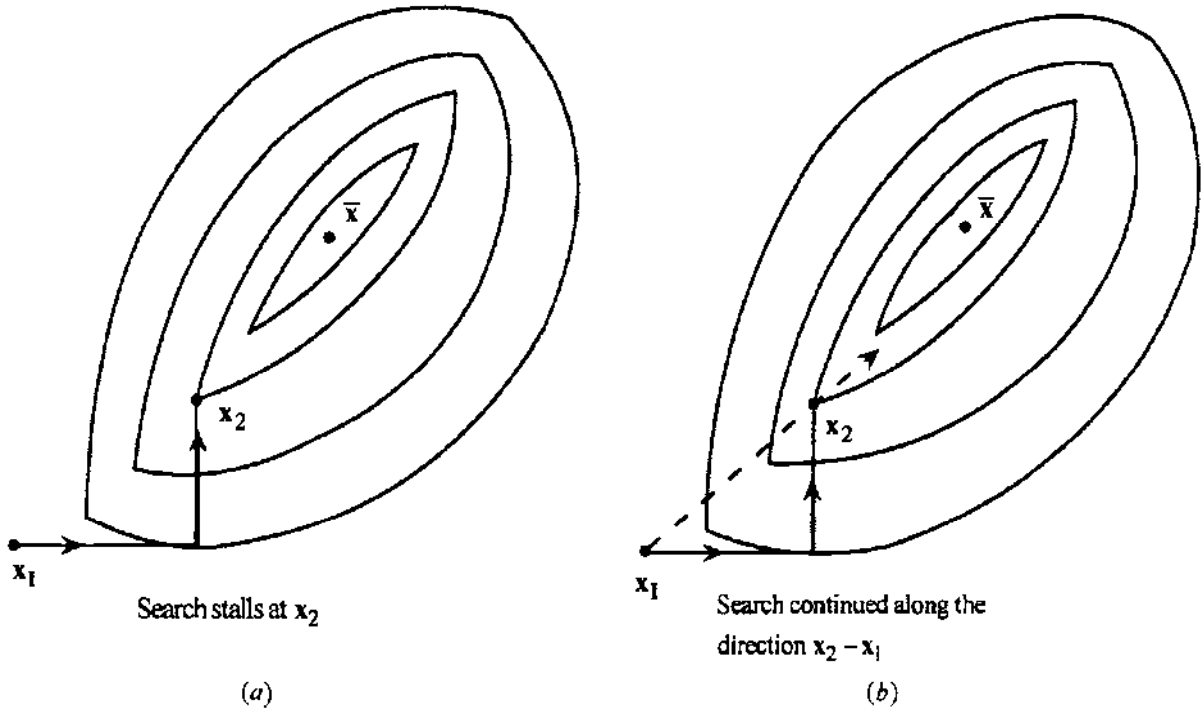


Figure 8.8 Effect of a sharp-edged valley.

the point y. Another exploratory search starting from y gives the point x_3. The next pattern search is conducted along the direction $x_3 - x_2$, yielding y'. The process is then repeated.

Summary of the Method of Hooke and Jeeves Using Line Searches

As originally proposed by Hooke and Jeeves, the method does not perform any line search but rather takes discrete steps along the search directions, as we discuss later. Here we present a continuous version of the method using line searches along the coordinate directions $d_1, \ldots, d_n$ and the pattern direction.

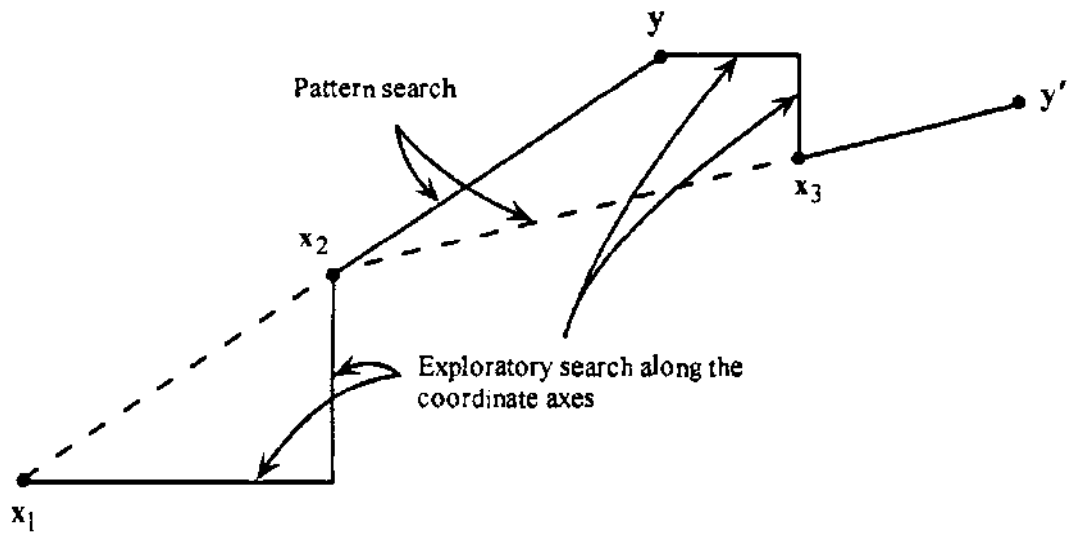


Figure 8.9 Method of Hooke and Jeeves.

Initialization Step Choose a scalar $\varepsilon > 0$ to be used in terminating the algorithm. Choose a starting point $\mathbf{x}_1$, let $\mathbf{y}_1 = \mathbf{x}_1$, let $k = j = 1$, and go to the Main Step.

Main Step

1. Let λ_j be an optimal solution to the problem to minimize $f(\mathbf{y}_j + \lambda \mathbf{d}_j)$ subject to $\lambda \in R$, and let $\mathbf{y}_{j+1} = \mathbf{y}_j + \lambda_j \mathbf{d}_j$. If $j < n$, replace j by $j + 1$, and repeat Step 1. Otherwise, if $j = n$, let $\mathbf{x}_{k+1} = \mathbf{y}_{n+1}$. If $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| < \varepsilon$, stop; otherwise, go to Step 2.

2. Let $\mathbf{d} = \mathbf{x}_{k+1} - \mathbf{x}_k$, and let $\hat{\lambda}$ be an optimal solution to the problem to minimize $f(\mathbf{x}_{k+1} + \lambda \mathbf{d})$ subject to $\lambda \in R$. Let $\mathbf{y}_1 = \mathbf{x}_{k+1} + \hat{\lambda}\mathbf{d}$, let $j = 1$, replace k by $k + 1$, and go to Step 1.

8.5.2 Example

Consider the following problem:

$$\text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2.$$

Note that the optimal solution is (2.00, 1.00) with objective value equal to zero. Table 8.7 summarizes the computations for the method of Hooke and Jeeves, starting from the initial point (0.00, 3.00). At each iteration, an exploratory search along the coordinate directions gives the points $\mathbf{y}_2$ and $\mathbf{y}_3$, and a pattern search along the direction $\mathbf{d} = \mathbf{x}_{k+1} - \mathbf{x}_k$ gives the point $\mathbf{y}_1$, except at iteration $k = 1$, where $\mathbf{y}_1 = \mathbf{x}_1$. Note that four iterations were required to move from the initial point to the optimal point (2.00, 1.00) whose objective value is zero. At this point, $\|\mathbf{x}_5 - \mathbf{x}_4\| = 0.045$, and the procedure is terminated.

Figure 8.10 illustrates the points generated by the method of Hooke and Jeeves using line searches. Note that the pattern search has substantially improved the convergence behavior by moving along a direction that is almost parallel to the valley shown by dashed lines.

Convergence of the Method of Hooke and Jeeves

Suppose that f is differentiable, and let the solution set $\Omega = \{\bar{\mathbf{x}} : \nabla f(\bar{\mathbf{x}}) = \mathbf{0}\}$. Note that each iteration of the method of Hooke and Jeeves consists of an application of the cyclic coordinate method, in addition to a pattern search. Let the cyclic coordinate search be denoted by the map **B** and the pattern search be denoted by the map **C**. Using an argument similar to that of Theorem 7.3.5, it follows that **B** is closed. If the minimum of f along any line is unique and letting $\alpha = f$, then

Table 8.7 Summary of Computations for the Method of Hooke and Jeeves Using Line Searches

Iteration k	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	j	$\mathbf{y}_j$	$\mathbf{d}_j$	λ_j	$\mathbf{y}_{j+1}$	$\mathbf{d}$	$\hat{\lambda}$	$\mathbf{y}_3 + \hat{\lambda}\mathbf{d}$
1	(0.00, 3.00)	1	(0.00, 3.00)	(1.0, 0.0)	3.13	(3.13, 3.00)	—	—	—
	52.00	2	(3.13, 3.00)	(0.0, 1.0)	−1.44	(3.13, 1.56)	(3.13, 1.44)	0.10	(2.82, 1.70)
2	(3.13, 1.56)	1	(2.82, 1.70)	(1.0, 0.0)	−0.12	(2.70, 1.70)	—	—	—
	1.63	2	(2.70, 1.70)	(0.0, 1.0)	−0.35	(2.70, 1.35)	(−0.43, −0.21)	1.50	(2.06, 1.04)
3	(2.70, 1.35)	1	(2.06, 1.04)	(1.0, 0.0)	−0.02	(2.04, 1.04)	—	—	—
	0.24	2	(2.04, 1.04)	(0.0, 1.0)	−0.02	(2.04, 1.02)	(−0.66, −0.33)	0.06	(2.00, 1.00)
4	(2.04, 1.02)	1	(2.00, 1.00)	(1.0, 0.0)	0.00	(2.00, 1.00)	—	—	—
	0.000003	2	(2.00, 1.00)	(0.0, 1.0)	0.00	(2.00, 1.00)			
5	(2.00, 1.00)								
	0.00								

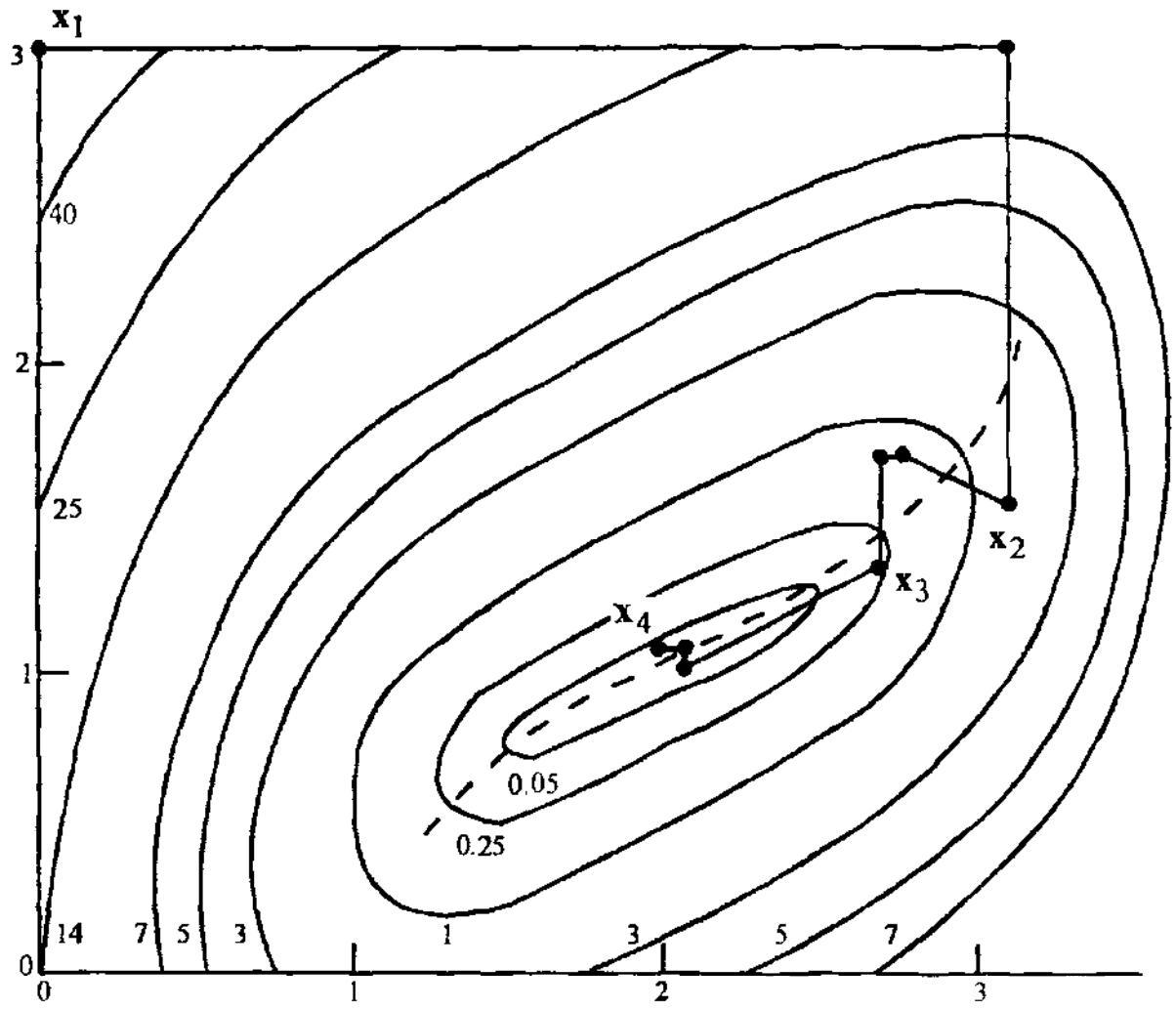


Figure 8.10 Method of Hooke and Jeeves using line searches. Method of Hooke and Jeeves with Discrete Steps

$\alpha(\mathbf{y}) < \alpha(\mathbf{x})$ for $\mathbf{x} \notin \Omega$. By the definition of $\mathbf{C}$, $\alpha(\mathbf{z}) \leq \alpha(\mathbf{y})$ for $\mathbf{z} \in \mathbf{C}(\mathbf{y})$. Assuming that $\Lambda = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}_1)\}$, where $\mathbf{x}_1$ is the starting point, is compact, convergence of the procedure is established by Theorem 7.3.4.

Method of Hooke and Jeeves with Discrete Steps

As mentioned earlier, the method of Hooke and Jeeves, as originally proposed, does not perform line searches but, instead, adopts a simple scheme involving functional evaluations. A summary of the method is given below.

Initialization Step Let $\mathbf{d}_1, \ldots, \mathbf{d}_n$ be the coordinate directions. Choose a scalar $\varepsilon > 0$ to be used for terminating the algorithm. Furthermore, choose an initial step size, $\Delta \geq \varepsilon$, and an acceleration factor, $\alpha > 0$. Choose a starting point $\mathbf{x}_1$, let $\mathbf{y}_1 = \mathbf{x}_1$, let $k = j = 1$, and go to the Main Step.

Main Step

1. If $f(\mathbf{y}_j + \Delta \mathbf{d}_j) < f(\mathbf{y}_j)$, the trial is termed a *success*; let $\mathbf{y}_{j+1} = \mathbf{y}_j + \Delta \mathbf{d}_j$, and go to Step 2. If, however, $f(\mathbf{y}_j + \Delta \mathbf{d}_j) \geq f(\mathbf{y}_j)$, the

trial is deemed a *failure*. In this case, if $f(\mathbf{y}_j - \Delta \mathbf{d}_j) < f(\mathbf{y}_j)$, let $\mathbf{y}_{j+1} = \mathbf{y}_j - \Delta \mathbf{d}_j$, and go to Step 2; if $f(\mathbf{y}_j - \Delta \mathbf{d}_j) \geq f(\mathbf{y}_j)$, let $\mathbf{y}_{j+1} = \mathbf{y}_j$, and go to Step 2.

2. If $j < n$, replace j by $j + 1$, and repeat Step 1. Otherwise, go to Step 3 if $f(\mathbf{y}_{n+1}) < f(\mathbf{x}_k)$, and go to Step 4 if $f(\mathbf{y}_{n+1}) \geq f(\mathbf{x}_k)$.
3. Let $\mathbf{x}_{k+1} = \mathbf{y}_{n+1}$, and let $\mathbf{y}_1 = \mathbf{x}_{k+1} + \alpha(\mathbf{x}_{k+1} - \mathbf{x}_k)$. Replace k by $k + 1$, let $j = 1$, and go to Step 1.
4. If $\Delta \leq \varepsilon$, stop; $\mathbf{x}_k$ is the prescribed solution. Otherwise, replace Δ by $\Delta/2$. Let $\mathbf{y}_1 = \mathbf{x}_k$, $\mathbf{x}_{k+1} = \mathbf{x}_k$, replace k by $k + 1$, let $j = 1$, and repeat Step 1.

The reader may note that steps 1 and 2 above describe an exploratory search. Furthermore, Step 3 is an acceleration step along the direction $\mathbf{x}_{k+1} - \mathbf{x}_k$. Note that a decision whether to accept or reject the acceleration step is not made until after an exploratory search is performed. In Step 4, the step size Δ is reduced. The procedure could easily be modified so that different step sizes are used along the different directions. This is sometimes adopted for the purpose of scaling.

8.5.3 Example

Consider the following problem:

$$\text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2.$$

We solve the problem using the method of Hooke and Jeeves with discrete steps. The parameters α and Δ are chosen as 1.0 and 0.2, respectively. Figure 8.11 shows the path taken by the algorithm starting from (0.0, 3.0). The points generated are numbered sequentially, and the acceleration step that is rejected is shown by the dashed lines. From this particular starting point, the optimal solution is easily reached.

To give a more comprehensive illustration, Table 8.8 summarizes the computations starting from the new initial point (2.0, 3.0). Here (S) denotes that the trial is a success and (F) denotes that the trial is a failure. At the first iteration, and at subsequent iterations whenever $f(\mathbf{y}_3) \geq f(\mathbf{x}_k)$, the vector $\mathbf{y}_1$ is taken as $\mathbf{x}_k$. Otherwise, $\mathbf{y}_1 = 2\mathbf{x}_{k+1} - \mathbf{x}_k$. Note that at the end of iteration $k = 10$, the point (1.70, 0.80) is reached having an objective value 0.02. The procedure is stopped here with the termination parameter $\varepsilon = 0.1$. If a greater degree of accuracy is required, Δ should be reduced to 0.05.

Figure 8.12 illustrates the path taken by the method. The points generated are again numbered sequentially, and dashed lines represent rejected acceleration steps.

Table 8.8 Summary of Computations for the Method of Hooke and Jeeves with Discrete Steps

Iteration k	Δ	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	j	$\mathbf{y}_j$ $f(\mathbf{y}_j)$	$\mathbf{d}_j$	$\mathbf{y}_j + \Delta\mathbf{d}_j$ $f(\mathbf{y}_j + \Delta\mathbf{d}_j)$	$\mathbf{y}_j - \Delta\mathbf{d}_j$ $f(\mathbf{y}_j - \Delta\mathbf{d}_j)$
1	0.2	(2.00, 3.00) 16.00	1	(2.00, 3.00) 16.00	(1.0, 0.0)	(2.20, 3.00) 14.44(S)	
			2	(2.20, 3.00) 14.44	(0.0, 1.0)	(2.20, 3.20) 17.64(F)	(2.20, 2.80) 11.56(S)
2	0.2	(2.20, 2.80) 11.56	1	(2.40, 2.60) 7.87	(1.0, 0.0)	(2.60, 2.60) 6.89(S)	
			2	(2.60, 2.60) 6.89	(0.0, 1.0)	(2.60, 2.80) 9.13(F)	(2.60, 2.40) 4.97(S)
3	0.2	(2.60, 2.40) 4.97	1	(3.00, 2.00) 2.00	(1.0, 0.0)	(3.20, 2.00) 2.71(F)	(2.80, 2.00) 1.85(S)
			2	(2.80, 2.00) 1.85	(0.0, 1.0)	(2.80, 2.20) 2.97(F)	(2.80, 1.80) 1.05(S)
4	0.2	(2.80, 1.80) 1.05	1	(3.00, 1.20) 1.36	(1.0, 0.0)	(3.20, 1.20) 2.71(F)	(2.80, 1.20) 0.57(S)
			2	(2.80, 1.20) 0.57	(0.0, 1.0)	(2.80, 1.40) 0.41(S)	—
5	0.2	(2.80, 1.40) 0.41	1	(2.80, 1.00) 1.05	(1.0, 0.0)	(3.00, 1.00) 2.00(F)	(2.60, 1.00) 0.49(S)
			2	(2.60, 1.00) 0.49	(0.0, 1.0)	(2.60, 1.20) 0.17(S)	—

(continued)

6	0.2	(2.60, 1.20) 0.17	1	(2.40, 1.00) 0.19	(1.0, 0.0)	(2.60, 1.00) 0.49(F)	(2.20, 1.00) 0.04(S)
			2	(2.20, 1.00) 0.04	(0.0, 1.0)	(2.20, 1.20) 0.04(F)	(2.20, 0.80) 0.36(F)
7	0.2	(2.20, 1.00) 0.04	1	(1.80, 0.80) 0.04	(1.0, 0.0)	(2.00, 0.80 0.16(F)	(1.60, 0.80) 0.03(S)
			2	(1.60, 0.80) 0.03	(0.0, 1.0)	(1.60, 1.00) 0.19(F)	(1.60, 0.60) 0.19(F)
8	0.2	(1.60, 0.80) 0.03	1	(1.00, 0.60) 0.67	(1.0, 0.0)	(1.20, 0.60) 0.41(S)	—
			2	(1.20, 0.60) 0.41	(0.0, 1.0)	(1.20, 0.80) 0.57(F)	(1.20, 0.40) 0.57(F)
9	0.1	(1.60, 0.80) 0.03	1	(1.60, 0.80) 0.03	(1.0, 0.0)	(1.70, 0.80) 0.02(S)	—
			2	(1.70, 0.80) 0.02	(0.0, 1.0)	(1.70, 0.90) 0.02(F)	(1.70, 0.70) 0.10(F)
10	0.1	(1.70, 0.80) 0.02	1	(1.80, 0.80) 0.04	(1.0, 0.0)	(1.90, 0.80) 0.09(F)	(1.70, 0.80) 0.02(S)
			2	(1.70, 0.80) 0.02	(0.0, 1.0)	(1.70, 0.90) 0.02(F)	(1.70, 0.70) 0.10(F)

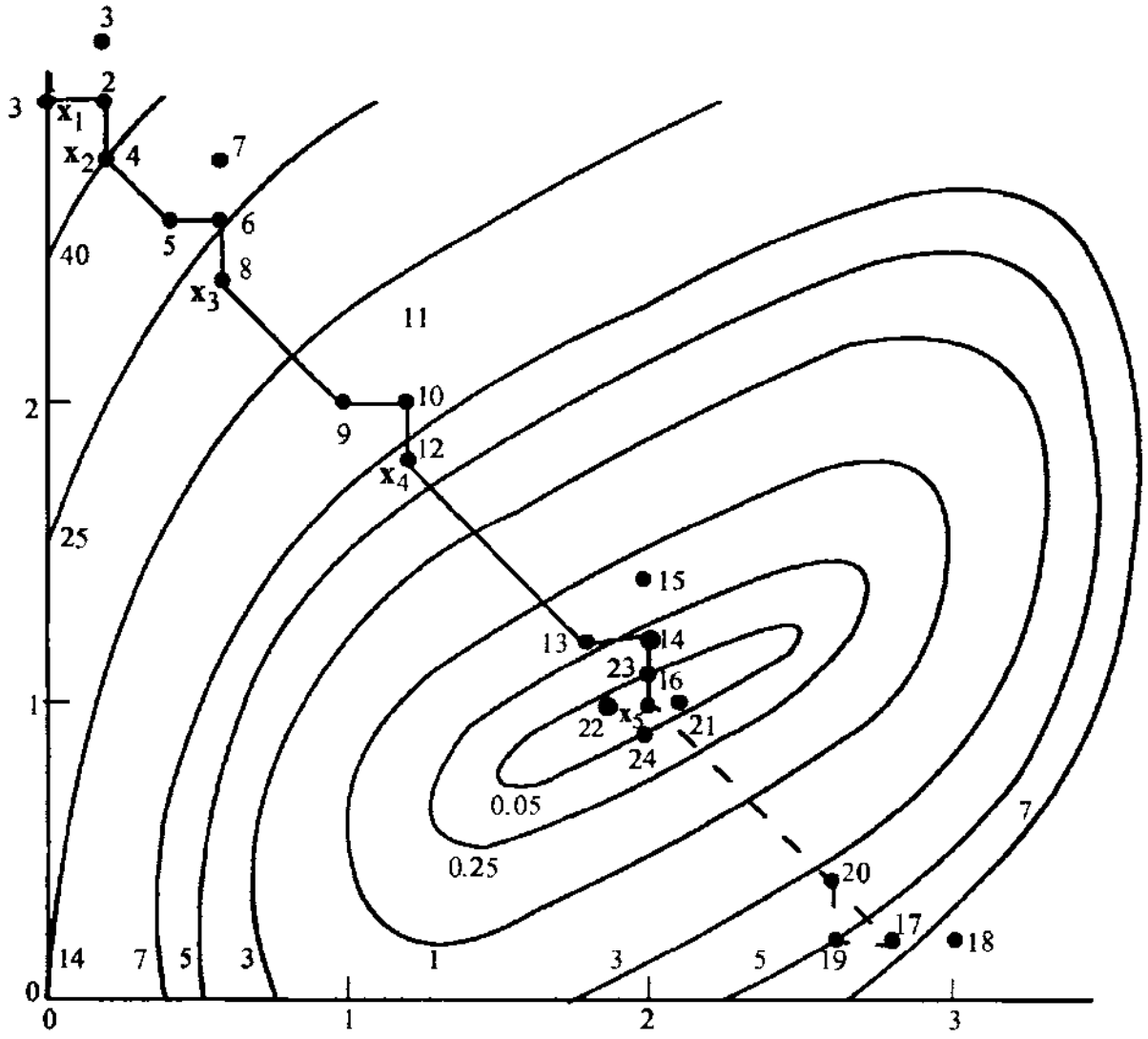


Figure 8.11 Method of Hooke and Jeeves using discrete steps starting from (0.0, 3.0). (The numbers denote the order in which points are generated.)

Method of Rosenbrock

As originally proposed, the method of Rosenbrock does not employ line searches but rather takes discrete steps along the search directions. We present here a continuous version of the method that utilizes line searches. At each iteration, the procedure searches iteratively along n linearly independent and orthogonal directions. When a new point is reached at the end of an iteration, a new set of orthogonal vectors is constructed. In Figure 8.13 the new directions are denoted by $\bar{\mathbf{d}}_1$ and $\bar{\mathbf{d}}_2$.

Construction of the Search Directions

Let $\mathbf{d}_1,...,\mathbf{d}_n$ be linearly independent vectors, each with a norm equal to 1. Furthermore, suppose that these vectors are mutually orthogonal; that is, $\mathbf{d}_i^t\mathbf{d}_j = 0$ for $i \neq j$. Starting from the current vector $\mathbf{x}_k$, the objective function f is

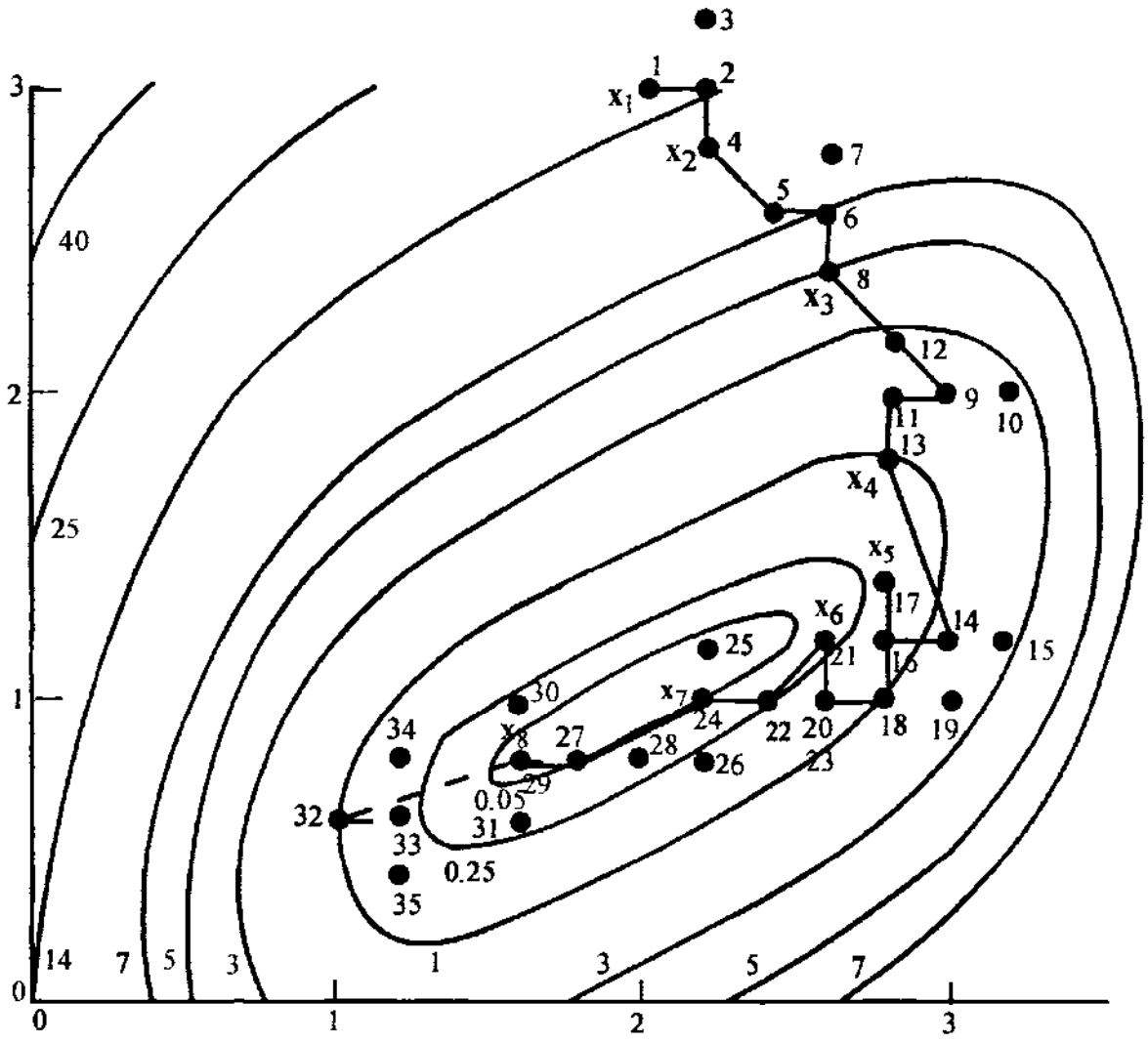


Figure 8.12 Method of Hooke and Jeeves using line searches. (The numbers denote the order in which points are generated.)

minimized along each of the directions iteratively, resulting in the point $\mathbf{x}_{k+1}$. In particular, $\mathbf{x}_{k+1} - \mathbf{x}_k = \sum_{j=1}^{n} \lambda_j \mathbf{d}_j$, where λ_j is the distance moved along $\mathbf{d}_j$. The new collection of directions $\bar{\mathbf{d}}_1, \ldots, \bar{\mathbf{d}}_n$ is formed by the *Gram-Schmidt procedure*, or *orthogonalization procedure*, as follows:

$$\mathbf{a}_j = \begin{cases} \mathbf{d}_j & \text{if } \lambda_j = 0 \\ \sum_{l=j}^{n} \lambda_l \mathbf{d}_l & \text{if } \lambda_j \neq 0 \end{cases}$$

$$\mathbf{b}_j = \begin{cases} \mathbf{a}_j, & j = 1 \\ \mathbf{a}_j - \sum_{i=1}^{j-1} (\mathbf{a}_j^t \bar{\mathbf{d}}_i) \bar{\mathbf{d}}_i, & j \geq 2 \end{cases} \tag{8.9}$$

$$\bar{\mathbf{d}}_j = \frac{\mathbf{b}_j}{\|\mathbf{b}_j\|}.$$

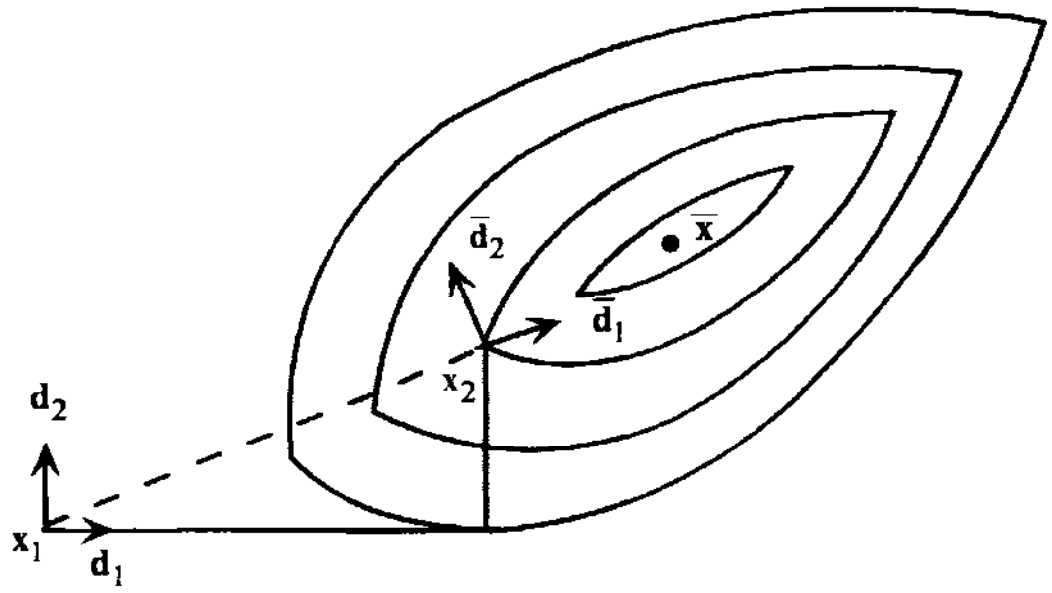


Figure 8.13 Rosenbrock's procedure using discrete steps.

Lemma 8.5.4 shows that the new directions established by the Rosenbrock procedure are indeed linearly independent and orthogonal.

8.5.4 Lemma

Suppose that the vectors $\mathbf{d}_1,...,\mathbf{d}_n$ are linearly independent and mutually orthogonal. Then the directions $\bar{\mathbf{d}}_1,...,\bar{\mathbf{d}}_n$ defined by (8.9) are also linearly independent and mutually orthogonal for any set of $\lambda_1,...,\lambda_n$. Furthermore, if $\lambda_j = 0$, then $\bar{\mathbf{d}}_j = \mathbf{d}_j$.

Proof

We first show that $\mathbf{a}_1,...,\mathbf{a}_n$ are linearly independent. Suppose that $\sum_{j=1}^{n} \mu_j \mathbf{a}_j = \mathbf{0}$. Let $I = \{j : \lambda_j = 0\}$, and let $J(j) = \{i : i \notin I, i \leq j\}$. Noting (8.9), we get

$$\mathbf{0} = \sum_{j=1}^{n} \mu_j \mathbf{a}_j = \sum_{j \in I} \mu_j \mathbf{d}_j + \sum_{j \notin I} \mu_j \left(\sum_{l=j}^{n} \lambda_l \mathbf{d}_l \right)$$

$$= \sum_{j \in I} \mu_j \mathbf{d}_j + \sum_{j \notin I} \left(\lambda_j \sum_{i \in J(j)} \mu_i \right) \mathbf{d}_j.$$

Since $\mathbf{d}_1,...,\mathbf{d}_n$ are linearly independent, $\mu_j = 0$ for $j \in I$ and $\lambda_j \sum_{i \in J(j)} \mu_i = 0$ for $j \notin I$. But $\lambda_j \neq 0$ for $j \notin I$, and hence, $\sum_{i \in J(j)} \mu_i = 0$ for each $j \notin I$. By the definition of $J(j)$, we therefore have $\mu_1 = \cdots = \mu_n = 0$, and hence, $\mathbf{a}_1,...,\mathbf{a}_n$ are linearly independent.

To show that $\mathbf{b}_1,...,\mathbf{b}_n$ are linearly independent, we use the following induction argument. Since $\mathbf{b}_1 = \mathbf{a}_1 \neq \mathbf{0}$, it suffices to show that if $\mathbf{b}_1,...,\mathbf{b}_k$ are linearly independent, then $\mathbf{b}_1,...,\mathbf{b}_k$, $\mathbf{b}_{k+1}$ are also linearly independent. Suppose that $\sum_{j=1}^{k+1}\alpha_j\mathbf{b}_j = 0$. Using the definition of $\mathbf{b}_{k+1}$, in (8.9), we get

$$\begin{aligned}\mathbf{0} &= \sum_{j=1}^{k}\alpha_j\mathbf{b}_j + \alpha_{k+1}\mathbf{b}_{k+1} \\ &= \sum_{j=1}^{k}\left[\alpha_j - \frac{\alpha_{k+1}\left(\mathbf{a}_{k+1}^t\bar{\mathbf{d}}_j\right)}{\left\|\mathbf{b}_j\right\|}\right]\mathbf{b}_j + \alpha_{k+1}\mathbf{a}_{k+1}.\end{aligned} \tag{8.10}$$

From (8.9) it follows that each vector $\mathbf{b}_j$ is a linear combination of $\mathbf{a}_1,...,\mathbf{a}_j$. Since $\mathbf{a}_1,...,\mathbf{a}_{k+1}$ are linearly independent, it follows from (8.10) that $\alpha_{k+1} = 0$. Since $\mathbf{b}_1,...,\mathbf{b}_k$ are assumed linearly independent by the induction hypotheses, from (8.10) we get $\alpha_j - \alpha_{k+1}(\mathbf{a}_{k+1}^t\bar{\mathbf{d}}_j)/\left\|\mathbf{b}_j\right\| = 0$ for $j = 1,...,k$. Since $\alpha_{k+1} = 0$, $\alpha_j = 0$ for each j. This shows that $\mathbf{b}_1,...,\mathbf{b}_{k+1}$ are linearly independent. By the definition of $\bar{\mathbf{d}}_j$, linear independence of $\bar{\mathbf{d}}_1,...,\bar{\mathbf{d}}_n$ is immediate.

Now we establish the orthogonality of $\mathbf{b}_1,...,\mathbf{b}_n$ and hence the orthogonality of $\bar{\mathbf{d}}_1,...,\bar{\mathbf{d}}_n$. From (8.9), $\mathbf{b}_1^t\mathbf{b}_2 = 0$; thus, it suffices to show that if $\mathbf{b}_1,...,\mathbf{b}_k$ are mutually orthogonal, then $\mathbf{b}_1,...,\mathbf{b}_k$, $\mathbf{b}_{k+1}$ are also mutually orthogonal. From (8.10) and noting that $\mathbf{b}_j^t\bar{\mathbf{d}}_i = 0$ for $i \neq j$, it follows that

$$\begin{aligned}\mathbf{b}_j^t\mathbf{b}_{k+1} &= \mathbf{b}_j^t\left[\mathbf{a}_{k+1} - \sum_{i=1}^{k}(\mathbf{a}_{k+1}^t\bar{\mathbf{d}}_i)\bar{\mathbf{d}}_i\right] \\ &= \mathbf{b}_j^t\mathbf{a}_{k+1} - (\mathbf{a}_{k+1}^t\bar{\mathbf{d}}_j)\mathbf{b}_j^t\bar{\mathbf{d}}_j = 0.\end{aligned}$$

Thus, $\mathbf{b}_1,...,\mathbf{b}_{k+1}$ are mutually orthogonal.

To complete the proof, we show that $\bar{\mathbf{d}}_j = \mathbf{d}_j$ if $\lambda_j = 0$. From (8.9), if $\lambda_j = 0$, we get

$$\mathbf{b}_j = \mathbf{d}_j - \sum_{i=1}^{j-1}\frac{1}{\left\|\mathbf{b}_i\right\|}(\mathbf{d}_j^t\mathbf{b}_i)\bar{\mathbf{d}}_i. \tag{8.11}$$

Note that $\mathbf{b}_i$ is a linear combination of $\mathbf{a}_1,...,\mathbf{a}_i$, so that $\mathbf{b}_i = \sum_{r=1}^{i}\beta_{ir}\mathbf{a}_r$. From (8.9), it thus follows that

$$\mathbf{b}_i = \sum_{r \in \mathscr{R}} \beta_{ir} \mathbf{d}_r + \sum_{r \in \bar{\mathscr{R}}} \beta_{ir} \left(\sum_{s=r}^{n} \lambda_s \mathbf{d}_s \right), \tag{8.12}$$

where $\mathscr{R} = \{r : r \leq i,\ \lambda_r = 0\}$ and $\bar{\mathscr{R}} = \{r : r \leq i,\ \lambda_r \neq 0\}$. Consider $i < j$ and note that $\mathbf{d}_j^t \mathbf{d}_v = 0$ for $v \neq j$. For $r \in \mathscr{R}$, $r \leq i < j$ and hence $\mathbf{d}_j^t \mathbf{d}_r = 0$. For $r \notin \mathscr{R}$, $\mathbf{d}_j^t (\sum_{s=r}^{n} \lambda_s \mathbf{d}_s) = \lambda_j \mathbf{d}_j^t \mathbf{d}_j = \lambda_j$. By assumption, $\lambda_j = 0$, and thus multiplying (8.12) by $\mathbf{d}_j^t$, we get $\mathbf{d}_j^t \mathbf{b}_i = 0$ for $i < j$. From (8.11) it follows that $\mathbf{b}_j = \mathbf{d}_j$, and hence, $\bar{\mathbf{d}}_j = \mathbf{d}_j$. This completes the proof.

From Lemma 8.5.4, if $\lambda_j = 0$, then the new direction $\bar{\mathbf{d}}_j$ is equal to the old direction $\mathbf{d}_j$. Hence, we only need to compute new directions for those indices with $\lambda_j \neq 0$.

Summary of the Method of Rosenbrock Using Line Searches

We now summarize Rosenbrock's method using line searches for minimizing a function f of several variables. As we shall show shortly, if f is differentiable, then the method converges to a point with zero gradient.

Initialization Step Let $\varepsilon > 0$ be the termination scalar. Choose $\mathbf{d}_1, \ldots, \mathbf{d}_n$ as the coordinate directions. Choose a starting point $\mathbf{x}_1$, let $\mathbf{y}_1 = \mathbf{x}_1$, $k = j = 1$, and go to the Main Step.

Main Step

1. Let λ_j be an optimal solution to the problem to minimize $f(\mathbf{y}_j + \lambda \mathbf{d}_j)$ subject to $\lambda \in R$, and let $\mathbf{y}_{j+1} = \mathbf{y}_j + \lambda_j \mathbf{d}_j$. If $j < n$, replace j by $j + 1$, and repeat Step 1. Otherwise, go to Step 2.
2. Let $\mathbf{x}_{k+1} = \mathbf{y}_{n+1}$. If $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| < \varepsilon$, then stop; otherwise, let $\mathbf{y}_1 = \mathbf{x}_{k+1}$, replace k by $k + 1$, let $j = 1$, and go to Step 3.
3. Form a new set of linearly independent orthogonal search directions by (8.9). Denote these new directions by $\mathbf{d}_1, \ldots, \mathbf{d}_n$ and go to Step 1.

8.5.5 Example

Consider the following problem:

$$\text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2.$$

We solve this problem by the method of Rosenbrock using line searches. Table 8.9 summarizes the computations starting from the point (0.00, 3.00). The

Table 8.9 Summary of Computations for the Method of Rosenbrock Using Line Searches

Iteration k	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	j	$\mathbf{y}_j$ $f(\mathbf{y}_j)$	$\mathbf{d}_j$	λ_j	$\mathbf{y}_{j+1}$ $f(\mathbf{y}_{j+1})$
1	(0.00, 3.00) 52.00	1	(0.00, 3.00) 52.00	(1.00, 0.00)	3.13	(3.13, 3.00) 9.87
		2	(3.13, 3.00) 9.87	(0.00, 1.00)	–1.44	(3.13, 1.56) 1.63
2	(3.13, 1.56) 1.63	1	(3.13, 1.56) 1.63	(0.91, –0.42)	–0.34	(2.82, 1.70) 0.79
		2	(2.82, 1.70) 0.79	(–0.42, –0.91)	0.51	(2.16, 1.24) 0.16
3	(2.61, 1.24) 0.16	1	(2.61, 1.24) 0.16	(–0.85, –0.52)	0.38	(2.29, 1.04) 0.05
		2	(2.29, 1.04) 0.05	(0.52, –0.85)	–0.10	(2.24, 1.13) 0.004
4	(2.24, 1.13) 0.004	1	(2.24, 1.13) 0.004	(–0.96, –0.28)	0.04	(2.20, 1.12) 0.003
		2	(2.20, 1.12) 0.003	(0.28, –0.96)	0.02	(2.21, 1.10) 0.002

point $\mathbf{y}_2$ is obtained by optimizing the function along the direction $\mathbf{d}_1$ starting from $\mathbf{y}_1$, and $\mathbf{y}_3$ is obtained by optimizing the function along the direction $\mathbf{d}_2$ starting from $\mathbf{y}_2$. After the first iteration, we have $\lambda_1 = 3.13$ and $\lambda_2 = -1.44$. Using (8.9), the new search directions are (0.91, –0.42) and (–0.42, –0.91). After four iterations, the point (2.21, 1.10) is reached, and the corresponding objective function value is 0.002. We now have $\|\mathbf{x}_4 - \mathbf{x}_3\| = 0.15$, and the procedure is stopped.

In Figure 8.14 the progress of the method is shown. It may be interesting to compare this figure with Figure 8.15, which is given later for the method of Rosenbrock using discrete steps.

Convergence of the Method of Rosenbrock

Note that according to Lemma 8.5.4, the search directions employed by the method are linearly independent and mutually orthogonal, and each has norm 1. Thus, at any given iteration, the matrix $\mathbf{D}$ denoting the search directions satisfies $\mathbf{D}'\mathbf{D} = \mathbf{I}$. Thus, $\det[\mathbf{D}] = 1$ and hence Assumption 1 of Theorem 7.3.5 holds true. By this theorem it follows that the method of Rosenbrock using line searches converges to a stationary point if the following assumptions are true:

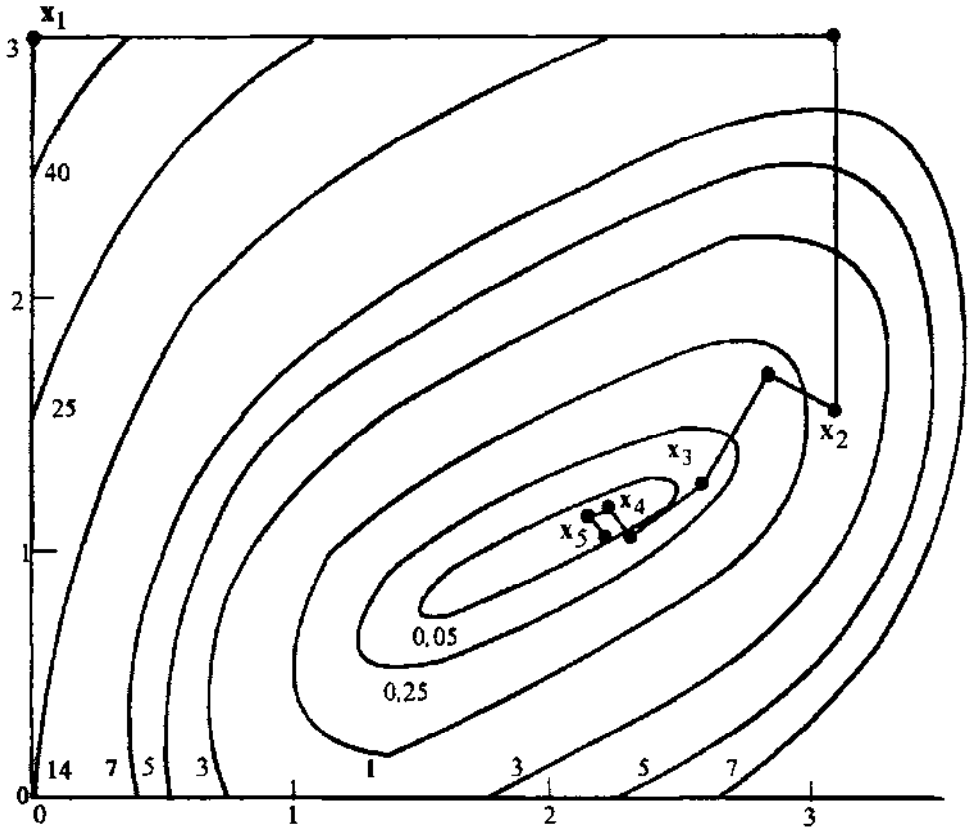


Figure 8.14 Method of Rosenbrock using line searches.

1. The minimum of f along any line in R^n is unique.
2. The sequence of points generated by the algorithm is contained in a compact subset of R^n.

Rosenbrock's Method with Discrete Steps

As mentioned earlier, the method proposed by Rosenbrock avoids line searches. Instead, functional values are made at specific points. Furthermore, an acceleration feature is incorporated by suitably increasing or decreasing the step lengths as the method proceeds. A summary of the method is given below.

Initialization Step Let $\varepsilon > 0$ be the termination scalar, let $\alpha > 1$ be a chosen expansion factor, and let $\beta \in (-1, 0)$ be a selected contraction factor. Choose $\mathbf{d}_1, \ldots, \mathbf{d}_n$ as the coordinate directions, and let $\bar{\Delta}_1, \ldots, \bar{\Delta}_n > 0$ be the initial step sizes along these directions. Choose a starting point $\mathbf{x}_1$, let $\mathbf{y}_1 = \mathbf{x}_1$, $k = j = 1$, let $\Delta_j = \bar{\Delta}_j$ for each j, and go to the Main Step.

Main Step

1. If $f(\mathbf{y}_j + \Delta_j \mathbf{d}_j) < f(\mathbf{y}_j)$, the jth trial is deemed a *success*; set $\mathbf{y}_{j+1} = \mathbf{y}_j + \Delta_j \mathbf{d}_j$, and replace Δ_j by $\alpha \Delta_j$. If, on the other hand, $f(\mathbf{y}_j + \Delta_j \mathbf{d}_j) \geq f(\mathbf{y}_j)$, the trial is considered a *failure*; set $\mathbf{y}_{j+1} = \mathbf{y}_j$, and

replace Δ_j by $\beta\Delta_j$. If $j < n$, replace j by $j + 1$, and repeat Step 1. Otherwise, if $j = n$, go to Step 2.

2. If $f(\mathbf{y}_{n+1}) < f(\mathbf{y}_1)$, that is, if any of the n trials of Step 1 were successful, let $\mathbf{y}_1 = \mathbf{y}_{n+1}$, set $j = 1$, and repeat Step 1. Now consider the case $f(\mathbf{y}_{n+1}) = f(\mathbf{y}_1)$, that is, when each of the last n trials of Step 1 was a failure. If $f(\mathbf{y}_{n+1}) < f(\mathbf{x}_k)$, that is, if at least one successful trial was encountered in Step 1 during iteration k, go to Step 3. If $f(\mathbf{y}_{n+1}) = f(\mathbf{x}_k)$, that is, if no successful trial is encountered, stop with $\mathbf{x}_k$ as an estimate of the optimal solution if $|\Delta_j| \le \varepsilon$ for j; otherwise, let $\mathbf{y}_1 = \mathbf{y}_{n+1}$, let $j = 1$, and go to Step 1.
3. Let $\mathbf{x}_{k+1} = \mathbf{y}_{n+1}$. If $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| < \varepsilon$, stop with $\mathbf{x}_{k+1}$ as an estimate of the optimal solution. Otherwise, compute $\lambda_1, \ldots, \lambda_n$ from the relationship $\mathbf{x}_{k+1} - \mathbf{x}_k = \sum_{j=1}^{n} \lambda_j \mathbf{d}_j$, form a new set of search directions by (8.9) and denote these directions by $\mathbf{d}_1, \ldots, \mathbf{d}_n$, let $\Delta_j = \bar{\Delta}_j$ for each j, let $\mathbf{y}_1 = \mathbf{x}_{k+1}$, replace k by $k + 1$, let $j = 1$, and go to Step 1.

Note that discrete steps are taken along the n search directions in Step 1. If a success occurs along $\mathbf{d}_j$, then Δ_j is replaced by $\alpha\Delta_j$; and if a failure occurs along $\mathbf{d}_j$, then Δ_j is replaced by $\beta\Delta_j$. Since $\beta < 0$, a failure results in reversing the jth search direction during the next pass through Step 1. Note that Step 1 is repeated until a failure occurs along each of the search directions, in which case, if at least one success was obtained during a previous loop at this iteration, a new set of search directions is formed by the Gram–Schmidt procedure. If the loops through the search directions continue to result in failures, the step length shrinks to zero.

8.5.6 Example

Consider the following problem:

$$\text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2.$$

We solve this problem by the method of Rosenbrock using discrete steps with $\bar{\Delta}_1 = \bar{\Delta}_2 = 0.1$, $\alpha = 2.0$, and $\beta = -0.5$. Table 8.10 summarizes the computations starting from (0.00, 3.00), where (S) denotes a success and (F) denotes a failure. Note that within each iteration the directions $\mathbf{d}_1$ and $\mathbf{d}_2$ are fixed. After seven passes through Step 1 of Rosenbrock's method, we move from $\mathbf{x}_1' = (0.00, 3.00)$ to $\mathbf{x}_2' = (3.10, 1.45)$. At this point, a change of directions is required. In particular, $(\mathbf{x}_2 - \mathbf{x}_1) = \lambda_1\mathbf{d}_1 + \lambda_2\mathbf{d}_2$, where $\lambda_1 = 3.10$ and $\lambda_2 = -1.55$. Using (8.9), the reader can easily verify that the new search directions are

given by (0.89, –0.45) and (–0.45, –0.89), which are used in the second iteration. The procedure is terminated during the second iteration.

Figure 8.15 displays the progress of Rosenbrock's method, where the points generated are numbered sequentially.

8.6 Multidimensional Search Using Derivatives

In the preceding section we described several minimization procedures that use only functional evaluations during the course of optimization. We now discuss some methods that use derivatives in determining the search directions. In particular, we discuss the steepest descent method and the method of Newton.

Method of Steepest Descent

The method of steepest descent, proposed by Cauchy in 1847, is one of the most fundamental procedures for minimizing a differentiable function of several variables. Recall that a vector **d** is called a **direction of descent** of a function f at x if there exists a $\delta > 0$ such that $f(\mathbf{x} + \lambda\mathbf{d}) < f(\mathbf{x})$ for all $\lambda \in (0, \delta)$. In particular, if $\lim_{\lambda \to 0^+} [f(\mathbf{x} + \lambda\mathbf{d}) - f(\mathbf{x})]/\lambda < 0$, then **d** is a direction of descent. The method of steepest descent moves along the direction **d** with $\|\mathbf{d}\| = 1$, which minimizes the above limit. Lemma 8.6.1 shows that if f is differentiable at **x** with a nonzero gradient, then $-\nabla f(\mathbf{x})/\|\nabla f(\mathbf{x})\|$ is indeed the direction of steepest descent. For this reason, in the presence of differentiability, the method of steepest descent is sometimes called the *gradient method*; it is also referred to as *Cauchy's method.*

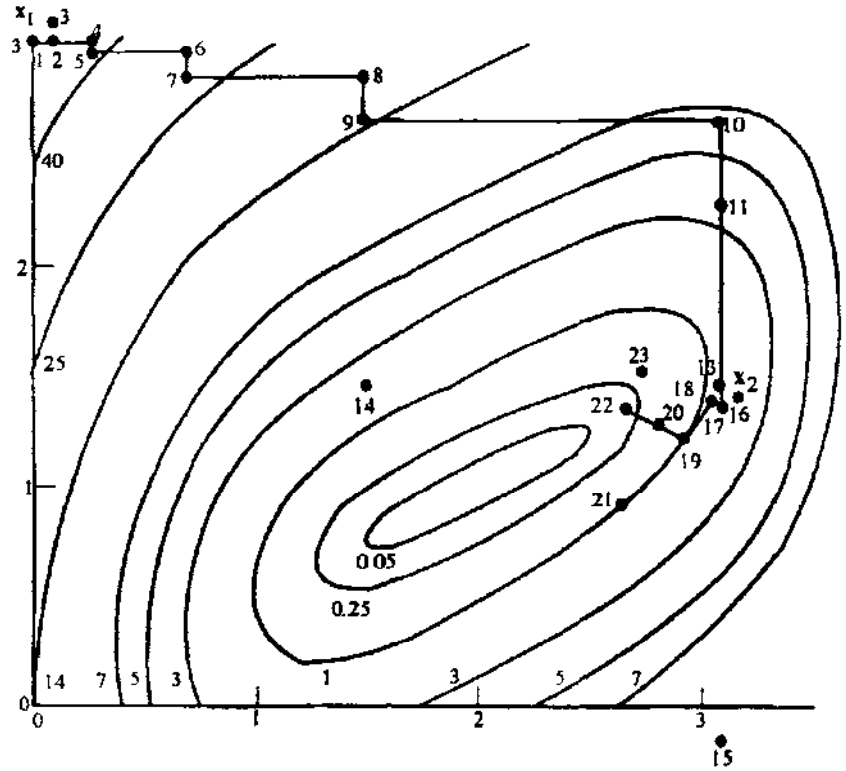


Figure 8.15 Rosenbrock's procedure using discrete steps. (The numbers denote the order in which points are generated.)

8.6.1 Lemma

Suppose that f: $R^n \to R$ is differentiable at $\mathbf{x}$, and suppose that $\nabla f(\mathbf{x}) \neq \mathbf{0}$. Then the optimal solution to the problem to minimize $f'(\mathbf{x}; \mathbf{d})$ subject to $\|\mathbf{d}\| \leq 1$ is given by $\bar{\mathbf{d}} = -\nabla f(\mathbf{x})/\|\nabla f(\mathbf{x})\|$; that is, $-\nabla f(\mathbf{x})/\|\nabla f(\mathbf{x})\|$ is the direction of steepest descent of f at $\mathbf{x}$.

Table 8.10 Summary of Computations for Rosenbrock's Method Using Discrete Steps

Iteration k	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	j	$\mathbf{y}_j$ $f(\mathbf{y}_j)$	Δ_j	$\mathbf{d}_j$	$\mathbf{y}_j + \Delta_j \mathbf{d}_j$ $f(\mathbf{y}_j + \Delta_j \mathbf{d}_j)$
1	(0.00, 3.00) 52.00	1	(0.00, 3.00) 52.00	0.10	(1.00, 0.00)	(0.10, 3.00) 47.84(S)
		2	(0.10, 3.00) 47.84	0.10	(0.00, 1.00)	(0.10, 3.10) 50.24(F)
		1	(0.10, 3.00) 47.84	0.20	(1.00, 0.00)	(0.30, 3.00) 40.84(S)
		2	(0.30, 3.00) 40.84	−0.05	(0.00, 1.00)	(0.30, 2.95) 39.71(S)
		1	(0.30, 2.95) 39.71	0.40	(1.00, 0.00)	(0.70, 2.95) 29.90(S)
		2	(0.70, 2.95) 29.90	−0.10	(0.00, 1.00)	(0.70, 2.85) 27.86(S)
		1	(0.70, 2.85) 27.86	0.80	(1.00, 0.00)	(1.50, 2.85) 17.70(S)
		2	(1.50, 2.85) 17.70	−0.20	(0.00, 1.00)	(1.50, 2.65) 14.50(S)
		1	(1.50, 2.65) 14.50	1.60	(1.00, 0.00)	(3.10, 2.65) 6.30(S)
		2	(3.10, 2.65) 6.30	−0.40	(0.00, 1.00)	(3.10, 2.25) 3.42(S)
		1	(3.10, 2.25) 3.42	3.20	(1.00, 0.00)	(6.30, 2.25) 345.12(F)
		2	(3.10, 2.25) 3.42	−0.80	(0.00, 1.00)	(3.10, 1.45) 1.50(S)

(*continued*)

Table **8.10** (*continued*)

Iteration k	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	j	$\mathbf{y}_j$ $f(\mathbf{y}_j)$	Δ_j	$\mathbf{d}_j$	$\mathbf{y}_j + \Delta_j \mathbf{d}_j$ $f(\mathbf{y}_j + \Delta_j \mathbf{d}_j)$
		1	(3.10, 1.45) 1.50	−1.60	(1.00, 0.00)	(1.50, 1.45) 2.02(F)
		2	(3.10, 1.45) 1.50	−1.60	(0.00, 1.00)	(3.10, −0.15) 13.02(F)
2	(3.10, 1.45) 1.50	1	(3.10, 1.45) 1.50	0.10	(0.89, −0.45)	(3.19, 1.41) 2.14(F)
		2	(3.10, 1.45) 1.50	0.10	(−0.45, −0.89)	(3.06, 1.36) 1.38(S)
		1	(3.06, 1.36) 1.38	−0.05	(0.89, −0.45)	(3.02, 1.38) 1.15(S)
		2	(3.02, 1.38) 1.15	0.20	(−0.45, −0.89)	(2.93, 1.20) 1.03(S)
		1	(2.93, 1.20) 1.03	−0.10	(0.89, −0.45)	(2.84, 1.25) 0.61(S)
		2	(2.84, 1.25) 0.61	0.40	(−0.45, −0.89)	(2.66, 0.89) 0.96(F)
		1	(2.84, 1.25) 0.61	−0.20	(0.89, −0.45)	(2.66, 1.34) 0.19(S)
		2	(2.66, 1.34) 0.19	−0.20	(−0.45, −0.89)	(2.75, 1.52) 0.40(F)

Proof

From the differentiability of f at $\mathbf{x}$, it follows that

$$f'(\mathbf{x}; \mathbf{d}) = \lim_{\lambda \to 0^+} \frac{f(\mathbf{x} + \lambda \mathbf{d}) - f(\mathbf{x})}{\lambda} = \nabla f(\mathbf{x})^t \mathbf{d}.$$

Thus, the problem reduces to minimizing $\nabla f(\mathbf{x})^t \mathbf{d}$ subject to $\|\mathbf{d}\| \leq 1$. By the Schwartz inequality, for $\|\mathbf{d}\| \leq 1$ we have

$$\nabla f(\mathbf{x})^t \mathbf{d} \geq -\|\nabla f(\mathbf{x})\| \|\mathbf{d}\| \geq -\|\nabla f(\mathbf{x})\|,$$

with equality holding throughout if and only if $\mathbf{d} = \bar{\mathbf{d}} \equiv -\nabla f(\mathbf{x}) / \|\nabla f(\mathbf{x})\|$. Thus, $\bar{\mathbf{d}}$ is the optimal solution, and the proof is complete.

Summary of the Steepest Descent Algorithm

Given a point $\mathbf{x}$, the steepest descent algorithm proceeds by performing a line search along the direction $-\nabla f(\mathbf{x})/\|\nabla f(\mathbf{x})\|$ or, equivalently, along the direction $-\nabla f(\mathbf{x})$. A summary of the method is given below.

Initialization Step Let $\varepsilon > 0$ be the termination scalar. Choose a starting point $\mathbf{x}_1$, let $k = 1$, and go to the Main Step.

Main Step

If $\|\nabla f(\mathbf{x}_k)\| < \varepsilon$, stop; otherwise, let $\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$, and let λ_k be an optimal solution to the problem to minimize $f(\mathbf{x}_k + \lambda \mathbf{d}_k)$ subject to $\lambda \geq 0$. Let $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$, replace k by $k + 1$, and repeat the Main Step.

8.6.2 Example

Consider the following problem:

$$\text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2.$$

We solve this problem using the method of steepest descent, starting with the point (0.00, 3.00). A summary of the computations is given in Table 8.11. After seven iterations, the point $\mathbf{x}_8 = (2.28, 1.15)^t$ is reached. The algorithm is terminated since $\|\nabla f(\mathbf{x}_8)\| = 0.09$ is small. The progress of the method is shown in Figure 8.16. Note that the minimizing point for this problem is (2.00, 1.00).

Convergence of the Steepest Descent Method

Let $\Omega = \{\bar{\mathbf{x}} : \nabla f(\bar{\mathbf{x}}) = \mathbf{0}\}$, and let f be the descent function. The algorithmic map is $\mathbf{A} = \mathbf{MD}$, where $\mathbf{D}(\mathbf{x}) = [\mathbf{x}, \nabla f(\mathbf{x})]$ and $\mathbf{M}$ is the line search map over the closed interval $[0, \infty)$. Assuming that f is continuously differentiable, $\mathbf{D}$ is continuous. Furthermore, $\mathbf{M}$ is closed by Theorem 8.4.1. Therefore, the algorithmic map $\mathbf{A}$ is closed by Corollary 2 to Theorem 7.3.2. Finally, if $\mathbf{x} \notin \Omega$, then $\nabla f(\mathbf{x})^t \mathbf{d} < 0$, where $\mathbf{d} = -\nabla f(\mathbf{x})$. By Theorem 4.1.2, $\mathbf{d}$ is a descent direction, and hence $f(\mathbf{y}) < f(\mathbf{x})$ for $\mathbf{y} \in \mathbf{A}(\mathbf{x})$. Assuming that the sequence generated by the algorithm is contained in a compact set, then by Theorem 7.2.3, the steepest descent algorithm converges to a point with zero gradient.

Zigzagging of the Steepest Descent Method

The method of steepest descent usually works quite well during early stages of the optimization process, depending on the point of initialization. However, as a stationary point is approached, the method usually behaves poorly, taking small,

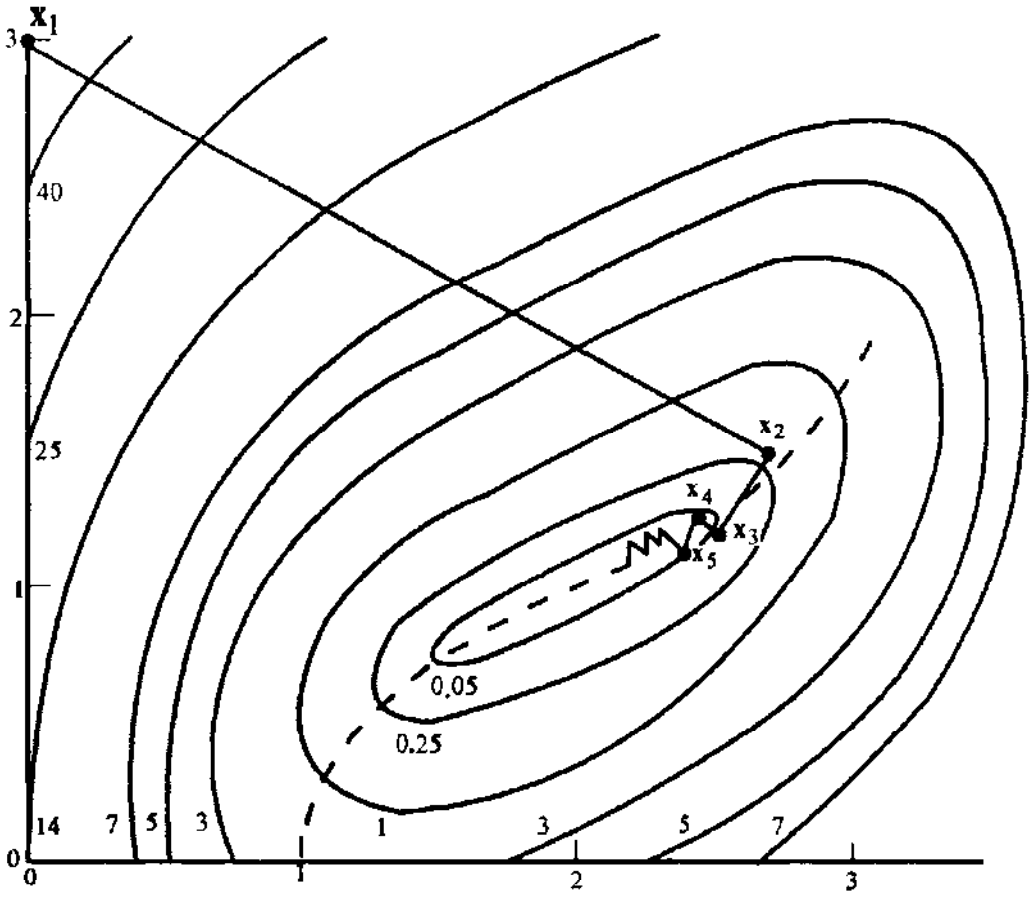


Figure 8.16 Method of steepest descent.

Table 8.11 Summary of Computations for the Method of Steepest Descent

Iteration k	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	$\nabla f(\mathbf{x}_k)$	$\|\nabla f(\mathbf{x}_k)\|$	$\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$	λ_k	$\mathbf{x}_{k+1}$
1	(0.00, 3.00) 52.00	(–44.00, 24.00)	50.12	(44.00, –24.00)	0.062	(2.70, 1.51)
2	(2.70, 1.51) 0.34	(0.73, 1.28)	1.47	(–0.73, –1.28)	0.24	(2.52, 1.20)
3	(2.52, 1.20) 0.09	(0.80, –0.48)	0.93	(–0.80, 0.48)	0.11	(2.43, 1.25)
4	(2.43, 1.25) 0.04	(0.18, 0.28)	0.33	(–0.18, –0.28)	0.31	(2.37, 1.16)
5	(2.37, 1.16) 0.02	(0.30, –0.20)	0.36	(–0.30, 0.20)	0.12	(2.33, 1.18)
6	(2.33, 1.18) 0.01	(0.08, 0.12)	0.14	(–0.08, –0.12)	0.36	(2.30, 1.14)
7	(2.30, 1.14) 0.009	(0.15, –0.08)	0.17	(–0.15, 0.08)	0.13	(2.28, 1.15)
8	(2.28, 1.15) 0.007	(0.05, 0.08)	0.09			

nearly orthogonal steps. This *zigzagging* phenomenon was encountered in Example 8.6.2 and is illustrated in Figure 8.16, in which zigzagging occurs along the valley shown by the dashed lines.

Zigzagging and poor convergence of the steepest descent algorithm at later stages can be explained intuitively by considering the following expression of the function f:

$$f(\mathbf{x}_k + \lambda\mathbf{d}) = f(\mathbf{x}_k) + \lambda\nabla f(\mathbf{x}_k)'\mathbf{d} + \lambda\|\mathbf{d}\|\alpha(\mathbf{x}_k; \lambda\mathbf{d}),$$

where $\alpha(\mathbf{x}_k; \lambda\mathbf{d}) \to 0$ as $\lambda\mathbf{d} \to \mathbf{0}$, and $\mathbf{d}$ is a search direction with $\|\mathbf{d}\| = 1$. If $\mathbf{x}_k$ is close to a stationary point with zero gradient and f is continuously differentiable, then $\|\nabla f(\mathbf{x}_k)\|$ will be small, making the coefficient of λ in the term $\lambda\nabla f(\mathbf{x}_k)'\mathbf{d}$ of a small order of magnitude. Since the steepest descent method employs the linear approximation of f to find a direction of movement, where the term $\lambda\|\mathbf{d}\|\alpha(\mathbf{x}_k; \lambda\mathbf{d})$ is essentially ignored, we should expect that the directions generated at late stages will not be very effective if the latter term contributes significantly to the description of f, even for relatively small values of λ.

As we shall learn in the remainder of the chapter, there are some ways to overcome the difficulties of zigzagging by *deflecting the gradient*. Rather than moving along $\mathbf{d} = -\nabla f(\mathbf{x})$, we can move along $\mathbf{d} = -\mathbf{D}\nabla f(\mathbf{x})$ or along $\mathbf{d} = -\nabla f(\mathbf{x}) + \mathbf{g}$, where $\mathbf{D}$ is an appropriate matrix and $\mathbf{g}$ is an appropriate vector. These correction procedures will be discussed in more detail shortly.

Convergence Rate Analysis for the Steepest Descent Algorithm

In this section we give a more formalized analysis of the zigzagging phenomenon and the empirically observed slow convergence rate of the steepest descent algorithm. This analysis will also afford insights into possible ways of alleviating this poor algorithmic performance.

Toward this end, let us begin by considering a bivariate quadratic function $f(x_1, x_2) = (1/2)(x_1^2 + \alpha x_2^2)$, where $\alpha > 1$. Note that the Hessian matrix to this function is $\mathbf{H} = \text{diag}\{1, \alpha\}$, with eigenvalues 1 and α. Let us define the *condition number* of a positive definite matrix to be the ratio of its largest to smallest eigenvalues. Hence, the condition number of $\mathbf{H}$ for our example is α. The contours of f are plotted in Figure 8.17. Observe that as α increases, a phenomenon that is known as *ill-conditioning*, or a *worsening of the condition number* results, whereby the contours become increasingly skewed and the graph of the function becomes increasingly steep in the x_2 direction relative to the x_1 direction.

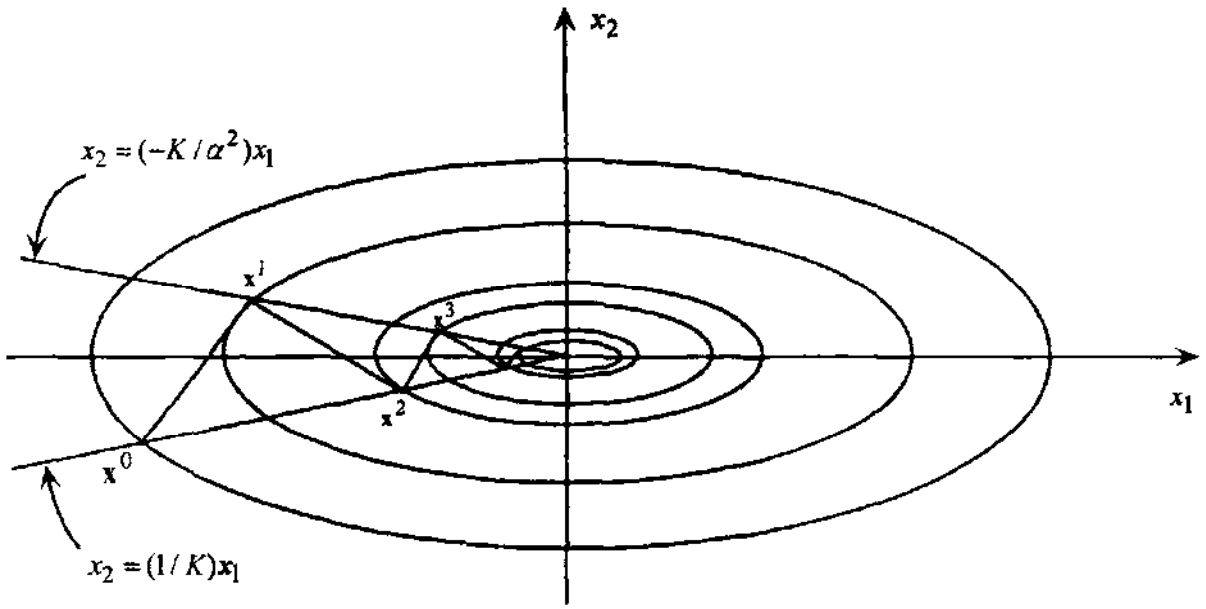


Figure 8.17 Convergence rate analysis of the steepest descent algorithm.

Now, given a starting point $\mathbf{x} = (x_1, x_2)^t$, let us apply an iteration of the steepest descent algorithm to obtain a point $\mathbf{x}_{new} = (x_{1new}, x_{2new})^t$. Note that if $x_1 = 0$ or $x_2 = 0$, the procedure converges to the optimal minimizing solution $\mathbf{x}^* = (0, 0)^t$ in one step. Hence, suppose that $x_1 \neq 0$ and $x_2 \neq 0$. The steepest descent direction is given by $\mathbf{d} = -\nabla f(\mathbf{x}) = -(x_1, \alpha x_2)^t$, resulting in $\mathbf{x}_{new} = \mathbf{x} + \lambda\mathbf{d}$, where λ solves the line search problem to minimize $\theta(\lambda) \equiv f(\mathbf{x} + \lambda\mathbf{d}) = (1/2)[x_1^2(1 - \lambda)^2 + \alpha x_2^2(1 - \alpha\lambda)^2]$ subject to $\lambda \geq 0$. Using simple calculus, we obtain

$$\lambda = \frac{x_1^2 + \alpha^2 x_2^2}{x_1^2 + \alpha^3 x_2^2},$$

so

$$\mathbf{x}_{new} = \left[\frac{\alpha^2 x_1 x_2^2(\alpha - 1)}{x_1^2 + \alpha^3 x_2^2}, \frac{x_1^2 x_2(1 - \alpha)}{x_1^2 + \alpha^3 x_2^2}\right]. \tag{8.13}$$

Observe that $x_{1new}/x_{2new} = -\alpha^2(x_2/x_1)$. Hence, if we begin with a solution $\mathbf{x}^0$ having $x_1^0/x_2^0 = K \neq 0$ and generate a sequence of iterates $\{\mathbf{x}^k\}$, $k = 1, 2, \ldots$, using the steepest descent algorithm, then the sequence of values $\{x_1^k/x_2^k\}$ alternate between the values K and $-\alpha^2/K$ as the sequence $\{\mathbf{x}^k\}$ converges to $\mathbf{x}^* = (0, 0)^t$. For our example this means that the sequence zigzags between the pair of straight lines $x_2 = (1/K)x_1$ and $x_2 = (-K/\alpha^2)x_1$, as shown in Figure

8.17. Note that as the condition number α increases, this zigzagging phenomenon becomes more pronounced. On the other hand, if $\alpha = 1$, then the contours of f are circular, and we obtain $\mathbf{x}^1 = \mathbf{x}^*$ in a single iteration.

To study the rate of convergence, let us examine the rate at which $\{f(\mathbf{x}^k)\}$ converges to the value zero. From (8.13) it is easily verified that

$$\frac{f(\mathbf{x}^{k+1})}{f(\mathbf{x}^k)} = \frac{K_k^2\alpha(\alpha-1)^2}{(K_k^2+\alpha^3)(K_k^2+\alpha)}, \qquad \text{where } K_k \equiv \frac{x_1^k}{x_2^k}. \tag{8.14}$$

Indeed, the expression in (8.14) can be seen to be maximized when $K_k^2 = \alpha^2$ (see Exercise 8.19), so that we obtain

$$\frac{f(\mathbf{x}^{k+1})}{f(\mathbf{x}^k)} \leq \frac{(\alpha-1)^2}{(\alpha+1)^2}. \tag{8.15}$$

Note from (8.15) that $\{f(\mathbf{x}^k)\} \to 0$ at a geometric or linear rate bounded by the ratio $(\alpha-1)^2/(\alpha+1)^2 < 1$. In fact, if we initialize the process with $x_1^0/x_2^0 = K = \alpha$, then, since $K_k^2 = (x_1^k/x_2^k)^2 = \alpha^2$ from above (see Figure 8.17), we get from (8.14) that the convergence ratio $f(\mathbf{x}^{k+1})/f(\mathbf{x}^k)$ is precisely $(\alpha-1)^2/(\alpha+1)^2$. Hence, as α approaches infinity, this ratio approaches 1 from below, and the rate of convergence becomes increasingly slower.

The foregoing analysis can be extended to a general quadratic function $f(\mathbf{x}) = \mathbf{c}^t\mathbf{x} + (1/2)\mathbf{x}^t\mathbf{H}\mathbf{x}$, where $\mathbf{H}$ is an $n \times n$, symmetric, positive definite matrix. The unique minimizer $\mathbf{x}^*$ for this function is given by the solution to the system $\mathbf{H}\mathbf{x}^* = -\mathbf{c}$ obtained by setting $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Also, given an iterate $\mathbf{x}_k$, the optimal step length λ and the revised iterate $\mathbf{x}_{k+1}$ are given by the following generalization of (8.13), where $\mathbf{g}_k \equiv \nabla f(\mathbf{x}_k) = \mathbf{c} + \mathbf{H}\mathbf{x}_k$:

$$\lambda = \frac{\mathbf{g}_k^t\mathbf{g}_k}{\mathbf{g}_k^t\mathbf{H}\mathbf{g}_k} \qquad \text{and} \qquad \mathbf{x}_{k+1} = \mathbf{x}_k - \lambda\mathbf{g}_k. \tag{8.16}$$

Now, to evaluate the rate of convergence, let us employ a convenient measure for convergence given by the following *error function*:

$$e(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^t\mathbf{H}(\mathbf{x}-\mathbf{x}^*) = f(\mathbf{x}) + \frac{1}{2}\mathbf{x}^{*t}\mathbf{H}\mathbf{x}^*, \tag{8.17}$$

where we have used the fact that $\mathbf{H}\mathbf{x}^* = -\mathbf{c}$. Note that $e(\mathbf{x})$ differs from $f(\mathbf{x})$ by only a constant and equals zero if and only if $\mathbf{x} = \mathbf{x}^*$. In fact, it can be shown, analogous to (8.15), that (see Exercise 8.21)

$$e(\mathbf{x}_{k+1}) = \left[1 - \frac{(\mathbf{g}_k^t \mathbf{g}_k)^2}{(\mathbf{g}_k^t \mathbf{H} \mathbf{g}_k)(\mathbf{g}_k^t \mathbf{H}^{-1} \mathbf{g}_k)}\right] e(\mathbf{x}_k) \leq \frac{(\alpha - 1)^2}{(\alpha + 1)^2} e(\mathbf{x}_k), \qquad (8.18)$$

where α is the condition number of $\mathbf{H}$. Hence, $\{e(\mathbf{x}_k)\} \to 0$ at a linear or geometric convergence rate bounded above by $(\alpha - 1)^2/(\alpha + 1)^2$; so, as before, we can expect the convergence to become increasingly slower as α increases, depending on the initial solution $\mathbf{x}_0$.

For continuously twice differentiable nonquadratic functions f: $R^n \to R$, a similar result is known to hold. In such a case, if $\mathbf{x}^*$ is a local minimum to which a sequence $\{\mathbf{x}_k\}$ generated by the steepest descent algorithm converges, and if $\mathbf{H}(\mathbf{x}^*)$ is positive definite with a condition number α, then the corresponding sequence of objective values $\{f(\mathbf{x}_k)\}$ can be shown to converge linearly to the value $f(\mathbf{x}^*)$ at a rate bounded above by $(\alpha - 1)^2/(\alpha + 1)^2$.

Convergence Analysis of the Steepest Descent Algorithm Using Armijo's Inexact Line Search

In Section 8.3 we introduced Armijo's rule for selecting an acceptable, inexact step length during a line search process. It is instructive to observe how such a criterion still guarantees algorithmic convergence. Below, we present a convergence analysis for an inexact steepest descent algorithm applied to a function f: $R^n \to R$ whose gradient function $\nabla f(\mathbf{x})$ is *Lipschitz continuous with constant* $G > 0$ on $S(\mathbf{x}_0) \equiv \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ for some given $\mathbf{x}_0 \in R^n$. That is, we have $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq G\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}$, $\mathbf{y} \in S(\mathbf{x}_0)$. For example, if the Hessian of f at any point has a norm bounded above by a constant G on $\text{conv}S(\mathbf{x}_0)$ (see Appendix A for the norm of a matrix), then such a function has Lipschitz continuous gradients. This follows from the mean value theorem, noting that for any $\mathbf{x} \neq \mathbf{y} \in S(\mathbf{x}_0)$, $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| = \|H(\hat{\mathbf{x}})(\mathbf{x} - \mathbf{y})\| \leq G\|\mathbf{x} - \mathbf{y}\|$.

The procedure we analyze is the often-used variant of Armijo's rule described in Section 8.3 with parameters $0 < \varepsilon < 1$, $\alpha = 2$, and a fixed-step-length parameter $\bar{\lambda}$, wherein either $\bar{\lambda}$ itself is chosen, if acceptable, or is sequentially halved until an acceptable step length results. This procedure is embodied in the following result.

8.6.3 Theorem

Let f: $R^n \to R$ be such that its gradient $\nabla f(\mathbf{x})$ is Lipschitz continuous with constant $G > 0$ on $S(\mathbf{x}_0) = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$ for some given $\mathbf{x}_0 \in R^n$. Pick some fixed-step-length parameter $\bar{\lambda} > 0$, and let $0 < \varepsilon < 1$. Given any iterate

$\mathbf{x}_k$, define the search direction $\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$, and consider Armijo's function $\hat{\theta}(\lambda) = \theta(0) + \lambda\varepsilon\theta'(0)$, $\lambda \geq 0$, where $\theta(\lambda) = f(\mathbf{x}_k + \lambda\mathbf{d}_k)$, $\lambda \geq 0$, is the line search function. If $\mathbf{d}_k = 0$, then stop. Otherwise, find the smallest integer $t \geq 0$ for which $\theta(\bar{\lambda}/2^t) \leq \hat{\theta}(\bar{\lambda}/2^t)$ and define the next iterate as $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k\mathbf{d}_k$, where $\lambda_k \equiv \bar{\lambda}/2^t$. Now suppose that starting with some iterate $\mathbf{x}_0$, this procedure produces a sequence of iterates $\mathbf{x}_0$, $\mathbf{x}_1$, $\mathbf{x}_2$,... . Then either the procedure terminates finitely with $\nabla f(\mathbf{x}_k) = \mathbf{0}$ for some K, or else an infinite sequence $\{\mathbf{x}_k\}$ is generated such that the corresponding sequence $\{\nabla f(\mathbf{x}_k)\} \to \mathbf{0}$.

Proof

The case of finite termination is clear. Hence, suppose that an infinite sequence $\{\mathbf{x}_k\}$ is generated. Note that the Armijo criterion $\theta(\bar{\lambda}/2^t) \leq \hat{\theta}(\bar{\lambda}/2^t)$ is equivalent to $\theta(\bar{\lambda}/2^t) \equiv f(\mathbf{x}_{k+1}) \leq \hat{\theta}(\bar{\lambda}/2^t) = \theta(0) + (\bar{\lambda}\varepsilon/2^t)\nabla f(\mathbf{x}_k)^t \mathbf{d}_k = f(\mathbf{x}_k) - (\bar{\lambda}\varepsilon/2^t)\|\nabla f(\mathbf{x}_k)\|^2$. Hence, $t \geq 0$ is the smallest integer for which

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq \frac{-\bar{\lambda}\varepsilon}{2^t}\|\nabla f(\mathbf{x}_k)\|^2. \tag{8.19}$$

Now, using the mean value theorem, we have, for some strict convex combination $\tilde{\mathbf{x}}$ of $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$, that

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &= \lambda_k \mathbf{d}_k^t \nabla f(\tilde{\mathbf{x}}) \\ &= -\lambda_k \nabla f(\mathbf{x}_k)^t [\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_k) + \nabla f(\tilde{\mathbf{x}})] \\ &= -\lambda_k \|\nabla f(\mathbf{x}_k)\|^2 + \lambda_k \nabla f(\mathbf{x}_k)^t [\nabla f(\mathbf{x}_k) - \nabla f(\tilde{\mathbf{x}})] \\ &\leq -\lambda_k \|\nabla f(\mathbf{x}_k)\|^2 + \lambda_k \|\nabla f(\mathbf{x}_k)\| \|\nabla f(\mathbf{x}_k) - \nabla f(\tilde{\mathbf{x}})\|. \end{aligned}$$

But by the Lipschitz continuity of ∇f, noting from (8.19) that the descent nature of the algorithm guarantees that $\mathbf{x}_k \in S(\mathbf{x}_0)$ for all k, we have $\|\nabla f(\mathbf{x}_k) - \nabla f(\tilde{\mathbf{x}})\| \leq G\|\mathbf{x}_k - \tilde{\mathbf{x}}\| \leq G\|\mathbf{x}_k - \mathbf{x}_{k+1}\| = G\lambda_k \|\nabla f(\mathbf{x}_k)\|$. Substituting this above, we obtain

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\lambda_k \|\nabla f(\mathbf{x}_k)\|^2 (1 - \lambda_k G) = \frac{-\bar{\lambda}}{2^t}\|\nabla f(\mathbf{x}_k)\|^2 \left(1 - \frac{\bar{\lambda}G}{2^t}\right). \tag{8.20}$$

Consequently, from (8.20), we know that (8.19) will hold true when t is increased to no larger an integer value than is necessary to make $1 - (\bar{\lambda}G/2^t) \geq$

ε, for then (8.20) will imply (8.19). But this means that $1 - (\bar{\lambda}G/2^{t-1}) < \varepsilon$; that is, $\bar{\lambda}\varepsilon/2^t > \varepsilon(1-\varepsilon)/2G$. Substituting this in (8.19), we get

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) < \frac{-\varepsilon(1-\varepsilon)}{2G}\|\nabla f(\mathbf{x}_k)\|^2.$$

Hence, noting that $\{f(\mathbf{x}_k)\}$ is a monotone decreasing sequence and so has a limit, taking limits as $t \to \infty$, we get

$$0 \leq \frac{-\varepsilon(1-\varepsilon)}{2G} \lim_{k\to\infty} \|\nabla f(\mathbf{x}_k)\|^2,$$

which implies that $\{\nabla f(\mathbf{x}_k)\} \to 0$. This completes the proof.

Method of Newton

In Section 8.2 we discussed Newton's method for minimizing a function of a single variable. The method of Newton is a procedure that deflects the steepest descent direction by premultiplying it by the inverse of the Hessian matrix. This operation is motivated by finding a suitable direction for the quadratic approximation to the function rather than by finding a linear approximation to the function, as in the gradient search. To motivate the procedure, consider the following approximation q at a given point $\mathbf{x}_k$:

$$q(\mathbf{x}) = f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^t(\mathbf{x}-\mathbf{x}_k) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_k)^t \mathbf{H}(\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k),$$

where $\mathbf{H}(\mathbf{x}_k)$ is the Hessian matrix of f at $\mathbf{x}_k$. A necessary condition for a minimum of the quadratic approximation q is that $\nabla q(\mathbf{x}) = \mathbf{0}$, or $\nabla f(\mathbf{x}_k) + \mathbf{H}(\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k) = \mathbf{0}$. Assuming that the inverse of $\mathbf{H}(\mathbf{x}_k)$ exists, the successor point $\mathbf{x}_{k+1}$ is given by

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k). \tag{8.21}$$

Equation (8.21) gives the recursive form of the points generated by Newton's method for the multidimensional case. Assuming that $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$, that $\mathbf{H}(\bar{\mathbf{x}})$ is positive definite at a local minimum $\bar{\mathbf{x}}$, and that f is continuously twice differentiable, it follows that $\mathbf{H}(\mathbf{x}_k)$ is positive definite at points close to $\bar{\mathbf{x}}$, and hence the successor point $\mathbf{x}_{k+1}$ is well defined.

It is interesting to note that Newton's method can be interpreted as a *steepest descent algorithm with affine scaling*. Specifically, given a point $\mathbf{x}_k$ at iteration k, suppose that $\mathbf{H}(\mathbf{x}_k)$ is positive definite and that we have a Cholesky factorization (see Appendix A.2) of its inverse given by $\mathbf{H}(\mathbf{x}_k)^{-1} = \mathbf{L}\mathbf{L}^t$, where $\mathbf{L}$ is a lower triangular matrix with positive diagonal elements. Now, consider

the affine scaling transformation $\mathbf{x} = \mathbf{L}\mathbf{y}$. This transforms the function $f(\mathbf{x})$ to the function $F(\mathbf{y}) \equiv f[\mathbf{L}\mathbf{y}]$, and the current point in the $\mathbf{y}$ space is $\mathbf{y}_k = \mathbf{L}^{-1}\mathbf{x}_k$. Hence, we have $\nabla F(\mathbf{y}_k) = \mathbf{L}^t \nabla f[\mathbf{L}\mathbf{y}_k] = \mathbf{L}^t \nabla f(\mathbf{x}_k)$. A unit step size along the negative gradient direction in the $\mathbf{y}$ space will then take us to the point $\mathbf{y}_{k+1} = \mathbf{y}_k - \mathbf{L}^t \nabla f(\mathbf{x}_k)$. Translating this to the corresponding movement in the $\mathbf{x}$ space by premultiplying throughout by $\mathbf{L}$ produces precisely Equation (8.21) and hence yields a steepest descent interpretation of Newton's method. Observe that this comment alludes to the benefits of using an appropriate scaling transformation. Indeed, if the function f was quadratic in the above analysis, then a unit step along the steepest descent direction in the transformed space would be an optimal step along that direction, which would moreover take us directly to the optimal solution in one iteration starting from any given solution.

We also comment here that (8.21) can be viewed as an application of the *Newton–Raphson method* to the solution of the system of equations $\nabla f(\mathbf{x}) = \mathbf{0}$. Given a well-determined system of nonlinear equations, each iteration of the Newton–Raphson method adopts a first-order Taylor series approximation to this equation system at the current iterate and solves the resulting linear system to determine the next iterate. Applying this to the system $\nabla f(\mathbf{x}) = \mathbf{0}$ at an iterate $\mathbf{x}_k$, the first-order approximation to $\nabla f(\mathbf{x})$ is given by $\nabla f(\mathbf{x}_k) + \mathbf{H}(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k)$. Setting this equal to zero and solving produces the solution $\mathbf{x} = \mathbf{x}_{k+1}$ as given by (8.21).

8.6.4 Example

Consider the following problem:

$$\text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2.$$

The summary of the computations using Newton's method is given in Table 8.12. At each iteration, $\mathbf{x}_{k+1}$ is given by $\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$. After six iterations, the point $\mathbf{x}_7 = (1.83,\ 0.91)^t$ is reached. At this point, $\|\nabla f(\mathbf{x}_7)\| = 0.04$, and the procedure is terminated. The points generated by the method are shown in Figure 8.18.

In Example 8.6.4 the value of the objective function decreased at each iteration. However, this will not generally be the case, so f cannot be used as a descent function. Theorem 8.6.5 indicates that Newton's method indeed converges, provided that we start from a point close enough to an optimal point.

Order-Two Convergence of the Method of Newton

In general, the points generated by the method of Newton may not converge. The reason for this is that $\mathbf{H}(\mathbf{x}_k)$ may be singular, so that $\mathbf{x}_{k+1}$ is not

Table 8.12 Summary of Computations for the Method of Newton

Iteration k	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	$\nabla f(\mathbf{x}_k)$	$\mathbf{H}(\mathbf{x}_k)$	$\mathbf{H}(\mathbf{x}_k)^{-1}$	$-\mathbf{H}(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k)$	$\mathbf{x}_{k+1}$
1	(0.00, 3.00) 52.00	(44.0, 24.0)	$\begin{bmatrix} 50.0 & -4.0 \\ -4.0 & 8.0 \end{bmatrix}$	$\frac{1}{384}\begin{bmatrix} 8.0 & 4.0 \\ 4.0 & 50.0 \end{bmatrix}$	(0.67, –2.67)	(0.67, 0.33)
2	(0.67, 0.33) 3.13	(–9.39, –0.04)	$\begin{bmatrix} 23.23 & -4.0 \\ -4.0 & 8.0 \end{bmatrix}$	$\frac{1}{169.84}\begin{bmatrix} 8.0 & 4.0 \\ 4.0 & 23.23 \end{bmatrix}$	(0.44, 0.23)	(1.11, 0.56)
3	(1.11, 0.56) 0.63	(–2.84, –0.04)	$\begin{bmatrix} 11.50 & -4.0 \\ -4.0 & 8.0 \end{bmatrix}$	$\frac{1}{76}\begin{bmatrix} 8.0 & 4.0 \\ 4.0 & 11.50 \end{bmatrix}$	(0.30, 0.14)	(1.41, 0.70)
4	(1.41, 0.70) 0.12	(–0.80, –0.04)	$\begin{bmatrix} 6.18 & -4.0 \\ -4.0 & 8.0 \end{bmatrix}$	$\frac{1}{33.44}\begin{bmatrix} 8.0 & 4.0 \\ 4.0 & 6.18 \end{bmatrix}$	(0.20, 0.10)	(1.61, 0.80)
5	(1.61, 0.80) 0.02	(–0.22, –0.04)	$\begin{bmatrix} 3.83 & -4.0 \\ -4.0 & 8.0 \end{bmatrix}$	$\frac{1}{14.64}\begin{bmatrix} 8.0 & 4.0 \\ 4.0 & 3.83 \end{bmatrix}$	(0.13, 0.07)	(1.74, 0.87)
6	(1.74, 0.87) 0.005	(–0.07, 0.00)	$\begin{bmatrix} 2.81 & -4.0 \\ -4.0 & 8.0 \end{bmatrix}$	$\frac{1}{6.48}\begin{bmatrix} 8.0 & 4.0 \\ 4.0 & 2.81 \end{bmatrix}$	(0.09, 0.04)	(1.83, 0.91)
7	(1.83, 0.91) 0.0009	(0.0003, –0.04)				

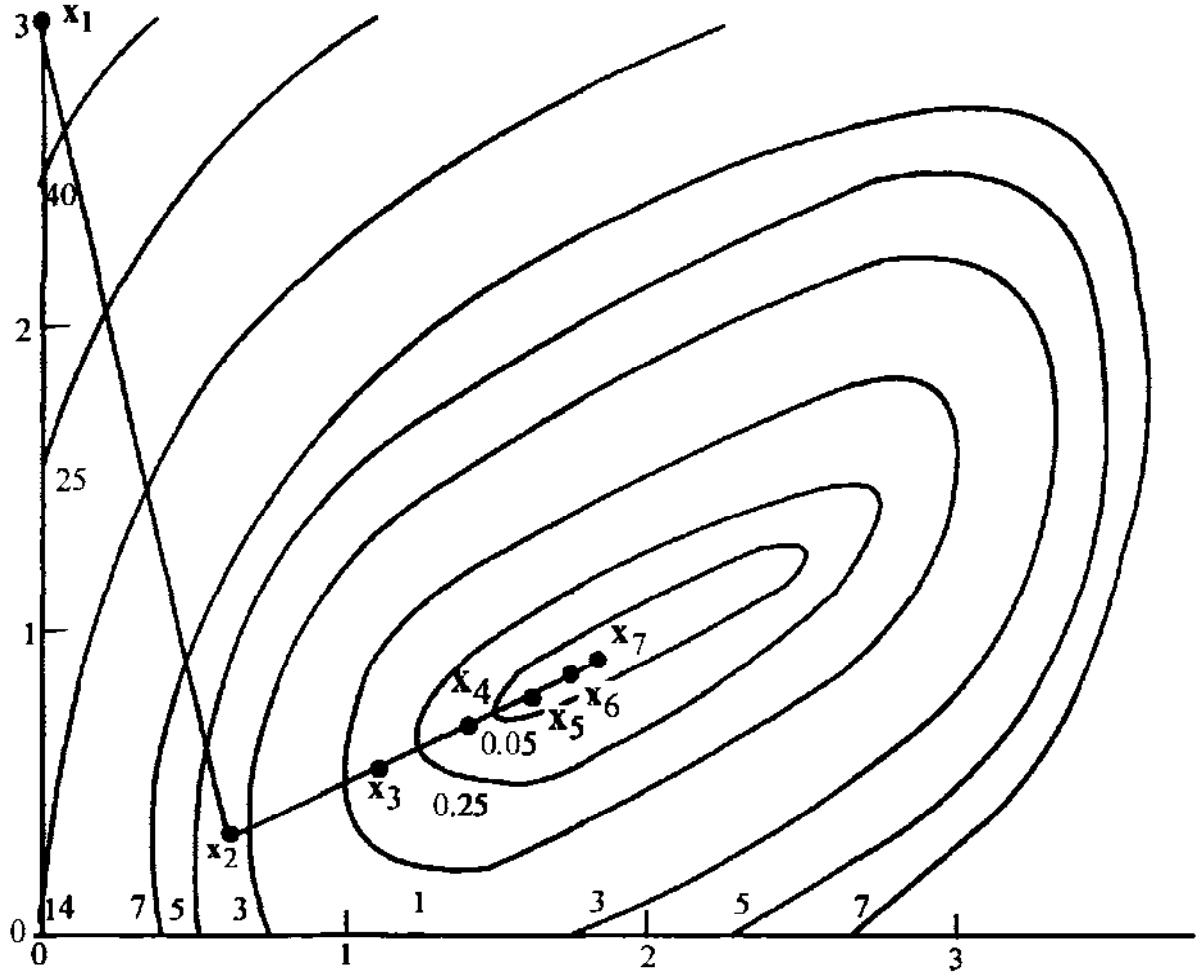


Figure 8.18 Method of Newton.

well defined. Even if $\mathbf{H}(\mathbf{x}_k)^{-1}$ exists, $f(\mathbf{x}_{k+1})$ is not necessarily less than $f(\mathbf{x}_k)$. However, if the starting point is close enough to a point $\bar{\mathbf{x}}$ such that $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$ and $\mathbf{H}(\bar{\mathbf{x}})$ is of full rank, then the method of Newton is well defined and converges to $\bar{\mathbf{x}}$. This is proved in Theorem 8.6.5 by showing that all the assumptions of Theorem 7.2.3 hold true, where the descent function α is given by $\alpha(\mathbf{x}) = \|\mathbf{x} - \bar{\mathbf{x}}\|$.

8.6.5 Theorem

Let $f\colon R^n \to R$ be continuously twice differentiable. Consider Newton's algorithm defined by the map $\mathbf{A}(\mathbf{x}) = \mathbf{x} - \mathbf{H}(\mathbf{x})^{-1}\nabla f(\mathbf{x})$. Let $\bar{\mathbf{x}}$ be such that $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$ and $\mathbf{H}(\bar{\mathbf{x}})^{-1}$ exists. Let the starting point $\mathbf{x}_1$ be sufficiently close to $\bar{\mathbf{x}}$ so that this proximity implies that there exist k_1, $k_2 > 0$ with $k_1 k_2 \|\mathbf{x}_1 - \bar{\mathbf{x}}\| < 1$ such that

1. $\left\|\mathbf{H}(\bar{\mathbf{x}})^{-1}\right\|^{\dagger} \leq k_1$

and by the Taylor series expansion of ∇f,

2. $\left\|\nabla f(\bar{\mathbf{x}}) - \nabla f(\mathbf{x}) - \mathbf{H}(\mathbf{x})(\bar{\mathbf{x}} - \mathbf{x})\right\| \leq k_2 \left\|\bar{\mathbf{x}} - \mathbf{x}\right\|^2$

for each $\mathbf{x}$ satisfying $\left\|\mathbf{x} - \bar{\mathbf{x}}\right\| \leq \left\|\mathbf{x}_1 - \bar{\mathbf{x}}\right\|$. Then the algorithm converges superlinearly to $\bar{\mathbf{x}}$ with at least an order-two or quadratic rate of convergence.

Proof

Let the solution set $\Omega = \{\bar{\mathbf{x}}\}$ and let $X = \{\mathbf{x} : \left\|\mathbf{x} - \bar{\mathbf{x}}\right\| \leq \left\|\mathbf{x}_1 - \bar{\mathbf{x}}\right\|\}$. We prove convergence by using Theorem 7.2.3. Note that X is compact and that the map A given via (8.21) is closed on X. We now show that $\alpha(\mathbf{x}) = \left\|\mathbf{x} - \bar{\mathbf{x}}\right\|$ is indeed a descent function. Let $\mathbf{x} \in X$, and suppose that $\mathbf{x} \neq \bar{\mathbf{x}}$. Let $\mathbf{y} \in \mathrm{A}(\mathbf{x})$. Then, by the definition of A and since $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$, we get

$$\begin{aligned}\mathbf{y} - \bar{\mathbf{x}} &= (\mathbf{x} - \bar{\mathbf{x}}) - \mathbf{H}(\mathbf{x})^{-1}[\nabla f(\mathbf{x}) - \nabla f(\bar{\mathbf{x}})] \\ &= \mathbf{H}(\mathbf{x})^{-1}[\nabla f(\bar{\mathbf{x}}) - \nabla f(\mathbf{x}) - \mathbf{H}(\mathbf{x})(\bar{\mathbf{x}} - \mathbf{x})].\end{aligned}$$

Noting 1 and 2, it then follows that

$$\begin{aligned}\left\|\mathbf{y} - \bar{\mathbf{x}}\right\| &= \left\|\mathbf{H}(\mathbf{x})^{-1}[\nabla f(\bar{\mathbf{x}}) - \nabla f(\mathbf{x}) - \mathbf{H}(\mathbf{x})(\bar{\mathbf{x}} - \mathbf{x})]\right\| \\ &\leq \left\|\mathbf{H}(\mathbf{x})^{-1}\right\| \left\|\nabla f(\bar{\mathbf{x}}) - \nabla f(\mathbf{x}) - \mathbf{H}(\mathbf{x})(\bar{\mathbf{x}} - \mathbf{x})\right\| \\ &\leq k_1 k_2 \left\|\mathbf{x} - \bar{\mathbf{x}}\right\|^2 \leq k_1 k_2 \left\|\mathbf{x}_1 - \bar{\mathbf{x}}\right\| \left\|\mathbf{x} - \bar{\mathbf{x}}\right\| \\ &< \left\|\mathbf{x} - \bar{\mathbf{x}}\right\|.\end{aligned}$$

This shows that α is indeed a descent function. By the corollary to Theorem 7.2.3, we have convergence to $\bar{\mathbf{x}}$. Moreover, for any iterate $\mathbf{x}_k \in X$, the new iterate $\mathbf{y} = \mathbf{x}_{k+1}$ produced by the algorithm satisfies $\left\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}\right\| \leq k_1 k_2 \left\|\mathbf{x}_k - \bar{\mathbf{x}}\right\|^2$ from above. Since $\{\mathbf{x}_k\} \to \bar{\mathbf{x}}$, we have at least an order-two rate of convergence.

8.7 Modification of Newton's Method: Levenberg–Marquardt and Trust Region Methods

In Theorem 8.6.5 we have seen that if Newton's method is initialized close enough to a local minimum $\bar{\mathbf{x}}$ with a positive definite Hessian $\mathbf{H}(\bar{\mathbf{x}})$, then it converges quadratically to this solution. In general, we have observed that the

---

† See Appendix A.1 for the norm of a matrix.

method may not be defined because of the singularity of $\mathbf{H}(\mathbf{x}_k)$ at a given point $\mathbf{x}_k$, or the search direction $\mathbf{d}_k = -\mathbf{H}(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k)$ may not be a descent direction; or even if $\nabla f(\mathbf{x}_k)^t \mathbf{d}_k < 0$, a unit step size might not give a descent in f. To safeguard against the latter, we could perform a line search given that $\mathbf{d}_k$ is a descent direction. However, for the more critical issue of having a well-defined algorithm that converges to a point of zero gradient irrespective of the starting solution (i.e., enjoys *global convergence*), the following modifications can be adopted.

We first discuss a modification of Newton's method that guarantees convergence regardless of the starting point. Given $\mathbf{x}$, consider the direction $\mathbf{d} = -\mathbf{B}\nabla f(\mathbf{x})$, where $\mathbf{B}$ is a symmetric positive definite matrix to be determined later. The successor point is $\mathbf{y} = \mathbf{x} + \hat{\lambda}\mathbf{d}$, where $\hat{\lambda}$ is an optimal solution to the problem to minimize $f(\mathbf{x} + \lambda\mathbf{d})$ subject to $\lambda \geq 0$.

We now specify the matrix $\mathbf{B}$ as $(\varepsilon\mathbf{I} + \mathbf{H})^{-1}$, where $\mathbf{H} = \mathbf{H}(\mathbf{x})$. The scalar $\varepsilon \geq 0$ is determined as follows. Fix $\delta > 0$, and let $\varepsilon \geq 0$ be the smallest scalar that would make all the eigenvalues of the matrix $(\varepsilon\mathbf{I} + \mathbf{H})$ greater than or equal to δ. Since the eigenvalues of $\varepsilon\mathbf{I} + \mathbf{H}$ are all positive, $\varepsilon\mathbf{I} + \mathbf{H}$ is positive definite and invertible. In particular, $\mathbf{B} = (\varepsilon\mathbf{I} + \mathbf{H})^{-1}$ is also positive definite. Since the eigenvalues of a matrix depend continuously on its elements, ε is a continuous function of $\mathbf{x}$, and hence the point-to-point map $\mathbf{D}$: $R^n \to R^n \times R^n$ defined by $\mathbf{D}(\mathbf{x}) = (\mathbf{x}, \mathbf{d})$ is continuous. Thus, the algorithmic map is $\mathbf{A} = \mathbf{MD}$, where $\mathbf{M}$ is the usual line search map over $\{\lambda : \lambda \geq 0\}$.

Let $\Omega = \{\bar{\mathbf{x}} : \nabla f(\bar{\mathbf{x}}) = \mathbf{0}\}$, and let $\mathbf{x} \notin \Omega$. Since $\mathbf{B}$ is positive definite, $\mathbf{d} = -\mathbf{B}\nabla f(\mathbf{x}) \neq \mathbf{0}$; and, by Theorem 8.4.1, it follows that $\mathbf{M}$ is closed at $(\mathbf{x}, \mathbf{d})$. Furthermore, since $\mathbf{D}$ is a continuous function, by Corollary 2 to Theorem 7.3.2, $\mathbf{A} = \mathbf{MD}$ is closed over the complement of Ω.

To invoke Theorem 7.2.3, we need to specify a continuous descent function. Suppose that $\mathbf{x} \notin \Omega$, and let $\mathbf{y} \in \mathbf{A}(\mathbf{x})$. Note that $\nabla f(\mathbf{x})^t \mathbf{d} = -\nabla f(\mathbf{x})^t \mathbf{B}\nabla f(\mathbf{x}) < 0$ since $\mathbf{B}$ is positive definite and $\nabla f(\mathbf{x}) \neq \mathbf{0}$. Thus, $\mathbf{d}$ is a descent direction of f at $\mathbf{x}$, and by Theorem 4.1.2, $f(\mathbf{y}) < f(\mathbf{x})$. Therefore, f is indeed a descent function. Assuming that the sequence generated by the algorithm is contained in a compact set, by Theorem 7.2.3 it follows that the algorithm converges.

It should be noted that if the smallest eigenvalue of $\mathbf{H}(\bar{\mathbf{x}})$ is greater than or equal to δ, then, as the points $\{\mathbf{x}_k\}$ generated by the algorithm approach $\bar{\mathbf{x}}$, ε_k will be equal to zero. Thus, $\mathbf{d}_k = -\mathbf{H}(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k)$, and the algorithm reduces to that of Newton and, hence, this method also enjoys an order-two rate of convergence.

This underscores the importance of selecting δ properly. If δ is chosen to be too small to ensure the asymptotic quadratic convergence rate because of the reduction of the method to Newton's algorithm, ill-conditioning might occur at points where the Hessian is (near) singular. On the other hand, if δ is chosen to be very large, which would necessitate using a large value of ε and would make $\mathbf{B}$ diagonally dominant, the method would behave similar to the steepest descent algorithm, and only a linear convergence rate would be realized.

The foregoing algorithmic scheme of determining the new iterate $\mathbf{x}_{k+1}$ from an iterate $\mathbf{x}_k$ according to the solution of the system

$$[\varepsilon_k \mathbf{I} + \mathbf{H}(\mathbf{x}_k)](\mathbf{x}_{k+1} - \mathbf{x}_k) = -\nabla f(\mathbf{x}_k) \tag{8.22}$$

in lieu of (8.21) is generally known as a *Levenberg–Marquardt method*, following a similar scheme proposed for solving nonlinear least squares problems. A typical operational prescription for such a method is as follows. (The parameters 0.25, 0.75, 2, 4, etc., used below have been found to work well empirically, and the method is relatively insensitive to these parameter values.)

Given an iterate $\mathbf{x}_k$ and a parameter $\varepsilon_k > 0$, first ascertain the positive definiteness of $\varepsilon_k \mathbf{I} + \mathbf{H}(\mathbf{x}_k)$ by attempting to construct its Cholesky factorization $\mathbf{L}\mathbf{L}^t$ (see Appendix A.2). If this is unsuccessful, then multiply ε_k by a factor of 4 and repeat until such a factorization is available. Then solve the system (8.22) via $\mathbf{L}\mathbf{L}^t(\mathbf{x}_{k+1} - \mathbf{x}_k) = -\nabla f(\mathbf{x}_k)$, exploiting the triangularity of $\mathbf{L}$ to obtain $\mathbf{x}_{k+1}$. Compute $f(\mathbf{x}_{k+1})$ and determine R_k as the ratio of the actual decrease $f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})$ in f to its predicted decrease $q(\mathbf{x}_k) - q(\mathbf{x}_{k+1})$ as foretold by the quadratic approximation q to f at $\mathbf{x} = \mathbf{x}_k$. Note that the closer R_k is to unity, the more reliable is the quadratic approximation, and the smaller we can afford ε to be. With this motivation, if $R_k < 0.25$, put $\varepsilon_{k+1} = 4\varepsilon_k$; if $R_k > 0.75$, put $\varepsilon_{k+1} = \varepsilon_k/2$; otherwise, put $\varepsilon_{k+1} = \varepsilon_k$. Furthermore, in case $R_k \leq 0$ so that no improvement in f is realized, reset $\mathbf{x}_{k+1} = \mathbf{x}_k$; or else, retain the computed $\mathbf{x}_{k+1}$. Increment k by 1 and reiterate until convergence to a point of zero gradient is obtained.

A scheme of this type bears a close resemblance and relationship to *trust region methods*, or *restricted step methods*, for minimizing f. Note that the main difficulty with Newton's method is that the region of trust within which the quadratic approximation at a given point $\mathbf{x}_k$ can be considered to be sufficiently reliable might not include a point in the solution set. To circumvent this problem, we can consider the *trust region subproblem*:

$$\text{Minimize } \{q(\mathbf{x}) : \mathbf{x} \in \Omega_k\}, \tag{8.23}$$

where q is the quadratic approximation to f at $\mathbf{x} = \mathbf{x}_k$ and Ω_k is a trust region defined by $\Omega_k = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_k\| \leq \Delta_k\}$ for some *trust region parameter* $\Delta_k > 0$.

(Here $\|\cdot\|$ is the ℓ_2 norm; when the ℓ_∞ norm is used instead, the method is also known as the *box-step*, or *hypercube, method.*) Now let $\mathbf{x}_{k+1}$ solve (8.23) and, as before, define R_k as the ratio of the actual to the predicted descent. If R_k is too small relative to unity, then the trust region needs to be reduced; but if it is sufficiently respectable in value, the trust region can actually be expanded. The following is a typical prescription for defining Δ_{k+1} for the next iteration, where again, the method is known to be relatively insensitive to the specified parameter choices. If $R_k < 0.25$, put $\Delta_{k+1} = \|\mathbf{x}_{k+1} - \mathbf{x}_k\|/4$. If $R_k > 0.75$ and $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| = \Delta_k$, that is, the trust region constraint is binding in (8.23), then put $\Delta_{k+1} = 2\Delta_k$. Otherwise, retain $\Delta_{k+1} = \Delta_k$. Furthermore, in case $R_k \leq 0$ so that f did not improve at this iteration, reset $\mathbf{x}_{k+1}$ to $\mathbf{x}_k$ itself. Then increment k by 1 and repeat until a point with a zero gradient obtains. If this does not occur finitely, it can be shown that if the sequence $\{\mathbf{x}_k\}$ generated is contained in a compact set, and if f is continuously twice differentiable, then there exists an accumulation point $\bar{\mathbf{x}}$ of this sequence for which $\nabla f(\bar{\mathbf{x}}) = 0$ and $\mathrm{H}(\bar{\mathbf{x}})$ is positive semidefinite. Moreover, if $\mathrm{H}(\bar{\mathbf{x}})$ is positive definite, then for k sufficiently large, the trust region bound is inactive, and hence, the method reduces to Newton's method with a second-order rate of convergence (see the Notes and References section for further details).

There are two noteworthy points in relation to the foregoing discussion. First, wherever the actual Hessian has been employed above in the quadratic representation of f, an approximation to this Hessian can be used in practice, following quasi-Newton methods as discussed in the next section. Second, observe that by writing $\delta = \mathbf{x} - \mathbf{x}_k$ and equivalently, squaring both sides of the constraint defining Ω_k, we can write (8.23) explicitly as follows:

$$\text{Minimize}\left\{\nabla f(\mathbf{x}_k)^t \delta + \frac{1}{2}\delta^t \mathrm{H}(\mathbf{x}_k)\delta : \frac{1}{2}\|\delta\|^2 \leq \frac{1}{2}\Delta_k^2\right\}. \tag{8.24}$$

The KKT conditions for (8.24) require a nonnegative Lagrange multiplier λ and a primal feasible solution δ such that the following holds true in addition to the complementary slackness condition:

$$[\mathrm{H}(\mathbf{x}_k) + \lambda \mathrm{I}]\delta = -\nabla f(\mathbf{x}_k).$$

Note the resemblance of this to the Levenberg–Marquardt method given by (8.22). In particular, if $\Delta_k = -[\mathrm{H}(\mathbf{x}_k) + \varepsilon_k \mathrm{I}]^{-1}\nabla f(\mathbf{x}_k)$ in (8.24), where $\mathbf{H}(\mathbf{x}_k) + \varepsilon_k \mathrm{I}$ is positive definite, then, indeed, it is readily verified that $\delta = \mathbf{x}_{k+1} - \mathbf{x}_k$ given by (8.22) and $\lambda = \varepsilon_k$ satisfy the saddle point optimality conditions for (8.24) (see Exercise 8.29). Hence, the Levenberg–Marquardt scheme described above can be viewed as a trust region type of method as well.

Finally, let us comment on a *dog-leg trajectory* proposed by Powell, which more directly follows the philosophy described above of compromising between a steepest descent step and Newton's step, depending on the trust region size Δ_k. Referring to Figure 8.19, let $\mathbf{x}_{k+1}^{SD}$ and $\mathbf{x}_{k+1}^{N}$, respectively, denote the new iterate obtained via a steepest descent step, (8.16), and a Newton step, (8.21) ($\mathbf{x}_{k+1}^{SD}$ is sometimes also called the *Cauchy point*). The piecewise linear curve defined by the line segments joining $\mathbf{x}_k$ to $\mathbf{x}_{k+1}^{SD}$ and $\mathbf{x}_{k+1}^{SD}$ to $\mathbf{x}_{k+1}^{N}$ is called the *dog-leg trajectory*. It can be shown that along this trajectory, the distance from $\mathbf{x}_k$ increases monotonically while the objective value of the quadratic model falls. The proposed new iterate $\mathbf{x}_{k+1}$ is taken as the (unique) point at which the circle with radius Δ_k and centered at $\mathbf{x}_k$ intercepts this trajectory, if at all, as shown in Figure 8.19, and is taken as the Newton iterate $\mathbf{x}_{k+1}^{N}$ otherwise. Hence, when Δ_k is small relative to the dog-leg trajectory, the method behaves as a steepest descent algorithm; and with a relatively larger Δ_k, it reduces to Newton's method. Again, under suitable assumptions, as above, second-order convergence to a stationary point can be established. Moreover, the algorithmic step is simple and obviates (8.22) or (8.23). We refer the reader to the Notes and References section for further reading on this subject.

8.8 Methods Using Conjugate Directions: Quasi-Newton and Conjugate Gradient Methods

In this section we discuss several procedures that are based on the important concept of conjugacy. Some of these procedures use derivatives, whereas others use only functional evaluations. The notion of conjugacy defined below is very useful in unconstrained optimization. In particular, if the objective function is quadratic, then, by searching along conjugate directions, in any order, the minimum point can be obtained in, at most, n steps.

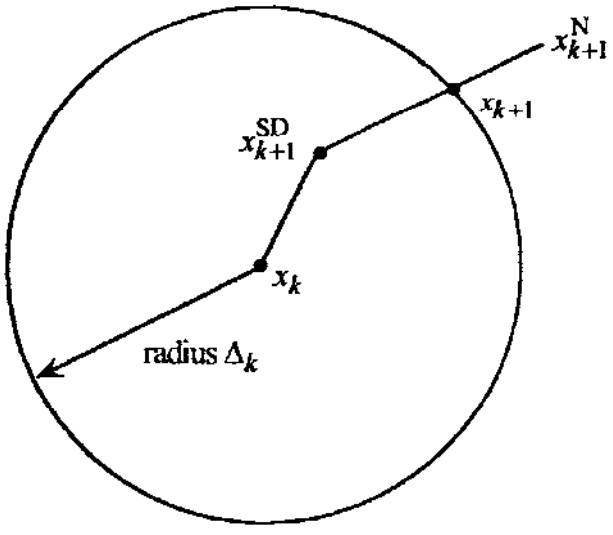


Figure 8.19 Dog-leg trajectory.

8.8.1 Definition

Let $\mathbf{H}$ be an $n \times n$ symmetric matrix. The vectors $\mathbf{d}_1,...,\mathbf{d}_n$ are called **H**-*conjugate* or simply *conjugate* if they are linearly independent and if $\mathbf{d}_i^t\mathbf{H}\mathbf{d}_j = 0$ for $i \neq j$.

It is instructive to observe the significance of conjugacy to the minimization of quadratic functions. Consider the quadratic function $f(\mathbf{x}) = \mathbf{c}^t\mathbf{x} + (1/2)\mathbf{x}^t\mathbf{H}\mathbf{x}$, where $\mathbf{H}$ is an $n \times n$ symmetric matrix, and suppose that $\mathbf{d}_1,...,\mathbf{d}_n$ are **H**–conjugate directions. By the linear independence of these direction vectors, given a starting point $\mathbf{x}_1$, any point $\mathbf{x}$ can be uniquely represented as $\mathbf{x} = \mathbf{x}_1 + \sum_{j=1}^{n}\lambda_j\mathbf{d}_j$. Using this substitution we can rewrite $f(\mathbf{x})$ as the following function of λ:

$$\mathbf{c}^t\mathbf{x}_1 + \sum_{j=1}^{n}\lambda_j\mathbf{c}^t\mathbf{d}_j + \frac{1}{2}\left(\mathbf{x}_1 + \sum_{j=1}^{n}\lambda_j\mathbf{d}_j\right)^t\mathbf{H}\left(\mathbf{x}_1 + \sum_{j=1}^{n}\lambda_j\mathbf{d}_j\right).$$

Using the **H**-conjugacy of $\mathbf{d}_1,...,\mathbf{d}_n$, this simplifies equivalently to minimizing

$$F(\lambda) \equiv \sum_{j=1}^{n}\left[\mathbf{c}^t(\mathbf{x}_1 + \lambda_j\mathbf{d}_j) + \frac{1}{2}(\mathbf{x}_1 + \lambda_j\mathbf{d}_j)^t\mathbf{H}(\mathbf{x}_1 + \lambda_j\mathbf{d}_j)\right].$$

Observe that F is separable in $\lambda_1,...,\lambda_n$ and can be minimized by minimizing each term in $[\cdot]$ independently and then composing the net result. Note that the minimization of each such term corresponds to minimizing f from $\mathbf{x}_1$ along the direction $\mathbf{d}_j$. (In particular, if **H** is positive definite, the minimizing value of λ_j is given by $\lambda_j^* = -[\mathbf{c}^t\mathbf{d}_j + \mathbf{x}_1^t\mathbf{H}\mathbf{d}_j]/\mathbf{d}_j^t\mathbf{H}\mathbf{d}_j$ for $j = 1,..., n$. Alternatively, the foregoing derivation readily reveals that the same minimizing step lengths λ_j^*, $j = 1,..., n$, result if we *sequentially* minimize f from $\mathbf{x}_1$ along the directions $\mathbf{d}_1,...,\mathbf{d}_n$ in any order, leading to an optimal solution.

The following example illustrates the notion of conjugacy and highlights the foregoing significance of optimizing along conjugate directions for quadratic functions.

8.8.2 Example

Consider the following problem:

$$\text{Minimize } -12x_2 + 4x_1^2 + 4x_2^2 + 4x_1x_2.$$

Note that the Hessian matrix **H** is given by

$$\mathbf{H} = \begin{bmatrix} 8 & -4 \\ -4 & 8 \end{bmatrix}.$$

We now generate two conjugate directions, d_1 and d_2. Suppose that we choose $d_1^t = (1, 0)$. Then $d_2^t = (a, b)$ must satisfy $0 = d_1^t H d_2 = 8a - 4b$. In particular, we may choose $a = 1$ and $b = 2$ so that $d_2^t = (1, 2)$. It may be noted that the conjugate directions are not unique.

If we minimize the objective function f starting from $x_1^t = (-1/2, 1)$ along the direction d_1, we get the point $x_2^t = (1/2, 1)$. Now, starting from x_2 and minimizing along d_2, we get $x_3^t = (1, 2)$. Note that x_3 is the minimizing point.

The contours of the objective function and the path taken to reach the optimal point are shown in Figure 8.20. The reader can easily verify that starting from any point and minimizing along d_1 and d_2, the optimal point is reached in, at most, two steps. For example, the dashed lines in Figure 8.20 exhibit the path obtained by sequentially minimizing along another pair of conjugate directions. Furthermore, if we had started at x_1 and then minimized along d_2 first and next along d_1, the optimizing step lengths along these respective directions would have remained the same as for the first case, taking the iterates from x_1 to $x_2' = (0, 2)^t$ to x_3.

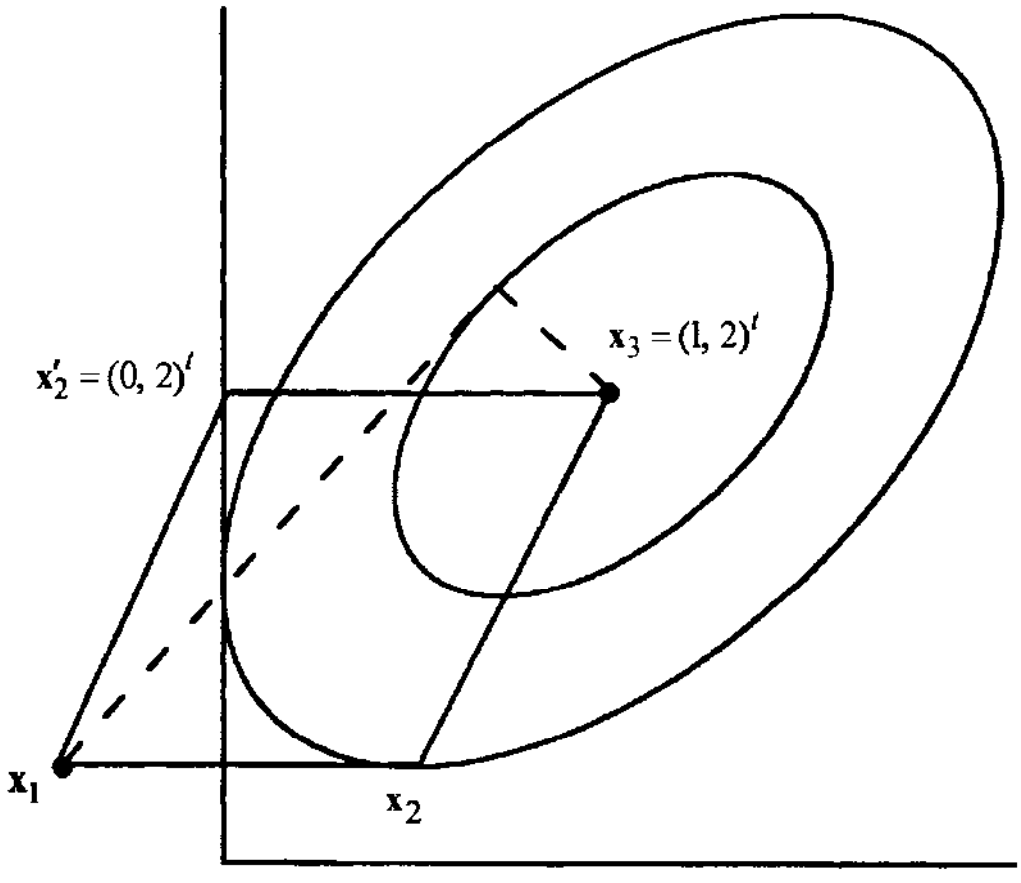


Figure 8.20 Illustration of conjugate directions.

Optimization of Quadratic Functions: Finite Convergence

Example 8.8.2 demonstrates that a quadratic function can be minimized in, at most, n steps whenever we search along conjugate directions of the Hessian matrix. This result is generally true for quadratic functions, as shown by Theorem 8.8.3. This, coupled with the fact that a general function can be closely represented by its quadratic approximation in the vicinity of the optimal point, makes the notion of conjugacy very useful for optimizing both quadratic and nonquadratic functions. Note also that this result shows that if we start at $\mathbf{x}_1$, then at each step $k = 1,..., n$, the point $\mathbf{x}_{k+1}$ obtained minimizes f over the linear subspace containing $\mathbf{x}_1$ that is spanned by the vectors $\mathbf{d}_1,...,\mathbf{d}_k$. Moreover, the gradient $\nabla f(\mathbf{x}_{k+1})$, if nonzero, is orthogonal to this subspace. This is sometimes called the *expanding subspace property* and is illustrated in Figure 8.21 for $k = 1, 2$.

8.8.3 Theorem

Let $f(\mathbf{x}) = \mathbf{c}^t\mathbf{x} + (1/2)\mathbf{x}^t\mathbf{H}\mathbf{x}$, where $\mathbf{H}$ is an $n \times n$ symmetric matrix. Let $\mathbf{d}_1,...,\mathbf{d}_n$ be $\mathbf{H}$-conjugate, and let $\mathbf{x}_1$ be an arbitrary starting point. For $k = 1,..., n$, let λ_k be an optimal solution to the problem to minimize $f(\mathbf{x}_k + \lambda\mathbf{d}_k)$ subject to $\lambda \in R$, and let $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k\mathbf{d}_k$. Then, for $k = 1,..., n$, we must have:

1. $\nabla f(\mathbf{x}_{k+1})^t\mathbf{d}_j = 0$ for $j = 1,..., k$.
2. $\nabla f(\mathbf{x}_1)^t\mathbf{d}_k = \nabla f(\mathbf{x}_k)^t\mathbf{d}_k$.
3. $\mathbf{x}_{k+1}$ is an optimal solution to the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} - \mathbf{x}_1 \in L(\mathbf{d}_1,...,\mathbf{d}_k)$, where $L(\mathbf{d}_1,...,\mathbf{d}_k)$ is the *linear subspace* formed by $\mathbf{d}_1,...,\mathbf{d}_k$; that is, $L(\mathbf{d}_1,...,\mathbf{d}_k) = \{\sum_{j=1}^{k}\mu_j\mathbf{d}_j : \mu_j \in R$ for each $j\}$. In particular, $\mathbf{x}_{n+1}$ is a minimizing point of f over R^n.

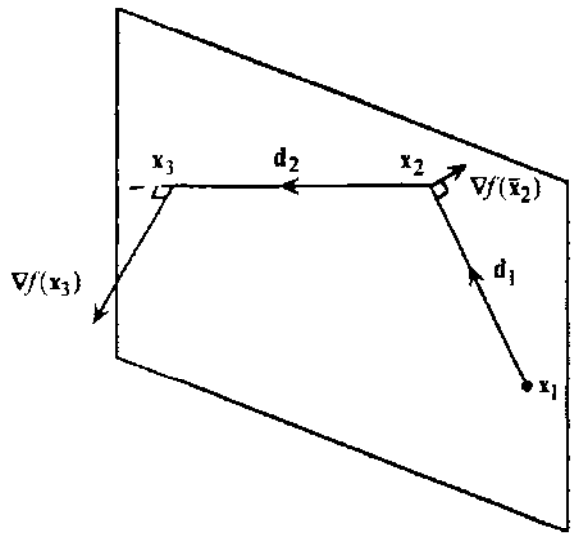


Figure 8.21 Expanding subspace property.

Proof

To prove Part 1, first note that $f(\mathbf{x}_j + \lambda \mathbf{d}_j)$ achieves a minimum at λ_j only if $\nabla f(\mathbf{x}_j + \lambda_j \mathbf{d}_j)^t \mathbf{d}_j = 0$; that is, $\nabla f(\mathbf{x}_{j+1})^t \mathbf{d}_j = 0$. Thus, Part 1 holds true for $j = k$. For $j < k$, note that

$$\begin{aligned} \nabla f(\mathbf{x}_{k+1}) = \mathbf{c} + \mathbf{H}\mathbf{x}_{j+1} &= \mathbf{c} + \mathbf{H}\mathbf{x}_{j+1} + \mathbf{H}\left(\sum_{i=j+1}^{k} \lambda_i \mathbf{d}_i\right) \\ &= \nabla f(\mathbf{x}_{j+1}) + \mathbf{H}\left(\sum_{i=j+1}^{k} \lambda_i \mathbf{d}_i\right). \end{aligned} \tag{8.25}$$

By conjugacy, $\mathbf{d}_i^t \mathbf{H} \mathbf{d}_j = 0$ for $i = j + 1, \ldots, k$. Thus, from (8.25) it follows that $\nabla f(\mathbf{x}_{k+1})^t \mathbf{d}_j = 0$, and Part 1 holds true.

Replacing k by $k - 1$ and letting $j = 0$ in (8.25), we get

$$\nabla f(\mathbf{x}_k) = \nabla f(\mathbf{x}_1) + \mathbf{H}\left(\sum_{i=1}^{k-1} \lambda_i \mathbf{d}_i\right) \qquad \text{for } k \geq 2.$$

Multiplying by $\mathbf{d}_k^t$ and noting that $\mathbf{d}_k^t \mathbf{H} \mathbf{d}_i = 0$ for $i = 1, \ldots, k - 1$ shows that Part 2 holds true for $k \geq 2$. Part 2 holds true trivially for $k = 1$.

To show Part 3, since $\mathbf{d}_i^t \mathbf{H} \mathbf{d}_j = 0$ for $i \neq j$, we get

$$\begin{aligned} f(\mathbf{x}_{k+1}) = f[\mathbf{x}_1 + (\mathbf{x}_{k+1} - \mathbf{x}_1)] &= f\left(\mathbf{x}_1 + \sum_{j=1}^{k} \lambda_j \mathbf{d}_j\right) \\ &= f(\mathbf{x}_1) + \nabla f(\mathbf{x}_1)^t \left(\sum_{j=1}^{k} \lambda_j \mathbf{d}_j\right) + \frac{1}{2} \sum_{j=1}^{k} \lambda_j^2 \mathbf{d}_j^t \mathbf{H} \mathbf{d}_j. \end{aligned} \tag{8.26}$$

Now suppose that $\mathbf{x} - \mathbf{x}_1 \in L(\mathbf{d}_1, \ldots, \mathbf{d}_k)$, so that $\mathbf{x}$ can be written as $\mathbf{x}_1 + \sum_{j=1}^{k} \mu_j \mathbf{d}_j$. As in (8.26), we get

$$f(\mathbf{x}) = f(\mathbf{x}_1) + \nabla f(\mathbf{x}_1)^t \left(\sum_{j=1}^{k} \mu_j \mathbf{d}_j\right) + \frac{1}{2} \sum_{j=1}^{k} \mu_j^2 \mathbf{d}_j^t \mathbf{H} \mathbf{d}_j. \tag{8.27}$$

To complete the proof, we need to show that $f(\mathbf{x}) \geq f(\mathbf{x}_{k+1})$. By contradiction, suppose that $f(\mathbf{x}) < f(\mathbf{x}_{k+1})$. Then by (8.26) and (8.27), we must have

$$\nabla f(\mathbf{x}_1)^t\left(\sum_{j=1}^{k}\mu_j\mathbf{d}_j\right)+\frac{1}{2}\sum_{j=1}^{k}\mu_j^2\mathbf{d}_j^t\mathbf{H}\mathbf{d}_j$$
$$< \nabla f(\mathbf{x}_1)^t\left(\sum_{j=1}^{k}\lambda_j\mathbf{d}_j\right)+\frac{1}{2}\sum_{j=1}^{k}\lambda_j^2\mathbf{d}_j^t\mathbf{H}\mathbf{d}_j. \tag{8.28}$$

By the definition of λ_j, note that $f(\mathbf{x}_j + \lambda_j\mathbf{d}_j) \le f(\mathbf{x}_j + \mu_j\mathbf{d}_j)$ for each j. Therefore,

$$f(\mathbf{x}_j)+\lambda_j\nabla f(\mathbf{x}_j)^t\mathbf{d}_j+\frac{1}{2}\lambda_j^2\mathbf{d}_j^t\mathbf{H}\mathbf{d}_j \le f(\mathbf{x}_j)+\mu_j\nabla f(\mathbf{x}_j)^t\mathbf{d}_j+\frac{1}{2}\mu_j^2\mathbf{d}_j^t\mathbf{H}\mathbf{d}_j.$$

By Part 2, $\nabla f(\mathbf{x}_j)^t\mathbf{d}_j = \nabla f(\mathbf{x}_1)^t\mathbf{d}_j$, and substituting this in the inequality above, we get

$$\lambda_j\nabla f(\mathbf{x}_1)^t\mathbf{d}_j+\frac{1}{2}\lambda_j^2\mathbf{d}_j^t\mathbf{H}\mathbf{d}_j \le \mu_j\nabla f(\mathbf{x}_1)^t\mathbf{d}_j+\frac{1}{2}\mu_j^2\mathbf{d}_j^t\mathbf{H}\mathbf{d}_j. \tag{8.29}$$

Summing (8.29) for $j = 1,..., k$ contradicts (8.28). Thus, $\mathbf{x}_{k+1}$ is a minimizing point over the manifold $\mathbf{x}_1 + L(\mathbf{d}_1,...,\mathbf{d}_k)$. In particular, since $\mathbf{d}_1,...,\mathbf{d}_n$ are linearly independent, $L(\mathbf{d}_1,...,\mathbf{d}_n) = R^n$, and hence, $\mathbf{x}_{n+1}$ is a minimizing point of f over R^n. This completes the proof.

Generating Conjugate Directions

In the remainder of this section we describe several methods for generating conjugate directions for quadratic forms. These methods lead naturally to powerful algorithms for minimizing both quadratic and nonquadratic functions. In particular, we discuss the classes of quasi-Newton and conjugate gradient methods.

Quasi-Newton Methods: Method of Davidon–Fletcher–Powell

This method was proposed by Davidon [1959] and later developed by Fletcher and Powell [1963]. The Davidon–Fletcher–Powell (DFP) method falls under the general class of *quasi-Newton procedures*, where the search directions are of the form $\mathbf{d}_j = -\mathbf{D}_j\nabla f(\mathbf{y})$, in lieu of $-\mathbf{H}^{-1}(\mathbf{y})\nabla f(\mathbf{y})$, as in Newton's method. The negative gradient direction is thus deflected by premultiplying it by $-\mathbf{D}_j$, where $\mathbf{D}_j$ is an $n \times n$ positive definite symmetric matrix that approximates the inverse of the Hessian matrix. The positive definiteness property ensures that $\mathbf{d}_j$ is a descent direction whenever $\nabla f(\mathbf{y}) \neq \mathbf{0}$, since then, $\mathbf{d}_j^t\nabla f(\mathbf{y}) < 0$. For the purpose of the next step, $\mathbf{D}_{j+1}$ is formed by adding to $\mathbf{D}_j$ two symmetric

matrices, each of rank one. Thus, this scheme is sometimes referred to as a *rank-two correction procedure.* For quadratic functions, this update scheme is shown later to produce the exact representation of the actual inverse Hessian within n steps. The DFP process is also called a *variable metric method* because it can be interpreted as adopting the steepest descent step in the transformed space based on the Cholesky factorization of the positive definite matrix $\mathbf{D}_j$, as discussed in Section 8.7, where this transformation varies with $\mathbf{D}_j$ from iteration to iteration. The quasi-Newton methods in which the quadratic approximation is permitted to be possibly indefinite are more generally called *secant methods.*

Summary of the Davidon–Fletcher–Powell (DFP) Method

We now summarize the Davidon–Fletcher–Powell (DFP) method for minimizing a differentiable function of several variables. In particular, if the function is quadratic, then, as shown later, the method yields conjugate directions and terminates in one complete iteration, that is, after searching along each of the conjugate directions as described below.

Initialization Step Let $\varepsilon > 0$ be a termination tolerance. Choose an initial point $\mathbf{x}_1$ and an initial symmetric positive definite matrix $\mathbf{D}_1$. Let $\mathbf{y}_1 = \mathbf{x}_1$, let $k = j = 1$, and go to the Main Step.

Main Step

1. If $\|\nabla f(\mathbf{y}_j)\| < \varepsilon$, stop; otherwise, let $\mathrm{d}_j = -\mathbf{D}_j \nabla f(\mathbf{y}_j)$ and let λ_j be an optimal solution to the problem to minimize $f(\mathbf{y}_j + \lambda \mathrm{d}_j)$ subject to $\lambda \geq 0$. Let $\mathbf{y}_{j+1} = \mathbf{y}_j + \lambda_j \mathrm{d}_j$. If $j < n$, go to Step 2. If $j = n$, let $\mathbf{y}_1 = \mathbf{x}_{k+1} = \mathbf{y}_{n+1}$, replace k by $k + 1$, let $j = 1$, and repeat Step 1.
2. Construct $\mathbf{D}_{j+1}$ as follows:

$$\mathbf{D}_{j+1} = \mathbf{D}_j + \frac{\mathrm{p}_j \mathrm{p}_j^t}{\mathrm{p}_j^t \mathbf{q}_j} - \frac{\mathbf{D}_j \mathbf{q}_j \mathbf{q}_j^t \mathbf{D}_j}{\mathbf{q}_j^t \mathbf{D}_j \mathbf{q}_j}, \tag{8.30}$$

where

$$\mathrm{p}_j \equiv \lambda_j \mathrm{d}_j = \mathbf{y}_{j+1} - \mathbf{y}_j \tag{8.31}$$

$$\mathbf{q}_j \equiv \nabla f(\mathbf{y}_{j+1}) - \nabla f(\mathbf{y}_j). \tag{8.32}$$

Replace j by $j + 1$, and go to Step 1.

We remark here that the inner loop of the foregoing algorithm resets the procedure every n steps (whenever $j = n$ at Step 1). Any variant that resets every $n' < n$ inner iteration steps is called a *partial quasi-Newton method.* This strategy

can be useful from the viewpoint of conserving storage when $n' << n$, since then, the inverse Hessian approximation can be stored implicitly by, instead, storing only the generating vectors $\mathbf{p}_j$ and $\mathbf{q}_j$ themselves within the inner loop iterations.

8.8.4 Example

Consider the following problem:

$$\text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2.$$

A summary of the computations using the DFP method is given in Table 8.13. At each iteration, for $j = 1, 2$, d_j is given by $-\mathbf{D}_j \nabla f(\mathbf{y}_j)$, where $\mathbf{D}_1$ is the identity matrix and $\mathbf{D}_2$ is computed from (8.30), (8.31), and (8.32). At Iteration $k = 1$, we have $\mathbf{p}_1 = (2.7, -1.49)^t$ and $\mathbf{q}_1 = (44.73, -22.72)^t$ in (8.30). At Iteration 2 we have $\mathbf{p}_1 = (-0.1, 0.05)^t$ and $\mathbf{q}_1 = (-0.7, 0.8)^t$, and finally, at Iteration 3 we have $\mathbf{p}_1 = (-0.02, 0.02)^t$ and $\mathbf{q}_1 = (-0.14, 0.24)^t$. The point $\mathbf{y}_{j+1}$ is computed by optimizing along the direction d_j starting from $\mathbf{y}_j$ for $j = 1, 2$. The procedure is terminated at the point $\mathbf{y}_2 = (2.115, 1.058)^t$ in the fourth iteration, since $\|\nabla f(\mathbf{y}_2)\| = 0.006$ is quite small. The path taken by the method is depicted in Figure 8.22.

Lemma 8.8.5 shows that each matrix $\mathbf{D}_j$ is positive definite and d_j is a direction of descent.

8.8.5 Lemma

Let $\mathbf{y}_1 \in R^n$, and let $\mathbf{D}_1$ be an initial positive definite symmetric matrix. For $j = 1, \ldots, n$, let $\mathrm{y}_{j+1} = \mathbf{y}_j + \lambda_j \mathrm{d}_j$, where $\mathrm{d}_j = -\mathbf{D}_j \nabla f(\mathbf{y}_j)$, and λ_j solves the problem to minimize $f(\mathbf{y}_j + \lambda \mathbf{d}_j)$ subject to $\lambda \geq 0$. Furthermore, for $j = 1, \ldots, n-1$, let $\mathbf{D}_{j+1}$ be given by (8.30), (8.31), and (8.32). If $\nabla f(\mathbf{y}_j) \neq \mathbf{0}$ for $j = 1, \ldots, n$, $\mathbf{D}_1, \ldots, \mathbf{D}_n$ are symmetric and positive definite so that $\mathrm{d}_1 \ldots, \mathrm{d}_n$ are descent directions.

Proof

We prove the result by induction. For $j = 1$, $\mathbf{D}_1$ is symmetric and positive definite by assumption. Furthermore, $\nabla f(\mathbf{y}_1)^t \mathrm{d}_1 = -\nabla f(\mathbf{y}_1)^t \mathbf{D}_1 \nabla f(\mathbf{y}_1) < 0$, since $\mathbf{D}_1$ is positive definite. By Theorem 4.1.2, d_1 is a descent direction. We

Table 8.13 Summary of Computations for the Davidon–Fletcher–Powell Method

Iteration k	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	j	$\mathbf{y}_j$ $f(\mathbf{y}_j)$	$\nabla f(\mathbf{y}_j)$	$\lVert\nabla f(\mathbf{y}_j)\rVert$	$\mathbf{D}_j$	$\mathbf{d}_j$	λ_j	$\mathbf{y}_{j+1}$
1	(0.00, 3.00) 52.00	1	(0.00, 3.00) 52.00	(−44.00, 24.00)	50.12	$\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$	(44.00, −24.00)	0.062	(2.70, 1.51)
		2	(2.70, 1.51) 0.34	(0.73, 1.28)	1.47	$\begin{bmatrix}0.25 & 0.38\\ 0.38 & 0.81\end{bmatrix}$	(−0.67, −1.31)	0.22	(2.55, 1.22)
2	(2.55, 1.22) 0.1036	1	(2.55, 1.22) 0.1036	(0.89, −0.44)	0.99	$\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$	(−0.89, 0.44)	0.11	(2.45, 1.27)
		2	(2.45, 1.27) 0.0490	(0.18, 0.36)	0.40	$\begin{bmatrix}0.65 & 0.45\\ 0.45 & 0.46\end{bmatrix}$	(−0.28, −0.25)	0.64	(2.27, 1.11)
3	(2.27, 1.11) 0.008	1	(2.27, 1.11) 0.008	(0.18, −0.20)	0.27	$\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$	(−0.18, 0.20)	0.10	(2.25, 1.13)
		2	(2.25, 1.13) 0.004	(0.04, 0.04)	0.06	$\begin{bmatrix}0.80 & 0.38\\ 0.38 & 0.31\end{bmatrix}$	(−0.05, −0.03)	2.64	(2.12, 1.05)
4	(2.12, 1.05) 0.0005	1	(2.12, 1.05) 0.0005	(0.05, −0.08)	0.09	$\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$	(−0.05, 0.08)	0.10	(2.115, 1.058)
		2	(2.115, 1.058) 0.0002	(0.004, 0.004)	0.006				

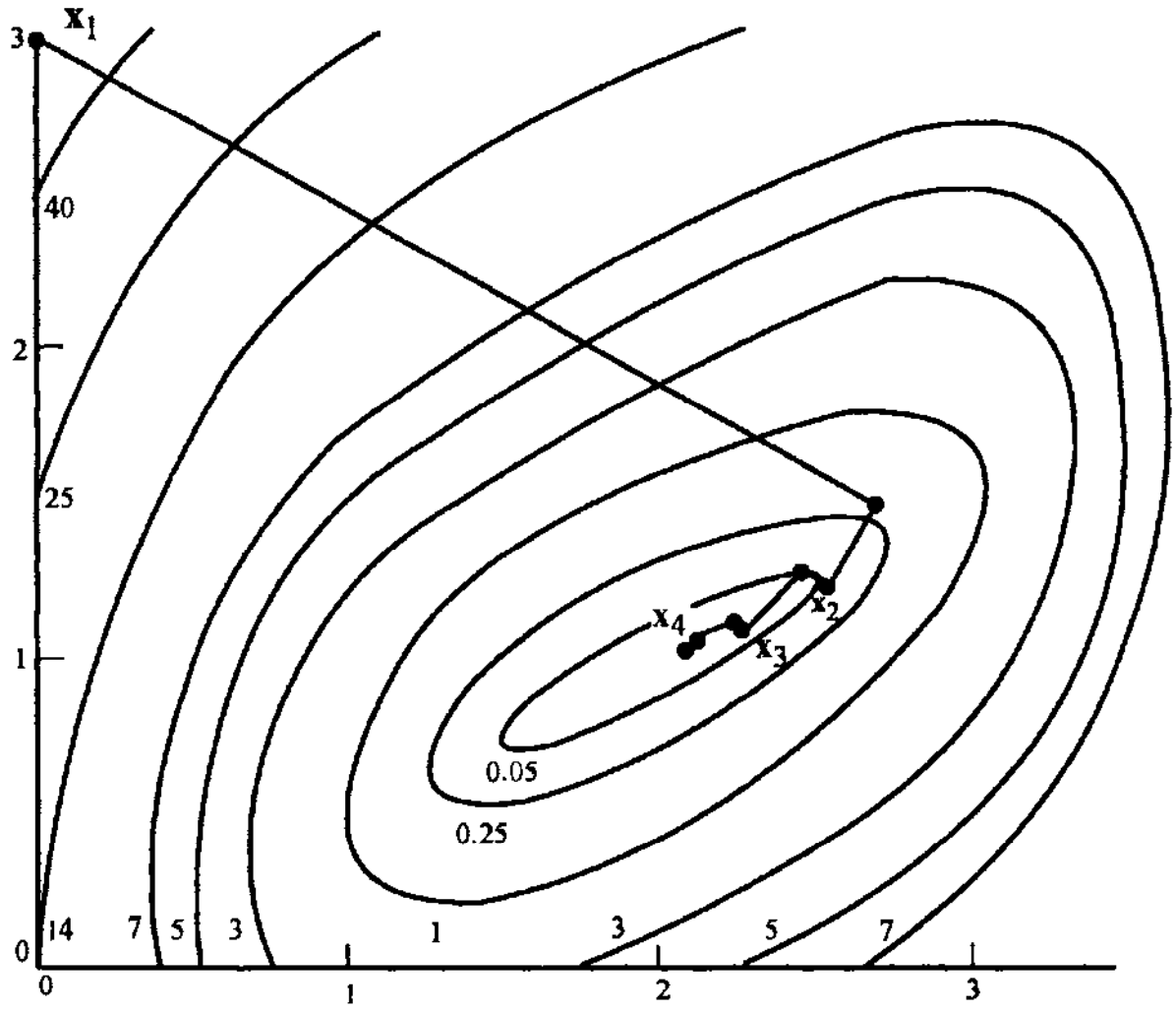


Figure 8.22 Davidon–Fletcher–Powell method.

shall assume that the result holds true for $j \leq n - 1$ and then show that it holds for $j + 1$. Let $\mathbf{x}$ be a nonzero vector in R^n; then, by (8.30), we have

$$\mathbf{x}^t\mathbf{D}_{j+1}\mathbf{x} = \mathbf{x}^t\mathbf{D}_j\mathbf{x} + \frac{(\mathbf{x}^t\mathbf{p}_j)^2}{\mathbf{p}_j^t\mathbf{q}_j} - \frac{(\mathbf{x}^t\mathbf{D}_j\mathbf{q}_j)^2}{\mathbf{q}_j^t\mathbf{D}_j\mathbf{q}_j}. \tag{8.33}$$

Since $\mathbf{D}_j$ is a symmetric positive definite matrix, there exists a positive definite symmetric matrix $\mathbf{D}_j^{1/2}$ such that $\mathbf{D}_j = \mathbf{D}_j^{1/2}\mathbf{D}_j^{1/2}$. Let $\mathbf{a} = \mathbf{D}_j^{1/2}\mathbf{x}$ and $\mathbf{b} = \mathbf{D}_j^{1/2}\mathbf{q}_j$. Then $\mathbf{x}^t\mathbf{D}_j\mathbf{x} = \mathbf{a}^t\mathbf{a}$, $\mathbf{q}_j^t\mathbf{D}_j\mathbf{q}_j = \mathbf{b}^t\mathbf{b}$, and $\mathbf{x}^t\mathbf{D}_j\mathbf{q}_j = \mathbf{a}^t\mathbf{b}$. Substituting in (8.33), we get

$$\mathbf{x}^t\mathbf{D}_{j+1}\mathbf{x} = \frac{(\mathbf{a}^t\mathbf{a})(\mathbf{b}^t\mathbf{b}) - (\mathbf{a}^t\mathbf{b})^2}{\mathbf{b}^t\mathbf{b}} + \frac{(\mathbf{x}^t\mathbf{p}_j)^2}{\mathbf{p}_j^t\mathbf{q}_j}. \tag{8.34}$$

By the Schwartz inequality, $(\mathbf{a}^t\mathbf{a})(\mathbf{b}^t\mathbf{b}) \geq (\mathbf{a}^t\mathbf{b})^2$. Thus, to show that $\mathbf{x}^t\mathbf{D}_{j+1}\mathbf{x} \geq 0$, it suffices to show that $\mathbf{p}_j^t\mathbf{q}_j > 0$ and that $\mathbf{b}^t\mathbf{b} > 0$. From (8.31) and (8.32) it follows that

$$\mathbf{p}_j^t\mathbf{q}_j = \lambda_j \mathbf{d}_j^t[\nabla f(\mathbf{y}_{j+1}) - \nabla f(\mathbf{y}_j)].$$

The reader may note that $\mathbf{d}_j^t\nabla f(\mathbf{y}_{j+1}) = 0$, and by definition, $\mathbf{d}_j = -\mathbf{D}_j\nabla f(\mathbf{y}_j)$. Substituting these in the above equation, it follows that

$$\mathbf{p}_j^t\mathbf{q}_j = \lambda_j \nabla f(\mathbf{y}_j)^t \mathbf{D}_j \nabla f(\mathbf{y}_j). \tag{8.35}$$

Note that $\nabla f(\mathbf{y}_j) \neq \mathbf{0}$ by assumption, and that $\mathbf{D}_j$ is positive definite, so that $\nabla f(\mathbf{y}_j)^t \mathbf{D}_j \nabla f(\mathbf{y}_j) > 0$. Furthermore, $\mathbf{d}_j$ is a descent direction and, hence, $\lambda_j > 0$. Therefore, from (8.35), $\mathbf{p}_j^t\mathbf{q}_j > 0$. Furthermore, $\mathbf{q}_j \neq \mathbf{0}$ and, hence, $\mathbf{b}^t\mathbf{b} = \mathbf{q}_j^t\mathbf{D}_j\mathbf{q}_j > 0$.

We now show that $\mathbf{x}^t\mathbf{D}_{j+1}\mathbf{x} > 0$. By contradiction, suppose that $\mathbf{x}^t\mathbf{D}_{j+1}\mathbf{x} = 0$. This is possible only if $(\mathbf{a}^t\mathbf{a})(\mathbf{b}^t\mathbf{b}) = (\mathbf{a}^t\mathbf{b})^2$ and $\mathbf{p}_j^t\mathbf{x} = 0$. First, note that $(\mathbf{a}^t\mathbf{a})(\mathbf{b}^t\mathbf{b}) = (\mathbf{a}^t\mathbf{b})^2$ only if $\mathbf{a} = \lambda\mathbf{b}$; that is, $\mathbf{D}_j^{1/2}\mathbf{x} = \lambda\mathbf{D}_j^{1/2}\mathbf{q}_j$. Thus, $\mathbf{x} = \lambda\mathbf{q}_j$. Since $\mathbf{x} \neq \mathbf{0}$, we have $\lambda \neq 0$. Now $0 = \mathbf{p}_j^t\mathbf{x} = \lambda\mathbf{p}_j^t\mathbf{q}_j$ contradicts the fact that $\mathbf{p}_j^t\mathbf{q}_j > 0$ and $\lambda \neq 0$. Therefore, $\mathbf{x}^t\mathbf{D}_{j+1}\mathbf{x} > 0$, so that $\mathbf{D}_{j+1}$ is positive definite.

Since $\nabla f(\mathbf{y}_{j+1}) \neq \mathbf{0}$ and since $\mathbf{D}_{j+1}$ is positive definite, $\nabla f(\mathbf{y}_{j+1})^t \mathbf{d}_{j+1} = -\nabla f(\mathbf{y}_{j+1})^t \mathbf{D}_{j+1}\nabla f(\mathbf{y}_{j+1}) < 0$. By Theorem 4.1.2, then, $\mathbf{d}_{j+1}$ is a descent direction. This completes the proof.

Quadratic Case

If the objective function f is quadratic, then by Theorem 8.8.6, the directions $\mathbf{d}_1,...,\mathbf{d}_n$ generated by the DFP method are conjugate. Therefore, by Part 3 of Theorem 8.8.3, the method stops after one complete iteration with an optimal solution. Furthermore, the matrix $\mathbf{D}_{n+1}$ obtained at the end of the iteration is precisely the inverse of the Hessian matrix H.

8.8.6 Theorem

Let H be an $n \times n$ symmetric positive definite matrix, and consider the problem to minimize $f(\mathbf{x}) = \mathbf{c}^t\mathbf{x} + (1/2)\mathbf{x}^t\mathbf{H}\mathbf{x}$ subject to $\mathbf{x} \in R^n$. Suppose that the

problem is solved by the DFP method, starting with an initial point $\mathbf{y}_1$ and a symmetric positive definite matrix $\mathbf{D}_1$. In particular, for $j = 1,\ldots, n$, let λ_j be an optimal solution to the problem to minimize $f(\mathbf{y}_j + \lambda \mathrm{d}_j)$ subject to $\lambda \geq 0$, and let $\mathbf{y}_{j+1} = \mathbf{y}_j + \lambda_j \mathrm{d}_j$, where $\mathrm{d}_j = -\mathbf{D}_j \nabla f(\mathbf{y}_j)$ and $\mathbf{D}_j$ is determined by (8.30), (8.31), and (8.32). If $\nabla f(\mathbf{y}_j) \neq \mathbf{0}$ for each j, then the directions $\mathrm{d}_1,\ldots,\mathrm{d}_n$ are $\mathbf{H}$-conjugate and $\mathbf{D}_{n+1} = \mathbf{H}^{-1}$. Furthermore, $\mathbf{y}_{n+1}$ is an optimal solution to the problem.

Proof

We first show that for any j with $1 \leq j \leq n$, we must have the following conditions:

1. $\mathrm{d}_1,\ldots,\mathrm{d}_j$ are linearly independent.
2. $\mathrm{d}_i^t \mathbf{H} \mathrm{d}_k = 0$ for $i \neq k$; $i, k \leq j$. (8.36)
3. $\mathbf{D}_{j+1}\mathbf{H}\mathrm{p}_k = \mathrm{p}_k$ or, equivalently, $\mathbf{D}_{j+1}\mathbf{H}\mathrm{d}_k = \mathrm{d}_k$ for $1 \leq k \leq j$, where $\mathrm{p}_k = \lambda_k \mathrm{d}_k$.

We prove this result by induction. For $j = 1$, parts 1 and 2 are obvious. To prove Part 3, first note that for any k, we have

$$\mathbf{H}\mathrm{p}_k = \mathbf{H}(\lambda_k \mathrm{d}_k) = \mathbf{H}(\mathbf{y}_{k+1} - \mathbf{y}_k) = \nabla f(\mathbf{y}_{k+1}) - \nabla f(\mathbf{y}_k) = \mathbf{q}_k. \tag{8.37}$$

In particular, $\mathbf{H}\mathrm{p}_1 = \mathbf{q}_1$. Thus, letting $j = 1$ in (8.30), we get

$$\mathbf{D}_2\mathbf{H}\mathrm{p}_1 = \left(\mathbf{D}_1 + \frac{\mathrm{p}_1\mathrm{p}_1^t}{\mathrm{p}_1^t\mathbf{q}_1} - \frac{\mathbf{D}_1\mathbf{q}_1\mathbf{q}_1^t\mathbf{D}_1}{\mathbf{q}_1^t\mathbf{D}_1\mathbf{q}_1}\right)\mathbf{q}_1 = \mathrm{p}_1$$

so that Part 3 holds true for $j = 1$.

Now suppose that Parts 1, 2, and 3 hold true for $j \leq n-1$. To show that they also hold true for $j + 1$, first recall by Part 1 of Theorem 8.8.3 that $\mathrm{d}_i^t \nabla f(\mathbf{y}_{j+1}) = 0$ for $i \leq j$. By the induction hypothesis of Part 3, $\mathrm{d}_i = \mathbf{D}_{j+1}\mathbf{H}\mathrm{d}_i$ for $i \leq j$. Thus, for $i \leq j$ we have

$$0 = \mathrm{d}_i^t \nabla f(\mathbf{y}_{j+1}) + \mathrm{d}_i^t \mathbf{H}\mathbf{D}_{j+1}\nabla f(\mathrm{y}_{j+1}) = -\mathrm{d}_i^t \mathbf{H}\mathrm{d}_{j+1}.$$

In view of the induction hypothesis for Part 2, the above equation shows that Part 2 also holds true for $j + 1$.

Now, we show that Part 3 holds true for $j + 1$. Letting $k \leq j + 1$ yields

$$\mathbf{D}_{j+2}\mathbf{H}\mathrm{p}_k = \left(\mathbf{D}_{j+1} + \frac{\mathrm{p}_{j+1}\mathrm{p}_{j+1}^t}{\mathrm{p}_{j+1}^t\mathbf{q}_{j+1}} - \frac{\mathbf{D}_{j+1}\mathbf{q}_{j+1}\mathbf{q}_{j+1}^t\mathbf{D}_{j+1}}{\mathbf{q}_{j+1}^t\mathbf{D}_{j+1}\mathbf{q}_{j+1}}\right)\mathbf{H}\mathrm{p}_k. \tag{838}$$

Noting (8.37) and letting $k = j + 1$ in (8.38), it follows that $\mathbf{D}_{j+2}\mathbf{H}\mathrm{p}_{j+1} = \mathrm{p}_{j+1}$. Now let $k \leq j$. Since Part 2 holds true for $j + 1$,

$$\mathrm{p}_{j+1}^t \mathbf{H}\mathrm{p}_k = \lambda_k \lambda_{j+1} \mathrm{d}_{j+1}^t \mathbf{H}\mathrm{d}_k = 0. \tag{8.39}$$

Noting the induction hypothesis for Part 3, (8.37), and the fact that Part 2 holds true for $j + 1$, we get

$$\mathbf{q}_{j+1}^t \mathbf{D}_{j+1}\mathbf{H}\mathrm{p}_k = \mathbf{q}_{j+1}^t \mathrm{p}_k = \mathrm{p}_{j+1}^t \mathbf{H}\mathrm{p}_k = \lambda_{j+1}\lambda_k \mathrm{d}_{j+1}^t \mathbf{H}\mathrm{d}_k = 0. \tag{8.40}$$

Substituting (8.39) and (8.40) in (8.38), and noting the induction hypothesis for Part 3, we get

$$\mathbf{D}_{j+2}\mathbf{H}\mathrm{p}_k = \mathbf{D}_{j+1}\mathbf{H}\mathrm{p}_k = \mathrm{p}_k.$$

Thus, Part 3 holds true for $j + 1$.

To complete the induction argument, we only need to show that Part 1 holds true for $j + 1$. Suppose that $\sum_{i=1}^{j+1} \alpha_i \mathrm{d}_i = \mathbf{0}$. Multiplying by $\mathrm{d}_{j+1}^t \mathbf{H}$ and noting that Part 2 holds true for $j + 1$, it follows that $\alpha_{j+1} \mathrm{d}_{j+1}^t \mathbf{H}\mathrm{d}_{j+1} = 0$. By assumption, $\nabla f(\mathrm{y}_{j+1}) \neq \mathbf{0}$, and by Lemma 8.8.5, $\mathbf{D}_{j+1}$ is positive definite, so that $\mathrm{d}_{j+1} = -\mathbf{D}_{j+1}\nabla f(\mathbf{y}_{j+1}) \neq \mathbf{0}$. Since $\mathbf{H}$ is positive definite, $\mathrm{d}_{j+1}^t \mathbf{H}\mathrm{d}_{j+1} \neq 0$, and hence, $\alpha_{j+1} = 0$. This in turn implies that $\sum_{i=1}^{j} \alpha_i \mathrm{d}_i = \mathbf{0}$; and since $\mathrm{d}_1, \ldots, \mathrm{d}_j$ are linearly independent by the induction hypothesis, $\alpha_i = 0$ for $i = 1, \ldots, j$. Thus, $\mathrm{d}_1, \ldots, \mathrm{d}_{j+1}$ are linearly independent and Part 1 holds true for $j + 1$. Thus, Parts 1, 2, and 3 hold true. In particular, the conjugacy of $\mathrm{d}_1, \ldots, \mathrm{d}_n$ follows from Parts 1 and 2 by letting $j = n$.

Now, let $j = n$ in Part 3. Then $\mathbf{D}_{n+1}\mathbf{H}\mathrm{d}_k = \mathrm{d}_k$ for $k = 1, \ldots, n$. If we let $\mathbf{D}$ be the matrix whose columns are $\mathrm{d}_1, \ldots, \mathrm{d}_n$, then $\mathbf{D}_{n+1}\mathbf{H}\mathbf{D} = \mathbf{D}$. Since $\mathbf{D}$ is invertible, $\mathbf{D}_{n+1}\mathbf{H} = \mathbf{I}$, which is possible only if $\mathbf{D}_{n+1} = \mathbf{H}^{-1}$. Finally, $\mathbf{y}_{n+1}$ is an optimal solution by Theorem 8.8.3.

Insightful Derivation of the DFP Method

At each step of the DFP method we have seen that given some approximation $\mathbf{D}_j$ to the inverse Hessian matrix, we computed the search direction $\mathrm{d}_j = -\mathbf{D}_j \nabla f(\mathbf{y}_j)$ by deflecting the negative gradient of f at the current solution $\mathbf{y}_j$ using this approximation $\mathbf{D}_j$ in the spirit of Newton's method. We then performed a line search along this direction, and based on the resulting solution $\mathbf{y}_{j+1}$ and the gradient $\nabla f(\mathbf{y}_{j+1})$ at this point, we obtained an updated approximation $\mathbf{D}_{j+1}$ according to (8.30), (8.31), and (8.32). As seen in Theorem

8.8.6, if f is a quadratic function given by $f(\mathbf{x}) = \mathbf{c}^t\mathbf{x} + (1/2)\mathbf{x}^t\mathbf{H}\mathbf{x}$; $\mathbf{x} \in R^n$, where $\mathbf{H}$ is symmetric and positive definite; and if $\nabla f(\mathbf{y}_j) \neq \mathbf{0}, j = 1,..., n$, then we indeed obtain $\mathbf{D}_{n+1} = \mathbf{H}^{-1}$. In fact, observe from Parts 1 and 3 of Theorem 8.8.6 that for each $j \in \{1,..., n\}$, the vectors $\mathbf{p}_1,...,\mathbf{p}_j$ are linearly independent eigenvectors of $\mathbf{D}_{j+1}\mathbf{H}$ having eigenvalues equal to 1. Hence, at each step of the method, the revised approximation accumulates one additional linearly independent eigenvector, with a unit eigenvalue for the product $\mathbf{D}_{j+1}\mathbf{H}$, until $\mathbf{D}_{n+1}\mathbf{H}$ finally has all its n eigenvalues equal to 1, giving $\mathbf{D}_{n+1}\mathbf{H}\mathbf{P} = \mathbf{P}$, where $\mathbf{P}$ is the nonsingular matrix of the eigenvectors of $\mathbf{D}_{n+1}\mathbf{H}$. Hence, $\mathbf{D}_{n+1}\mathbf{H} = \mathbf{I}$, or $\mathbf{D}_{n+1} = \mathbf{H}^{-1}$.

Based on the foregoing observation, let us derive the update scheme (8.30) for the DFP method and use this derivation to motivate other more prominent updates. Toward this end, suppose that we have some symmetric, positive definite approximation $\mathbf{D}_j$ of the inverse Hessian matrix for which $\mathbf{p}_1,...,\mathbf{p}_{j-1}$ are the eigenvectors of $\mathbf{D}_j\mathbf{H}$ with unit eigenvalues. (For $j = 1$, no such vector exists.) Adopting the inductive scheme of Theorem 8.8.6, assume that these eigenvectors are linearly independent and are H-conjugate. Now, given the current point $\mathbf{y}_j$, we conduct a line search along the direction $\mathbf{d}_j = -\mathbf{D}_j\nabla f(\mathbf{y}_j)$ to obtain the new point $\mathbf{y}_{j+1}$ and, accordingly, we define

$$\begin{aligned} \mathbf{p}_j &= (\mathbf{y}_{j+1} - \mathbf{y}_j) \\ \mathbf{q}_j &= \nabla f(\mathbf{y}_{j+1}) - \nabla f(\mathbf{y}_j) = \mathbf{H}(\mathbf{y}_{j+1} - \mathbf{y}_j) = \mathbf{H}\mathbf{p}_j. \end{aligned} \tag{8.41}$$

Following the argument in the proof of Theorem 8.8.6, the vectors $\mathbf{p}_k = \lambda_k\mathbf{d}_k$, $k = 1,..., j$, are easily shown to be linearly independent and H-conjugate. We now want to construct a matrix

$$\mathbf{D}_{j+1} = \mathbf{D}_j + \mathbf{C}_j,$$

where $\mathbf{C}_j$ is some symmetric correction matrix, which ensures that $\mathbf{p}_1,...,\mathbf{p}_j$ are eigenvectors of $\mathbf{D}_{j+1}\mathbf{H}$ having unit eigenvalues. Hence, we want $\mathbf{D}_{j+1}\mathbf{H}\mathbf{p}_k = \mathbf{p}_k$ or, from (8.41), that $\mathbf{D}_{j+1}\mathbf{q}_k = \mathbf{p}_k$ for $k = 1,...,j$. For $1 \leq k < j$, this translates to requiring that $\mathbf{p}_k = \mathbf{D}_j\mathbf{q}_k + \mathbf{C}_j\mathbf{q}_k = \mathbf{D}_j\mathbf{H}\mathbf{p}_k + \mathbf{C}_j\mathbf{q}_k = \mathbf{p}_k + \mathbf{C}_j\mathbf{q}_k$, or that

$$\mathbf{C}_j\mathbf{q}_k = \mathbf{0} \qquad \text{for } k = 1,..., j-1. \tag{8.42}$$

For $k = j$, the aforementioned condition

$$\mathbf{D}_{j+1}\mathbf{q}_j = \mathbf{p}_j \tag{8.43}$$

is called the *quasi-Newton condition*, or the *secant equation*, the latter term leading to the alternative name *secant updates* for this type of scheme. This condition translates to the requirement that

$$\mathbf{C}_j\mathbf{q}_j = \mathbf{p}_j - \mathbf{D}_j\mathbf{q}_j. \tag{8.44}$$

Now if $\mathbf{C}_j$ had a symmetric rank-one term $\mathbf{p}_j\mathbf{p}_j^t/\mathbf{p}_j^t\mathbf{q}_j$, then $\mathbf{C}_j\mathbf{q}_j$ operating on this term would yield $\mathbf{p}_j$, as required in (8.44). Similarly, if $\mathbf{C}_j$ had a symmetric rank-one term, $-(\mathbf{D}_j\mathbf{q}_j)(\mathbf{D}_j\mathbf{q}_i)^t/(\mathbf{D}_j\mathbf{q}_j)^t\mathbf{q}_j$, then $\mathbf{C}_j\mathbf{q}_j$ operating on this term would yield $-\mathbf{D}_j\mathbf{q}_j$, as required in (8.44). This therefore leads to the *rank-two DFP update* (8.30) via the correction term,

$$\mathbf{C}_j = \frac{\mathbf{p}_j\mathbf{p}_j^t}{\mathbf{p}_j^t\mathbf{q}_j} - \frac{\mathbf{D}_j\mathbf{q}_j\mathbf{q}_j^t\mathbf{D}_j}{\mathbf{q}_j^t\mathbf{D}_j\mathbf{q}_j} \equiv \mathbf{C}_j^{\mathrm{DFP}}, \tag{8.45}$$

which satisfies the quasi-Newton condition (8.43) via (8.44). (Note that as in Lemma 8.8.5, $\mathbf{D}_{j+1} = \mathbf{D}_j + \mathbf{C}_j$ is symmetric and positive definite.) Moreover, (8.42) also holds since for any $k \in \{1,\ldots, j-1\}$, we have from (8.45) and (8.41) that

$$\mathbf{C}_j\mathbf{q}_k = \mathbf{C}_j\mathbf{H}\mathbf{p}_k = \frac{\mathbf{p}_j\mathbf{p}_j^t\mathbf{H}\mathbf{p}_k}{\mathbf{p}_j^t\mathbf{q}_j} - \frac{\mathbf{D}_j\mathbf{q}_j\mathbf{p}_j^t\mathbf{H}\mathbf{D}_j\mathbf{H}\mathbf{p}_k}{\mathbf{q}_j^t\mathbf{D}_j\mathbf{q}_j} = \mathbf{0}$$

since $\mathbf{p}_j^t\mathbf{H}\mathbf{p}_k = 0$ in the first term and $\mathbf{p}_j^t\mathbf{H}\mathbf{D}_j\mathbf{H}\mathbf{p}_k = \mathbf{p}_j^t\mathbf{H}\mathbf{p}_k = 0$ in the second term as well. Hence, following this sequence of corrections, we shall ultimately obtain $\mathbf{D}_{n+1}\mathbf{H} = \mathbf{I}$ or $\mathbf{D}_{n+1} = \mathbf{H}^{-1}$.

Broyden Family and Broyden–Fletcher–Goldfarb–Shanno (BFGS) Updates

The reader might have observed in the foregoing derivation of $\mathbf{C}_j^{\mathrm{DFP}}$ that there was a degree of flexibility in prescribing the correction matrix $\mathbf{C}_j$, the restriction being to satisfy the quasi-Newton condition (8.44) along with (8.42) and to maintain symmetry and positive definiteness of $\mathbf{D}_{j+1} = \mathbf{D}_j + \mathbf{C}_j$. In light of this, the Broyden updates suggest the use of the correction matrix $\mathbf{C}_j = \mathbf{C}_j^B$ given by the following family parameterized by ϕ:

$$\mathbf{C}_j^B = \mathbf{C}_j^{\mathrm{DFP}} + \frac{\phi\tau_j\mathbf{v}_j\mathbf{v}_j^t}{\mathbf{p}_j^t\mathbf{q}_j}, \tag{8.46}$$

where $\mathbf{v}_j \equiv \mathbf{p}_j - (1/\tau_j)\mathbf{D}_j\mathbf{q}_j$ and where τ_j is chosen so that the quasi-Newton condition (8.44) holds by virtue of $\mathbf{v}_j^t\mathbf{q}_j$ being zero. This implies that $[\mathbf{p}_j - \mathbf{D}_j\mathbf{q}_j/\tau_j]^t\mathbf{q}_j = 0$, or that

$$\tau_j = \frac{\mathbf{q}_j^t\mathbf{D}_j\mathbf{q}_j}{\mathbf{p}_j^t\mathbf{q}_j} > 0. \tag{8.47}$$

Note that for $1 \leq k < j$, we have

$$\mathbf{v}_j^t\mathbf{q}_k = \mathbf{p}_j^t\mathbf{q}_k - \frac{1}{\tau_j}\mathbf{q}_j^t\mathbf{D}_j\mathbf{q}_k = \mathbf{p}_j^t\mathbf{H}\mathbf{p}_k - \frac{1}{\tau_j}\mathbf{p}_j^t\mathbf{H}\mathbf{D}_j\mathbf{H}\mathbf{p}_k = 0$$

because $\mathbf{p}_j^t\mathbf{H}\mathbf{p}_k = 0$ by conjugacy and $\mathbf{p}_j^t[\mathbf{D}_j\mathbf{H}\mathbf{p}_k] = \mathbf{p}_j^t\mathbf{H}\mathbf{p}_k = 0$, since $\mathbf{p}_k$ is an eigenvector of $\mathbf{D}_j\mathbf{H}$ having a unit eigenvalue. Hence, (8.42) also continues to hold true. Moreover, it is clear that $\mathbf{D}_{j+1} = \mathbf{D}_j + \mathbf{C}_j^B$ continues to be symmetric and, at least for $\phi \geq 0$, positive definite. Hence, the correction matrix (8.46)–(8.47) in this case yields a valid sequence of updates satisfying the assertion of Theorem 8.8.6.

For the value $\phi = 1$, the Broyden family yields a very useful special case, which coincides with that derived independently by Broyden, Fletcher, Goldfarb, and Shanno. This update, known as the *BFGS update*, or the *positive definite secant update*, has been consistently shown in many computational studies to dominate other updating schemes in its overall performance. In contrast, the DFP update has been observed to exhibit numerical difficulties, sometimes having the tendency to produce near-singular Hessian approximations. The additional correction term in (8.46) seems to alleviate this propensity.

To derive this update correction $\mathbf{C}_j^{\text{BFGS}}$, say, we simply substitute (8.47) into (8.46), and simplify (8.46) using $\phi = 1$ to get

$$\mathbf{C}_j^{\text{BFGS}} \equiv \mathbf{C}_j^B(\phi = 1) = \frac{\mathbf{p}_j\mathbf{p}_j^t}{\mathbf{p}_j^t\mathbf{q}_j}\left(1 + \frac{\mathbf{q}_j^t\mathbf{D}_j\mathbf{q}_j}{\mathbf{p}_j^t\mathbf{q}_j}\right) - \frac{\mathbf{D}_j\mathbf{q}_j\mathbf{p}_j^t + \mathbf{p}_j\mathbf{q}_j^t\mathbf{D}_j}{\mathbf{p}_j^t\mathbf{q}_j}. \tag{8.48}$$

Since with $\phi = 0$ we have $\mathbf{C}_j^B = \mathbf{C}_j^{\text{DFP}}$, we can write (8.46) as

$$\mathbf{C}_j^B = (1-\phi)\mathbf{C}_j^{\text{DFP}} + \phi\mathbf{C}_j^{\text{BFGS}}. \tag{8.49}$$

The above discussion assumes the use of a constant value ϕ in (8.46). This is known as a *pure Broyden update*. However, for the analytical results to hold true, it is not necessary to work with a constant value of ϕ. A variable value ϕ_j can be chosen from one iteration to the next if so desired. However, there is a

value of ϕ in (8.46) that will make $\mathbf{d}_{j+1} = -\mathbf{D}_{j+1}\nabla f(\mathbf{y}_{j+1})$ identically zero (see Exercise 8.35), namely,

$$\phi = \frac{-[\nabla f(\mathbf{y}_j)^t \mathbf{D}_j \nabla f(\mathbf{y}_j)]\mathbf{q}_j^t \mathbf{D}_j \mathbf{q}_j}{\nabla f(\mathbf{y}_{j+1})^t \mathbf{D}_j \nabla f(\mathbf{y}_{j+1})}. \tag{8.50}$$

Hence, the algorithm stalls and, in particular, $\mathbf{D}_{j+1}$ becomes singular and loses positive definiteness. Such a value of ϕ is said to be *degenerate*, and should be avoided. For this reason, as a safeguard, ϕ is usually taken to be nonnegative, although sometimes, admitting negative values seems to be computationally attractive. In this connection, note that for a general differentiable function, if *perfect line searches* are performed (i.e., either an exact minimum, or in the nonconvex case, the first local minimum along a search direction is found), then it can be shown that the sequence of iterates generated by the Broyden family is invariant with respect to the choice of the parameter ϕ as long as nondegenerate ϕ values are chosen (see the Notes and References section). Hence, the choice of ϕ becomes critical only with inexact line searches. Also, if inaccurate line searches are used, then maintaining the positive definiteness of the Hessian approximations becomes a matter of concern. In particular, this motivates the following strategy.

Updating Hessian Approximations

Note that in a spirit similar to the foregoing derivations, we could alternatively have started with a symmetric, positive definite approximation $\mathbf{B}_1$ to the Hessian $\mathbf{H}$ itself, and then updated this to produce a sequence of symmetric, positive definite approximations according to $\mathbf{B}_{j+1} = \mathbf{B}_j + \bar{\mathbf{C}}_j$ for $j = 1,..., n$. Again, for each $j = 1,..., n$, we would like $\mathbf{p}_1,...,\mathbf{p}_j$ to be eigenvectors of $\mathbf{H}^{-1}\mathbf{B}_{j+1}$ having eigenvalues of 1, so that for $j = n$ we would obtain $\mathbf{H}^{-1}\mathbf{B}_{n+1} = \mathbf{I}$ or $\mathbf{B}_{n+1} = \mathbf{H}$ itself. Proceeding inductively as before, given that $\mathbf{p}_1,...,\mathbf{p}_{j-1}$ are eigenvectors of $\mathbf{H}^{-1}\mathbf{B}_j$ associated with unit eigenvalues, we need to construct a correction matrix $\bar{\mathbf{C}}_j$ such that $\mathbf{H}^{-1}(\mathbf{B}_j + \bar{\mathbf{C}}_j)\mathbf{p}_k = \mathbf{p}_k$ for $k = 1, ..., j$. In other words, multiplying throughout by $\mathbf{H}$ and noting that $\mathbf{q}_k = \mathbf{H}\mathbf{p}_k$ for $k = 1,..., j$ by (8.41), if we are given that

$$\mathbf{B}_j\mathbf{p}_k = \mathbf{q}_k \qquad \text{for } k = 1,..., j-1 \tag{8.51}$$

we are required to ensure that $(\mathbf{B}_j + \bar{\mathbf{C}}_j)\mathbf{p}_k = \mathbf{q}_k$ for $k = 1,..., j$ or, using (8.51), that

$$\bar{\mathbf{C}}_j\mathbf{p}_k = \mathbf{0} \text{ for } 1 \le k \le j-1 \qquad \text{and} \qquad \bar{\mathbf{C}}_j\mathbf{p}_j = \mathbf{q}_j - \mathbf{B}_j\mathbf{p}_j. \tag{8.52}$$

Comparing (8.51) with the condition $\mathbf{D}_j\mathbf{q}_k = \mathrm{p}_k$ for $k = 1,\ldots, j-1$ and, similarly, comparing (8.52) with (8.42) and (8.44), we observe that the present analysis differs from the foregoing analysis involving an update of inverse Hessians in that the role of $\mathbf{D}_j$ and $\mathbf{B}_j$, and that of p_j and $\mathbf{q}_j$, are interchanged. By symmetry, we can derive a formula for $\bar{\mathbf{C}}_j$ simply by replacing $\mathbf{D}_j$ by $\mathbf{B}_j$ and by interchanging p_j and $\mathbf{q}_j$ in (8.45). An update obtained in this fashion is called a *complementary update*, or *dual update*, to the preceding one. Of course, the dual of the dual formula will naturally yield the original formula. The $\bar{\mathbf{C}}_j$ derived as the dual to $\mathbf{C}_j^{\mathrm{DFP}}$ was actually obtained independently by Broyden, Fletcher, Goldfarb, and Shanno in 1970, and the update is therefore known as the BFGS update. Hence, we have

$$\bar{\mathbf{C}}_j^{\mathrm{BFGS}} = \frac{\mathbf{q}_j\mathbf{q}_j^t}{\mathbf{q}_j^t\mathrm{p}_j} - \frac{\mathbf{B}_j\mathrm{p}_j\mathrm{p}_j^t\mathbf{B}_j}{\mathrm{p}_j^t\mathbf{B}_j\mathrm{p}_j}. \tag{8.53}$$

In Exercise 8.37 we ask the reader to derive (8.53) directly following the derivation of (8.41) through (8.45).

Note that the relationship between $\bar{\mathbf{C}}_j^{\mathrm{BFGS}}$ and $\mathbf{C}_j^{\mathrm{BFGS}}$ is as follows:

$$\mathbf{D}_{j+1} = \mathbf{D}_j + \mathbf{C}_j^{\mathrm{BFGS}} = \mathbf{B}_{j+1}^{-1} = (\mathbf{B}_j + \bar{\mathbf{C}}_j^{\mathrm{BFGS}})^{-1}. \tag{8.54}$$

That is, $\mathbf{D}_{j+1}\mathbf{q}_k = \mathrm{p}_k$ for $k = 1,\ldots, j$ implies that $\mathbf{D}_{j+1}^{-1}\mathrm{p}_k = \mathbf{q}_k$ or that $\mathbf{B}_{j+1} = \mathbf{D}_{j+1}^{-1}$ indeed satisfies (8.51) (written for $j+1$). In fact, the inverse relationship (8.54) between (8.48) and (8.53) can readily be verified (see Exercise 8.36) by using two sequential applications of the *Sherman–Morrison–Woodbury formula* given below, which is valid for any general $n \times n$ matrix $\mathbf{A}$ and $n \times 1$ vectors $\mathbf{a}$ and $\mathbf{b}$, given that the inverse exists (or equivalently, given that $1 + \mathbf{b}^t\mathbf{A}^{-1}\mathbf{a} \neq 0$):

$$(\mathbf{A} + \mathbf{a}\mathbf{b}^t)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{a}\mathbf{b}^t\mathbf{A}^{-1}}{1 + \mathbf{b}^t\mathbf{A}^{-1}\mathbf{a}}. \tag{8.55}$$

Note that if the Hessian approximations $\mathbf{B}_j$ are generated as above, then the search direction d_j at any step needs to be obtained by solving the system of equations $\mathbf{B}_j\mathrm{d}_j = -\nabla f(\mathbf{y}_j)$. This can be done more conveniently by maintaining a Cholesky factorization $\mathfrak{L}_j\mathfrak{D}_j\mathfrak{L}_j^t$ of $\mathbf{B}_j$, where $\mathfrak{L}_j$ is a lower triangular matrix and $\mathfrak{D}_j$ is a diagonal matrix. Besides the numerical benefits of adopting this procedure, it can also be helpful in that the condition number of $\mathfrak{D}_j$ can be useful in assessing the ill-conditioning status of $\mathbf{B}_j$, and the positive

definiteness of $\mathbf{B}_j$ can be verified by checking the positivity of the diagonal elements of $\mathcal{D}_j$. Hence, when an update of $\mathcal{D}_j$ reveals a loss of positive definiteness, alternative steps can be taken by restoring the diagonal elements of $\mathcal{D}_{j+1}$ to be positive.

Scaling of Quasi-Newton Algorithms

Let us conclude our discussion on quasi-Newton methods by making a brief but important comment on adopting a proper scaling of the updates generated by these methods. In our discussion leading to the derivation of (8.41)–(8.45), we learned that at each step j, the revised update $\mathbf{D}_{j+1}$ had an additional eigenvector associated with a unit eigenvalue for the matrix $\mathbf{D}_{j+1}\mathbf{H}$. Hence, if, for example, $\mathbf{D}_1$ is chosen such that the eigenvalues of $\mathbf{D}_1\mathbf{H}$ are all significantly larger than unity, then since these eigenvalues are transformed to unity one at a time as the algorithm proceeds, one can expect an unfavorable ratio of the largest to smallest eigenvalues of $\mathbf{D}_j\mathbf{H}$ at the intermediate steps. When minimizing nonquadratic functions and/or employing inexact line searches, in particular, such a phenomenon can result in ill-conditioning effects and exhibit poor convergence performance. To alleviate this, it is useful to multiply each $\mathbf{D}_j$ by some scale factor $s_j > 0$ before using the update formula. With exact line searches, this can be shown to preserve the conjugacy property in the quadratic case, although we may no longer have $\mathbf{D}_{n+1} = \mathbf{H}^{-1}$. However, the focus here is to improve the single-step rather than the n-step convergence behavior of the algorithm. Methods that automatically prescribe scale factors in a manner such that if the function is quadratic, then the eigenvalues of $s_j\mathbf{D}_j\mathbf{H}$ tend to be spread above and below unity are called *self-scaling methods*. We refer the reader to the Notes and References section for further reading on this subject.

Conjugate Gradient Methods

Conjugate gradient methods were proposed by Hestenes and Stiefel in 1952 for solving systems of linear equations. The use of this method for unconstrained optimization was prompted by the fact that the minimization of a positive definite quadratic function is equivalent to solving the linear equation system that results when its gradient is set at zero. Actually, conjugate gradient methods were first extended to solving nonlinear equation systems and general unconstrained minimization problems by Fletcher and Reeves in 1964. Although these methods are typically less efficient and less robust than quasi-Newton methods, they have very modest storage requirements (only three n-vectors are required for the method of Fletcher and Reeves described below) and are quite indispensable for large problems (n exceeding about 100) when quasi-Newton methods become impractical because of the size of the Hessian matrix. Some

very successful applications are reported by Fletcher [1987] in the context of atomic structures, where problems having 3000 variables were solved using only about 50 gradient evaluations, and by Reid [1971], who solved some linear partial differential equations having some 4000 variables in about 40 iterations. Moreover, conjugate gradient methods have the advantage of simplicity, being *gradient deflection methods* that deflect the negative gradient direction using the previous direction. This deflection can alternatively be viewed as an update of a *fixed*, symmetric, positive definite matrix, usually the identity matrix, in the spirit of quasi-Newton methods. For this reason they are sometimes referred to as *fixed-metric methods* in contrast to the term *variable-metric methods*, which applies to quasi-Newton procedures. Again, these are conjugate direction methods that converge in, at most, n iterations for unconstrained quadratic optimization problems in R^n when using exact line searches. In fact, for the latter case, they generate directions identical to the BFGS method, as shown later.

The basic approach of conjugate gradient methods for minimizing a differentiable function f: $R^n \rightarrow R$ is to generate a sequence of iterates $\mathbf{y}_j$ according to

$$\mathbf{y}_{j+1} = \mathbf{y}_j + \lambda_j \mathbf{d}_j \tag{8.56a}$$

where $\mathbf{d}_j$ is the search direction and λ_j is the step length that minimizes f along $\mathbf{d}_j$ from the point $\mathbf{y}_j$. For $j = 1$, the search direction $\mathbf{d}_1 = -\nabla f(\mathbf{y}_1)$ can be used, and for subsequent iterations, given $\mathbf{y}_{j+1}$ with $\nabla f(\mathbf{y}_{j+1}) \neq 0$ for $j \geq 1$, we use

$$\mathbf{d}_{j+1} = -\nabla f(\mathbf{y}_{j+1}) + \alpha_j \mathbf{d}_j, \tag{8.56b}$$

where α_j is a suitable deflection parameter that characterizes a particular conjugate gradient method. Note that we can write $\mathbf{d}_{j+1}$ in (8.56b) whenever $\alpha_j \geq 0$ as

$$\mathbf{d}_{j+1} = \frac{1}{\mu}[\mu[-\nabla f(\mathbf{y}_{j+1})] + (1-\mu)\mathbf{d}_j],$$

where $\mu = 1/(1 + \alpha_j)$, so $\mathbf{d}_{j+1}$ can then be essentially viewed as a convex combination of the current steepest descent direction and the direction used at the last iteration.

Now suppose that we assume f to be a quadratic function having a positive definite Hessian $\mathbf{H}$, and that we require $\mathbf{d}_{j+1}$ and $\mathbf{d}_j$ to be H-conjugate. From (8.56a) and (8.41), $\mathbf{d}_{j+1}^t \mathbf{H} \mathbf{d}_j = 0$ amounts to requiring that $0 = \mathbf{d}_{j+1}^t \mathbf{H} \mathbf{p}_j =$

$\mathbf{d}_{j+1}^t \mathbf{q}_j$. Using this in (8.56b) gives Hestenes and Stiefel's [1952] choice for α_j, used even in nonquadratic situations by assuming a local quadratic behavior, as

$$\alpha_j^{\mathrm{HS}} = \frac{\nabla f(\mathbf{y}_{j+1})^t \mathbf{q}_j}{\mathbf{d}_j^t \mathbf{q}_j} = \frac{\lambda_j \nabla f(\mathbf{y}_{j+1})^t \mathbf{q}_j}{\mathbf{p}_j^t \mathbf{q}_j}. \tag{8.57}$$

When exact line searches are performed, we have $\mathbf{d}_j^t \nabla f(\mathbf{y}_{j+1}) = 0 = \mathbf{d}_{j-1}^t \nabla f(\mathbf{y}_j)$, leading to $\mathbf{d}_j^t \mathbf{q}_j = -\mathbf{d}_j^t \nabla f(\mathbf{y}_j) = [\nabla f(\mathbf{y}_j) - \alpha_{j-1}\mathbf{d}_{j-1}]^t \nabla f(\mathbf{y}_j) = \left\|\nabla f(\mathbf{y}_j)\right\|^2$. Substituting this into (8.57) yields Polak and Ribiere's [1969] choice for α_j as

$$\alpha_j^{\mathrm{PR}} = \frac{\nabla f(\mathbf{y}_{j+1})^t \mathbf{q}_j}{\left\|\nabla f(\mathbf{y}_j)\right\|^2}. \tag{8.58}$$

Furthermore, if f is quadratic and if exact line searches are performed, we have, using (8.56) along with $\nabla f(\mathbf{y}_{j+1})^t \mathbf{d}_j = 0 = \nabla f(\mathbf{y}_j)^t \mathbf{d}_{j-1}$ as above, that

$$\begin{aligned}
\nabla f(\mathbf{y}_{j+1})^t \nabla f(\mathbf{y}_j) &= \nabla f(\mathbf{y}_{j+1})^t [\alpha_{j-1}\mathbf{d}_{j-1} - \mathbf{d}_j] \\
&= \alpha_{j-1} \nabla f(\mathbf{y}_{j+1})^t \mathbf{d}_{j-1} = \alpha_{j-1} [\nabla f(\mathbf{y}_j) + \lambda_j \mathbf{H}\mathbf{d}_j]^t \mathbf{d}_{j-1} \\
&= \alpha_{j-1} \lambda_j \mathbf{d}_j^t \mathbf{H} \mathbf{d}_{j-1} = 0
\end{aligned}$$

by the H-conjugacy of $\mathbf{d}_j$ and $\mathbf{d}_{j-1}$ (where $\mathbf{d}_0 \equiv 0$). Hence,

$$\nabla f(\mathbf{y}_{j+1})^t \nabla f(\mathbf{y}_j) = 0. \tag{8.59}$$

Substituting this into (8.58) and using (8.41) gives Fletcher and Reeves's [1964] choice of α_j as

$$\alpha_j^{\mathrm{FR}} = \frac{\left\|\nabla f(\mathbf{y}_{j+1})\right\|^2}{\left\|\nabla f(\mathbf{y}_j)\right\|^2}. \tag{8.60}$$

We now proceed to present and formally analyze the conjugate gradient method using Fletcher and Reeves's choice (8.60) for α_j. A similar discussion follows for other choices as well.

Summary of the Conjugate Gradient Method of Fletcher and Reeves

A summary of this conjugate gradient method for minimizing a general differentiable function is given below.

Initialization Step Choose a termination scalar $\varepsilon > 0$ and an initial point $\mathbf{x}_1$. Let $\mathbf{y}_1 = \mathbf{x}_1$, $\mathbf{d}_1 = -\nabla f(\mathbf{y}_j)$, $k = j = 1$, and go to the Main Step.

Main Step

1. If $\|\nabla f(\mathbf{y}_j)\| < \varepsilon$, stop. Otherwise, let λ_j be an optimal solution to the problem to minimize $f(\mathbf{y}_j + \lambda \mathbf{d}_j)$ subject to $\lambda \geq 0$, and let $\mathbf{y}_{j+1} = \mathbf{y}_j + \lambda_j \mathbf{d}_j$. If $j < n$, go to Step 2; otherwise, go to Step 3.
2. Let $\mathbf{d}_{j+1} = -\nabla f(\mathbf{y}_{j+1}) + \alpha_j \mathbf{d}_j$, where

$$\alpha_j = \frac{\|\nabla f(\mathbf{y}_{j+1})\|^2}{\|\nabla f(\mathbf{y}_j)\|^2}.$$

   Replace j by $j + 1$, and go to Step 1.
3. Let $\mathbf{y}_1 = \mathbf{x}_{k+1} = \mathbf{y}_{n+1}$, and let $\mathbf{d}_1 = -\nabla f(\mathbf{y}_1)$. Let $j = 1$, replace k by $k + 1$, and go to Step 1.

8.8.7 Example

Consider the following problem:

$$\text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2.$$

The summary of the computations using the method of Fletcher and Reeves is given in Table 8.14. At each iteration, $\mathbf{d}_1$ is given by $-\nabla f(\mathbf{y}_1)$, and $\mathbf{d}_2$ is given by $\mathbf{d}_2 = -\nabla f(\mathbf{y}_2) + \alpha_1 \mathbf{d}_1$, where $\alpha_1 = \|\nabla f(\mathbf{y}_2)\|^2 / \|\nabla f(\mathbf{y}_1)\|^2$. Furthermore, $\mathbf{y}_{j+1}$ is obtained by optimizing along $\mathbf{d}_j$, starting from $\mathbf{y}_j$. At Iteration 4, the point $\mathbf{y}_2 = (2.185, 1.094)^t$, which is very close to the optimal point (2.00, 1.00), is reached. Since the norm of the gradient is equal to 0.02, which is, say, sufficiently small, we stop here. The progress of the algorithm is shown in Figure 8.23.

Quadratic Case

If the function f is quadratic, Theorem 8.8.8 shows that the directions $\mathbf{d}_1, \ldots, \mathbf{d}_n$ generated are indeed conjugate, and hence, by Theorem 8.8.3, the conjugate gradient algorithm produces an optimal solution after one complete application of the Main Step, that is, after at most n line searches have been performed.

Table 8.14 Summary of Computations for the Method of Fletcher and Reeves

Iteration k	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	j	$\mathbf{y}_j$ $f(\mathbf{y}_j)$	$\nabla f(\mathbf{y}_j)$	$\|\nabla f(\mathbf{y}_j)\|$	α_{j-1}	$\mathbf{d}_j$	λ_j	$\mathbf{y}_{j+1}$
1	(0.00, 3.00) 52.00	1	(0.00, 3.00) 52.00	(–44.00, 24.00)	50.12	—	(44.00, –24.00)	0.062	(2.70, 1.51)
		2	(2.70, 1.51) 0.34	(0.73, 1.28)	1.47	0.0009	(–0.69, –1.30)	0.23	(2.54, 1.21)
2	(2.54, 1.21) 0.10	1	(2.54, 1.21) 0.10	(0.87, –0.48)	0.99	—	(–0.87, 0.48)	0.11	(2.44, 1.26)
		2	(2.44, 1.26) 0.04	(0.18, 0.32)	0.37	0.14	(–0.30, –0.25)	0.63	(2.25, 1.10)
3	(2.25, 1.10) 0.008	1	(2.25, 1.10) 0.008	(0.16, –0.20)	0.32	—	(–0.16, 0.20)	0.10	(2.23, 1.12)
		2	(2.23, 1.12) 0.003	(0.03, 0.04)	0.05	0.04	(–0.036, –0.032)	1.02	(2.19, 1.09)
4	(2.19, 1.09) 0.0017	1	(2.19, 1.09) 0.0017	(0.05, –0.04)	0.06	—	(–0.05, 0.04)	0.11	(2.185, 1.094)
		2	(2.185, 1.094) 0.0012	(0.002, 0.01)	0.02				

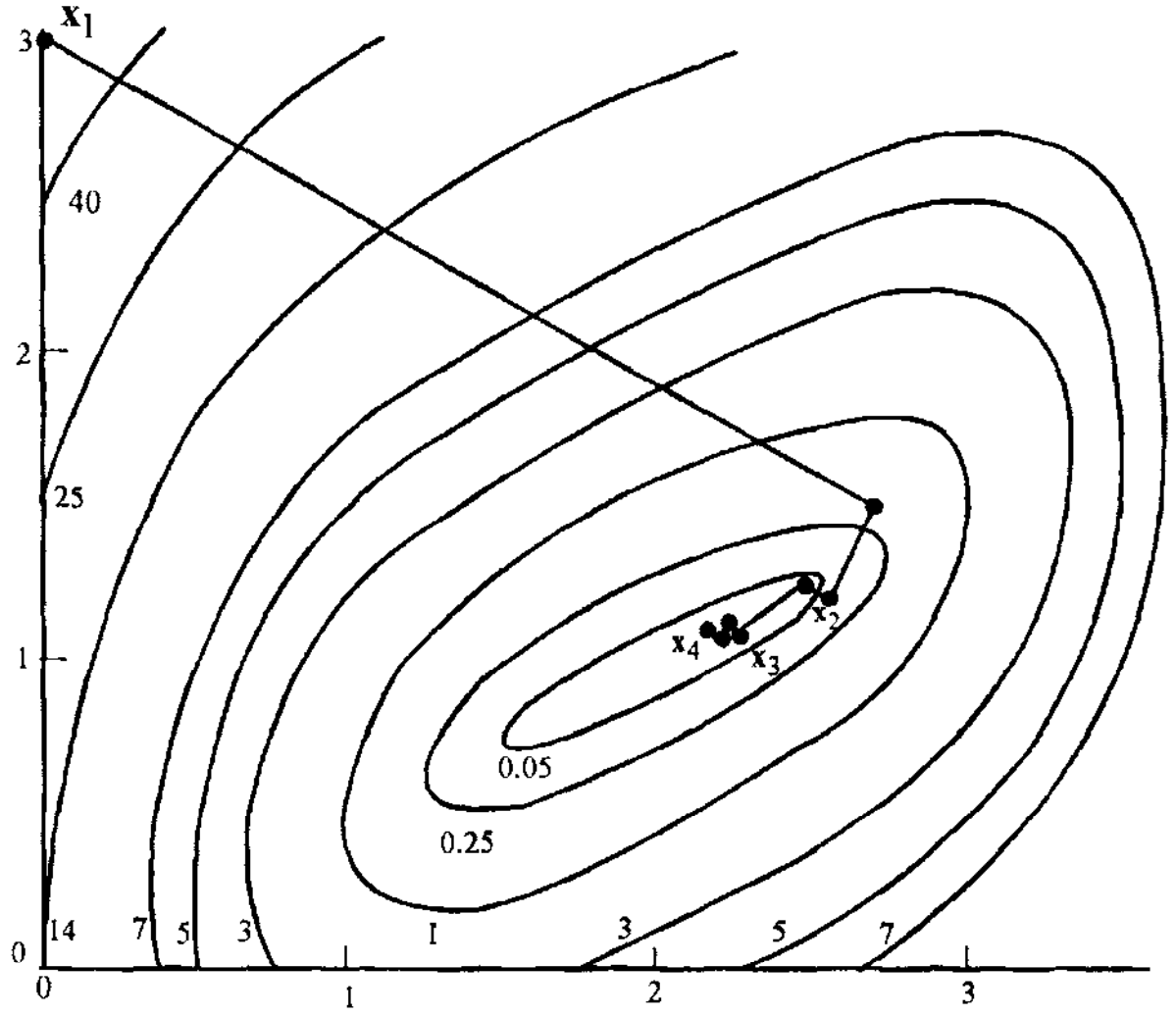


Figure 8.23 Method of Fletcher and Reeves.

8.8.8 Theorem

Consider the problem to minimize $f(\mathbf{x}) = \mathbf{c}^t\mathbf{x} + (1/2)\mathbf{x}^t\mathbf{H}\mathbf{x}$ subject to $\mathbf{x} \in R^n$. Suppose that the problem is solved by the conjugate gradient method, starting with $\mathbf{y}_1$ and letting $\mathbf{d}_1 = -\nabla f(\mathbf{y}_1)$. In particular, for $j = 1,\ldots, n$, let λ_j be an optimal solution to the problem to minimize $f(\mathbf{y}_j + \lambda\mathbf{d}_j)$ subject to $\lambda \geq 0$. Let $\mathbf{y}_{j+1} = \mathbf{y}_j + \lambda_j\mathbf{d}_j$, and let $\mathbf{d}_{j+1} = -\nabla f(\mathbf{y}_{j+1}) + \alpha_j\mathbf{d}_j$, where $\alpha_j = \left\|\nabla f(\mathbf{y}_{j+1})\right\|^2 / \left\|\nabla f(\mathbf{y}_j)\right\|^2$. If $\nabla f(\mathbf{y}_j) \neq \mathbf{0}$ for $j = 1,\ldots, n$, then the following statements are true:

1. $\mathbf{d}_1,\ldots,\mathbf{d}_n$ are **H**-conjugate.
2. $\mathbf{d}_1,\ldots,\mathbf{d}_n$ are descent directions.
3. $\alpha_j = \dfrac{\left\|\nabla f(\mathbf{y}_{j+1})\right\|^2}{\left\|\nabla f(\mathbf{y}_j)\right\|^2} = \dfrac{\mathbf{d}_j^t\mathbf{H}\nabla f(\mathbf{y}_{j+1})}{\mathbf{d}_j^t\mathbf{H}\mathbf{d}_j}$ for $j = 1,\ldots, n$.

Proof

First, suppose that Parts 1, 2, and 3 hold true for j. We show that they also hold true for $j + 1$. To show that Part 1 holds true for $j + 1$, we first demonstrate that $\mathbf{d}_k^t \mathbf{H} \mathbf{d}_{j+1} = 0$ for $k \leq j$. Since $\mathbf{d}_{j+1} = -\nabla f(\mathbf{y}_{j+1}) + \alpha_j \mathbf{d}_j$, noting the induction hypothesis of Part 3 and letting $k = j$, we get

$$\mathbf{d}_j^t \mathbf{H} \mathbf{d}_{j+1} = \mathbf{d}_j^t \mathbf{H} \left[-\nabla f(\mathbf{y}_{j+1}) + \frac{\mathbf{d}_j^t \mathbf{H} \nabla f(\mathbf{y}_{j+1})}{\mathbf{d}_j^t \mathbf{H} \mathbf{d}_j} \mathbf{d}_j \right] = 0. \tag{8.61}$$

Now let $k < j$. Since $\mathbf{d}_{j+1} = -\nabla f(\mathbf{y}_{j+1}) + \alpha_j \mathbf{d}_j$, and since $\mathbf{d}_k^t \mathbf{H} \mathbf{d}_j = 0$ by the induction hypothesis of Part 1,

$$\mathbf{d}_k^t \mathbf{H} \mathbf{d}_{j+1} = -\mathbf{d}_k^t \mathbf{H} \nabla f(\mathbf{y}_{j+1}). \tag{8.62}$$

Since $\nabla f(\mathbf{y}_{k+1}) = \mathbf{c} + \mathbf{H}\mathbf{y}_{k+1}$ and $\mathbf{y}_{k+1} = \mathbf{y}_k + \lambda_k \mathbf{d}_k$, note that

$$\begin{aligned} \mathbf{d}_{k+1} &= -\nabla f(\mathbf{y}_{k+1}) + \alpha_k \mathbf{d}_k \\ &= -[\nabla f(\mathbf{y}_k) + \lambda_k \mathbf{H} \mathbf{d}_k] + \alpha_k \mathbf{d}_k \\ &= -[-\mathbf{d}_k + \alpha_{k-1} \mathbf{d}_{k-1} + \lambda_k \mathbf{H} \mathbf{d}_k] + \alpha_k \mathbf{d}_k. \end{aligned}$$

By the induction hypothesis of Part 2, $\mathbf{d}_k$ is a descent direction and hence, $\lambda_k > 0$. Therefore,

$$\mathbf{d}_k^t \mathbf{H} = \frac{1}{\lambda_k} [-\mathbf{d}_{k+1}^t + (1 + \alpha_k) \mathbf{d}_k^t - \alpha_{k-1} \mathbf{d}_{k-1}^t]. \tag{8.63}$$

From (8.62) and (8.63) it follows that

$$\begin{aligned} \mathbf{d}_k^t \mathbf{H} \mathbf{d}_{j+1} &= -\mathbf{d}_k^t \mathbf{H} \nabla f(\mathbf{y}_{j+1}) \\ &= -\frac{1}{\lambda_k} [-\mathbf{d}_{k+1}^t \nabla f(\mathbf{y}_{j+1}) + (1 + \alpha_k) \mathbf{d}_k^t \nabla f(\mathbf{y}_{j+1}) - \alpha_{k-1} \mathbf{d}_{k-1}^t \nabla f(\mathbf{y}_{j+1})]. \end{aligned}$$

By part 1 of Theorem 8.8.3, and since $\mathbf{d}_1, \ldots, \mathbf{d}_j$ are assumed conjugate, $\mathbf{d}_{k+1}^t \nabla f(\mathbf{y}_{j+1}) = \mathbf{d}_k^t \nabla f(\mathbf{y}_{j+1}) = \mathbf{d}_{k-1}^t \nabla f(\mathbf{y}_{j+1}) = 0$. Thus, the above equation implies that $\mathbf{d}_k^t \mathbf{H} \mathbf{d}_{j+1} = 0$ for $k < j$. This, together with (8.61), shows that $\mathbf{d}_k^t \mathbf{H} \mathbf{d}_{j+1} = 0$ for all $k \leq j$.

To show that $\mathbf{d}_1, \ldots, \mathbf{d}_{j+1}$ are H-conjugate, it thus suffices to show that they are linearly independent. Suppose that $\sum_{i=1}^{j+1} \gamma_i \mathbf{d}_i = \mathbf{0}$. Then $\sum_{i=1}^{j} \gamma_i \mathbf{d}_i + \gamma_{j+1}[-\nabla f(\mathbf{y}_{j+1}) + \alpha_j \mathbf{d}_j] = \mathbf{0}$. Multiplying by $\nabla f(\mathbf{y}_{j+1})^t$, and noting part 1 of

Theorem 8.8.3, it follows that $\gamma_{j+1}\left\|\nabla f(\mathbf{y}_{j+1})\right\|^2 = 0$. Since $\nabla f(\mathbf{y}_{j+1}) \neq \mathbf{0}$, $\gamma_{j+1} = 0$. This implies that $\sum_{i=1}^{j}\gamma_i\mathbf{d}_i = \mathbf{0}$, and in view of the conjugacy of $\mathbf{d}_1,...,\mathbf{d}_j$, it follows that $\gamma_1 = \cdots = \gamma_j = 0$. Thus, $\mathbf{d}_1,...,\mathbf{d}_{j+1}$ are linearly independent and H-conjugate, so that Part 1 holds true for $j + 1$.

Now we show that Part 2 holds true for $j + 1$; that is, $\mathbf{d}_{j+1}$ is a descent direction. Note that $\nabla f(\mathbf{y}_{j+1}) \neq \mathbf{0}$ by assumption, and that $\nabla f(\mathbf{y}_{j+1})^t\mathbf{d}_j = 0$ by Part 1 of Theorem 8.8.3. Then $\nabla f(\mathbf{y}_{j+1})^t\mathbf{d}_{j+1} = -\left\|\nabla f(\mathbf{y}_{j+1})\right\|^2 + \alpha_j\nabla f(\mathbf{y}_{j+1})^t\mathbf{d}_j$ $= -\left\|\nabla f(\mathbf{y}_{j+1})\right\|^2 < 0$. By Theorem 4.1.2, $\mathbf{d}_{j+1}$ is a descent direction.

Next we show that Part 3 holds true for $j + 1$. By letting $k = j + 1$ in (8.63) and multiplying by $\nabla f(\mathbf{y}_{j+2})$, it follows that

$$\begin{aligned}\lambda_{j+1}\mathbf{d}_{j+1}^t\mathbf{H}\nabla f(\mathbf{y}_{j+2}) &= [-\mathbf{d}_{j+2}^t + (1+\alpha_{j+1})\mathbf{d}_{j+1}^t - \alpha_j\mathbf{d}_j^t]\nabla f(\mathbf{y}_{j+2}) \\ &= [\nabla f(\mathbf{y}_{j+2})^t + \mathbf{d}_{j+1}^t - \alpha_j\mathbf{d}_j^t]\nabla f(\mathbf{y}_{j+2}).\end{aligned}$$

Since $\mathbf{d}_1,...,\mathbf{d}_{j+1}$ are H-conjugate, then, by Part 1 of Theorem 8.8.3, $\mathbf{d}_{j+1}^t\nabla f(\mathbf{y}_{j+2})$ $= \mathbf{d}_j^t\nabla f(\mathbf{y}_{j+2}) = 0$. The above equation then implies that

$$\left\|\nabla f(\mathbf{y}_{j+2})\right\|^2 = \lambda_{j+1}\mathbf{d}_{j+1}^t\mathbf{H}\nabla f(\mathbf{y}_{j+2}). \tag{8.64}$$

Multiplying $\nabla f(\mathbf{y}_{j+1}) = \nabla f(\mathbf{y}_{j+2}) - \lambda_{j+1}\mathbf{H}\mathbf{d}_{j+1}$ by $\nabla f(\mathbf{y}_{j+1})^t$, and noting that $\mathbf{d}_j^t\mathbf{H}\mathbf{d}_{j+1} = \mathbf{d}_{j+1}^t\nabla f(\mathbf{y}_{j+2}) = \mathbf{d}_j^t\nabla f(\mathbf{y}_{j+2}) = 0$, we get

$$\begin{aligned}\left\|\nabla f(\mathbf{y}_{j+1})\right\|^2 &= \nabla f(\mathbf{y}_{j+1})^t[\nabla f(\mathbf{y}_{j+2}) - \lambda_{j+1}\mathbf{H}\mathbf{d}_{j+1}] \\ &= (-\mathbf{d}_{j+1}^t + \alpha_j\mathbf{d}_j^t)[\nabla f(\mathbf{y}_{j+2}) - \lambda_{j+1}\mathbf{H}\mathbf{d}_{j+1}] \\ &= \lambda_{j+1}\mathbf{d}_{j+1}^t\mathbf{H}\mathbf{d}_{j+1}.\end{aligned} \tag{8.65}$$

From (8.64) and (8.65), it is obvious that part 3 holds true for $j + 1$.

We have thus shown that if Parts 1, 2, and 3 hold true for j, then they also hold true for $j + 1$. Note that Parts 1 and 2 trivially hold true for $j = 1$. In addition, using an argument similar to that used in proving that Part 3 holds true for $j + 1$, it can easily be demonstrated that it holds true for $j = 1$. This completes the proof.

The reader should note here that when the function f is quadratic and when exact line searches are performed, the choices of α_j, given variously by

(8.57), (8.58), and (8.60) all coincide, and thus Theorem 8.8.8 also holds true for the Hestenes and Stiefel (HS) and the Polak and Ribiere (PR) choices of α_j. However, for nonquadratic functions, the choice α_j^{PR} appears to be empirically superior to α_j^{FR}. This is understandable, since the reduction of (8.58) to (8.60) assumes f to be quadratic. In the same vein, when inexact line searches are performed, the choice α_j^{HS} appears to be preferable. Note that even when f is quadratic, if inexact line searches are performed, the conjugacy relationship holds true only between consecutive directions. We refer the reader to the Notes and References section for a discussion on some alternative *three-term recurrence relationships* for generating mutually conjugate directions in such a case.

Also note that we have used $\mathbf{d}_1 = -\mathbf{I}\nabla f(\mathbf{y}_1)$ in the foregoing analysis. In lieu of using the identity matrix here, we could have used some general *preconditioning matrix* $\mathbf{D}$, where $\mathbf{D}$ is symmetric and positive definite. This would have given $\mathbf{d}_1 = -\mathbf{D}\nabla f(\mathbf{y}_1)$, and (8.56b) would have become $\mathbf{d}_{j+1} = -\mathbf{D}\nabla f(\mathbf{y}_{j+1}) + \alpha_j \mathbf{d}_j$, where, for example, in the spirit of (8.57), we have

$$\alpha_j^{\mathrm{HS}} = \frac{\mathbf{q}_j \mathbf{D}\nabla f(\mathbf{y}_{j+1})}{\mathbf{q}_j' \mathbf{d}_j}.$$

This corresponds, essentially to making a change of variables $\mathbf{y}' = \mathbf{D}^{-1/2}\mathbf{y}$ and using the original conjugate gradient algorithm. Therefore, this motivates the choice of $\mathbf{D}$ from the viewpoint of improving the eigenstructure of the problem, as discussed earlier.

For quadratic functions f, the conjugate gradient step also has an interesting *pattern search* interpretation. Consider Figure 8.24 and suppose that the successive points $\mathbf{y}_j$, $\mathbf{y}_{j+1}$, and $\mathbf{y}_{j+2}$ are generated by the conjugate gradient algorithm. Now, suppose that at the point $\mathbf{y}_{j+1}$ obtained from $\mathbf{y}_j$ by minimizing along $\mathbf{d}_j$, we had instead minimized next along the steepest descent direction $-\nabla f(\mathbf{y}_{j+1})$ at $\mathbf{y}_{j+1}$, leading to the point $\mathbf{y}'_{j+1}$. Then it can be shown (see Exercise 8.38) that a pattern search step of minimizing the quadratic function f from $\mathbf{y}_j$ along the direction $\mathbf{y}'_{j+1} - \mathbf{y}_j$ would also have led to the same point $\mathbf{y}_{j+2}$. The method, which uses the latter kind of step in general (even for nonquadratic functions), is more popularly known as PARTAN (see Exercise 8.53). Note that the global convergence of PARTAN for general functions is tied into using the negative gradient direction as a spacer step in Theorem 7.3.4 and is independent of any restart conditions, although it is recommended that the method be restarted every n iterations to promote its behavior as a conjugate gradient method.

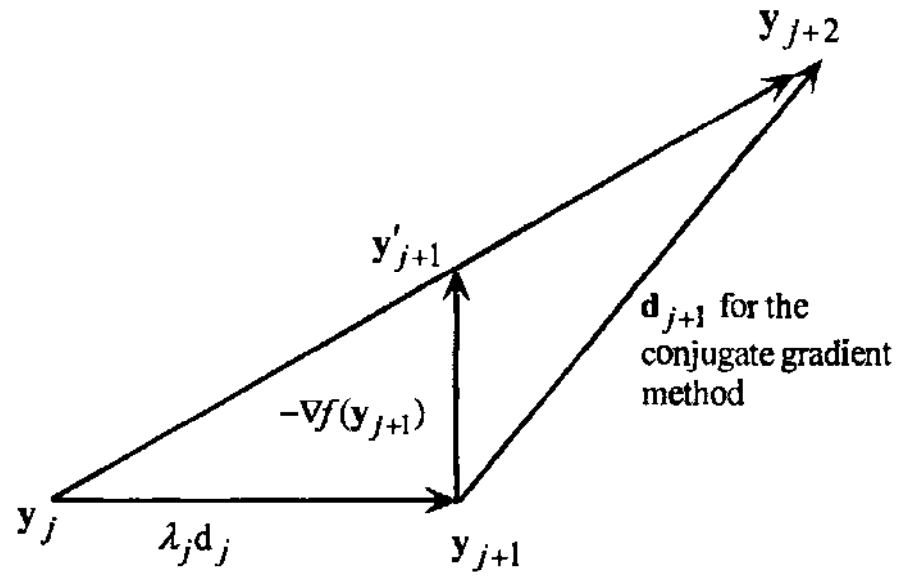


Figure 8.24 Equivalence between the conjugate gradient method and PARTAN.

Memoryless Quasi-Newton Methods

There is an interesting connection between conjugate gradient methods and a simplified variant of the BFGS quasi-Newton method. Suppose that we operate the latter method by updating the inverse Hessian approximation according to $\mathbf{D}_{j+1} = \mathbf{D}_j + \mathbf{C}_j^{\text{BFGS}}$, where the correction matrix $\mathbf{C}_j^{\text{BFGS}}$ is given in (8.48), but assuming that $\mathbf{D}_j \equiv \mathbf{I}$. Hence, we get

$$\mathbf{D}_{j+1} = \mathbf{I} + \frac{\mathbf{p}_j \mathbf{p}_j^t}{\mathbf{p}_j^t \mathbf{q}_j}\left(1 + \frac{\mathbf{q}_j^t \mathbf{q}_j}{\mathbf{p}_j^t \mathbf{q}_j}\right) - \frac{\mathbf{q}_j \mathbf{p}_j^t + \mathbf{p}_j \mathbf{q}_j^t}{\mathbf{p}_j^t \mathbf{q}_j}. \tag{8.66a}$$

We then move along the direction

$$\mathbf{d}_{j+1} = -\mathbf{D}_{j+1} \nabla f(\mathbf{y}_{j+1}). \tag{8.66b}$$

This is akin to "forgetting" the previous approximation $\mathbf{D}_j$ and, instead, updating the identity matrix as might be done at the first iteration of a quasi-Newton method: hence, the name *memoryless quasi-Newton method.* Observe that the storage requirements are similar to that of conjugate gradient methods and that inexact line searches can be performed as long as $\mathbf{p}_j^t \mathbf{q}_j = \lambda_j \mathbf{d}_j^t [\nabla f(\mathbf{y}_{j+1}) - \nabla f(\mathbf{y}_j)]$ remains positive and $\mathbf{d}_{j+1}$ continues to be a descent direction. Also, note that the loss of positive definiteness of the approximations $\mathbf{D}_j$ in the quasi-Newton method is now no longer of concern. In fact, this scheme has proved to be computationally very effective in conjunction with inexact line searches. We refer the reader to the Notes and References section for a discussion on conjugate gradient methods operated with inexact line searches.

Now, suppose that we do employ exact line searches. Then we have $\mathbf{p}_j^t \nabla f(\mathbf{y}_{j+1}) = \lambda_j \mathbf{d}_j^t \nabla f(\mathbf{y}_{j+1}) = 0$, so (8.66) gives

$$\mathbf{d}_{j+1} = -\nabla f(\mathbf{y}_{j+1}) + \frac{\mathbf{q}_j^t \nabla f(\mathbf{y}_{j+1})}{\mathbf{p}_j^t \mathbf{q}_j} \mathbf{p}_j = -\nabla f(\mathbf{y}_{j+1}) + \alpha_j^{\mathrm{HS}} \mathbf{d}_j$$

from (8.57). Hence, the BFGS memoryless update scheme is equivalent to the conjugate gradient method of Hestenes and Stiefel (or Polak and Ribiere) when exact line searches are employed. We mention here that although this memoryless update can be performed on any other member of the Broyden family as well (see Exercise 8.34), the equivalence with conjugate gradient methods results only for $\phi = 1$ (the BFGS update), as does the observed empirical effectiveness of this scheme (see Exercise 8.40).

Recommendations for Restarting Conjugate Gradient Methods

In several computational experiments using different conjugate gradient techniques, with or without exact line searches, it has been demonstrated time and again that the performance of conjugate gradient methods can be greatly enhanced by employing a proper restart criterion. In particular, a restart procedure suggested by Beale [1970c] and augmented by Powell [1977b] has proved to be very effective and is invariably implemented, as described below.

Consider the conjugate gradient method summarized formally above in the context of Fletcher and Reeves's choice of α_j. (Naturally, this strategy applies to any other admissible choice of α_j as well.) At some inner loop iteration j of this procedure, having found that $\mathbf{y}_{j+1} = \mathbf{y}_j + \lambda_j \mathbf{d}_j$ by searching along $\mathbf{d}_j$ from the point $\mathbf{y}_j$, suppose that we decide to reset. (In the previous description of the algorithm, this decision was made whenever $j = n$.) Let $\tau = j$ denote this restart iteration. For the next iteration, we find the search direction

$$\mathbf{d}_{\tau+1} = -\nabla f(\mathbf{y}_{\tau+1}) + \alpha_\tau \mathbf{d}_\tau \tag{8.67}$$

as usual. Then at Step 3, we replace $\mathbf{y}_1$ by $\mathbf{y}_{\tau+1}$, let $\mathbf{x}_{k+1} \equiv \mathbf{y}_{\tau+1}$, $\mathbf{d}_1 = \mathbf{d}_{\tau+1}$, and return to Step 1 to continue with the next set of inner loop iterations. However, instead of computing $\mathbf{d}_{j+1} = -\nabla f(\mathbf{y}_{j+1}) + \alpha_j \mathbf{d}_j$ for $j \geq 1$, we now use

$$\mathbf{d}_2 = -\nabla f(\mathbf{y}_2) + \alpha_1 \mathbf{d}_1 \tag{8.68a}$$

and

$$\mathbf{d}_{j+1} = -\nabla f(\mathbf{y}_{j+1}) + \alpha_j \mathbf{d}_j + \gamma_j \mathbf{d}_1 \quad \text{for } j \geq 2,$$

where

$$\gamma_j = \frac{\nabla f(\mathbf{y}_{j+1})^t \mathbf{q}_1}{\mathbf{d}_1^t \mathbf{q}_1} \tag{8.68b}$$

and where α_j is computed as before, depending on the method being used. Note that (8.68a) employs the usual conjugate gradient scheme, thereby yielding $\mathbf{d}_1$ and $\mathbf{d}_2$ as H-conjugate when f is quadratic. However, when f is quadratic with a positive definite Hessian $\mathbf{H}$ and $\mathbf{d}_1$ is chosen arbitrarily, then when $j = 2$, for example, the usual choice of α_2 would make $\mathbf{d}_3$ and $\mathbf{d}_2$ H-conjugate, but we would need something additional to make $\mathbf{d}_3$ and $\mathbf{d}_1$ H-conjugate. This is accomplished by the extra term $\gamma_2 \mathbf{d}_1$. Indeed, requiring that $\mathbf{d}_3^t \mathbf{H} \mathbf{d}_1 = 0$, where $\mathbf{d}_3$ is given by the expression in (8.68b), and noting that $\mathbf{d}_2^t \mathbf{H} \mathbf{d}_1 = 0$ yields $\gamma_2 = \nabla f(\mathbf{y}_3)^t \mathbf{H} \mathbf{d}_1 / \mathbf{d}_1^t \mathbf{H} \mathbf{d}_1 = \nabla f(\mathbf{y}_3)^t \mathbf{q}_1 / \mathbf{d}_1^t \mathbf{q}_1$. Proceeding inductively in this manner, the additional term in (8.68b) ensures the H-conjugacy of all directions generated (see Exercise 8.48).

The foregoing scheme was suggested by Beale with the motivation that whenever a restart is done using $\mathbf{d}_1 = -\nabla f(\mathbf{y}_1)$ instead of $\mathbf{d}_1 = \mathbf{d}_{\tau+1}$ as given by (8.67), we lose important second-order information inherent in $\mathbf{d}_\tau$. Additionally, Powell suggested that after finding $\mathbf{y}_{j+1}$, if any of the following three conditions holds true, then the algorithm should be restarted by putting $\tau = j$, computing $\mathbf{d}_{\tau+1}$ via (8.67), and resetting $\mathbf{d}_1 = \mathbf{d}_{\tau+1}$ and $\mathbf{y}_1 = \mathbf{y}_{\tau+1}$:

1. $j = n - 1$.
2. $\left|\nabla f(\mathbf{y}_{j+1})^t \nabla f(\mathbf{y}_j)\right| \geq 0.2 \left\|\nabla f(\mathbf{y}_{j+1})\right\|^2$ for some $j \geq 1$.
3. $-1.2 \left\|\nabla f(\mathbf{y}_{j+1})\right\|^2 \leq \mathbf{d}_{j+1}^t \nabla f(\mathbf{y}_{j+1}) \leq -0.8 \left\|\nabla f(\mathbf{y}_{j+1})\right\|^2$ is violated for some $j \geq 2$.

Condition 1 is the usual reset criterion by which, after searching along the direction $\mathbf{d}_{\tau+1} = \mathbf{d}_n$, we will have searched along n conjugate directions for the quadratic case. Condition 2 suggests a reset if a sufficient measure of orthogonality has been lost between $\nabla f(\mathbf{y}_j)$ and $\nabla f(\mathbf{y}_{j+1})$, motivated by the expanding subspace property illustrated in Figure 8.21. (Computationally, instead of using 0.2 here, any constant in the interval [0.1, 0.9] appears to give satisfactory performance.) Condition 3 checks for a sufficient descent along the direction $\mathbf{d}_{j+1}$ at the point $\mathbf{y}_{j+1}$, and it also checks for the relative accuracy of the identity $\mathbf{d}_{j+1}^t \nabla f(\mathbf{y}_{j+1}) = -\left\|\nabla f(\mathbf{y}_{j+1})\right\|^2$, which must hold true under exact line searches [whence, using (8.56b), we would have $\mathbf{d}_j^t \nabla f(\mathbf{y}_{j+1}) = 0$]. For similar ideas when employing inexact line searches, we refer the reader to the Notes and References section.

Convergence of Conjugate Direction Methods

As shown in Theorem 8.8.3, if the function under consideration is quadratic, then any conjugate direction algorithm produces an optimal solution in a finite number of steps. We now discuss the convergence of these methods if the function is not necessarily quadratic.

In Theorem 7.3.4 we showed that a composite algorithm $\mathbf{A} = \mathbf{CB}$ converges to a point in the solution set Ω if the following properties hold true:

1. $\mathbf{B}$ is closed at points not in Ω.
2. If $\mathbf{y} \in \mathbf{B}(\mathbf{x})$, then $f(\mathbf{y}) < f(\mathbf{x})$ for $\mathbf{x} \notin \Omega$.
3. If $\mathbf{z} \in \mathbf{C}(\mathbf{y})$, then $f(\mathbf{z}) \leq f(\mathbf{y})$.
4. The set $\Lambda = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}_1)\}$ is compact, where $\mathbf{x}_1$ is the starting solution.

For the conjugate direction (quasi-Newton or conjugate gradient) algorithms discussed in this chapter, the map $\mathbf{B}$ is of the following form. Given $\mathbf{x}$, then $\mathbf{y} \in \mathbf{B}(\mathbf{x})$ means that $\mathbf{y}$ is obtained by minimizing f starting from $\mathbf{x}$ along the direction $\mathbf{d} = -\mathbf{D}\nabla f(\mathbf{x})$, where $\mathbf{D}$ is a specified positive definite matrix. In particular, for the conjugate gradient methods, $\mathbf{D} = \mathbf{I}$, and for the quasi–Newton methods, $\mathbf{D}$ is an arbitrary positive definite matrix. Furthermore, starting from the point obtained by applying the map $\mathbf{B}$, the map $\mathbf{C}$ is defined by minimizing the function f along the directions specified by the particular algorithms. Thus, the map $\mathbf{C}$ satisfies Property 3.

Letting $\Omega = \{\mathbf{x} : \nabla f(\mathbf{x}) = \mathbf{0}\}$, we now show that the map $\mathbf{B}$ satisfies Properties 1 and 2. Let $\mathbf{x} \in \Omega$ and let $\mathbf{x}_k \to \mathbf{x}$. Furthermore, let $\mathbf{y}_k \in \mathbf{B}(\mathbf{x}_k)$ and let $\mathbf{y}_k \to \mathbf{y}$. We need to show that $\mathbf{y} \in \mathbf{B}(\mathbf{x})$. By the definition of $\mathbf{y}_k$, we have $\mathbf{y}_k = \mathbf{x}_k - \lambda_k \mathbf{D}\nabla f(\mathbf{x}_k)$ for $\lambda_k \geq 0$ such that

$$f(\mathbf{y}_k) \leq f[\mathbf{x}_k - \lambda \mathbf{D}\nabla f(\mathbf{x}_k)] \qquad \text{for all } \lambda \geq 0. \tag{8.69}$$

Since $\nabla f(\mathbf{x}) \neq \mathbf{0}$, then λ_k converges to $\bar{\lambda} = \|\mathbf{y} - \mathbf{x}\| / \|\mathbf{D}\nabla f(\mathbf{x})\| \geq 0$. Therefore, $\mathbf{y} = \mathbf{x} - \bar{\lambda}\mathbf{D}\nabla f(\mathbf{x})$. Taking the limit as $k \to \infty$ in (8.69), $f(\mathbf{y}) \leq f[\mathbf{x} - \lambda \mathbf{D}\nabla f(\mathbf{x})]$ for all $\lambda \geq 0$, so that $\mathbf{y}$ is indeed obtained by minimizing f starting from $\mathbf{x}$ in the direction $-\mathbf{D}\nabla f(\mathbf{x})$. Thus, $\mathbf{y} \in \mathbf{B}(\mathbf{x})$, and $\mathbf{B}$ is closed. Also, Part 2 holds true by noting that $-\nabla f(\mathbf{x})^t \mathbf{D}\nabla f(\mathbf{x}) < 0$, so that $-\mathbf{D}\nabla f(\mathbf{x})$ is a descent direction. Assuming that the set defined in Part 4 is compact, it follows that the conjugate direction algorithms discussed in this section converge to a point with zero gradient.

The role played by the map $\mathbf{B}$ described above is akin to that of a *spacer step*, as discussed in connection with Theorem 7.3.4. For algorithms that are designed empirically and that may not enjoy theoretical convergence, this can be alleviated by inserting such a spacer step involving a periodic minimization

along the negative gradient direction, for example, hence, achieving theoretical convergence.

We now turn our attention to addressing the rate of convergence or local convergence characteristics of the algorithms discussed in this section.

Convergence Rate Characteristics for Conjugate Gradient Methods

Consider the quadratic function $f(\mathbf{x}) = \mathbf{c}^t\mathbf{x} + (1/2)\mathbf{x}^t\mathbf{H}\mathbf{x}$, where $\mathbf{H}$ is an $n \times n$ symmetric, positive definite matrix. Suppose that the eigenvalues of $\mathbf{H}$ are grouped into two sets, of which one set is composed of some m relatively large and perhaps dispersed values, and the other set is a cluster of some $n - m$ relatively smaller eigenvalues. (Such a structure arises, for example, with the use of quadratic penalty functions for linearly constrained quadratic programs, as discussed in Chapter 9.) Let us assume that $(m + 1) < n$, and let α denote the ratio of the largest to the smallest eigenvalue in the latter cluster. Now, we know that a standard application of the conjugate gradient method will result in a finite convergence to the optimum in n, or fewer, steps. However, suppose that we operate the conjugate gradient algorithm by restarting with the steepest descent direction every $m + 1$ line searches or steps. Such a procedure is called a *partial conjugate gradient method.*

Starting with a solution $\mathbf{x}_1$, let $\{\mathbf{x}_k\}$ be the sequence thus generated, where for each $k \geq 1$, $\mathbf{x}_{k+1}$ is obtained after applying $m + 1$ conjugate gradient steps upon restarting with $\mathbf{x}_k$ as above. Let us refer to this as an $(m + 1)$-*step process.* As in Equation (8.17), let us define an error function $e(\mathbf{x}) = (1/2)(\mathbf{x}-\mathbf{x}^*)^t\mathbf{H}(\mathbf{x}-\mathbf{x}^*)$, which differs from $f(\mathbf{x})$ by a constant, and which is zero if and only if $\mathbf{x} = \mathbf{x}^*$. Then it can be shown (see the Notes and References section) that

$$e(\mathbf{x}_{k+1}) \leq \frac{(\alpha-1)^2}{(\alpha+1)^2} e(\mathbf{x}_k). \tag{8.70}$$

Hence, this establishes a linear rate of convergence for the above process as in the special case of the steepest descent method for which $m = 0$ [see Equation (8.18)]. However, the ratio α that governs the convergence rate is now independent of the m largest eigenvalues. Thus, the effect of the m largest eigenvalues is eliminated, but at the expense of an $(m + 1)$-step process versus the single-step process of the steepest descent method.

Next, consider the general nonquadratic case to which the usual n-step conjugate gradient process is applied. Intuitively, since the conjugate gradient method accomplishes in n steps what Newton's method does in a single step, by the local quadratic convergence rate of Newton's method, we might similarly expect that the n–step conjugate gradient process also converges quadratically; that is, $\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq \beta\|\mathbf{x}_k - \mathbf{x}^*\|^2$ for some $\beta > 0$. Indeed, it can be shown (see

the Notes and References section) that if the sequence $\{\mathbf{x}_k\} \rightarrow \mathbf{x}^*$, the function under consideration is twice continuously differentiable in some neighborhood of $\mathbf{x}^*$, and the Hessian matrix at $\mathbf{x}^*$ is positive definite, the n-step process converges superlinearly to $\mathbf{x}^*$. Moreover, if the Hessian matrix satisfies an appropriate Lipschitz condition in some neighborhood of $\mathbf{x}^*$, then the rate of superlinear convergence is n-step quadratic. Again, caution must be exercised in interpreting these results in comparison with, say, the linear convergence rate of steepest descent methods. That is, these are n-step asymptotic results, whereas the steepest descent method is a single-step procedure. Also, given that these methods are usually applied when n is relatively large, it is seldom practical to perform more than $5n$ iterations, or five n-step iterations. Fortunately, empirical results seem to indicate that this does not pose a problem because reasonable convergence is typically obtained within $2n$ iterations.

Convergence Rate Characteristics for Quasi-Newton Methods

The Broyden class of quasi-Newton methods can also be operated as *partial quasi-Newton methods* by restarting every $m + 1$ iterations with, say, the steepest descent direction. For the quadratic case, the local convergence properties of such a scheme resembles that for conjugate gradient methods as discussed above. Also, for nonquadratic cases, the n-step quasi-Newton algorithm has a local superlinear convergence rate behavior similar to that of the conjugate gradient method. Intuitively, this is because of the identical effect that the n-step process of either method has on quadratic functions. Again, the usual caution must be adopted in interpreting the value of an n-step superlinear convergence behavior. Additionally, we draw the reader's attention to Exercise 8.52 and to the section on scaling quasi-Newton methods, where we discuss the possible ill-conditioning effects resulting from the sequential transformation of the eigenvalues of $\mathbf{D}_{j+1}\mathbf{H}$ to unity for the quadratic case.

Quasi-Newton methods are also sometimes operated as a continuing updating process, without resets. Although the global convergence of such a scheme requires rather stringent conditions, the local convergence rate behavior is often asymptotically superlinear. For example, for the BFGS update scheme, which has been seen to exhibit a relatively superior empirical performance, as mentioned previously, the following result holds true (see the Notes and References section). Let $\mathbf{y}^*$ be such that the Hessian $\mathbf{H}(\mathbf{y}^*)$ is positive definite and that there exists an ε-neighborhood $N_\varepsilon(\mathbf{y}^*)$ of $\mathbf{y}^*$ such that the Lipschitz condition $\left\|\mathbf{H}(\mathbf{y}) - \mathbf{H}(\mathbf{y}^*)\right\| \leq L\left\|\mathbf{y} - \mathbf{y}^*\right\|$ holds true for $\mathbf{y} \in N_\varepsilon(\mathbf{y}^*)$, where L is a positive constant. Then, if a sequence $\{\mathbf{y}_k\}$ generated by a continually updated quasi-Newton process with a fixed step size of unity converges to such a $\mathbf{y}^*$, the asymptotic rate of convergence is superlinear. Similar superlinear convergence rate results are available for the DFP algorithm, with both exact line searches

and unit step size choices under appropriate conditions. We refer the reader to the Notes and References section for further reading on this subject.

8.9 Subgradient Optimization

Consider Problem P, defined as

$$\text{P: Minimize } \{f(\mathbf{x}) : \mathbf{x} \in X\}, \tag{8.71}$$

where f: $R^n \to R$ is a convex but not necessarily differentiable function and where X is a nonempty, closed, convex subset of R^n. We assume that an optimal solution exists, as it would be, for example, if X is bounded or if $f(\mathbf{x}) \to \infty$ whenever $\|\mathbf{x}\| \to \infty$.

For such a Problem P, we now describe a *subgradient optimization algorithm* that can be viewed as a direct generalization of the steepest descent algorithm in which the negative gradient direction is substituted by a negative subgradient-based direction. However, the latter direction need not necessarily be a descent direction, although, as we shall see, it does result in the new iterate approaching closer to an optimal solution for a sufficiently small step size. For this reason we do not perform a line search along the negative subgradient direction, but rather, we prescribe a step size at each iteration that guarantees that the sequence generated will eventually converge to an optimal solution. Also, given an iterate $\mathbf{x}_k \in X$ and adopting a step size λ_k along the direction $\mathbf{d}_k = -\xi_k / \|\xi_k\|$, where ξ_k belongs to the subdifferential $\partial f(\mathbf{x}_k)$ of f at $\mathbf{x}_k$ ($\xi_k \neq \mathbf{0}$, say), the resulting point $\bar{\mathbf{x}}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$ need not belong to X. Consequently, the new iterate $\mathbf{x}_{k+1}$ is obtained by *projecting* $\bar{\mathbf{x}}_{k+1}$ onto X, that is, finding the (unique) closest point in X to $\bar{\mathbf{x}}_{k+1}$. We denote this operation as $\mathbf{x}_{k+1} = P_X(\bar{\mathbf{x}}_{k+1})$, where

$$P_X(\bar{\mathbf{x}}) \equiv \text{argmin}\{\|\mathbf{x} - \bar{\mathbf{x}}\| : \mathbf{x} \in X\}. \tag{8.72}$$

The foregoing projection operation should be easy to perform if the method is to be computationally viable. For example, in the context of Lagrangian duality (Chapter 6), wherein subgradient methods and their variants are most frequently used, the set X might simply represent nonnegativity restrictions $\mathbf{x} \geq \mathbf{0}$ on the variables. In this case, we easily obtain $(\mathbf{x}_{k+1})_i = \max\{0, (\bar{\mathbf{x}}_{k+1})_i\}$ for each component $i = 1,\ldots, n$ in (8.72). In other contexts, the set $X = \{\mathbf{x} : \ell_i \leq x_i \leq u_i,\ i = 1,\ldots, n\}$ might represent simple finite lower and upper bounds on the variables. In this case, it is again easy to verify that

$$(\mathbf{x}_{k+1})_i = \begin{cases} (\bar{\mathbf{x}}_{k+1})_i & \text{if } \ell_i \leq (\bar{\mathbf{x}}_{k+1})_i \leq u_i \\ \ell_i & \text{if } (\bar{\mathbf{x}}_{k+1})_i < \ell_i \\ u_i & \text{if } (\bar{\mathbf{x}}_{k+1})_i > u_i \end{cases} \qquad \text{for } i = 1,\ldots, n. \tag{8.73}$$

Also, when an additional knapsack constraint $\alpha^t \mathbf{x} = \beta$ is introduced to define $X = \{\mathbf{x} : \alpha^t \mathbf{x} = \beta, \ \ell \le \mathbf{x} \le \mathbf{u}\}$, then, again, $P_X(\bar{\mathbf{x}})$ is relatively easy to obtain (see Exercise 8.60).

Summary of a (Rudimentary) Subgradient Algorithm

Initialization Step Select a starting solution $\mathbf{x}_1 \in X$, let the current upper bound on the optimal objective value be $UB_1 = f(\mathbf{x}_1)$, and let the current incumbent solution be $\mathbf{x}^* = \mathbf{x}_1$. Put $k = 1$, and go to the Main Step.

Main Step Given $\mathbf{x}_k$, find a subgradient $\xi_k \in \partial f(\mathbf{x}_k)$ of f at $\mathbf{x}_k$. If $\xi_k = 0$, then stop; $\mathbf{x}_k$ (or $\mathbf{x}^*$) solves Problem P. Otherwise, let $\mathbf{d}_k = -\xi_k / \|\xi_k\|$, select a step size $\lambda_k > 0$, and compute $\mathbf{x}_{k+1} = P_X(\bar{\mathbf{x}}_{k+1})$, where $\bar{\mathbf{x}}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$. If $f(\mathbf{x}_{k+1}) < UB_k$, put $UB_{k+1} = f(\mathbf{x}_{k+1})$ and $\mathbf{x}^* = \mathbf{x}_{k+1}$. Otherwise, let $UB_{k+1} = UB_k$. Increment k by 1 and repeat the Main Step.

Note that the stopping criterion $\xi_k = 0$ may never be realized, even if there exists an interior point optimum and we do find a solution $\mathbf{x}_k$ for which $0 \in \partial f(\mathbf{x}_k)$, because the algorithm arbitrarily selects the subgradient ξ_k. Hence, a practical stopping criterion based on a maximum limit on the number of iterations performed is used almost invariably. Note also that we can terminate the procedure whenever $\mathbf{x}_{k+1} = \mathbf{x}_k$ for any iteration. Alternatively, if the optimal objective value f^* is known, as in the problem of finding a feasible solution by minimizing the sum of (absolute) constraint violations, an ε stopping criterion $UB_k \le f^* + \varepsilon$ may be used for some tolerance $\varepsilon > 0$. (See the Notes and References section for a primal–dual scheme employing a termination criterion based on the duality gap.)

8.9.1 Example

Consider the following Problem P:

$$\text{Minimize } \{f(x, y) : -1 \le x \le 1, -1 \le y \le 1\}$$
$$\text{where } f(x, y) = \max\{-x, x + y, x - 2y\}.$$

By considering $f(x, y) \le c$, where c is a constant, and examining the region bounded by $-x \le c$, $x + y \le c$, and $x - 2y \le c$, we can plot the contours of f as shown in Figure 8.25. Note that the points of nondifferentiability are of the type $(t, 0)$, $(-t, 2t)$, and $(-t, -t)$ for $t \ge 0$. Also, the optimal solution is $(x, y) = (0, 0)$, at which all three linear functions defining f tie in value. Hence, although $(0, 0)^t \in$

$\partial f(\mathbf{0})$, we also evidently have $(-1, 0)^t$, $(1, 1)^t$, and $(1, -2)^t$ belonging to $\partial f(0)$.

Now consider the point $(x, y) = (1, 0)$. We have $f(1, 0) = 1$, as determined by the linear functions $x + y$ and $x - 2y$. (See Figure 8.25.) Hence, $\xi = (1, 1)^t \in \partial f(1,0)$. Consider the direction $-\xi = (-1, -1)^t$. Note that this is not a descent direction. However, as we begin to move along this direction, we do approach closer to the optimal solution $(0, 0)^t$. Figure 8.25 shows the ideal step that we could take along the direction $\mathbf{d} = -\xi$ to arrive closest to the optimal solution. However, suppose that we take a step length $\lambda = 2$ along $-\xi$. This brings us to the point $(1, 0) - 2(1, 1) = (-1, -2)$. The projection $P_X(-1, -2)$ of $(-1, -2)$ onto X is obtained via (8.73) as $(-1, -1)$. This constitutes one iteration of the foregoing algorithm.

The following result prescribes a step-size selection scheme that will guarantee convergence to an optimum.

8.9.2 Theorem

Let Problem P be as defined in (8.71) and assume that an optimum exists. Consider the foregoing subgradient optimization algorithm to solve Problem P, and suppose that the prescribed nonnegative step size sequence $\{\lambda_k\}$ satisfies the conditions $\{\lambda_k\} \to 0^+$ and $\sum_{k=0}^{\infty} \lambda_k = \infty$. Then, either the algorithm terminates finitely with an optimal solution, or else an infinite sequence is generated such that

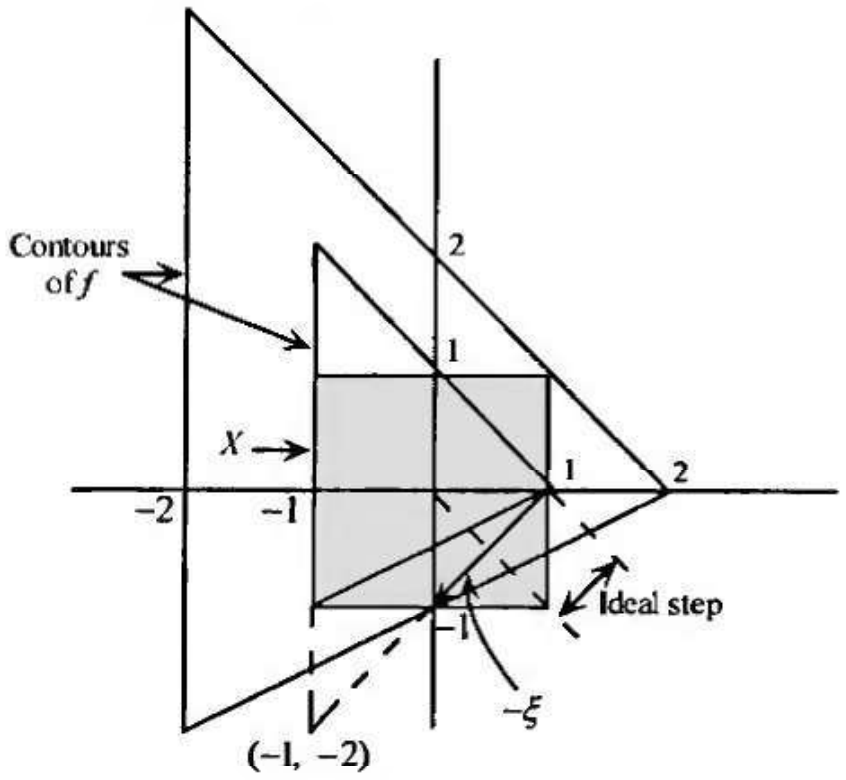


Figure 8.25 Contours of f, in Example 8.9.1.

$$\{\mathrm{UB}_k\} \to f^* \equiv \min\{f(\mathrm{x}) : \mathrm{x} \in X\}.$$

Proof

The case of finite termination follows from Theorem 3.4.3. Hence, suppose that an infinite sequence $\{\mathbf{x}_k\}$ is generated along with the accompanying sequence of upper bounds $\{\mathrm{UB}_k\}$. Since $\{\mathrm{UB}_k\}$ is monotone nonincreasing, it has a limit point $\bar{f}$. We show that this limit $\bar{f}$ equals f^* by exhibiting that for any given value $\alpha > f^*$, the sequence $\{\mathbf{x}_k\}$ enters the level set $S_\alpha = \{\mathbf{x} : f(\mathbf{x}) \le \alpha\}$. Hence, we cannot have $\bar{f} > f^*$, or else we would obtain a contradiction by taking $\alpha \in (f^*, \bar{f})$, so we must then have $\bar{f} = f^*$.

Toward this end, consider any $\hat{\mathbf{x}} \in X$ such that $f(\hat{\mathbf{x}}) < \alpha$. (For example, we can take $\hat{\mathbf{x}}$ as an optimal solution to Problem P.) Since $\hat{\mathbf{x}} \in \text{int } S_\alpha$ because f is continuous, there exists a $\rho > 0$ such that $\|\mathbf{x} - \hat{\mathbf{x}}\| \le \rho$ implies that $\mathbf{x} \in S_\alpha$. In particular, $\mathbf{x}_{Bk} = \hat{\mathbf{x}} + \rho\xi_k / \|\xi_k\|$ lies on the boundary of the ball centered at $\hat{\mathbf{x}}$ with radius ρ and hence lies in S_α for all k. But by the convexity of f, we have $f(\mathbf{x}_{Bk}) \ge f(\mathbf{x}_k) + (\mathbf{x}_{Bk} - \mathbf{x}_k)^t \xi_k$ for all k. Hence, on the contrary, if $\{\mathbf{x}_k\}$ never enters S_α, that is, $f(\mathbf{x}_k) > \alpha$ for all k, we shall have $(\mathbf{x}_{Bk} - \mathbf{x}_k)^t \xi_k \le f(\mathbf{x}_{Bk}) - f(\mathbf{x}_k) < 0$. Substituting for $\mathbf{x}_{Bk}$, this gives $(\hat{\mathbf{x}} - \mathbf{x}_k)^t \xi_k < -\rho\|\xi_k\|$. Hence, using $\mathbf{d}_k = -\xi_k / \|\xi_k\|$, we get

$$(\mathbf{x}_k - \hat{\mathbf{x}})^t \mathbf{d}_k < -\rho \qquad \text{for all } k. \tag{8.74}$$

Now we have

$$\begin{aligned}\|\bar{\mathbf{x}}_{k+1} - \hat{\mathbf{x}}\|^2 &= \|\bar{\mathbf{x}}_{k+1} - \mathbf{x}_{k+1} + \mathbf{x}_{k+1} - \hat{\mathbf{x}}\|^2 \\ &= \|\mathbf{x}_{k+1} - \hat{\mathbf{x}}\|^2 + \|\bar{\mathbf{x}}_{k+1} - \mathbf{x}_{k+1}\|^2 + 2(\bar{\mathbf{x}}_{k+1} - \mathbf{x}_{k+1})^t (\mathbf{x}_{k+1} - \hat{\mathbf{x}}).\end{aligned}$$

Hence, by Theorem 2.4.1,

$$\begin{aligned}\|\bar{\mathbf{x}}_{k+1} - \hat{\mathbf{x}}\|^2 &= \|\bar{\mathbf{x}}_{k+1} - \hat{\mathbf{x}}\|^2 - \|\bar{\mathbf{x}}_{k+1} - \mathbf{x}_{k+1}\|^2 - 2(\bar{\mathbf{x}}_{k+1} - \mathbf{x}_{k+1})^t (\mathbf{x}_{k+1} - \hat{\mathbf{x}}) \\ &\le \|\bar{\mathbf{x}}_{k+1} - \hat{\mathbf{x}}\|^2.\end{aligned}$$

Hence, we get

$$\begin{aligned}\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}\|^2 \le \|\bar{\mathbf{x}}_{k+1} - \hat{\mathbf{x}}\|^2 &= \|\mathbf{x}_k + \lambda_k \mathbf{d}_k - \hat{\mathbf{x}}\|^2 \\ &= \|\mathbf{x}_k - \hat{\mathbf{x}}\|^2 + \lambda_k^2 + 2\lambda_k \mathbf{d}_k^t (\mathbf{x}_k - \hat{\mathbf{x}}).\end{aligned}$$

Using (8.74), this gives

$$\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}\|^2 \leq \|\mathbf{x}_k - \hat{\mathbf{x}}\|^2 + \lambda_k(\lambda_k - 2\rho).$$

Since $\lambda_k \to 0^+$, there exists a K such that for $k \geq K$, we have $\lambda_k \leq \rho$. Hence,

$$\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}\|^2 \leq \|\mathbf{x}_k - \hat{\mathbf{x}}\|^2 - \rho\lambda_k \qquad \text{for all } k \geq K. \tag{8.75}$$

Summing the inequalities (8.75) written for $k = K, K+1, \ldots, K+r$, say, we get

$$\rho \sum_{k=K}^{K+r} \lambda_k \leq \|\mathbf{x}_K - \hat{\mathbf{x}}\|^2 - \|\mathbf{x}_{K+r+1} - \hat{\mathbf{x}}\|^2 \leq \|\mathbf{x}_K - \hat{\mathbf{x}}\|^2 \qquad \text{for all } r \geq 0.$$

Since the sum on the left-hand side diverges to infinity as $r \to \infty$, this leads to a contradiction, and the proof is complete.

Note that the proof of the theorem can easily be modified to show that for each $\alpha > f^*$, the sequence $\{\mathbf{x}_k\}$ enters S_α infinitely often or else, for some K', we would have $f(\mathbf{x}_k) > \alpha$ for all $k \geq K'$, leading to the same contradiction. Hence, whenever $\mathbf{x}_{k+1} = \mathbf{x}_k$ in the foregoing algorithm, $\mathbf{x}_k$ must be an optimal solution.

Furthermore, the above algorithm and proof can be extended readily to solve the problem of minimizing $f(\mathbf{x})$ subject to $\mathbf{x} \in X \cap Q$, where f and X are as above and where $Q = \{\mathbf{x} : g_i(\mathbf{x}) \leq 0, i = 1, \ldots, m\}$. Here, we assume that each g_i, $i = 1, \ldots, m$, is convex and that $X \cap \text{int}(Q) \neq \varnothing$, so that for each $\alpha > f^*$, by defining $S_\alpha \equiv \{\mathbf{x} \in Q : f(\mathbf{x}) \leq \alpha\}$, we have a point $\hat{\mathbf{x}} \in X \cap \text{int}(S_\alpha)$. Now, in the algorithm, if we let ξ_k be a subgradient of f whenever $\mathbf{x}_k \in Q$, and if we let ξ_k be a subgradient of the most violated constraint in Q if $\mathbf{x}_k \notin Q$ (noting that $\mathbf{x}_k$ always lies in X by virtue of the projection operation), we shall again have (8.74) holding true, and the remainder of the convergence proof would then follow as before.

Choice of Step Sizes

Theorem 8.2.9 guarantees that as long as the step sizes λ_k, $\forall k$, satisfy the conditions stated, convergence to an optimal solution will be obtained. Although this is true theoretically, it is unfortunately far from what is realized in practice. For example, choosing $\lambda_k = 1/k$ according to the divergent harmonic series $[\sum_{k=1}^{\infty}(1/k) = \infty]$, the algorithm can easily stall and be remote from optimality after thousands of iterations. A careful fine tuning of the choice of step sizes is usually required to obtain a satisfactory algorithmic performance.

To gain some insight into the choice of step sizes, let $\mathbf{x}_k$ be a nonoptimal iterate with $\xi_k \in \partial f(\mathbf{x}_k)$ and denote by $\mathbf{x}^*$ an optimal solution to Problem (8.71) having objective value $f^* = f(\mathbf{x}^*)$. By the convexity of f, we have $f(\mathbf{x}^*) \geq f(\mathbf{x}_k) + (\mathbf{x}^* - \mathbf{x}_k)^t \xi_k$, or $(\mathbf{x}^* - \mathbf{x}_k)^t(-\xi_k) \geq f(\mathbf{x}_k) - f^* > 0$. Hence, as observed in Example 8.9.1 (see Figure 8.25), although the direction $\mathbf{d}_k = -\xi_k / \|\xi_k\|$ need not be an improving direction, it does lead to points that are closer in Euclidean norm to $\mathbf{x}^*$ than was $\mathbf{x}_k$. In fact, this is the feature that drives the convergence of the algorithm and ensures an eventual improvement in objective function value.

Now, as in Figure 8.25, an ideal step size to adopt might be that which brings us closest to $\mathbf{x}^*$. This step size λ_k^* can be found by requiring that the vector $(\mathbf{x}_k + \lambda_k^* \mathbf{d}_k) - \mathbf{x}^*$ is orthogonal to $\mathbf{d}_k$, or that $\mathbf{d}_k^t[\mathbf{x}_k + \lambda_k^* \mathbf{d}_k - \mathbf{x}^*] = 0$. This gives

$$\lambda_k^* = (\mathbf{x}^* - \mathbf{x}_k)^t \mathbf{d}_k = \frac{(\mathbf{x}_k - \mathbf{x}^*)^t \xi_k}{\|\xi_k\|}. \tag{8.76}$$

Of course, the problem with trying to implement this step size λ_k^* is that $\mathbf{x}^*$ is unknown. However, by the convexity of f, we have $f^* = f(\mathbf{x}^*) \geq f(\mathbf{x}_k) + (\mathbf{x}^* - \mathbf{x}_k)^t \xi_k$. Hence, from (8.76), we have that $\lambda_k^* \geq [f(\mathbf{x}_k) - f^*]/\|\xi_k\|$. Since f^* is also usually unknown, we can recommend using an underestimate, $\bar{f}$, in lieu of f^*, noting that the foregoing relationship is a "greater than or equal to" type of inequality. This leads to a choice of step size

$$\lambda_k = \frac{\beta_k[f(\mathbf{x}_k) - \bar{f}]}{\|\xi_k\|}, \tag{8.77}$$

where $\beta_k > 0$. In fact, by selecting $\varepsilon_1 < \beta_k \leq 2 - \varepsilon_2$ for all k for some positive ε_1 and ε_2, and using f^* itself instead of $\bar{f}$ in (8.77), it can be shown that the generated sequence $\{\mathbf{x}_k\}$ converges to an optimum $\mathbf{x}^*$. (A linear or geometric convergence rate can be exhibited under some additional assumptions on f.)

A practical way of employing (8.77) that has been found empirically to be computationally attractive is as follows (this is called a *block halving scheme*). First, designate an upper limit N on the number of iterations to be performed. Next, select some $\bar{r} < N$ and divide the potential sequence of iterations 1,..., N into $T = \lceil N/\bar{r} \rceil$ blocks, with the first $T - 1$ blocks having $\bar{r}$ iterations, and the final block having the remaining ($\leq \bar{r}$) iterations. Also, for

each block t, select a parameter value $\beta(t)$, for $t = 1,...,T$. [Typical values are $N = 200$, $\bar{r} = 75$, $\beta(1) = 0.75$, $\beta(2) = 0.5$, and $\beta(3) = 0.25$, with $T = 3$.] Now, within each block t, compute the first step length using (8.77), with β_k equal to the corresponding $\beta(t)$ value. However, for the remaining iterations within the block, the *step length* is kept the same as for the initial iteration in that block, except that each time the objective function value fails to improve over some $\bar{v}$ (= 10, say) consecutive iterations, the step length is successively halved. [Alternatively, (8.77) can be used to compute the step length for each iteration, with β_k starting at $\beta(t)$ for block t, and with this β parameter being halved whenever the method experiences $\bar{v}$ consecutive failures as before.] Additionally, at the beginning of a new block, and also whenever the method experiences $\bar{v}$ consecutive failures, the process is reset to the incumbent solution before the modified step length is used. Although some fine tuning of the foregoing parameter values might be required, depending on the class of problems being solved, the prescribed values work well on reasonably well-scaled problems (see the Notes and References section for empirical evidence using such a scheme).

Subgradient Deflection, Cutting Plane, and Variable Target Value Methods

It has frequently been observed that the difficulty associated with subgradient methods is that as the iterates progress, the angle between the subgradient–based direction $\mathbf{d}_k$ and the direction $\mathbf{x}^* - \mathbf{x}_k$ toward optimality, although acute, tends to approach 90°. As a result, the step size needs to shrink considerably before a descent is realized, and this, in turn, causes the procedure to stall. Hence, it becomes almost imperative to adopt some suitable deflection or rotation scheme to accelerate the convergence behavior.

Toward this end, in the spirit of conjugate gradient methods, we could adopt a direction of search as $\mathbf{d}_1 = -\xi_1$ and $\mathbf{d}_k = -\xi_k + \phi_k \mathbf{d}^a_{k-1}$, where $\mathbf{d}^a_{k-1} \equiv \mathbf{x}_k - \mathbf{x}_{k-1}$ and ϕ_k is an appropriate parameter. (These directions can be normalized and then used in conjunction with the same block-halving step size strategy described above.) Various strategies prompted by theoretical convergence and/or practical efficiency can be designed by choosing ϕ_k appropriately (see the Notes and References section). A simple choice that works reasonably well in practice is the *average direction strategy*, for which $\phi_k = \|\xi_k\| / \|\mathbf{d}^a_{k-1}\|$, so that $\mathbf{d}_k$ bisects the angle between $-\xi_k$ and $\mathbf{d}^a_{k-1}$.

Another viable strategy is to imitate quasi-Newton procedures by using $\mathbf{d}_k = -\mathbf{D}_k \xi_k$, where $\mathbf{D}_k$ is a suitable, symmetric, positive definite matrix. This leads to the class of *space dilation* methods (see the Notes and References section). Alternatively, we could generate a search direction by finding the minimum norm subgradient as motivated by Theorem 6.3.11, but based on an

approximation to the subdifferential at $\mathbf{x}_k$ and not the actual subdifferential as in the theorem. The class of *bundle methods* (see the Notes and References section) are designed to iteratively refine such an approximation to the subdifferential until the least norm element yields a descent direction. Note that this desirable strict descent property comes at the expense of having to solve quadratic optimization subproblems, which detract from the simplicity of the foregoing types of subgradient methods.

Thus far, the basic algorithm scheme that we have adopted involves first finding a direction of motion $\mathbf{d}_k$ at a given iterate $\mathbf{x}_k$, followed by computing a prescribed step size λ_k in order to determine the next iterate according to

$$\mathbf{x}_{k+1} = P_X(\bar{\mathbf{x}}_{k+1}), \qquad \text{where } \bar{\mathbf{x}}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k.$$

There exists an alternative approach in which $\bar{\mathbf{x}}_{k+1}$ is determined directly via a projection of $\mathbf{x}_k$ onto the polyhedron defined by one or more cutting planes, thereby effectively yielding the direction and step size simultaneously. To motivate this strategy, let us first consider the case of a single cutting plane. Note that by the assumed convexity of f, we have $f(\mathbf{x}) \geq f(\mathbf{x}_k) + (\mathbf{x} - \mathbf{x}_k)^t \xi_k$, where $\xi_k \in \partial f(\mathbf{x}_k)$. Let f^* denote the optimal objective value, and assume for now that $f(\mathbf{x}_k) > f^*$, so that ξ_k is nonzero. Consider the *Polyak–Kelly cutting plane* generated from the foregoing convexity-based inequality by imposing the desired restriction that $f(\mathbf{x}) \leq f^*$ as given by

$$(\mathbf{x} - \mathbf{x}_k)^t \xi_k \leq f^* - f(\mathbf{x}_k). \tag{8.78}$$

Observe that the current iterate $\mathbf{x}_k$ violates this inequality since $f(\mathbf{x}_k) > f^*$, and hence, (8.78) constitutes a *cutting plane* that deletes $\mathbf{x}_k$. If we were to project the point $\mathbf{x}_k$ onto this cutting plane, we would effectively move from $\mathbf{x}_k$ a step length of $\tilde{\lambda}$, say, along the negative normalized gradient $\mathbf{d}_k \equiv -\xi_k / \|\xi_k\|$, such that $\mathbf{x}_k + \tilde{\lambda}\mathbf{d}_k$ satisfies (8.78) as an equality. This yields the projected solution

$$\tilde{\mathbf{x}}_{k+1} = \mathbf{x}_k + \tilde{\lambda}\mathbf{d}_k, \qquad \text{where } \mathbf{d}_k = \frac{-\xi_k}{\|\xi_k\|} \text{ and } \tilde{\lambda} = \frac{f(\mathbf{x}_k) - f^*}{\|\xi_k\|}. \tag{8.79}$$

Observe that the effective step length $\tilde{\lambda}$ in (8.79) is of the type given in (8.77), with f^* itself being used in lieu of the underestimate $\bar{f}$, and with $\beta_k \equiv 1$. This affords another interpretation for the step size (8.77).

Following this concept, we now concurrently examine a pair of Polyak–Kelly cutting planes based on the current and previous iterates $\mathbf{x}_k$ and $\mathbf{x}_{k-1}$,

respectively. These cuts are predicated on some underestimate $\bar{f}_k$ at the present iteration k that is less than the currently best known objective value. Imitating (8.78), this yields the set

$$G_k = \{\mathbf{x} : (\mathbf{x} - \mathbf{x}_j)^t \xi_j \leq \bar{f}_k - f(\mathbf{x}_j) \text{ for } j = k-1 \text{ and } k\}. \tag{8.80}$$

We then compose the next iterate via a projection onto the polyhedron G_k according to

$$\mathbf{x}_{k+1} = P_X(\bar{\mathbf{x}}_{k+1}), \qquad \text{where } \bar{\mathbf{x}}_{k+1} = P_{G_k}(\mathbf{x}_k). \tag{8.81}$$

Because of its simple two-constraint structure, the projection $P_{G_k}(\cdot)$ is relatively easy to compute in closed-form by examining the KKT conditions for the underlying projection problem (see Exercise 8.58). This process of determining the direction and step size simultaneously has been found to be computationally very effective, and can be proven to converge to an optimum under an appropriate prescription of $\bar{f}_k$, $\forall k$ (see below as well as the Notes and References section).

We conclude this section by providing an important relevant comment on selecting a suitable underestimating value $\bar{f}_k$, $\forall k$, that could be used in place of $\bar{f}$ within (8.77), or in the algorithmic process described by (8.80) and (8.81). Note that, in general, we typically do not have any prior information on such a suitable lower bound on the problem. It is of interest, therefore, to design algorithms that would prescribe an automatic scheme for generating and iteratively manipulating an estimate $\bar{f}_k$ for f^*, $\forall k$, that in concert with prescribed direction-finding and step-size schemes would ensure that $\{\bar{f}_k\} \to f^*$ and $\{\mathbf{x}_k\} \to \mathbf{x}^*$ (over some convergent subsequence) as $k \to \infty$. There exists a class of algorithms called *variable target value methods* that possesses this feature. Note that the estimate $\bar{f}_k$ at any iteration k in these procedures might not be a true underestimate for f^*. Rather, $\bar{f}_k$ merely serves as a current *target value* to be achieved, which happens to be less than the objective function value best known at present. The idea, then, is to decrease or increase $\bar{f}_k$ suitably, depending on whether or not a defined sufficient extent of progress is made by the algorithm, in a manner that finally induces convergence to an optimal solution. Approaches of this type have been designed to yield both theoretically convergent and practically effective schemes under various deflected subgradient and step-size schemes, including cutting plane projection methods as described above. We refer the reader to the Notes and References section for a further study on this subject.

Exercises

[8.1] Find the minimum of $6e^{-2\lambda} + 2\lambda^2$ by each of the following procedures:

a. Golden section method.
b. Dichotomous search method.
c. Newton's method.
d. Bisection search method.

[8.2] For the uniform search method, the dichotomous search method, the golden section method, and the Fibonacci search method, compute the number of functional evaluations required for $\alpha = 0.1, 0.01, 0.001$, and 0.0001, where α is the ratio of the final interval of uncertainty to the length of the initial interval of uncertainty.

[8.3] Consider the function f defined by $f(\mathbf{x}) = (x_1 + x_2^3)^2 + 2(x_1 - x_2 - 4)^4$. Given a point $\mathbf{x}_1$ and a nonzero direction vector $\mathbf{d}$, let $\theta(\lambda) = f(\mathbf{x}_1 + \lambda\mathbf{d})$.

a. Obtain an explicit expression for $\theta(\lambda)$.
b. For $\mathbf{x}_1 = (0, 0)^t$ and $\mathbf{d} = (1, 1)^t$, using the Fibonacci method, find the value of λ that solves the problem to minimize $\theta(\lambda)$ subject to $\lambda \in R$.
c. For $\mathbf{x}_1 = (5, 4)^t$ and $\mathbf{d} = (-2, 1)^t$, using the golden section method, find the value of λ that solves the problem to minimize $\theta(\lambda)$ subject to $\lambda \in R$.
d. Repeat parts b and c using the interval bisection method.

[8.4] Show that the method of Fibonacci approaches the golden section method as the number of funcitonal evaluations n approaches ∞.

[8.5] Consider the problem to minimize $f(\mathbf{x} + \lambda\mathbf{d})$ subject to $\lambda \in R$. Show that a necessary condition for a minimum at $\bar{\lambda}$ is that $\mathbf{d}^t\nabla f(\mathbf{y}) = 0$, where $\mathbf{y} = \mathbf{x} + \bar{\lambda}\mathbf{d}$. Under what assumptions is this condition sufficient for optimality?

[8.6] Suppose that θ is differentiable, and let $|\theta'| \leq a$. Furthermore, suppose that the uniform search method is used to minimize θ. Let $\hat{\lambda}$ be a grid point such that $\theta(\bar{\lambda}) - \theta(\hat{\lambda}) \geq \varepsilon > 0$ for each grid point $\bar{\lambda} \neq \hat{\lambda}$. If the grid length δ is such that $a\delta \leq \varepsilon$, show, without assuming strict quasiconvexity, that no point outside the interval $[\hat{\lambda} - \delta, \hat{\lambda} + \delta]$ could provide a functional value of less than $\theta(\hat{\lambda})$.

[8.7] Consider the problem to minimize $f(\mathbf{x} + \lambda\mathbf{d})$ subject to $\mathbf{x} + \lambda\mathbf{d} \in S$ and $\lambda \geq 0$, where S is a compact convex set and f is a convex function. Furthermore, suppose that $\mathbf{d}$ is an improving direction. Show that an optimal solution $\bar{\lambda}$ is

given by $\bar{\lambda} = \min\{\lambda_1, \lambda_2\}$, where λ_1 satisfies $\mathbf{d}^t \nabla f(\mathbf{x} + \lambda_1 \mathbf{d}) = 0$, and $\lambda_2 = \max\{\lambda : \mathbf{x} + \lambda \mathbf{d} \in S\}$.

[8.8] Define the *percentage test line search map* that determines the step length λ to within $100p\%$, $0 \le p \le 1$, of the ideal step λ^* according to $\mathbf{M}(\mathbf{x}, \mathbf{d}) = \{\mathbf{y} : \mathbf{y} = \mathbf{x} + \lambda \mathbf{d}$, where $0 \le \lambda < \infty$, and $|\lambda - \lambda^*| \le p\lambda^*\}$, where defining $\theta(\lambda) \equiv f(\mathbf{x} + \lambda \mathbf{d})$, we have $\theta'(\lambda^*) = 0$. Show that if $\mathbf{d} \neq \mathbf{0}$ and θ is continuously differentiable, then $\mathbf{M}$ is closed at $(\mathbf{x}, \mathbf{d})$. Explain how you can use this test in conjunction with the quadratic-fit method described in Section 8.3.

[8.9] Consider the problem to minimize $3\lambda - 2\lambda^2 + \lambda^3 + 2\lambda^4$ subject to $\lambda \ge 0$.

a. Write a necessary condition for a minimum. Can you make use of this condition to find the global minimum?
b. Is the function strictly quasiconvex over the region $\{\lambda : \lambda \ge 0\}$? Apply the Fibonacci search method to find the minimum.
c. Apply both the bisection search method and Newton's method to the above problem, starting from $\lambda_1 = 6$.

[8.10] Consider the following definitions:

A function $\theta : R \to R$ to be *minimized* is said to be *strongly unimodal* over the interval $[a, b]$ if there exists a $\bar{\lambda}$ that minimizes θ over the interval; and for any $\lambda_1, \lambda_2 \in [a, b]$ such that $\lambda_1 < \lambda_2$, we have

$$\lambda_2 \le \bar{\lambda} \quad \text{implies that } \theta(\lambda_1) > \theta(\lambda_2)$$
$$\lambda_1 \ge \bar{\lambda} \quad \text{implies that } \theta(\lambda_1) < \theta(\lambda_2).$$

A function $\theta : R \to R$ to be *minimized* is said to be *strictly unimodal* over the interval $[a, b]$ if there exists a $\bar{\lambda}$ that minimizes θ over the interval; and for $\lambda_1, \lambda_2 \in [a, b]$ such that $\theta(\lambda_1) \neq \theta(\bar{\lambda})$, $\theta(\lambda_2) \neq \theta(\bar{\lambda})$, and $\lambda_1 < \lambda_2$, we have

$$\lambda_2 \le \bar{\lambda} \quad \text{implies that } \theta(\lambda_1) > \theta(\lambda_2)$$
$$\lambda_1 \ge \bar{\lambda} \quad \text{implies that } \theta(\lambda_1) < \theta(\lambda_2).$$

a. Show that if θ is strongly unimodal over $[a, b]$, then θ is strongly quasiconvex over $[a, b]$. Conversely, show that if θ is strongly quasiconvex over $[a, b]$ and has a minimum in this interval, then it is strongly unimodal over the interval.
b. Show that if θ is strictly unimodal and continuous over $[a, b]$, then θ is strictly quasiconvex over $[a, b]$. Conversely, show that if θ is strictly quasiconvex over $[a, b]$ and has a minimum in this interval, then it is strictly unimodal over this interval.

[8.11] Let $\theta: R \to R$ and suppose that we have the three points (λ_1, θ_1), (λ_2, θ_2), and (λ_3, θ_3), where $\theta_j = \theta(\lambda_j)$ for $j = 1, 2, 3$. Show that the quadratic curve q passing through these points is given by

$$q(\lambda) = \frac{\theta_1(\lambda-\lambda_2)(\lambda-\lambda_3)}{(\lambda_1-\lambda_2)(\lambda_1-\lambda_3)} + \frac{\theta_2(\lambda-\lambda_1)(\lambda-\lambda_3)}{(\lambda_2-\lambda_1)(\lambda_2-\lambda_3)} + \frac{\theta_3(\lambda-\lambda_1)(\lambda-\lambda_2)}{(\lambda_3-\lambda_1)(\lambda_3-\lambda_2)}.$$

Furthermore, show that the derivative of q vanishes at the point $\bar{\lambda}$ given by

$$\bar{\lambda} = \frac{1}{2} \cdot \frac{b_{23}\theta_1 + b_{31}\theta_2 + b_{12}\theta_3}{a_{23}\theta_1 + a_{31}\theta_2 + a_{12}\theta_3},$$

where $a_{ij} = \lambda_i - \lambda_j$ and $b_{ij} = \lambda_i^2 - \lambda_j^2$. Find the quadratic curve passing through the points (1, 4), (3, 1), and (4, 7), and compute $\bar{\lambda}$. Show that if $(\lambda_1, \lambda_2, \lambda_3)$ satisfy the three-point pattern (TPP), then $\lambda_1 < \bar{\lambda} < \lambda_3$. Also:

a. Propose a method for finding λ_1, λ_2, λ_3 such that $\lambda_1 < \lambda_2 < \lambda_3$, $\theta_1 \geq \theta_2$, and $\theta_2 \leq \theta_3$.

b. Show that if θ is strictly quasiconvex, then the new interval of uncertainty defined by the revised λ_1 and λ_3 of the quadratic-fit line search indeed contains the minimum.

c. Use the procedure described in this exercise to minimize $-3\lambda - 2\lambda^2 + 2\lambda^3 + 3\lambda^4$ over $\lambda \geq 0$.

[8.12] Let $\theta: R \to R$ be a continuous strictly quasiconvex function. Let $0 \leq \lambda_1 < \lambda_2 < \lambda_3$ and denote $\theta_j = \theta(\lambda_j)$ for $j = 1, 2, 3$.

a. If $\theta_1 = \theta_2 = \theta_3$, show that this common value coincides with the value of $\min\{\theta(\lambda): \lambda \geq 0\}$.

b. Let $(\lambda_1, \lambda_2, \lambda_3) \in R^3$ represent a three-point pattern iterate generated by the quadratic-fit algorithm described in Section 8.3. Show that the function $\bar{\theta}(\lambda_1, \lambda_2, \lambda_3) \equiv \theta(\lambda_1) + \theta(\lambda_2) + \theta(\lambda_3)$ is a continuous function that satisfies the descent property $\bar{\theta}[(\lambda_1, \lambda_2, \lambda_3)_{\text{new}}] < \bar{\theta}(\lambda_1, \lambda_2, \lambda_3)$ whenever θ_1, θ_2, and θ_3 are not all equal to each other.

[8.13] Let θ be pseudoconvex and continuously twice differentiable. Consider the algorithm of Section 8.3 with the modification that in Case 3, when $\bar{\lambda} = \lambda_2$, we let $\lambda_{\text{new}} = (\lambda_1, \lambda_2, \bar{\lambda})$ if $\theta'(\lambda_2) > 0$, we let $\lambda_{\text{new}} = (\lambda_2, \bar{\lambda}, \lambda_3)$ if $\theta'(\lambda_2) < 0$, and we stop if $\theta'(\lambda_2) = 0$. Accordingly, if λ_1, λ_2, and λ_3 are not all distinct, let them be said to satisfy the *three-point pattern* (TPP) whenever $\theta'(\lambda_2) < 0$ if $\lambda_1 = \lambda_2 < \lambda_3$, $\theta'(\lambda_2) > 0$ if $\lambda_1 < \lambda_2 = \lambda_3$, and $\theta'(\lambda_2) = 0$ and

$\theta''(\lambda_2) \geq 0$ if $\lambda_1 = \lambda_2 = \lambda_3$. With this modification, suppose that we use the quadratic interpolation algorithm of Section 8.3 applied to θ given a starting TPP $(\lambda_1, \lambda_2, \lambda_3)$, where the quadratic fit matches the two function values and the derivative $\theta'(\lambda_2)$ whenever two of the three points λ_1, λ_2, λ_3 are coincident, and where at any iteration, if $\theta'(\lambda_2) = 0$, we put $\lambda^* = (\lambda_2, \lambda_2, \lambda_2)$ and terminate. Define the solution set $\Omega = \{(\lambda, \lambda, \lambda) : \theta'(\lambda) = 0\}$.

a. Let **A** define the algorithmic map that produces $\lambda_{\text{new}} \in \mathbf{A}(\lambda_1, \lambda_2, \lambda_3)$. Show that **A** is closed.

b. Show that the function $\bar{\theta}(\lambda_1, \lambda_2, \lambda_3) = \theta(\lambda_1) + \theta(\lambda_2) + \theta(\lambda_3)$ is a continuous descent function that satisfies $\bar{\theta}(\lambda_{\text{new}}) < \bar{\theta}(\lambda_1, \lambda_2, \lambda_3)$ if $\theta'(\lambda_2) \neq 0$.

c. Hence, show that the algorithm defined either terminates finitely or generates an infinite sequence whose accumulation points lie in Ω.

d. Comment on the convergence of the algorithm and the nature of the solution obtained if θ is strictly quasiconvex and twice continuously differentiable.

[8.14] In Section 8.2 we described Newton's method for finding a point where the derivative of a function vanishes.

a. Show how the method can be used to find a point where the value of a continuously differentiable function is equal to zero. Illustrate the method for $\theta(\lambda) = 2\lambda^3 - \lambda$, starting from $\lambda_1 = 5$.

b. Will the method converge for any starting point? Prove or give a counterexample.

[8.15] Show how the line search procedures of Section 8.1 can be used to find a point where a given function assumes the value zero. Illustrate by the function θ defined by $\theta(\lambda) = 2\lambda^2 - 5\lambda + 3$. (*Hint*: Consider the absolute value function $\hat{\theta} = |\theta|$.)

[8.16] In Section 8.2 we discussed the bisection search method for finding a point where the derivative of a pseudoconvex function vanishes. Show how the method can be used to find a point where a function is equal to zero. Explicitly state the assumptions that the function needs to satisfy. Illustrate by the function θ defined by $\theta(\lambda) = 2\lambda^3 - \lambda$ defined on the interval [0.5, 10.0].

[8.17] It can be verified that in Example 9.2.4, for a given value of μ, if $\mathbf{x}_\mu = (x_1, x_2)^t$, then x_1 satisfies

$$2(x_1 - 2)^3 + \frac{\mu x_1(8x_1^2 - 6x_1 + 1)}{4 + \mu} = 0.$$

For $\mu = 1$, 10, 100, and 1000, find the value of x_1 satisfying the above equation, using a suitable procedure.

[8.18] Consider applying the steepest descent method to minimize $f(\mathbf{x})$ versus the application of this method to minimize $F(\mathbf{x}) = \|\nabla f(\mathbf{x})\|^2$. Assuming that f is quadratic with a positive definite Hessian, compare the rates of convergence of the two schemes and, hence, justify why the equivalent minimization of F is not an attractive strategy.

[8.19] Show that as a function of K_k, the expression in Equation (8.14) is maximized when $K_k^2 = \alpha^2$.

[8.20] Solve the problem to maximize $3x_1 + x_2 + 6x_1x_2 - 2x_1^2 + 2x_2^2$ by the method of Hooke and Jeeves.

[8.21] Let $\mathbf{H}$ be an $n \times n$, symmetric, positive definite matrix with condition number α. Then the *Kantorovich inequality* asserts that for any $\mathbf{x} \in R^n$, we have

$$\frac{(\mathbf{x}^t\mathbf{x})^2}{(\mathbf{x}^t\mathbf{H}\mathbf{x})(\mathbf{x}^t\mathbf{H}^{-1}\mathbf{x})} \geq \frac{4\alpha}{(1+\alpha)^2}.$$

Justify this inequality, and use it to establish Equation (8.18).

[8.22] Consider the problem to minimize $(3 - x_1)^2 + 7(x_2 - x_1^2)^2$. Starting from the point (0, 0), solve the problem by the following procedures:

a. The cyclic coordinate method.
b. The method of Hooke and Jeeves.
c. The method of Rosenbrock.
d. The method of Davidon–Fletcher–Powell.
e. The method of Broyden–Fletcher–Goldfarb–Shanno (BFGS).

[8.23] Consider the following problem:

$$\text{Minimize } \sum_{i=2}^{n} [100(x_i - x_{i-1}^2)^2 + (1 - x_{i-1})^2].$$

For values of $n = 5$, 10 and 50, and starting from the solution $\mathbf{x}^0 = (-1.2, 1.0, -1.2, 1.0, \ldots)$, solve this problem using each of the following methods. (Write subroutines for evaluating the objective function, its gradient, and for performing a line search via the quadratic-fit method, and then use these subroutines to compose codes for the following methods. Also, you could use the previous iteration's step length as the initial step for establishing a three-point pattern (TPP) for the current iteration. Present a summary of comparative results.

a. The method of Hooke and Jeeves (use the line search variant and the same termination criteria as for the other methods, to facilitate comparisons).
b. Rosenbrock's method (again, use the line search variant as in Part a).
c. The steepest descent method.
d. Newton's method.
e. Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton method.
f. The conjugate gradient method of Hestenes and Stiefel.
g. The conjugate gradient method of Fletcher and Reeves.
h. The conjugate gradient method of Polyak and Ribiere.

[8.24] Consider the problem to minimize $(x_1 - x_2^3)^2 + 3(x_1 - x_2)^4$. Solve the problem using each of the following methods. Do the methods converge to the same point? If not, explain.

a. The cyclic coordinate method.
b. The method of Hooke and Jeeves.
c. The method of Rosenbrock.
d. The method of steepest descent.
e. The method of Fletcher and Reeves.
f. The method of Davidon–Fletcher–Powell.
g. The method of Broyden–Fletcher–Goldfarb–Shanno (BFGS).

[8.25] Consider the model $y = \alpha + \beta x + \gamma x^2 + \varepsilon$, where x is the independent variable, y is the observed dependent variable, α, β, and γ are unknown parameters, and ε is a random component representing the experimental error. The following table gives the values of x and the corresponding values of y. Formulate the problem of finding the best estimates of α, β, and γ as an unconstrained optimization problem by minimizing:

a. The sum of squared errors.
b. The sum of the absolute values of the errors.
c. The maximum absolute value of the error.

For each case, find α, β, and γ by a suitable method.

x	0	1	2	3	4	5
y	3	3	−10	−25	−50	−100

[8.26] Consider the following problem:

$$\begin{aligned} \text{Minimize} \quad & 2x_1 + x_2 \\ \text{subject to} \quad & x_1^2 + x_2^2 = 9 \\ & -2x_1 - 3x_2 \leq 6. \end{aligned}$$

a. Formulate the Lagrangian dual problem by incorporating both constraints into the objective function via the Lagrangian multipliers u_1 and u_2.

b. Using a suitable unconstrained optimization method, compute the gradient of the dual function θ at the point (1, 2).

c. Starting from the point $\bar{\mathbf{u}} = (1, 2)^t$, perform one iteration of the steepest ascent method for the dual problem. In particular, solve the following problem, where $\mathrm{d} = \nabla\theta(\bar{u})$:

$$\begin{aligned} &\text{Maximize } \theta(\bar{\mathrm{u}} + \lambda \mathrm{d}) \\ &\text{subject to } \bar{u}_2 + \lambda d_2 \geq 0 \\ &\qquad\qquad \lambda \geq 0. \end{aligned}$$

[8.27] Let f: $R^n \to R$ be differentiable at x and let the vectors $\mathrm{d}_1, \ldots, \mathrm{d}_n$ in R^n be linearly independent. Suppose that the minimum of $f(\mathrm{x} + \lambda \mathrm{d}_j)$ over $\lambda \in R$ occurs at $\lambda = 0$ for $j = 1, \ldots, n$. Show that $\nabla f(\mathrm{x}) = 0$. Does this imply that f has a local minimum at x?

[8.28] Let H be a symmetric $n \times n$ matrix, and let $\mathrm{d}_1, \ldots, \mathrm{d}_n$ be a set of characteristic vectors of H. Show that $\mathrm{d}_1, \ldots, \mathrm{d}_n$ are H-conjugate.

[8.29] Consider the problem in Equation (8.24) and suppose that $\varepsilon_k \geq 0$ is such that $\mathbf{H}(\mathrm{x}_k) + \varepsilon_k \mathbf{I}$ is positive definite. Let $\Delta_k = -[\mathbf{H}(\mathrm{x}_k) + \varepsilon_k \mathrm{I}]^{-1} \nabla f(\mathrm{x}_k)$. Show that $\delta = \mathrm{x}_{k+1} - \mathrm{x}_k$, given by (8.22), and the Lagrange multiplier $\lambda = \varepsilon_k$ satisfy the saddle point optimality conditions for (8.24). Hence, comment on the relationship between the Levenberg–Marquardt and trust region methods. Also comment on the case $\varepsilon_k = 0$.

[8.30] The following method for generating a set of conjugate directions for minimizing f: $R^n \to R$ is due to Zangwill [1967b]:

Initialization Step Choose a termination scalar $\varepsilon > 0$, and choose an initial point x_1. Let $\mathbf{y}_1 = \mathbf{x}_1$, let $\mathbf{d}_1 = -\nabla f(\mathrm{y}_1)$, let $k = j = 1$, and go to the Main Step.

Main Step

1. Let λ_j be an optimal solution to the problem to minimize $f(\mathbf{y}_j + \lambda \mathbf{d}_j)$ subject to $\lambda \in R$, and let $\mathbf{y}_{j+1} = \mathbf{y}_j + \lambda_j \mathrm{d}_j$. If $j = n$, go to Step 4; otherwise, go to Step 2.
2. Let $\mathrm{d} = -\nabla f(\mathbf{y}_{j+1})$, and let $\hat{\mu}$ be an optimal solution to the problem to minimize $f(\mathbf{y}_{j+1} + \mu \mathrm{d})$ subject to $\mu \geq 0$. Let $\mathbf{z}_1 = \mathbf{y}_{j+1} + \hat{\mu} \mathrm{d}$. Let $i = 1$, and go to Step 3.
3. If $\|\nabla f(\mathrm{z}_i)\| < \varepsilon$, stop with $\mathbf{z}_i$. Otherwise, let μ_i be an optimal solution to the problem to minimize $f(\mathrm{z}_i + \mu \mathrm{d}_i)$ subject to $\mu \in R$.

Let $\mathbf{z}_{i+1} = \mathbf{z}_i + \mu_i \mathbf{d}_i$. If $i < j$, replace i by $i + 1$, and repeat Step 3. Otherwise, let $\mathbf{d}_{j+1} = \mathbf{z}_{j+1} - \mathbf{y}_{j+1}$, replace j by $j + 1$, and go to Step 1.

4. Let $\mathbf{y}_1 = \mathbf{x}_{k+1} = \mathbf{y}_{n+1}$. Let $\mathbf{d}_1 = -\nabla f(\mathbf{y}_1)$, replace k by $k + 1$, let $j = 1$, and go to Step 1.

Note that the steepest descent search in Step 2 is used to ensure that $\mathbf{z}_1 - \mathbf{y}_1 \notin L(\mathbf{d}_1, ..., \mathbf{d}_j)$ for the quadratic case so that finite convergence is guaranteed. Illustrate using the problem to minimize $(x_1 - 2)^4 + (x_1 - 2x_2)^2$, starting from the point (0.0, 3.0).

[8.31] Suppose that f is continuously twice differentiable and that the Hessian matrix is invertible everywhere. Given $\mathbf{x}_k$, let $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$, where $\mathbf{d}_k = -\mathbf{H}(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$ and λ_k is an optimal solution to the problem to minimize $f(\mathbf{x}_k + \lambda \mathbf{d}_k)$ subject to $\lambda \in R$. Show that this modification of Newton's method converges to a point in the solution set $\Omega = \{\bar{\mathbf{x}} : \nabla f(\bar{\mathbf{x}})^t \mathbf{H}(\bar{\mathbf{x}})^{-1} \nabla f(\bar{\mathbf{x}}) = 0\}$. Illustrate by minimizing $(x_1 - 2)^4 + (x_1 - 2x_2)^2$ starting from the point (–2, 3).

[8.32] Let $\mathbf{a}_1, ..., \mathbf{a}_n$ be a set of linearly independent vectors in R^n, and let $\mathbf{H}$ be an $n \times n$ symmetric positive definite matrix.

a. Show that the vectors $\mathbf{d}_1, ..., \mathbf{d}_n$ defined below are $\mathbf{H}$-conjugate.

$$\mathbf{d}_k = \begin{cases} \mathbf{a}_k & \text{if } k = 1 \\ \mathbf{a}_k - \sum_{i=1}^{k-1} \left(\dfrac{\mathbf{d}_i^t \mathbf{H} \mathbf{a}_k}{\mathbf{d}_i^t \mathbf{H} \mathbf{d}_i} \right) \mathbf{d}_i & \text{if } k \geq 2. \end{cases}$$

b. Suppose that $\mathbf{a}_1, ..., \mathbf{a}_n$ are the unit vectors in R^n, and let $\mathbf{D}$ be the matrix whose columns are the vectors $\mathbf{d}_1, ..., \mathbf{d}_n$ defined in Part a. Show that $\mathbf{D}$ is upper triangular with all diagonal elements equal to 1.

c. Illustrate by letting $\mathbf{a}_1 = (1, 0, 0)^t$, $\mathbf{a}_2 = (1, -1, 4)^t$, $\mathbf{a}_3 = (2, -1, 6)^t$, and

$$\mathbf{H} = \begin{bmatrix} 2 & 0 & -1 \\ 0 & 3 & 2 \\ -1 & 2 & 2 \end{bmatrix}.$$

d. Illustrate by letting $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ be the unit vectors in R^3 and $\mathbf{H}$ as given in Part c.

[8.33] Consider the following problem:

$$\text{Minimize } 2x_1^2 + 3x_1x_2 + 4x_2^2 + 2x_3^2 - 2x_2x_3 + 5x_1 + 3x_2 - 4x_3.$$

Using Exercise 8.32 or any other method, generate three conjugate directions. Starting from the origin, solve the problem by minimizing along these directions.

[8.34] Show that analogous to (8.66), assuming exact line searches, a memoryless quasi-Newton update performed on a member of the Broyden family (taking $\mathbf{D}_j \equiv \mathbf{I}$) results in a direction $\mathbf{d}_{j+1} = -\mathbf{D}_{j+1}\nabla f(\mathbf{y}_{j+1})$, where

$$\mathbf{D}_{j+1} = \mathbf{I} - (1-\phi)\frac{\mathbf{q}_j\mathbf{q}_j^t}{\mathbf{q}_j^t\mathbf{q}_j} - \phi\frac{\mathbf{p}_j\mathbf{q}_j^t}{\mathbf{p}_j^t\mathbf{q}_j}.$$

Observe that the equivalence with conjugate gradient methods results only when $\phi = 1$ (BFGS update).

[8.35] Show that there exists a value of ϕ [as given by Equation (8.50)] for the Broyden correction formula (8.46) that will yield $\mathbf{d}_{j+1} = -\mathbf{D}_{j+1}\nabla f(\mathbf{y}_{j+1}) = \mathbf{0}$.

[*Hint*: Use $\mathbf{p}_j = \lambda_j\mathbf{d}_j - \lambda_j\mathbf{D}_j\nabla f(\mathbf{y}_j)$, $\mathbf{q}_j = \nabla f(\mathbf{y}_{j+1}) - \nabla f(\mathbf{y}_j)$, and $\mathbf{d}_j^t\nabla f(\mathbf{y}_{j+1})$ $= \mathbf{p}_j^t\nabla f(\mathbf{y}_{j+1}) = \nabla f(\mathbf{y}_j)^t\mathbf{D}_j\nabla f(\mathbf{y}_{j+1}) = 0$.]

[8.36] Use two sequential applications of the Sherman–Morrison–Woodbury formula given in Equation (8.55) to verify the inverse relationship (8.54) between (8.48) and (8.53).

[8.37] Derive the Hessian correction (8.53) for the BFGS update directly, following the scheme used for the update of the Hessian inverse via (8.41)–(8.45).

[8.38] Referring to Figure 8.24 and the associated discussion, verify that the minimization of the quadratic function f from $\mathbf{y}_j$ along the pattern direction $\mathbf{d}_P \equiv \mathbf{y}'_{j+1} - \mathbf{y}_j$ will produce the point $\mathbf{y}_{j+2}$. [*Hint*: Let $\mathbf{y}'_{j+2}$ denote the point thus obtained. Using the fact that $\nabla f(\mathbf{y}'_{j+1})^t\nabla f(\mathbf{y}_{j+1}) = 0$ and that ∇f is linear, since f is quadratic, show that $\nabla f(\mathbf{y}'_{j+2})$ is orthogonal to both $\nabla f(\mathbf{y}_{j+1})$ and to $\mathbf{d}_P$, so that $\mathbf{y}'_{j+2}$ is a minimizing point in the plane of Figure 8.24. Using Part 3 of Theorem 8.8.3, argue now that $\mathbf{y}'_{j+2} = \mathbf{y}_{j+2}$.]

[8.39] Consider the quadratic form $f(\mathbf{x}) = \mathbf{c}^t\mathbf{x} + (1/2)\mathbf{x}^t\mathbf{H}\mathbf{x}$, where $\mathbf{H}$ is a symmetric $n \times n$ matrix. In many applications it is desirable to obtain separability in the variables by eliminating the cross-product terms. This could be done by rotating the axes as follows. Let $\mathbf{D}$ be an $n \times n$ matrix whose columns $\mathbf{d}_1, \ldots, \mathbf{d}_n$ are $\mathbf{H}$-conjugate. Letting $\mathbf{x} = \mathbf{D}\mathbf{y}$, verify that the quadratic form is

equivalent to $\sum_{j=1}^{n}\alpha_j y_j + (1/2)\sum_{j=1}^{n}\beta_j y_j^2$, where $(\alpha_1,...,\alpha_n) = \mathbf{c}^t\mathbf{D}$, and $\beta_j = \mathbf{d}_j^t\mathbf{H}\mathbf{d}_j$ for $j = 1,..., n$. Furthermore, translating and rotating the axes could be accomplished by the transformation $\mathbf{x} = \mathbf{D}\mathbf{y} + \mathbf{z}$, where z is any vector satisfying $\mathbf{H}\mathbf{z} + \mathbf{c} = \mathbf{0}$, that is, $\nabla f(\mathbf{z}) = \mathbf{0}$. In this case show that the quadratic form is equivalent to $[\mathbf{c}^t\mathbf{z} + (1/2)\mathbf{z}^t\mathbf{H}\mathbf{z}] + (1/2)\sum_{j=1}^{n}\beta_j y_j^2$. Use the result of this exercise to draw accurate contours of the quadratic form $3x_1 - 6x_2 + 2x_1^2 + x_1x_2 + 2x_2^2$.

[8.40] Consider the problem to maximize $-2x_1^2 - 3x_2^2 + 3x_1x_2 - 2x_1 + 4x_2$. Starting from the origin, solve the problem by the Davidon–Fletcher–Powell method, with $\mathbf{D}_1$ as the identity. Also solve the problem by the Fletcher and Reeves conjugate gradient method. Note that the two procedures generate identical sets of directions. Show that, in general, if $\mathbf{D}_1 = \mathbf{I}$, then the two methods are identical for quadratic functions.

[8.41] Derive a quasi-Newton correction matrix $\mathbf{C}$ for the Hessian approximation $\mathbf{B}_k$ that achieves the minimum Frobenius norm (squared) $\sum_i\sum_j C_{ij}^2$, where C_{ij} are the elements of $\mathbf{C}$ (to be determined), subject to the quasi-Newton condition $(\mathbf{C} + \mathbf{B}_k)\mathbf{p}_k = \mathbf{q}_k$ and the symmetry condition $\mathbf{C} = \mathbf{C}^t$. [*Hint*: Set up the corresponding optimization problem after enforcing symmetry, and use the KKT conditions. This gives the *Powell–Symmetric Broyden* (PSB) *update*.]

[8.42] Solve the problem to minimize $2x_1 + 3x_2^2 + e^{2x_1^2 + x_2^2}$, starting with the point (1, 0) and using both the Fletcher and Reeves conjugate gradient method and the BFGS quasi-Newton method.

[8.43] A problem of the following structure frequently arises in the context of solving a more general nonlinear programming problem:

$$\begin{aligned} &\text{Minimize} && f(\mathbf{x}) \\ &\text{subject to} && a_i \le x_i \le b_i \qquad \text{for } i = 1,..., m. \end{aligned}$$

a. Investigate appropriate modifications of the unconstrained optimization methods discussed in this chapter so that lower and upper bounds on the variables could be handled.

b. Use the results of Part a to solve the following problem:

$$\begin{aligned} &\text{Minimize} && (x_1 - 2)^4 + (x_1 - 2x_2)^2 \\ &\text{subject to} && 4 \le x_1 \le 6 \\ & && 3 \le x_2 \le 5. \end{aligned}$$

[8.44] Consider the system of simultaneous equations

$$h_i(\mathbf{x}) = 0 \qquad \text{for } i = 1,..., \ell.$$

a. Show how to solve the above system by unconstrained optimization techniques. [*Hint*: Consider the problem to minimize $\sum_{i=1}^{\ell} |h_i(\mathbf{x})|^p$, where p is a positive integer.]

b. Solve the following system:

$$2(x_1 - 2)^4 + (2x_1 - x_2)^2 - 4 = 0$$
$$x_1^2 - 2x_2 + 1 = 0.$$

[8.45] Consider the problem to minimize $f(\mathbf{x})$ subject to $h_i(\mathbf{x}) = 0$ for $i = 1,...,$ ℓ. A point x is said to be a KKT point if there exists a vector $\mathbf{v} \in R^{\ell}$ such that

$$\nabla f(\mathbf{x}) + \sum_{i=1}^{\ell} v_i \nabla h_i(\mathbf{x}) = \mathbf{0}$$
$$h_i(\mathbf{x}) = 0 \qquad \text{for } i = 1,..., \ell.$$

a. Show how to solve the above system using unconstrained optimization techniques. (*Hint*: See Exercise 8.44.)

b. Find the KKT point for the following problem:

$$\text{Minimize } (x_1 - 3)^4 + (x_1 - 3x_2)^2$$
$$\text{subject to } 2x_1^2 - x_2 = 0.$$

[8.46] Consider the problem to minimize $f(\mathbf{x})$ subject to $g_i(\mathbf{x}) \leq 0$ for $i = 1,..., m$.

a. Show that the KKT conditions are satisfied at a point x if there exist u_i and s_i for $i = 1,..., m$ such that

$$\nabla f(\mathbf{x}) + \sum_{i=1}^{m} u_i^2 \nabla g_i(\mathbf{x}) = \mathbf{0}$$
$$g_i(\mathbf{x}) + s_i^2 = 0 \qquad \text{for } i = 1,..., m$$
$$u_i s_i = 0 \qquad \text{for } i = 1,..., m.$$

b. Show that unconstrained optimization techniques could be used to find a solution to the above system. (*Hint*: See Exercise 8.44.)

c. Use a suitable unconstrained optimization technique to find a KKT point to the following problem:

$$\text{Minimize } 3x_1^2 + 2x_2^2 - 2x_1x_2 + 4x_1 + 6x_2$$
$$\text{subject to } -2x_1 - 3x_2 + 6 \leq 0.$$

[8.47] Consider the problem to minimize $x_1^2 + x_2^2$ subject to $x_1 + x_2 - 4 = 0$.

a. Find the optimal solution to this problem, and verify optimality by the KKT conditions.

b. One approach to solving the problem is to transform it into a problem of the form to minimize $x_1^2 + x_2^2 + \mu(x_1 + x_2 - 4)^2$, where $\mu > 0$ is a large scalar. Solve the unconstrained problem for $\mu = 10$ by a conjugate gradient method, starting from the origin.

[8.48] Using induction, show that the inclusion of the extra term $\gamma_j \mathbf{d}_1$ in Equation (8.68b), where γ_j is as given therein, ensures the mutual **H**-conjugacy of the directions $\mathbf{d}_1, \ldots, \mathbf{d}_n$ thus generated.

[8.49] Let **H** be an $n \times n$ symmetric matrix, and let $f(\mathbf{x}) = \mathbf{c}^t\mathbf{x} + (1/2)\mathbf{x}^t\mathbf{H}\mathbf{x}$. Consider the following *rank-one correction algorithm* for minimizing f. First, let $\mathbf{D}_1$ be an $n \times n$ positive definite symmetric matrix, and let $\mathbf{x}_1$ be a given vector. For $j = 1, \ldots, n$, let λ_j be an optimal solution to the problem to minimize $f(\mathbf{x}_j + \lambda \mathbf{d}_j)$ subject to $\lambda \in R$, and let $\mathbf{x}_{j+1} = \mathbf{x}_j + \lambda_j \mathbf{d}_j$, where $\mathbf{d}_j = -\mathbf{D}_j \nabla f(\mathbf{x}_j)$ and $\mathbf{D}_{j+1}$ is given by

$$\mathbf{D}_{j+1} = \mathbf{D}_j + \frac{(\mathbf{p}_j - \mathbf{D}_j\mathbf{q}_j)(\mathbf{p}_j - \mathbf{D}_j\mathbf{q}_j)^t}{\mathbf{q}_j^t(\mathbf{p}_j - \mathbf{D}_j\mathbf{q}_j)}$$

$$\mathbf{p}_j = \mathbf{x}_{j+1} - \mathbf{x}_j$$

$$\mathbf{q}_j = \mathbf{H}\mathbf{p}_j.$$

a. Verify that the matrix added to $\mathbf{D}_j$ to obtain $\mathbf{D}_{j+1}$ is of rank 1.

b. For $j = 1, \ldots, n$, show that $\mathbf{p}_i = \mathbf{D}_{j+1}\mathbf{q}_i$ for $i \leq j$.

c. Supposing that **H** is invertible, does $\mathbf{D}_{n+1} = \mathbf{H}^{-1}$ hold?

d. Even if $\mathbf{D}_j$ is positive definite, show that $\mathbf{D}_{j+1}$ is not necessarily positive definite. This explains why a line search over the entire real line is used.

e. Are the directions $\mathbf{d}_1, \ldots, \mathbf{d}_n$ necessarily conjugate?

f. Use the above algorithm for minimizing $x_1 - 4x_2 + 2x_1^2 + 2x_1x_2 + 3x_2^2$.

g. Suppose that $\mathbf{q}_j$ is replaced by $\nabla f(\mathbf{x}_{j+1}) - \nabla f(\mathbf{x}_j)$. Develop a procedure similar to that of Davidon–Fletcher–Powell for minimizing a nonquadratic function, using the above scheme for updating $\mathbf{D}_j$. Use the procedure to minimize $(x_1 - 2)^4 + (x_1 - 2x_2)^2$.

[8.50] Consider the design of a conjugate gradient method in which $\mathbf{d}_{j+1} = -\nabla f(\mathbf{y}_{j+1}) + \alpha_j \mathbf{d}_j$ in the usual notation, and where, for a choice of a scale parameter s_{j+1}, we would like $s_{j+1}\mathbf{d}_{j+1}$ to coincide with the Newton direction $-\mathbf{H}^{-1}\nabla f(\mathbf{y}_{j+1})$, if at all possible. Equating $s_{j+1}[-\nabla f(\mathbf{y}_{j+1}) + \alpha_j \mathbf{d}_j] = -\mathbf{H}^{-1}\nabla f(\mathbf{y}_{j+1})$, transpose both sides and multiply these by $\mathbf{H}\mathbf{d}_j$, and use the quasi-Newton condition $\lambda_j \mathbf{H}\mathbf{d}_j = \mathbf{q}_j$ to derive

$$\alpha_j = \frac{\nabla f(\mathbf{y}_{j+1})^t \mathbf{q}_j - (1/s_{j+1})\nabla f(\mathbf{y}_{j+1})^t \mathbf{p}_j}{\mathbf{d}_j^t \mathbf{q}_j}.$$

a. Show that with exact line searches, the choice of s_{j+1} is immaterial. Moreover, show that as $s_{j+1} \to \infty$, $\alpha_j \to \alpha_j^{\mathrm{HS}}$ in (8.57). Motivate the choice of s_{j+1} by considering the situation in which the Newton direction $-\mathbf{H}^{-1}\nabla f(\mathbf{y}_{j+1})$ is, indeed, contained in the cone spanned by $-\nabla f(\mathbf{y}_{j+1})$ and $\mathbf{d}_j$ but is not coincident with $\mathbf{d}_j$. Hence, suggest a scheme for choosing a value for s_{j+1}.

b. Illustrate, using Example 8.8.2, by assuming that at the previous iteration, $\mathbf{y}_j = (-1/2, 1)^t$, $\mathbf{d}_j = (1, 0)^t$, $\lambda_j = 1/2$ (inexact step), so that $\mathbf{y}_{j+1} = (0, 1)^t$, and consider your suggested choice along with the choices (i) $s_{j+1} = \infty$, (ii) $s_{j+1} = 1$, and (iii) $s_{j+1} = 1/4$ at the next iteration. Obtain the corresponding directions $\mathbf{d}_{j+1} = -\nabla f(\mathbf{y}_{j+1}) + \alpha_j \mathbf{d}_j$. Which of these can potentially lead to optimality? (Choice (ii) is Perry's [1978] choice. Sherali and Ulular [1990] suggest the scaled version, prescribing a choice for s_{j+1}.)

[8.51] In this exercise we describe a modification of the *simplex method* of Spendley et al. [1962] for solving a problem of the form to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in R^n$. The version of the method described here is credited to Nelder and Mead [1965].

Initialization Step Choose the points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n+1}$ to form a simplex in R^n. Choose a reflection coefficient $\alpha > 0$, an expansion coefficient $\gamma > 1$, and a positive contraction coefficient $0 < \beta < 1$. Go to the Main Step.

Main Step

1. Let $\mathbf{x}_r, \mathbf{x}_s \in \{\mathbf{x}_1, \ldots, \mathbf{x}_{n+1}\}$ be such that

$$f(\mathbf{x}_r) = \min_{1 \le j \le n+1} f(\mathbf{x}_j) \text{ and } f(\mathbf{x}_s) = \max_{1 \le j \le n+1} f(\mathbf{x}_j).$$

Let $\bar{\mathbf{x}} = \frac{1}{n} \sum_{j=1, j \ne s}^{n+1} \mathbf{x}_j$, and go to Step 2.

2. Let $\hat{\mathbf{x}} = \bar{\mathbf{x}} + \alpha(\bar{\mathbf{x}} - \mathbf{x}_s)$. If $f(\mathbf{x}_r) > f(\hat{\mathbf{x}})$, let $\mathbf{x}_e = \bar{\mathbf{x}} + \gamma(\hat{\mathbf{x}} - \bar{\mathbf{x}})$, and go to Step 3. Otherwise, go to Step 4.
3. The point $\mathbf{x}_s$ is replaced by $\mathbf{x}_e$ if $f(\hat{\mathbf{x}}) > f(\mathbf{x}_e)$ and by $\hat{\mathbf{x}}$ if $f(\hat{\mathbf{x}}) \le f(\mathbf{x}_e)$ to yield a new set of $n + 1$ points. Go to Step 1.
4. If $\max_{1 \le j \le n+1}\{f(\mathbf{x}_j): j \ne s\} \ge f(\hat{\mathbf{x}})$, then $\mathbf{x}_s$ is replaced by $\hat{\mathbf{x}}$ to form a new set of $n + 1$ points, and we go to Step 1. Otherwise, go to Step 5.
5. Let $\mathbf{x}'$ be defined by $f(\mathbf{x}') = \min\{f(\hat{\mathbf{x}}), f(\mathbf{x}_s)\}$, and let $\mathbf{x}'' = \bar{\mathbf{x}} + \beta(\mathbf{x}' - \bar{\mathbf{x}})$. If $f(\mathbf{x}'') > f(\mathbf{x}')$, replace $\mathbf{x}_j$ by $\mathbf{x}_j + (1/2)(\mathbf{x}_r - \mathbf{x}_j)$ for $j = 1,..., n + 1$, and go to Step 1. If $f(\mathbf{x}'') \le f(\mathbf{x}')$, then $\mathbf{x}''$ replaces $\mathbf{x}_s$ to form a new set of $n + 1$ points. Go to Step 1.

a. Let $\mathbf{d}_j$ be an n-vector with the jth component equal to a and all other components equal to b, where

$$a = \frac{c}{n\sqrt{2}}(\sqrt{n+1} + n - 1), \qquad b = \frac{c}{n\sqrt{2}}(\sqrt{n+1} - 1),$$

and where c is a positive scalar. Show that the initial simplex defined by $\mathbf{x}_1, ..., \mathbf{x}_{n+1}$ could be chosen by letting $\mathbf{x}_{j+1} = \mathbf{x}_1 + \mathbf{d}_j$, where $\mathbf{x}_1$ is selected arbitrarily. (In particular, show that $\mathbf{x}_{j+1} - \mathbf{x}_1$ for $j = 1,..., n$ are linearly independent. What is the interpretation of c in terms of the geometry of this initial simplex?)

b. Solve the problem to minimize $2x_1^2 + 2x_1x_2 + x_3^2 + 3x_2^2 - 3x_1 - 10x_3$ using the simplex method described in this exercise.

[8.52] Consider the quadratic function $f(\mathbf{y}) = \mathbf{c}'\mathbf{y} + (1/2)\mathbf{y}'\mathbf{H}\mathbf{y}$, where $\mathbf{H}$ is an $n \times n$ symmetric, positive definite matrix. Suppose that we use some algorithm for which the iterate $\mathbf{y}_{j+1} = \mathbf{y}_j - \lambda_j \mathbf{D}_j \nabla f(\mathbf{y}_j)$ is generated by an exact line search along the direction $-\mathbf{D}_j \nabla f(\mathbf{y}_j)$ from the previous iterate $\mathbf{y}_j$, where $\mathbf{D}_j$ is some positive definite matrix. Then, if $\mathbf{y}^*$ is the minimizing solution for f, and if $e(\mathbf{y}) = (1/2)(\mathbf{y} - \mathbf{y}^*)'\mathbf{H}(\mathbf{y} - \mathbf{y}^*)$ is an error function, show that at every step j, we have

$$e(\mathbf{y}_{j+1}) \le \frac{(\alpha_j - 1)^2}{(\alpha_j + 1)^2} e(\mathbf{y}_j),$$

where α_j is the ratio of the largest to the smallest eigenvalue of $\mathbf{D}_j\mathbf{H}$.

[8.53] Consider the following method of *parallel tangents* credited to Shah et al. [1964] for minimizing a differentiable function f of several variables:

Initialization Step Choose a termination scalar $\varepsilon > 0$, and choose a starting point $\mathbf{x}_1$. Let $\mathbf{y}_0 = \mathbf{x}_1$, $k = j = 1$, and go to the Main Step.

Main Step

1. Let $\mathbf{d} = -\nabla f(\mathbf{x}_k)$ and let $\hat{\lambda}$ be an optimal solution to the problem to minimize $f(\mathbf{x}_k + \lambda\mathbf{d})$ subject to $\lambda \geq 0$. Let $\mathbf{y}_1 = \mathbf{x}_k + \hat{\lambda}\mathbf{d}$. Go to Step 2.
2. Let $\mathbf{d} = -\nabla f(\mathbf{y}_j)$, and let λ_j be an optimal solution to the problem to minimize $f(\mathbf{y}_j + \lambda\mathbf{d})$ subject to $\lambda \geq 0$. Let $\mathbf{z}_j = \mathbf{y}_j + \lambda_j\mathbf{d}$, and go to Step 3.
3. Let $\mathbf{d} = \mathbf{z}_j - \mathbf{y}_{j-1}$, and let μ_j be an optimal solution to the problem to minimize $f(\mathbf{z}_j + \mu\mathbf{d})$ subject to $\mu \in R$. Let $\mathbf{y}_{j+1} = \mathbf{z}_j + \mu_j\mathbf{d}$. If $j < n$, replace j by $j + 1$, and go to Step 2. If $j = n$, go to Step 4.
4. Let $\mathbf{x}_{k+1} = \mathbf{y}_{n+1}$. If $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| < \varepsilon$, stop. Otherwise, let $\mathbf{y}_0 = \mathbf{x}_{k+1}$, replace k by $k + 1$, let $j = 1$, and go to Step 1.

Using Theorem 7.3.4, show that the method converges. Solve the following problems using the method of parallel tangents:

a. Minimize $2x_1^2 + 3x_2^2 + 2x_1x_2 - 2x_1 - 6x_2$.

b. Minimize $x_1^2 + x_2^2 - 2x_1x_2 - 2x_1 - x_2$. (Note that the optimal solution for this problem is unbounded.)

c. Minimize $(x_1 - 3)^2 + (x_1 - 3x_2)^2$.

[8.54] Let f: $R^n \rightarrow R$ be differentiable. Consider the following procedure for minimizing f:

Initialization Step Choose a termination scalar $\varepsilon > 0$ and an initial step size $\Delta > 0$. Let m be a positive integer denoting the number of allowable failures before reducing the step size. Let $\mathbf{x}_1$ be the starting point and let the current upper bound on the optimal objective value be $\text{UB} = f(\mathbf{x}_1)$. Let $v = 0$, let $k = 1$, and go to the Main Step.

Main Step

1. Let $\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$, and let $\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta\mathbf{d}_k$. If $f(\mathbf{x}_{k+1}) < \text{UB}$, let $v = 0$, $\hat{\mathbf{x}} = \mathbf{x}_{k+1}$, $\text{UB} = f(\hat{\mathbf{x}})$, and go to Step 2. If, on the other hand, $f(\mathbf{x}_{k+1}) \geq \text{UB}$, replace v by $v + 1$. If $v = m$, go to Step 3; and if $v < m$, go to Step 2.

2. Replace k by $k + 1$, and go to Step 1.
3. Replace k by $k + 1$. If $\Delta < \varepsilon$, stop with $\hat{\mathbf{x}}$ as an estimate of the optimal solution. Otherwise, replace Δ by $\Delta/2$, let $v = 0$, let $\mathbf{x}_k = \hat{\mathbf{x}}$, and go to Step 1.

a. Can you prove convergence of the above algorithm for $\varepsilon = 0$?

b. Apply the above algorithm for the three problems in Exercise 8.53.

[8.55] The method of Rosenbrock can be described by the map $\mathbf{A}$: $R^n \times U \times R^n \to R^n \times U \times R^n$. Here $U = \{\mathbf{D} : \mathbf{D}$ is an $n \times n$ matrix satisfying $\mathbf{D}^t\mathbf{D} = \mathbf{I}\}$. The algorithmic map $\mathbf{A}$ operates on the triple $(\mathbf{x}, \mathbf{D}, \lambda)$, where $\mathbf{x}$ is the current vector, $\mathbf{D}$ is the $n \times n$ matrix whose columns are the directions of the previous iteration, and λ is the vector whose components $\lambda_1, ..., \lambda_n$ give the distances moved along the directions $\mathbf{d}_1, ..., \mathbf{d}_n$. The map $\mathbf{A} = \mathbf{A}_3\mathbf{A}_2\mathbf{A}_1$ is a composite map whose components are discussed in detail below.

1. $\mathbf{A}_1$ is the point-to-point map defined by $\mathbf{A}_1(\mathbf{x}, \mathbf{D}, \lambda) = (\mathbf{x}, \bar{\mathbf{D}})$, where $\bar{\mathbf{D}}$ is the matrix whose columns are the new directions defined by (8.9).
2. The point-to-set map $\mathbf{A}_2$ is defined by $(\mathbf{x}, \mathbf{y}, \bar{\mathbf{D}}) \in \mathbf{A}_2(\mathbf{x}, \bar{\mathbf{D}})$ if minimizing f, starting from $\mathbf{x}$, in the directions $\bar{\mathbf{d}}_1, ... \bar{\mathbf{d}}_n$ leads to $\mathbf{y}$. By Theorem 7.3.5, the map $\mathbf{A}_2$ is closed.
3. $\mathbf{A}_3$ is the point-to-point map defined by $\mathbf{A}_3(\mathbf{x}, \mathbf{y}, \bar{\mathbf{D}}) = (\mathbf{y}, \bar{\mathbf{D}}, \bar{\lambda})$, where $\bar{\lambda} = (\bar{\mathbf{D}})^{-1}(\mathbf{y} - \mathbf{x})$.

a. Show that the map $\mathbf{A}_1$ is closed at $(\mathbf{x}, \mathbf{D}, \lambda)$ if $\lambda_j \neq 0$ for $j = 1, ..., n$.

b. Is the map $\mathbf{A}_1$ closed if $\lambda_j = 0$ for some j? (*Hint*: Consider the sequence $\mathbf{D}_k = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\lambda_k = \begin{bmatrix} 1/k \\ 1 \end{bmatrix}$.)

c. Show that $\mathbf{A}_3$ is closed.

d. Verify that the function f could be used as a descent function.

e. Discuss the applicability of Theorem 7.2.3 to prove convergence of Rosenbrock's procedure. (This exercise illustrates that some difficulties could arise in viewing the algorithmic map as a composition of several maps. In Section 8.5 a proof of convergence was provided without decomposing the map $\mathbf{A}$.)

[8.56] Consider the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in R^n$, and consider the following algorithm credited to Powell [1964] (and modified by Zangwill [1967b] as in Part c).

Initialization Step Choose a termination scalar $\varepsilon > 0$. Choose an initial point $\mathbf{x}_1$, let $\mathbf{d}_1, ..., \mathbf{d}_n$ be the coordinate directions, and let $k = j = i = 1$. Let $\mathbf{z}_1 = \mathbf{y}_1 = \mathbf{x}_1$, and go to the Main Step.

Main Step

1. Let λ_i be an optimal solution to the problem to minimize $f(\mathbf{z}_i + \lambda \mathbf{d}_i)$ subject to $\lambda \in R$, and let $\mathbf{z}_{i+1} = \mathbf{z}_i + \lambda_i \mathbf{d}_i$. If $i < n$, replace i by $i + 1$, and repeat Step 1. Otherwise, go to Step 2.
2. Let $\mathbf{d} = \mathbf{z}_{n+1} - \mathbf{z}_1$, and let $\hat{\lambda}$ be an optimal solution to the problem to minimize $f(\mathbf{z}_{n+1} + \lambda \mathbf{d})$ subject to $\lambda \in R$. Let $\mathbf{y}_{j+1} = \mathbf{z}_{n+1} + \hat{\lambda}\mathbf{d}$. If $j < n$, replace $\mathbf{d}_\ell$ by $\mathbf{d}_\ell = \mathbf{d}_{\ell+1}$ for $\ell = 1, \ldots, n-1$, let $\mathbf{d}_n = \mathbf{d}$, let $\mathbf{z}_1 = \mathbf{y}_{j+1}$, let $i = 1$, replace j by $j + 1$, and go to Step 1. Otherwise, $j = n$, and go to Step 3.
3. Let $\mathbf{x}_{k+1} = \mathbf{y}_{n+1}$. If $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| < \varepsilon$, stop. Otherwise, let $i = j = 1$, let $\mathbf{z}_1 = \mathbf{y}_1 = \mathbf{x}_{k+1}$, replace k by $k + 1$, and go to Step 1.

a. Suppose that $f(\mathbf{x}) = \mathbf{c}^t\mathbf{x} + (1/2)\mathbf{x}^t\mathbf{H}\mathbf{x}$, where $\mathbf{H}$ is an $n \times n$ symmetric matrix. After one pass through the main step, show that if $\mathbf{d}_1, \ldots, \mathbf{d}_n$ are linearly independent, then they are also $\mathbf{H}$-conjugate, so that by Theorem 8.8.3, an optimal solution is produced in one iteration.

b. Consider the following problem credited to Zangwill [1967b]:

$$\text{Minimize } (x_1 - x_2 + x_3)^2 + (-x_1 + x_2 + x_3)^2 + (x_1 + x_2 - x_3)^2.$$

Apply Powell's method discussed in this exercise, starting from the point (1/2, 1, 1/2). Note that the procedure generates a set of dependent directions and hence will not yield the optimal point (0, 0, 0).

c. Zangwill [1967b] proposed a slight modification of Powell's method to guarantee linear independence of the direction vectors. In particular, in Step 2, the point $\mathbf{z}_1$ is obtained from $\mathbf{y}_{j+1}$ by a spacer step application, such as one iteration of the cyclic coordinate method. Show that this modification indeed guarantees linear independence, and hence, by Part a, finite convergence for a quadratic function is assured.

d. Apply Zangwill's modified method to solve the problem of Part b.

e. If the function is not quadratic, consider the introduction of a spacer step so that in Step 3, $\mathbf{z}_1 = \mathbf{y}_1$ is obtained by the application of one iteration of the cyclic coordinate method starting from $\mathbf{x}_{k+1}$. Use Theorem 7.3.4 to prove convergence.

[8.57] Solve the Lagrangian dual problem of Example 6.4.1 using the subgradient algorithm. Resolve using the deflected subgradient strategy suggested in Section 8.9.

[8.58] Consider the problem of finding $\bar{\mathbf{x}} = P_G(\mathbf{x})$, where $G = \{\mathbf{y} : \xi_j^t\mathbf{y} \le \beta_j, \text{ for } j = 1, 2\}$.

a. Formulate this as a linearly constrained quadratic optimization problem and write the KKT conditions for this problem. Explain why these KKT

conditions are both necessary and sufficient for optimality for this problem.

b. Prescribe a closed-form solution to these conditions, enumerating cases as necessary. Illustrate geometrically each such case identified.

c. Identify the above analysis with the main computation in the Polyak–Kelly cutting plane algorithm as embodied by Equations (8.80) and (8.81).

[**8.59**] Solve the example of Exercise 6.30 using the subgradient optimization algorithm starting with the point (0, 4). Re-solve using the deflected subgradient strategy suggested in Section 8.9.

[**8.60**] Consider the problem of finding the projection $\mathbf{x}^* = P_X(\bar{\mathbf{x}})$ of the point $\bar{\mathbf{x}}$ onto $X = \{\mathbf{x} : \alpha^t\mathbf{x} = \beta, \ell \le \mathbf{x} \le \mathrm{u}\}$, where $\mathbf{x}, \bar{\mathbf{x}}, \mathbf{x}^* \in R^n$. The following *variable dimension* algorithm projects the current point successively onto the equality constraint and reduces the problem to an equivalent one in a lower-dimensional space, or else stops. Justify the various steps of this algorithm. Illustrate by projecting the point $(-2, 3, 1, 2)^t$ onto $\{\mathbf{x} : x_1 + x_2 + x_3 + x_4 = 1, 0 \le x_i \le 1 \text{ for all } i\}$. (This method is a generalization of the procedures that appear in Bitran and Hax [1976] and in Sherali and Shetty [1980b].)

Initialization Set $(\bar{\mathbf{x}}^0, I^0, \ell^0, \mathrm{u}^0, \beta^0) = (\bar{\mathbf{x}}, I, \ell, \mathrm{u}, \beta)$, where $I = \{i : \alpha_i \neq 0\}$. For $i \notin I$, put $x_i^* = \bar{x}_i$ if $\ell_i \le \bar{x}_i \le u_i$, $x_i^* = \ell_i$, if $\bar{x}_i < \ell_i$ and $x_i^* = u_i$ if $\bar{x}_i > u_i$. Let $k = 0$.

Step 1 Compute the projection $\hat{\mathbf{x}}^k$ of $\bar{\mathbf{x}}^k$ onto the equality constraint in the subspace I^k according to

$$\hat{x}_i^k = \bar{x}_i^k + \frac{\beta^k - \sum_{i \in I^k} \alpha_i \bar{x}_i^k}{\sum_{i \in I^k} \alpha_i^2} \alpha_i \qquad \text{for each } i \in I^k.$$

If $\ell_i^k \le \hat{x}_i^k \le u_i^k$ for all $i \in I^k$, put $x_i^* = \hat{x}_i^k$ for all $i \in I^k$, and stop. Otherwise, proceed to Step 2.

Step 2 Define $J_1 = \{i \in I^k : \hat{x}_i^k \le \ell_i^k\}$, $J_2 = \{i \in I^k : \hat{x}_i^k \ge u_i^k\}$, and compute

$$\gamma = \beta^k + \sum_{i \in J_1} \alpha_i(\ell_i^k - \hat{x}_i^k) + \sum_{i \in J_2} \alpha_i(u_i^k - \hat{x}_i^k).$$

If $\gamma = \beta^k$, then put $x_i^* = \ell_i^k$ for $i \in J_1$, $x_i^* = u_i^k$ for $i \in J_2$, and $x_i^* = \hat{x}_i^k$ for $i \in I^k - J_1 \cup J_2$, and stop. Otherwise, define

$$J_3 = \{i \in J_1 : \alpha_i > 0\} \quad \text{and} \quad J_4 = \{i \in J_2 : \alpha_i < 0\} \quad \text{if } \gamma > \beta^k$$

$$J_3 = \{i \in J_1 : \alpha_i < 0\} \quad \text{and} \quad J_4 = \{i \in J_2 : \alpha_i > 0\} \quad \text{if } \gamma < \beta^k.$$

Set $x_i^* = \ell_i^k$ if $i \in J_3$, and $x_i^* = u_i^k$ if $i \in J_4$. (*Note*: $J_3 \cup J_4 \neq \emptyset$.) Update $I^{k+1} = I^k - J_3 \cup J_4$. If $I^{k+1} = \emptyset$, then stop. Otherwise, update ($\bar{x}_i^{k+1} = \hat{x}_i^k$ for $i \in I^{k+1}$), ($\ell_i^{k+1} = \max\{\ell_i^k, \hat{x}_i^k\}$ if $\alpha_i(\beta^k - \gamma) > 0$, and $\ell_i^{k+1} = \ell_i^k$ otherwise, for $i \in I^{k+1}$), ($u_i^{k+1} = \min\{u_i^k, \hat{x}_i^k\}$ if $\alpha_i(\beta^k - \gamma) < 0$, and $u_i^{k+1} = u_i^k$ otherwise, for $i \in I^{k+1}$),. and $\beta^{k+1} = \beta^k - \sum_{i \in J_3} \alpha_i \ell_i^k - \sum_{i \in J_4} \alpha_i u_i^k$. Increment k by 1 and go to Step 1.

Notes and References

We have discussed several iterative procedures for solving an unconstrained optimization problem. Most of the procedures involve a line search of the type described in Sections 8.1 through 8.3 and, by and large, the effectiveness of the search direction and the efficiency of the line search method greatly affect the overall performance of the solution technique. The Fibonacci search procedure discussed in Section 8.1 is credited to Kiefer [1953]. Several other search procedures, including the golden section method, are discussed in Wilde [1964] and Wilde and Beightler [1967]. These references also show that the Fibonacci search procedure is the best for unimodal functions in that it reduces the maximum interval of uncertainty with the least number of observations.

Another class of procedures uses curve fitting, as discussed in Section 8.3 and illustrated by Exercises 8.11 through 8.13. If a function f of one variable is to be minimized, the procedures involve finding an approximating quadratic or cubic function q. In the quadratic case, the function is selected such that given three points λ_1, λ_2, and λ_3, the functional values of f and q are equal at these points. In the cubic case, given two points λ_1 and λ_2, q is selected such that the functional values and derivatives of both functions are the same at these points. In any case, the minimum of q is determined, and this point replaces one of the initial points. Refer to Davidon [1959], Fletcher and Powell [1963], Kowalik and Osborne [1968], Luenberger [1973a/1984], Pierre [1969], Powell [1964], and Swann [1964] for more detailed discussions, particularly on precautions to be taken to ensure convergence. Some limited computational studies on the efficiency of this approach may be found in Himmelblau [1972b] and Murtagh and Sargent [1970]. See Armijo [1966] and Luenberger [1973a/1984] for further discussions on inexact line searches.

Among the gradient-free methods, the method of Rosenbrock [1960], discussed in Section 8.4, and the method of Zangwill [1967b], discussed in Exercises 8.30 and 8.56, are generally considered quite efficient. As originally proposed, the Rosenbrock method and the procedure of Hooke and Jeeves [1961] do not use line searches but employ instead discrete steps along the search directions. Incorporating a line search within Rosenbrock's procedure

Appendix A
Mathematical Review

In this appendix we review notation, basic definitions, and results related to vectors, matrices, and real analysis that are used throughout the text. For more details, see Bartle [1976], Berge [1963], Berge and Ghoulia-Houri [1965], Buck [1965], Cullen [1972], Flet [1966], and Rudin [1964].

A.1 Vectors and Matrices

Vectors

An n-vector $\mathbf{x}$ is an array of n scalars $x_1, x_2, ..., x_n$. Here x_j is called the jth *component*, or *element*, of the vector $\mathbf{x}$. The notation $\mathbf{x}$ represents a *column vector*, whereas the notation $\mathbf{x}^t$ represents the transposed *row vector*. Vectors are denoted by lowercase boldface letters, such as $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{x}$, and $\mathbf{y}$. The collection of all n-vectors forms the *n-dimensional Euclidean space*, which is denoted by R^n.

Special Vectors

The *zero vector*, denoted by 0, is a vector consisting entirely of zeros. The *sum vector* is denoted by $\mathbf{1}$ or $\mathbf{e}$ and has each component equal to 1. The ith *coordinate vector*, also referred to as the ith *unit vector*, is denoted by $\mathbf{e}_i$ and consists of zeros except for a 1 at the ith position.

Vector Addition and Multiplication by a Scalar

Let $\mathbf{x}$ and $\mathbf{y}$ be two n-vectors. The *sum* of $\mathbf{x}$ and $\mathbf{y}$ is written as the vector $\mathbf{x} + \mathbf{y}$. The jth component of the vector $\mathbf{x} + \mathbf{y}$ is $x_j + y_j$. The *product* of a vector $\mathbf{x}$ and a scalar α is denoted by $\alpha\mathbf{x}$ and is obtained by multiplying each element of $\mathbf{x}$ by α.

Linear and Affine Independence

A collection of vectors $\mathbf{x}_1, ..., \mathbf{x}_k$ in R^n is considered *linearly independent* if $\sum_{j=1}^{k} \lambda_j \mathbf{x}_j = 0$ implies that $\lambda_j = 0$ for all $j = 1, ..., k$. A collection of vectors $\mathbf{x}_0$, $\mathbf{x}_1, ..., \mathbf{x}_k$ in R^n is considered to be *affinely independent* if $(\mathbf{x}_1 - \mathbf{x}_0), ..., (\mathbf{x}_k - \mathbf{x}_0)$ are linearly independent.

Linear, Affine, and Convex Combinations and Hulls

A vector y in R^n is said to be a *linear combination* of the vectors $\mathbf{x}_1,...,\mathbf{x}_k$ in R^n if y can be written as $\mathbf{y} = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j$ for some scalars $\lambda_1,...,\lambda_k$. If, in addition, $\lambda_1,...,\lambda_k$ are restricted to satisfy $\sum_{j=1}^{k} \lambda_j = 1$, then y is said to be an *affine combination* of $\mathbf{x}_1,...,\mathbf{x}_k$. Furthermore, if we also restrict $\lambda_1,...,\lambda_k$ to be nonnegative, then this is known as a *convex combination* of $\mathbf{x}_1,...,\mathbf{x}_k$. The *linear*, *affine*, or *convex hull* of a set $S \subseteq R^n$ is, respectively, the set of all linear, affine, or convex combinations of points within S.

Spanning Vectors

A collection of vectors $\mathbf{x}_1,...,\mathbf{x}_k$ in R^n, where $k \geq n$, is said to *span* R^n if any vector in R^n can be represented as a linear combination of $\mathbf{x}_1,...,\mathbf{x}_k$. The *cone spanned* by a collection of vectors $\mathbf{x}_1,...,\mathbf{x}_k$, for any $k \geq 1$, is the set of nonnegative linear combinations of these vectors.

Basis

A collection of vectors $\mathbf{x}_1,...,\mathbf{x}_k$ in R^n is called a *basis* of R^n if it spans R^n and if the deletion of any of the vectors prevents the remaining vectors from spanning R^n. It can be shown that $\mathbf{x}_1,...,\mathbf{x}_k$ form a basis of R^n if and only if $\mathbf{x}_1,...,\mathbf{x}_k$ are linearly independent and if, in addition, $k = n$.

Inner Product

The *inner product* of two vectors x and y in R^n is defined by $\mathbf{x}^t\mathbf{y} = \sum_{j=1}^{n} x_j y_j$. If the inner product of two vectors is equal to zero, then the two vectors are said to be *orthogonal.*

Norm of a Vector

The *norm* of a vector x in R^n is denoted by $\|\mathbf{x}\|$ and defined by $\|\mathbf{x}\| = (\mathbf{x}^t\mathbf{x})^{1/2} = (\sum_{j=1}^{n} x_j^2)^{1/2}$. This is also referred to as the ℓ_2 norm, or *Euclidean norm.*

Schwartz Inequality

Let x and y be two vectors in R^n, and let $|\mathbf{x}^t\mathbf{y}|$ denote the absolute value of $\mathbf{x}^t\mathbf{y}$. Then the following inequality, referred to as the *Schwartz inequality*, holds true:

$$|\mathbf{x}^t\mathbf{y}| \leq \|\mathbf{x}\|\|\mathbf{y}\|.$$

Matrices

A *matrix* is a rectangular array of numbers. If the matrix has m rows and n columns, it is called an $m \times n$ *matrix*. Matrices are denoted by boldface capital letters, such as **A**, **B**, and **C**. The entry in row i and column j of a matrix **A** is denoted by a_{ij}, its ith row is denoted by $\mathbf{A}_i$, and its jth column is denoted by $\mathbf{a}_j$.

Special Matrices

An $m \times n$ matrix whose elements are all equal to zero is called a *zero matrix* and is denoted by **0**. A square $n \times n$ matrix is called the *identity matrix* if $a_{ij} = 0$ for $i \neq j$ and $a_{ii} = 1$ for $i = 1,..., n$. The $n \times n$ identity matrix is denoted by **I** and sometimes by $\mathbf{I}_n$ to highlight its dimension. An $n \times n$ *permutation matrix* **P** is one that has the same rows of $\mathbf{I}_n$ but which are permuted in some order. An *orthogonal matrix* **Q** having dimension $m \times n$ is one that satisfies $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_n$ or $\mathbf{Q}\mathbf{Q}' = \mathbf{I}_m$. In particular, if **Q** is square, $\mathbf{Q}^{-1} = \mathbf{Q}'$. Note that a permutation matrix **P** is an orthogonal square matrix.

Addition of Matrices and Scalar Multiplication of a Matrix

Let **A** and **B** be two $m \times n$ matrices. The *sum* of **A** and **B**, denoted by **A** + **B**, is the matrix whose (i, j)th entry is $a_{ij} + b_{ij}$. The *product* of a matrix **A** by a scalar α is the matrix whose (i, j)th entry is αa_{ij}.

Matrix Multiplication

Let **A** be an $m \times n$ matrix and **B** be an $n \times p$ matrix. Then the *product* **AB** is defined to be the $m \times p$ matrix **C** whose (i, j)th entry c_{ij} is given by

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \quad \text{for } i = 1,..., m, \text{ and } j = 1,..., p.$$

Transposition

Let **A** be an $m \times n$ matrix. The *transpose* of **A**, denoted by $\mathbf{A}'$, is the $n \times m$ matrix whose (i, j)th entry is equal to a_{ji}. A square matrix **A** is said to be *symmetric* if $\mathbf{A} = \mathbf{A}'$. It is said to be *skew symmetric* if $\mathbf{A}' = -\mathbf{A}$.

Partitioned Matrices

A matrix can be partitioned into submatrices. For example, the $m \times n$ matrix **A** could be partitioned as follows:

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} \end{array}\right],$$

where $\mathbf{A}_{11}$ is $m_1 \times n_1$, $\mathbf{A}_{12}$ is $m_1 \times n_2$, $\mathbf{A}_{21}$ is $m_2 \times n_1$, $\mathbf{A}_{22}$ is $m_2 \times n_2$, $m = m_1 + m_2$, and $n = n_1 + n_2$.

Determinant of a Matrix

Let $\mathbf{A}$ be an $n \times n$ matrix. The *determinant* of $\mathbf{A}$, denoted by $\det[\mathbf{A}]$, is defined iteratively as follows:

$$\det[\mathbf{A}] = \sum_{i=1}^{n} a_{i1} \det[\mathbf{A}_{i1}].$$

Here $\mathbf{A}_{i1}$ is the *cofactor* of a_{i1}, defined as $(-1)^{i+1}$ times the submatrix of $\mathbf{A}$ formed by deleting the ith row and the first column, and the determinant of any scalar is the scalar itself. Similar to the use of the first column above, the determinant can be expressed in terms of any row or column.

Inverse of a Matrix

A square matrix $\mathbf{A}$ is said to be *nonsingular* if there is a matrix $\mathbf{A}^{-1}$, called the *inverse matrix*, such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. The inverse of a square matrix, if it exists, is unique. Furthermore, a square matrix has an inverse if and only if its determinant is not equal to zero.

Rank of a Matrix

Let $\mathbf{A}$ be an $m \times n$ matrix. The *rank* of $\mathbf{A}$ is the maximum number of linearly independent rows or, equivalently, the maximum number of linearly independent columns of the matrix $\mathbf{A}$. If the rank of $\mathbf{A}$ is equal to $\min\{m, n\}$, $\mathbf{A}$ is said to have *full rank*.

Norm of a Matrix

Let $\mathbf{A}$ be an $n \times n$ matrix. Most commonly, the *norm* of $\mathbf{A}$, denoted by $\|\mathbf{A}\|$, is defined by

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|$$

where $\|\mathbf{A}\mathbf{x}\|$ and $\|\mathbf{x}\|$ are the usual Euclidean (ℓ_2) norms of the corresponding vectors. Hence, for any vector $\mathbf{z}$, $\|\mathbf{A}\mathbf{z}\| \leq \|\mathbf{A}\| \, \|\mathbf{z}\|$. A similar use of an ℓ_p norm $\|\cdot\|_p$ induces a corresponding matrix norm $\|\mathbf{A}\|_p$. In particular, the above matrix norm, sometimes denoted $\|\mathbf{A}\|_2$, is equal to the [maximum eigenvalue of $\mathbf{A}'\mathbf{A}$]$^{1/2}$. Also, the *Frobenius norm* of $\mathbf{A}$ is given by

$$\|\mathbf{A}\|_F = \left[\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2\right]^{1/2}$$

and is simply the ℓ_2 norm of the vector whose elements are all the elements of **A**.

Eigenvalues and Eigenvectors

Let **A** be an $n \times n$ matrix. A scalar λ and a nonzero vector **x** satisfying the equation $\mathbf{Ax} = \lambda\mathbf{x}$ are called, respectively, an *eigenvalue* and an *eigenvector* of **A**. To compute the eigenvalues of **A**, we solve the equation $\det[\mathbf{A} - \lambda\mathbf{I}] = 0$. This yields a polynomial equation in λ that can be solved for the eigenvalues of **A**. If **A** is symmetric, then it has n (possibly nondistinct) eigenvalues. The eigenvectors associated with distinct eigenvalues are necessarily orthogonal, and for any collection of some p coincident eigenvalues, there exists a collection of p orthogonal eigenvectors. Hence, given a symmetric matrix **A**, we can construct an orthogonal basis **B** for R^n, that is, a basis having orthogonal column vectors, each representing an eigenvector of **A**. Furthermore, let us assume that each column of **B** has been normalized to have a unit norm. Hence, $\mathbf{B}^t\mathbf{B} = \mathbf{I}$, so that $\mathbf{B}^{-1} = \mathbf{B}^t$. Such a matrix is said to be an *orthogonal matrix* or an *orthonormal matrix*.

Now, consider the (pure) *quadratic form* $\mathbf{x}^t\mathbf{Ax}$, where **A** is an $n \times n$ symmetric matrix. Let $\lambda_1,\ldots,\lambda_n$ be the eigenvalues of **A**, let $\mathbf{\Lambda} = \text{diag}\{\lambda_1,\ldots,\lambda_n\}$ be a *diagonal matrix* comprised of diagonal elements $\lambda_1,\ldots,\lambda_n$ and zeros elsewhere, and let **B** be the orthogonal eigenvector matrix comprised of the orthogonal, normalized eigenvectors $\mathbf{b}_1,\ldots,\mathbf{b}_n$ as its columns. Define the linear transformation $\mathbf{x} = \mathbf{By}$ that writes any vector **x** in terms of the eigenvectors of **A**. Under this transformation, the given quadratic form becomes

$$\mathbf{x}^t\mathbf{Ax} = \mathbf{y}^t\mathbf{B}^t\mathbf{ABy} = \mathbf{y}^t\mathbf{B}^t\mathbf{\Lambda By} = \mathbf{y}^t\mathbf{\Lambda y} = \sum_{i=1}^{n}\lambda_i y_i^2.$$

This is called a *diagonalization process*.

Observe also that we have $\mathbf{AB} = \mathbf{B\Lambda}$, so that because **B** is orthogonal, we get $\mathbf{A} = \mathbf{B\Lambda B}^t = \sum_{i=1}^{n}\lambda_i\mathbf{b}_i\mathbf{b}_i^t$. This representation is called the *spectral decomposition* of **A**. For an $m \times n$ matrix **A**, a related factorization $\mathbf{A} = \mathbf{U\Sigma V}^t$, where **U** is an $m \times m$ orthogonal matrix, **V** is an $n \times n$ orthogonal matrix, and **Σ** is an $m \times n$ matrix having elements $\Sigma_{ij} = 0$ for $i \neq j$, and $\Sigma_{ij} \geq 0$ for $i = j$, is known as a *singular-value decomposition* (SVD) of **A**. Here, the columns of **U** and **V** are normalized eigenvectors of $\mathbf{AA}^t$ and $\mathbf{A}^t\mathbf{A}$, respectively. The Σ_{ij} values are the (absolute) square roots of the eigenvalues of $\mathbf{AA}^t$ if $m \leq n$ or of $\mathbf{A}^t\mathbf{A}$ if $m \geq n$. The number of nonzero Σ_{ij} values equals the rank of **A**.

Definite and Semidefinite Matrices

Let **A** be an $n \times n$ symmetric matrix. Here **A** is said to be *positive definite* if $\mathbf{x}^t\mathbf{A}\mathbf{x} > 0$ for all nonzero **x** in R^n and is said to be *positive semidefinite* if $\mathbf{x}^t\mathbf{A}\mathbf{x} \geq 0$ for all **x** in R^n. Similarly, if $\mathbf{x}^t\mathbf{A}\mathbf{x} < 0$ for all nonzero **x** in R^n, then **A** is called *negative definite*; and if $\mathbf{x}^t\mathbf{A}\mathbf{x} \leq 0$ for all **x** in R^n, then **A** is called *negative semidefinite*. A matrix that is neither positive semidefinite nor negative semidefinite is called *indefinite*. By the foregoing diagonalization process, the matrix **A** is positive definite, positive semidefinite, negative definite, and negative semidefinite if and only if its eigenvalues are positive, nonnegative, negative, and nonpositive, respectively. (Note that the superdiagonalization algorithm discussed in Chapter 3 is a more efficient method for ascertaining definiteness properties.) Also, by the definition of $\mathbf{\Lambda}$ and **B** above, if **A** is positive definite, then its *square root* $\mathbf{A}^{1/2}$ is the matrix satisfying $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$ and is given by $\mathbf{A}^{1/2} = \mathbf{B}\mathbf{\Lambda}^{1/2}\mathbf{B}^t$.

A.2 Matrix Factorizations

Let **B** be a nonsingular $n \times n$ matrix, and consider the system of equations $\mathbf{B}\mathbf{x} = \mathbf{b}$. The solution given by $\mathbf{x} = \mathbf{B}^{-1}\mathbf{b}$ is seldom computed by finding the inverse $\mathbf{B}^{-1}$ directly. Instead, a factorization or decomposition of **B** into multiplicative components is usually employed whereby $\mathbf{B}\mathbf{x} = \mathbf{b}$ is solved in a numerically stable fashion, often through the solution of triangular systems via back-substitution. This becomes particularly relevant in ill-conditioned situations when **B** is nearly singular or when we wish to verify positive definiteness of **B** as in quasi-Newton or Levenberg-Marquardt methods. Several useful factorizations are discussed below. For more details, including schemes for updating such factors in the context of iterative methods, we refer the reader to Bartels et al. [1970], Bazaraa et al. [2005], Dennis and Schnabel [1983], Dongarra et al. [1979], Gill et al. [1974, 1976], Golub and Van Loan [1983/1989], Murty [1983], and Stewart [1973], along with the many accompanying references cited therein. Standard software such as LINPACK, MATLAB, and the Harwell Library routines are also available to perform these factorizations efficiently.

LU and PLU Factorization for a Basis B

In the LU factorization, we reduce **B** to an upper triangular form **U** through a series of permutations and Gaussian pivot operations. At the ith stage of this process, having reduced **B** to $\mathbf{B}^{(i-1)}$, say, which is upper triangular in columns $1,\ldots, i-1$ (where $\mathbf{B}^0 \equiv \mathbf{B}$), we first premultiply $\mathbf{B}^{(i-1)}$ by a *permutation matrix* $\mathbf{P}_i$ to exchange row i with that row in $\{i, i+1,\ldots, n\}$ of $\mathbf{B}^{(i-1)}$ that has the largest absolute-valued element in column i. This is done to ensure that the (i, i)th element of $\mathbf{P}_i\mathbf{B}^{(i-1)}$ is significantly nonzero. Using this as a pivot element, we

perform row operations to zero out the elements in rows $i + 1,..., n$ of column i. This *triangularization* can be represented as a premultiplication with a suitable *Gaussian pivot matrix* $\mathbf{G}_i$, which is a *unit lower triangular matrix*, having ones on the diagonal and suitable possibly nonzero elements in rows $i + 1,..., n$ of column i. This gives $\mathbf{B}^{(i)} = (\mathbf{G}_i\mathbf{P}_i)\mathbf{B}^{(i-1)}$. Hence, we get, after some $r \leq (n - 1)$ such operations,

$$(\mathbf{G}_r\mathbf{P}_r)\cdots(\mathbf{G}_2\mathbf{P}_2)(\mathbf{G}_1\mathbf{P}_1)\mathbf{B} = \mathbf{U}. \tag{A.1}$$

The system $\mathbf{Bx} = \mathbf{b}$ can now be solved by computing $\bar{\mathbf{b}} = (\mathbf{G}_r\mathbf{P}_r)\cdots(\mathbf{G}_1\mathbf{P}_1)\mathbf{b}$ and then solving the triangular system $\mathbf{Ux} = \bar{\mathbf{b}}$ by back-substitution. If no permutations are performed, $\mathbf{G}_r\cdots\mathbf{G}_1$ is lower triangular, and denoting its (lower triangular) inverse as $\mathbf{L}$, we have the factored form $\mathbf{B} = \mathbf{LU}$ for $\mathbf{B}$, hence its name. Also, if $\mathbf{P}^t$ is a permutation matrix that is used to *a priori* rearrange the rows of $\mathbf{B}$ and we then apply the Gaussian triangularization operation to derive $\mathbf{L}^{-1}\mathbf{P}^t\mathbf{B} = \mathbf{U}$, we can write $\mathbf{B} = (\mathbf{P}^t)^{-1}\mathbf{LU} = \mathbf{PLU}$, noting that $\mathbf{P}^t = \mathbf{P}^{-1}$. Hence, this factorization is sometimes called a *PLU decomposition*. If $\mathbf{B}$ is sparse, $\mathbf{P}^t$ can be used to make $\mathbf{P}^t\mathbf{B}$ nearly upper triangular (assuming that the columns of $\mathbf{B}$ have been appropriately permuted) and then only a few and sparse Gaussian pivot operations will be required to obtain $\mathbf{U}$. This method is therefore very well suited for sparse matrices.

QR and *QRP* Factorization for a Basis B

This factorization is most suitable and is used frequently for solving *dense equation systems*. Here the matrix $\mathbf{B}$ is reduced to an upper triangular form $\mathbf{R}$ by premultiplying it with a sequence of square, symmetric orthogonal matrices $\mathbf{Q}_i$. Given $\mathbf{B}^{(i-1)} \equiv \mathbf{Q}_{i-1}\cdots\mathbf{Q}_1\mathbf{B}$ that is upper triangular in columns $1,..., i - 1$ (where $\mathbf{B}^{(0)} = \mathbf{B}$), we construct a matrix $\mathbf{Q}_i$ so that $\mathbf{Q}_i\mathbf{B}^{(i-1)} = \mathbf{B}^{(i)}$ is upper triangular in column i as well, while columns $1,..., i - 1$ remain unaffected. The matrix $\mathbf{Q}_i$ is a square, symmetric orthogonal matrix of the form $\mathbf{Q}_i \equiv \mathbf{I} - \gamma_i\mathbf{q}_i\mathbf{q}_i^t$, where $\mathbf{q}_i = (0,...,0, q_{ii},..., q_{ni})^t$ and $\gamma_i \in R^1$ are suitably chosen to perform the foregoing operation. Such a matrix $\mathbf{Q}_i$ is called a *Householder transformation matrix*. If the elements in rows $i,..., n$ of column i of $\mathbf{B}^{(i-1)}$ are denoted by $(\alpha_i,...,\alpha_n)^t$, then we have $q_{ii} = \alpha_i + \theta_i$, $q_{ji} = \alpha_j$ for $j = i + 1,..., n$, $\gamma_i = 1/\theta_i q_{ii}$, where $\theta_i = \text{sign}(\alpha_i)[\alpha_i^2 + \cdots + \alpha_n^2]^{1/2}$, and where $\text{sign}(\alpha_i) = 1$ if $\alpha_i > 0$ and -1 otherwise. Defining $\mathbf{Q} = \mathbf{Q}_{n-1}\cdots\mathbf{Q}_1$, we see that $\mathbf{Q}$ is also a symmetric orthogonal matrix and that $\mathbf{QB} = \mathbf{R}$, or that $\mathbf{B} = \mathbf{QR}$, since $\mathbf{Q} = \mathbf{Q}^t = \mathbf{Q}^{-1}$; that is, $\mathbf{Q}$ is an *involutory matrix*.

Now, to solve $\mathbf{Bx} = \mathbf{b}$, we equivalently solve $\mathbf{QRx} = \mathbf{b}$ or $\mathbf{Rx} = \mathbf{Qb}$ by finding $\bar{\mathbf{b}} = \mathbf{Qb}$ first and then solving the upper triangular system $\mathbf{Rx} = \bar{\mathbf{b}}$ via back-substitution. Note that since $\|\mathbf{Qv}\| = \|\mathbf{v}\|$ for any vector v, we have $\|\mathbf{R}\| = \|\mathbf{QR}\| = \|\mathbf{B}\|$, so that $\mathbf{R}$ preserves the relative magnitudes of the elements in $\mathbf{B}$, maintaining stability. This is its principal advantage.

Also, a permutation matrix $\mathbf{P}$ is sometimes used to postmultiply $\mathbf{B}^{(i-1)}$ before applying $\mathbf{Q}_i$ to it, so as to move a column that has the largest value of the sum of squares below row $i-1$ into the ith column position (see the computation of θ_i above). Since the product of permutation matrices is also a permutation matrix, and since a permutation matrix is orthogonal, this leads to the decomposition $\mathbf{B} = \mathbf{QRP}$ via the operation sequence $\mathbf{Q}_{n-1}\cdots\mathbf{Q}_1\mathbf{B}\mathbf{P}_1\mathbf{P}_2\cdots\mathbf{P}_{n-1} = \mathbf{R}$.

Cholesky Factorization $\mathbf{LL}^t$ and $\mathbf{LDL}^t$ for Symmetric, Positive Definite Matrices B

The Cholesky factorization of a symmetric, positive definite matrix $\mathbf{B}$ represents this matrix as $\mathbf{B} = \mathbf{LL}^t$, where $\mathbf{L}$ is a lower triangular matrix of the form

$$\mathbf{L} = \begin{bmatrix} \ell_{11} & & & & \\ \ell_{21} & \ell_{22} & & \mathbf{0} & \\ \ell_{31} & \ell_{32} & \ell_{33} & & \\ \vdots & \vdots & & \ddots & \\ \ell_{n1} & \ell_{n2} & & & \ell_{nn} \end{bmatrix},$$

so that

$$\mathbf{LL}^t = \begin{bmatrix} \ell_{11}^2 & & & & \text{(symmetric)} \\ \ell_{21}\ell_{11} & (\ell_{21}^2 + \ell_{22}^2) & & & \\ \ell_{31}\ell_{11} & (\ell_{21}\ell_{31} + \ell_{22}\ell_{32}) & (\ell_{31}^2 + \ell_{32}^2 + \ell_{33}^2) & & \\ \vdots & \vdots & \vdots & & \ddots \\ \ell_{n1}\ell_{11} & (\ell_{21}\ell_{n1} + \ell_{22}\ell_{n2}) & (\ell_{31}\ell_{n1} + \ell_{32}\ell_{n2} + \ell_{33}\ell_{n3}) & \cdots & (\ell_{n1}^2 + \cdots + \ell_{nn}^2) \end{bmatrix}.$$

By equating the elements of $\mathbf{B}$ directly to those in $\mathbf{LL}^t$, we obtain the system of equations

$$\begin{aligned} \ell_{11}^2 = b_{11},\quad & \ell_{21}\ell_{11} = b_{21}, && \ell_{31}\ell_{11} = b_{31}, && \cdots, && \ell_{n1}\ell_{11} = b_{n1} \\ & \ell_{21}^2 + \ell_{22}^2 = b_{22}, && \ell_{21}\ell_{31} + \ell_{22}\ell_{32} = b_{32}, && \cdots, && \ell_{21}\ell_{n1} + \ell_{22}\ell_{n2} = b_{n2} \\ & && \ell_{31}^2 + \ell_{32}^2 + \ell_{33}^2 = b_{33}, && \cdots, && \ell_{31}\ell_{n1} + \ell_{32}\ell_{n2} + \ell_{33}\ell_{n3} = b_{n3} \\ & && && \ddots && \vdots \\ & && && && \ell_{n1}^2 + \cdots + \ell_{nn}^2 = b_{nn}. \end{aligned}$$

These equations can be used sequentially to compute the unknowns ℓ_{ij} in the order $\ell_{11}, \ell_{21},\ldots,\ell_{n1}, \ell_{22}, \ell_{32},\ldots,\ell_{n2}, \ell_{33},\ldots,\ell_{n3},\ldots,\ell_{nn}$, by using the equation

for b_{ij} to compute ℓ_{ij} for $j = 1,..., n$, $i = j,..., n$. Note that these equations are well-defined for a symmetric, positive definite matrix **B** and that $\mathbf{LL}^t$ is positive definite if and only if $\ell_{ii} > 0$ for all $i = 1,..., n$.

The equation system $\mathbf{Bx} = \mathbf{b}$ can now be solved via $\mathbf{L}(\mathbf{L}^t\mathbf{x}) = \mathbf{b}$ through the solution of two triangular systems of equations. We first find **y** to satisfy $\mathbf{Ly} = \mathbf{b}$ and then compute **x** via the system $\mathbf{L}^t\mathbf{x} = \mathbf{y}$.

Sometimes the Cholesky factorization is represented as $\mathbf{B} = \mathbf{LDL}^t$, where **L** is a lower triangular matrix (usually having ones along its diagonal) and **D** is a diagonal matrix, both having positive diagonal entries. Writing $\mathbf{B} = \mathbf{LDL}^t = (\mathbf{LD}^{1/2})(\mathbf{LD}^{1/2})^t \equiv \mathbf{L}'\mathbf{L}'^t$, we see that the two representations are related equivalently. The advantage of the representation $\mathbf{LDL}^t$ is that **D** can be used to avoid the square root operation associated with the diagonal system of equations, and this improves the accuracy of computations. (For example, the diagonal components of **L** can be made unity.)

Also, if **B** is a general basis matrix, then since $\mathbf{BB}^t$ is symmetric and positive definite, it has a Cholesky factorization $\mathbf{BB}^t = \mathbf{LL}^t$. In such a case, **L** is referred to as the *Cholesky factor associated with* **B**. Note that we can determine **L** in this case by finding the *QR decomposition* for $\mathbf{B}^t$ so that $\mathbf{BB}^t = \mathbf{R}^t\mathbf{Q}^t\mathbf{QR} = \mathbf{R}^t\mathbf{R}$, and therefore, $\mathbf{L} \equiv \mathbf{R}^t$. Whenever this is done, note that the matrix **Q** or its components $\mathbf{Q}_i$ need not be stored, since we are only interested in the resulting upper triangular matrix **R**.

A.3 Sets and Sequences

A *set* is a collection of elements or objects. A set may be specified by listing its elements or by specifying the properties that the elements must satisfy. For example, the set $S = \{1, 2, 3, 4\}$ can be represented alternatively as $S = \{x : 1 \leq x \leq 4, x \text{ integer}\}$. If **x** is a member of S, we write $\mathbf{x} \in S$, and if **x** is not a member of S, we write $\mathbf{x} \notin S$. Sets are denoted by capital letters, such as S, X, and A. The *empty set*, denoted by $\varnothing$, has no elements.

Unions, Intersections, and Subsets

Given two sets, S_1 and S_2, the set consisting of elements that belong to either S_1 or S_2, or both, is called the *union* of S_1 and S_2 and is denoted by $S_1 \cup S_2$. The elements belonging to both S_1 and S_2 form the *intersection* of S_1 and S_2, denoted $S_1 \cap S_2$. If S_1 is a *subset* of S_2, that is, if each element of S_1 is also an element of S_2, we write $S_1 \subseteq S_2$ or $S_2 \supseteq S_1$. Thus, we write $S \subseteq R^n$ to denote

that all elements in S are points in R^n. A *strict containment* $S_1 \subseteq S_2$, $S_1 \neq S_2$, is denoted by $S_1 \subset S_2$.

Closed and Open Intervals

Let a and b be two real numbers. The *closed interval* $[a, b]$ denotes all real numbers satisfying $a \leq x \leq b$. Real numbers satisfying $a \leq x < b$ are represented by $[a, b)$, while those satisfying $a < x \leq b$ are denoted by $(a, b]$. Finally, the set of points x with $a < x < b$ is represented by the *open interval* (a, b).

Greatest Lower Bound and Least Upper Bound

Let S be a set of real numbers. Then the *greatest lower bound*, or the *infimum*, of S is the largest possible scalar α satisfying $\alpha \leq x$ for each $x \in S$. The infimum is denoted by $\inf \{x : x \in S\}$. The *least upper bound*, or the *supremum*, of S is the smallest possible scalar α satisfying $\alpha \geq x$ for each $x \in S$. The supremum is denoted by $\sup \{x : x \in S\}$.

Neighborhoods

Given a point $\mathbf{x} \in R^n$ and an $\varepsilon > 0$, the *ball* $N_\varepsilon(\mathbf{x}) = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\| \leq \varepsilon\}$ is called an *ε-neighborhood* of $\mathbf{x}$. The inequality in the definition of $N_\varepsilon(\mathbf{x})$ is sometimes replaced by a strict inequality.

Interior Points and Open Sets

Let S be a subset of R^n, and let $\mathbf{x} \in S$. Then $\mathbf{x}$ is called an *interior point* of S if there is an ε-neighborhood of $\mathbf{x}$ that is contained in S, that is, if there exists an $\varepsilon > 0$ such that $\|\mathbf{y} - \mathbf{x}\| \leq \varepsilon$ implies that $\mathbf{y} \in S$. The set of all such points is called the *interior* of S and is denoted by $\text{int } S$. Furthermore, S is called *open* if $S = \text{int } S$.

Relative Interior

Let $S \subset R^n$, and let $\text{aff}(S)$ denote the *affine hull* of S. Although $\text{int}(S) = \varnothing$, the interior of S as viewed in the space of its affine hull may be nonempty. This is called the *relative interior* of S and is denoted by $\text{relint}(S)$. Specifically, $\text{relint}(S) = \{\mathbf{x} \in S : N_\varepsilon(\mathbf{x}) \cap \text{aff}(S) \subset S \text{ for some } \varepsilon > 0\}$. Note that if $S_1 \subseteq S_2$, $\text{relint}(S_1)$ is not necessarily contained within $\text{relint}(S_2)$, although $\text{int}(S_1) \subseteq \text{int}(S_2)$. For example, if $S_1 = \{\mathbf{x} : \alpha' \mathbf{x} = \beta\}$, $\alpha \neq 0$ and $S_2 = \{\mathbf{x} : \alpha' \mathbf{x} \leq \beta\}$, $S_1 \subseteq S_2$, $\text{int}(S_1) = \varnothing \subseteq \text{int}(S_2) = \{\mathbf{x} : \alpha' \mathbf{x} < \beta\}$, but $\text{relint}(S_1) = S_1 \not\subseteq \text{relint}(S_2) = \text{int}(S_2)$.

Bounded Sets

A set $S \subset R^n$ is said to be *bounded* if it can be contained within a ball of finite radius.

Closure Points and Closed Sets

Let S be a subset of R^n. The *closure* of S, denoted cl S, is the set of all points that are arbitrarily close to S. In particular, $\mathbf{x} \in \text{cl } S$ if for each $\varepsilon > 0$, $S \cap N_\varepsilon(\mathbf{x}) \neq \varnothing$, where $N_\varepsilon(\mathbf{x}) = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\| \leq \varepsilon\}$. The set S is said to be *closed* if $S = \text{cl } S$.

Boundary Points

Let S be a subset of R^n. Then $\mathbf{x}$ is called a *boundary point* of S if for each $\varepsilon > 0$, $N_\varepsilon(\mathbf{x})$ contains a point in S and a point not in S, where $N_\varepsilon(\mathbf{x}) = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\| \leq \varepsilon\}$. The set of all boundary points is called the *boundary* of S and is denoted by ∂S.

Sequences and Subsequences

A *sequence* of vectors $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots$, is said to *converge* to the *limit point* $\bar{\mathbf{x}}$ if $\|\mathbf{x}_k - \bar{\mathbf{x}}\| \to 0$ as $k \to \infty$; that is, if for any given $\varepsilon > 0$, there is a positive integer N such that $\|\mathbf{x}_k - \bar{\mathbf{x}}\| < \varepsilon$ for all $k \geq N$. The sequence is usually denoted by $\{\mathbf{x}_k\}$, and the limit point $\bar{\mathbf{x}}$ is represented by either $\mathbf{x}_k \to \bar{\mathbf{x}}$ as $k \to \infty$ or by $\lim_{k\to\infty} \mathbf{x}_k = \bar{\mathbf{x}}$. Any converging sequence has a unique limit point.

By deleting certain elements of a sequence $\{\mathbf{x}_k\}$, we obtain a *subsequence*. A subsequence is usually denoted as $\{\mathbf{x}_k\}_{\mathscr{K}}$ where $\mathscr{K}$ is a subset of all positive integers. To illustrate, let $\mathscr{K}$ be the set of all even positive integers. Then $\{\mathbf{x}_k\}_{\mathscr{K}}$ denotes the subsequence $\{\mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_6, \ldots\}$.

Given a subsequence $\{\mathbf{x}_k\}_{\mathscr{K}}$, the notation $\{\mathbf{x}_{k+1}\}_{\mathscr{K}}$ denotes the subsequence obtained by adding 1 to the indices of all elements in the subsequence $\{\mathbf{x}_k\}_{\mathscr{K}}$. To illustrate, if $\mathscr{K} = \{3, 5, 10, 15, \ldots\}$, then $\{\mathbf{x}_{k+1}\}_{\mathscr{K}}$ denotes the subsequence $\{\mathbf{x}_4, \mathbf{x}_6, \mathbf{x}_{11}, \mathbf{x}_{16}, \ldots\}$.

A sequence $\{\mathbf{x}_k\}$ is called a *Cauchy sequence* if for any given $\varepsilon > 0$, there is a positive integer N such that $\|\mathbf{x}_k - \mathbf{x}_m\| < \varepsilon$ for all $k, m \geq N$. A sequence in R^n has a limit if and only if it is Cauchy.

Let $\{x_n\}$ be a bounded sequence in R. The *limit superior* of $\{x_n\}$, denoted limsup(x_n) or $\overline{\lim}(x_n)$, equals the infimum of all numbers $q \in R$ for which at most a finite number of the elements of $\{x_n\}$ (strictly) exceed q. Similarly, the *limit inferior* of $\{x_n\}$ is given by liminf $(x_n) \equiv \underline{\lim}(x_n) \equiv \sup\{q :$ at most a finite

number of elements of $\{x_n\}$ are (strictly) less than $q\}$. A bounded sequence always has a unique $\overline{\lim}$ and $\underline{\lim}$.

Compact Sets

A set S in R^n is said to be *compact* if it is closed and bounded. For every sequence $\{\mathbf{x}_k\}$ in a compact set S, there is a convergent subsequence with a limit in S.

A.4 Functions

A *real-valued function* f defined on a subset S of R^n associates with each point $\mathbf{x}$ in S a real number $f(\mathbf{x})$. The notation $f\colon S \to R$ denotes that the domain of f is S and that the range is a subset of the real numbers. If f is defined everywhere on R^n or if the domain is not important, the notation $f\colon R^n \to R$ is used. A collection of real-valued functions $f_1,\ldots,f_m$ can be viewed as a single *vector function* f whose jth component is f_j.

Continuous Functions

A function $f\colon S \to R$ is said to be *continuous at* $\bar{\mathbf{x}} \in S$ if for any given $\varepsilon > 0$, there is a $\delta > 0$ such that $\mathbf{x} \in S$ and $\|\mathbf{x} - \bar{\mathbf{x}}\| < \delta$ imply that $|f(\mathbf{x}) - f(\bar{\mathbf{x}})| < \varepsilon$. Equivalently, f is continuous at $\bar{\mathbf{x}} \in S$, if for any sequence $\{\mathbf{x}_n\} \to \bar{\mathbf{x}}$ such that $\{f(\mathbf{x}_n)\} \to \bar{f}$, we have that $f(\bar{\mathbf{x}}) = \bar{f}$ as well. A vector-valued function is said to be continuous at $\bar{\mathbf{x}}$ if each of its components is continuous at $\bar{\mathbf{x}}$.

Upper and Lower Semicontinuity

Let S be a nonempty set in R^n. A function $f\colon S \to R$ is said to be *upper semicontinuous* at $\bar{\mathbf{x}} \in S$ if for each $\varepsilon > 0$ there exists a $\delta > 0$ such that $\mathbf{x} \in S$ and $\|\mathbf{x} - \bar{\mathbf{x}}\| < \delta$ imply that $f(\mathbf{x}) - f(\bar{\mathbf{x}}) < \varepsilon$. Similarly, a function $f\colon R^n \to R$ is called *lower semicontinuous at* $\bar{\mathbf{x}} \in S$ if for each $\varepsilon > 0$ there exists a $\delta > 0$ such that $\mathbf{x} \in S$ and $\|\mathbf{x} - \bar{\mathbf{x}}\| < \delta$ imply that $f(\mathbf{x}) - f(\bar{\mathbf{x}}) > -\varepsilon$. Equivalently, then f is *upper semicontinuous at* $\bar{\mathbf{x}} \in S$, if, for any sequence $\{\mathbf{x}_n\} \to \bar{\mathbf{x}}$ such that $\{f(\mathbf{x}_n)\} \to \bar{f}$, we have $f(\bar{\mathbf{x}}) \geq \bar{f}$. Similarly, if $f(\bar{\mathbf{x}}) \leq \bar{f}$ for any such sequence, then f is said to be *lower semicontinuous at* $\bar{\mathbf{x}}$. Hence, a function is *continuous at* $\bar{\mathbf{x}}$ if and only if it is both upper and lower semicontinuous at $\bar{\mathbf{x}}$. A vector-valued function is called upper or lower semicontinuous if each of its components is upper or lower semicontinuous, respectively.

Minima and Maxima of Semicontinuous Functions

Let S be a nonempty compact set in R^n and suppose that $f\colon R^n \to R$. If f is lower semicontinuous, then it assumes a minimum over S; that is, there exists an $\bar{\mathbf{x}} \in S$

such that $f(\mathbf{x}) \leq f(\bar{\mathbf{x}})$ for each $\mathbf{x} \in S$. Similarly, if f is upper semicontinuous, then it assumes a maximum over S. Since a continuous function is both lower and upper semicontinuous, it achieves both a minimum and a maximum over any nonempty compact set.

Differentiable Functions

Let S be a nonempty set in R^n, $\bar{\mathbf{x}} \in \text{int } S$ and let $f: S \to R$. Then f is said to be *differentiable at* $\bar{\mathbf{x}}$ if there is a vector $\nabla f(\bar{\mathbf{x}})$ in R^n called the *gradient* of f at $\bar{\mathbf{x}}$ and a function β satisfying $\beta(\bar{\mathbf{x}}; \mathbf{x}) \to 0$ as $\mathbf{x} \to \bar{\mathbf{x}}$ such that

$$f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) + \|\mathbf{x} - \bar{\mathbf{x}}\| \beta(\bar{\mathbf{x}}; \mathbf{x}) \qquad \text{for each } \mathbf{x} \in S.$$

The gradient vector consists of the partial derivatives, that is,

$$\nabla f(\bar{\mathbf{x}})^t = \left(\frac{\partial f(\bar{\mathbf{x}})}{\partial x_1}, \frac{\partial f(\bar{\mathbf{x}})}{\partial x_2}, \ldots, \frac{\partial f(\bar{\mathbf{x}})}{\partial x_n} \right).$$

Furthermore, f is called *twice differentiable at* $\bar{\mathbf{x}}$ if, in addition to the gradient vector, there exist an $n \times n$ symmetric matrix $\mathrm{H}(\bar{\mathbf{x}})$, called the *Hessian matrix* of f at $\bar{\mathbf{x}}$, and a function β satisfying $\beta(\bar{\mathbf{x}}; \mathbf{x}) \to 0$ as $\mathbf{x} \to \bar{\mathbf{x}}$ such that

$$f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^t \mathrm{H}(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) + \|\mathbf{x} - \bar{\mathbf{x}}\|^2 \beta(\bar{\mathbf{x}}; \mathbf{x})$$

$$\text{for each } \mathbf{x} \in S.$$

The element in row i and column j of the Hessian matrix is the second partial $\partial^2 f(\bar{\mathbf{x}})/\partial x_i \, \partial x_j$.

A vector-valued function is differentiable if each of its components is differentiable and is twice differentiable if each of its components is twice differentiable.

In particular, for a differentiable vector function $\mathbf{h}$: $R^n \to R^\ell$ where $\mathbf{h}(\mathbf{x}) = (h_1(\mathbf{x}), \ldots, h_\ell(\mathbf{x}))^t$, the *Jacobian* of $\mathbf{h}$, denoted by the gradient notation $\nabla \mathbf{h}(\mathbf{x})$, is given by the $\ell \times n$ matrix

$$\nabla \mathbf{h}(\mathbf{x}) = \begin{bmatrix} \nabla h_1(\mathbf{x})^t \\ \vdots \\ \nabla h_\ell(\mathbf{x})^t \end{bmatrix}_{\ell \times n},$$

whose rows correspond to the transpose of the gradients of $h_1, \ldots, h_\ell$, respectively.

Mean Value Theorem

Let S be a nonempty open convex set in R^n, and let f: $S \to R$ be differentiable. The mean value theorem can be stated as follows. For every $\mathbf{x}_1$ and $\mathbf{x}_2$ in S, we must have

$$f(\mathbf{x}_2) = f(\mathbf{x}_1) + \nabla f(\mathbf{x})^t(\mathbf{x}_2 - \mathbf{x}_1),$$

where $\mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ for some $\lambda \in (0, 1)$.

Taylor's Theorem

Let S be a nonempty open convex set in R^n, and let f: $S \to R$ be twice differentiable. The second-order form of *Taylor's theorem* can be stated as follows. For every $\mathbf{x}_1$ and $\mathbf{x}_2$ in S, we must have

$$f(\mathbf{x}_2) = f(\mathbf{x}_1) + \nabla f(\mathbf{x}_1)^t(\mathbf{x}_2 - \mathbf{x}_1) + \frac{1}{2}(\mathbf{x}_2 - \mathbf{x}_1)^t \mathbf{H}(\mathbf{x})(\mathbf{x}_2 - \mathbf{x}_1),$$

where $\mathbf{H}(\mathbf{x})$ is the Hessian of f at $\mathbf{x}$, and where $\mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ for some $\lambda \in (0, 1)$.

Appendix B

Summary of Convexity, Optimality Conditions, and Duality

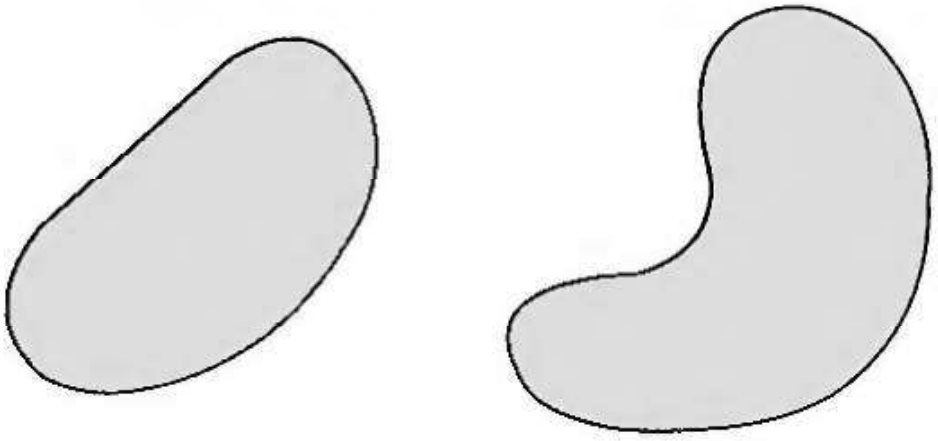
This appendix gives a summary of the relevant results from Chapters 2 through 6 on convexity, optimality conditions, and duality. *It is intended to provide the minimal background needed for an adequate coverage of Chapters 8 through 11, excluding convergence analysis.*

B.1 Convex Sets

A set S in R^n is said to be *convex* if for each $\mathbf{x}_1, \mathbf{x}_2 \in S$, the *line segment* $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ for $\lambda \in [0, 1]$ belongs to S. Points of the form $\mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ for $\lambda \in [0, 1]$ are called *convex combinations* of $\mathbf{x}_1$ and $\mathbf{x}_2$. Figure B.1 illustrates an example of a convex set and an example of a nonconvex set.

We present below some examples of convex sets frequently encountered in mathematical programming.

1. ***Hyperplane***: $S = \{\mathbf{x} : \mathbf{p}^t\mathbf{x} = \alpha\}$, where $\mathbf{p}$ is a nonzero vector in R^n, called the *normal* to the hyperplane, and α is a scalar.
2. ***Half-space***: $S = \{\mathbf{x} : \mathbf{p}^t\mathbf{x} \leq \alpha\}$, where $\mathbf{p}$ is a nonzero vector in R^n and α is a scalar.
3. ***Open half-space***: $S = \{\mathbf{x} : \mathbf{p}^t\mathbf{x} < \alpha\}$, where $\mathbf{p}$ is a nonzero vector in R^n and α is a scalar.
4. ***Polyhedral set***: $S = \{\mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$, where $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{b}$ is an m-vector .
5. ***Polyhedral cone***: $S = \{\mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{0}\}$, where $\mathbf{A}$ is an $m \times n$ matrix.
6. *Cone spanned by a finite number of vectors*: $S = \{\mathbf{x} : \mathbf{x} = \sum_{j=1}^{m} \lambda_j \mathbf{a}_j, \lambda_j \geq 0 \text{ for } j = 1,..., m\}$, where $\mathbf{a}_1,...,\mathbf{a}_m$ are given vectors in R^n.
7. ***Neighborhood***: $S = \{\mathbf{x} : \|\mathbf{x} - \bar{\mathbf{x}}\| \leq \varepsilon\}$, where $\bar{\mathbf{x}}$ is a fixed vector in $R^n$and $\varepsilon > 0$.



A convex set A nonconvex set

Figure B.1 Convexity.

Given two nonempty convex sets S_1 and S_2 in R^n such that $S_1 \cap S_2 = \emptyset$, there exists a hyperplane $H = \{\mathbf{x} : \mathbf{p}^t\mathbf{x} = \alpha\}$ that separates them; that is,

$$\mathbf{p}^t\mathbf{x} \leq \alpha \text{ for all } \mathbf{x} \in S_1 \qquad \text{and} \qquad \mathbf{p}^t\mathbf{x} \geq \alpha \text{ for all } \mathbf{x} \in S_2.$$

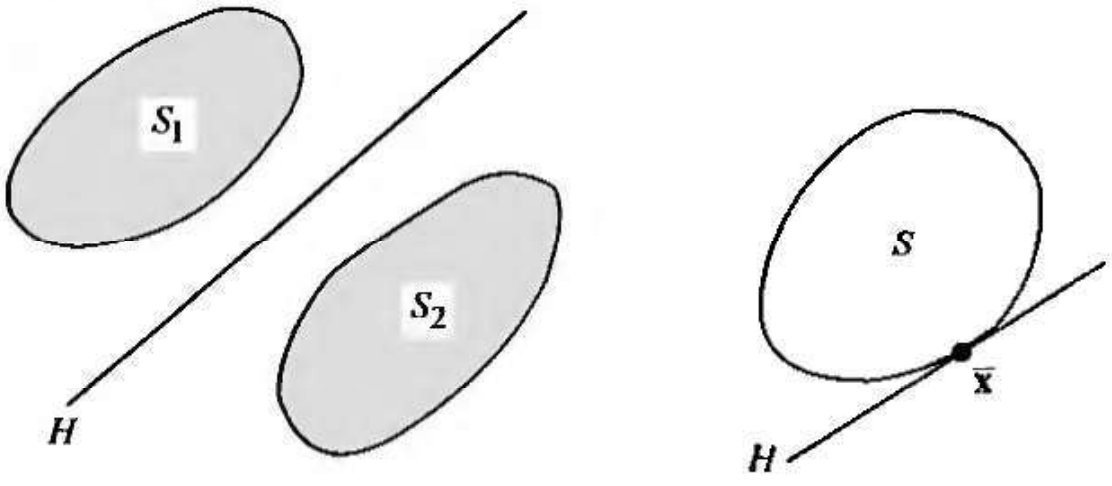
Here H is called a *separating hyperplane* whose normal is the nonzero vector p.

Closely related to the above concept is the notion of a *supporting hyperplane.* Let S be a nonempty convex set in R^n, and let $\bar{\mathbf{x}}$ be a boundary point. Then there exists a hyperplane $H = \{\mathbf{x} : \mathbf{p}^t\mathbf{x} = \alpha\}$ that supports S at $\bar{\mathbf{x}}$; that is,

$$\mathbf{p}^t\bar{\mathbf{x}} = \alpha \qquad \text{and} \qquad \mathbf{p}^t\mathbf{x} \leq \alpha \text{ for all } \mathbf{x} \in S.$$

In Figure B.2 we illustrate the concepts of separating and supporting hyperplanes.

The following two theorems are used in proving optimality conditions and duality relationships and in developing termination criteria for algorithms.



Separating hyperplane Supporting hyperplane

Figure B.2 Separating and supporting hyperplanes.

Farkas's Theorem

Let **A** be an $m \times n$ matrix and let **c** be an n-vector. Then exactly one of the following two systems has a solution:

System 1 $\mathbf{Ax} \leq \mathbf{0}, \mathbf{c}^t\mathbf{x} > 0$ for some $\mathbf{x} \in R^n$.

System 2 $\mathbf{A}^t\mathbf{y} = \mathbf{c}, \mathbf{y} \geq \mathbf{0}$ for some $\mathbf{y} \in R^m$.

Gordan's Theorem

Let **A** be an $m \times n$ matrix. Then exactly one of the following systems has a solution.

System 1 $\mathbf{Ax} < \mathbf{0}$ for some $\mathbf{x} \in R^n$.

System 2 $\mathbf{A}^t\mathbf{y} = \mathbf{0}, \mathbf{y} \geq \mathbf{0}$ for some nonzero $\mathbf{y} \in R^m$.

An important concept in convexity is that of an extreme point. Let S be a non-empty convex set in R^n. A vector $\mathbf{x} \in S$ is called an *extreme point* of S if $\mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ with $\mathbf{x}_1, \mathbf{x}_2 \in S$, and $\lambda \in (0, 1)$ implies that $\mathbf{x} = \mathbf{x}_1 = \mathbf{x}_2$. In other words, **x** is an extreme point if it cannot be represented as a strict convex combination of two distinct points in S. In particular, for the set $S = \{\mathbf{x} : \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$, where **A** is an $m \times n$ matrix of rank m and **b** is an m-vector, **x** is an *extreme point* of S if and only if the following conditions hold true. The matrix **A** can be decomposed into [**B**, **N**], where **B** is an $m \times m$ invertible matrix and $\mathbf{x}^t = (\mathbf{x}_B^t, \mathbf{x}_N^t)$, where $\mathbf{x}_B = \mathbf{B}^{-1}\mathbf{b} \geq \mathbf{0}$ and $\mathbf{x}_N = \mathbf{0}$.

Another concept that is used in the case of an unbounded convex set is that of a direction of the set. Specifically, if S is an unbounded closed convex set, a vector **d** is a *direction* of S if $\mathbf{x} + \lambda\mathbf{d} \in S$ for each $\lambda \geq 0$ and for each $\mathbf{x} \in S$.

B.2 Convex Functions and Extensions

Let S be a nonempty convex set in R^n. The function f: $S \rightarrow R$ is said to be *convex* on S if

$$f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] \leq \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2)$$

for each $\mathbf{x}_1, \mathbf{x}_2 \in S$ and for each $\lambda \in [0, 1]$. The function f is said to be *strictly convex* on S if the above inequality holds as a strict inequality for each distinct $\mathbf{x}_1, \mathbf{x}_2 \in S$ and for each $\lambda \in (0, 1)$. The function f is said to be *concave* (*strictly concave*) if $-f$ is convex (strictly convex). Figure B.3 shows some examples of convex and concave functions.

Following are some examples of convex functions. By taking the negatives of these functions, we get some examples of concave functions.

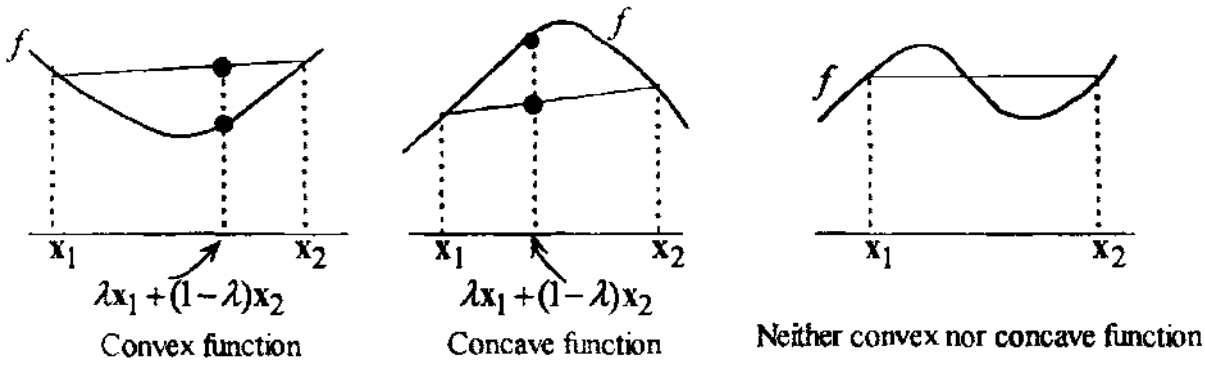


Figure B.3 Convex and concave functions.

1. $f(x) = 3x + 4$.
2. $f(x) = |x|$.
3. $f(x) = x^2 - 2x$.
4. $f(x) = -x^{1/2}$ for $x \geq 0$.
5. $f(x_1, x_2) = 2x_1^2 + x_2^2 - 2x_1x_2$.
6. $f(x_1, x_2, x_3) = x_1^4 + 2x_2^2 + 3x_3^2 - 4x_1 - 4x_2x_3$.

In many cases, the assumption of convexity of a function can be relaxed to the weaker notions of quasiconvex and pseudoconvex functions.

Let S be a nonempty convex set in R^n. The function $f: S \to R$ is said to be *quasiconvex* on S if for each $\mathbf{x}_1, \mathbf{x}_2 \in S$, the following inequality holds true:

$$f[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2] \leq \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\} \qquad \text{for each } \lambda \in (0, 1).$$

The function f is said to be *strictly quasiconvex* on S if the above inequality holds as a strict inequality, provided that $f(\mathbf{x}_1) \neq f(\mathbf{x}_2)$. The function f is said to be *strongly quasiconvex* on S if the above inequality holds as a strict inequality for $\mathbf{x}_1 \neq \mathbf{x}_2$.

Let S be a nonempty open convex set in R^n. The function $f: S \to R$ is said to be *pseudoconvex* if for each $\mathbf{x}_1, \mathbf{x}_2 \in S$ with $\nabla f(\mathbf{x}_1)^t(\mathbf{x}_2 - \mathbf{x}_1) \geq 0$, we have $f(\mathbf{x}_2) \geq f(\mathbf{x}_1)$. The function f is said to be *strictly pseudoconvex* on S if whenever $\mathbf{x}_1$ and $\mathbf{x}_2$ are distinct points in S with $\nabla f(\mathbf{x}_1)^t(\mathbf{x}_2 - \mathbf{x}_1) \geq 0$, we have $f(\mathbf{x}_2) > f(\mathbf{x}_1)$.

The above generalizations of convexity extend to the concave case by replacing f by $-f$. Figure B.4 illustrates these concepts. Figure B.5 summarizes the relationships among different types of convexity.

We now give a summary of important properties for various types of convex functions. Here $f: S \to R$, where S is a nonempty convex set in R^n.

Figure B.4 Quasiconvexity and pseudoconvexity.

Strictly Convex Functions

1. The function f is continuous on the interior of S.
2. The set $\{(\mathbf{x}, y) : \mathbf{x} \in S, y \geq f(\mathbf{x})\}$ is convex.
3. The set $\{\mathbf{x} \in S : f(\mathbf{x}) \leq \alpha\}$ is convex for each real α.
4. A differentiable function f is strictly convex on S if and only if $f(\mathbf{x}) > f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}})$ for each distinct $\mathbf{x}, \bar{\mathbf{x}} \in S$.
5. Let f be twice differentiable. Then if the Hessian $\mathbf{H}(\mathbf{x})$ is positive definite for each $\mathbf{x} \in S$, f is strictly convex on S. Furthermore, if f is strictly convex on S, then the Hessian $\mathbf{H}(\mathbf{x})$ is positive semidefinite for each $\mathbf{x} \in S$.
6. Every local minimum of f over a convex set $X \subseteq S$ is the unique global minimum.
7. If $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$, then $\bar{\mathbf{x}}$ is the unique global minimum of f over S.
8. The maximum of f over a nonempty compact polyhedral set $X \subseteq S$ is achieved at an extreme point of X.

Convex Functions

1. The function f is continuous on the interior of S.
2. The function f is convex on S if and only if the set $\{(\mathbf{x}, y) : \mathbf{x} \in S, y \geq f(\mathbf{x})\}$ is convex.
3. The set $\{\mathbf{x} \in S : f(\mathbf{x}) \leq \alpha\}$ is convex for each real α.
4. A differentiable function f is convex on S if and only if $f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}})$ for each $\mathbf{x}, \bar{\mathbf{x}} \in S$.
5. A twice differentiable function f is convex on S if and only if the Hessian $\mathbf{H}(\mathbf{x})$ is positive semidefinite for each $\mathbf{x} \in S$.
6. Every local minimum of f over a convex set $X \subseteq S$ is a global minimum.
7. If $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$, then $\bar{\mathbf{x}}$ is a global minimum of f over S.

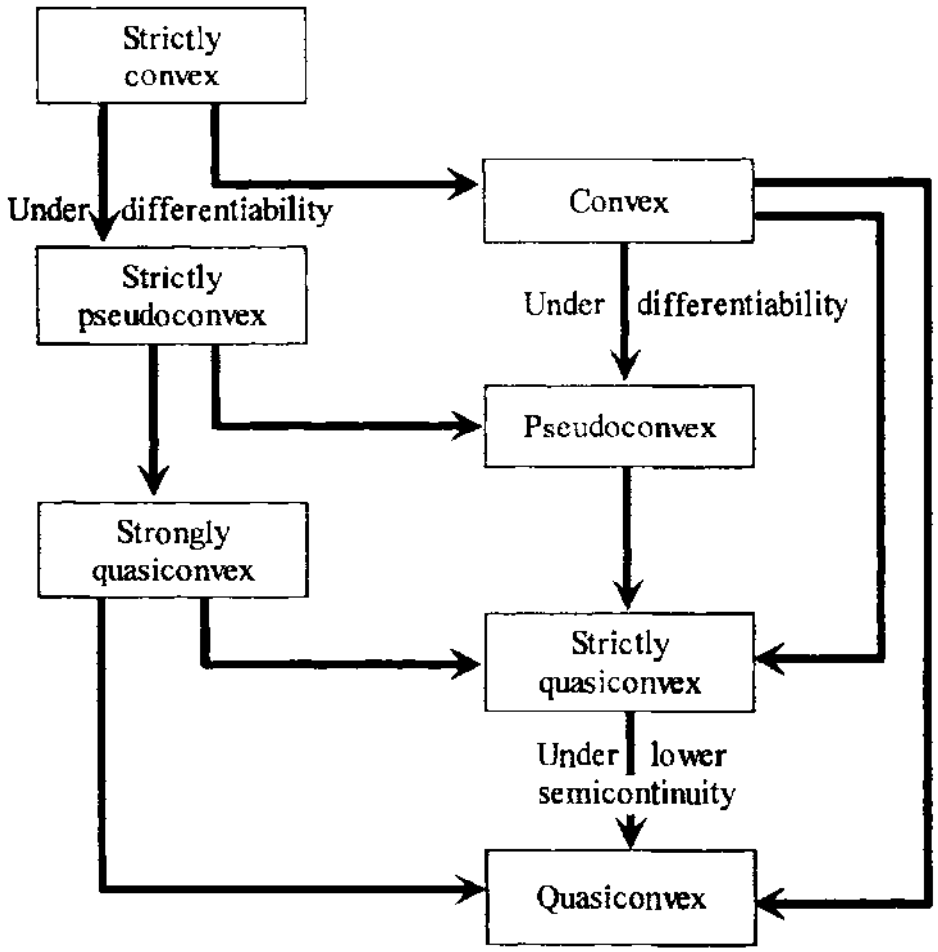


Figure B.5 Relationship among various types of convexity.

8. A maximum of f over a nonempty compact polyhedral set $X \subseteq S$ is achieved at an extreme point of X.

Pseudoconvex Functions

1. The set $\{\mathbf{x} \in S : f(\mathbf{x}) \leq \alpha\}$ is convex for each real α.
2. Every local minimum of f over a convex set $X \subseteq S$ is a global minimum.
3. If $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$, then $\bar{\mathbf{x}}$ is a global minimum of f over S.
4. A maximum of f over a nonempty compact polyhedral set $X \subseteq S$ is achieved at an extreme point of X.
5. This characterization and the next relate to twice differentiable functions f defined on an open convex set $S \subseteq R^n$, with Hessian $\mathbf{H}(\mathbf{x})$.

   The function f is pseudoconvex on S if $\mathbf{H}(\mathbf{x}) + r(\mathbf{x})\nabla f(\mathbf{x})\nabla f(\mathbf{x})^t$ is positive semidefinite for all $\mathbf{x} \in S$, where $r(\mathbf{x}) = (1/2)[\delta - f(\mathbf{x})]$ for some $\delta > f(\mathbf{x})$. Moreover, this condition is both necessary and sufficient if f is quadratic.
6. Define the $(n + 1) \times (n + 1)$ *bordered Hessian* $\mathbf{B}(\mathbf{x})$ of f as follows, where $\mathbf{H}(\mathbf{x})$ is "bordered" by an additional row and column:

$$\mathbf{B}(\mathbf{x}) = \begin{bmatrix} \mathbf{H}(\mathbf{x}) & \nabla f(\mathbf{x}) \\ \nabla f(\mathbf{x})^t & \mathbf{0} \end{bmatrix}.$$

Given any $k \in \{1,..., n\}$,and $\gamma = \{i_1,..., i_k\}$ composed of some k distinct indices $1 \leq i_1 < i_2 < \cdots < i_k \leq n$, the *principal submatrix* $\mathbf{B}_{\gamma,k}(\mathbf{x})$ is a $(k + 1) \times (k + 1)$ submatrix of $\mathbf{B}(\mathbf{x})$ formed by picking the elements of $\mathbf{B}(\mathbf{x})$ that intersect in the rows $i_1,..., i_k$, $(n + 1)$ and the columns $i_1,..., i_k$, $(n + 1)$ of $\mathbf{B}(\mathbf{x})$. The *leading principal submatrix* of $\mathbf{B}(\mathbf{x})$ is denoted by $\mathbf{B}_k(\mathbf{x})$ and equals $\mathbf{B}_{\gamma,k}$ for $\gamma \equiv \{1,..., k\}$. Similarly, let $\mathbf{H}_{\gamma,k}(\mathbf{x})$ and $\mathbf{H}_k(\mathbf{x})$ be the $k \times k$ principal submatrix and the leading principal submatrix, respectively, of $\mathbf{H}(\mathbf{x})$. Then f is pseudoconvex on S if for each $\mathbf{x} \in S$, we have (i) $\det \mathbf{B}_{\gamma,k}(\mathbf{x}) \leq 0$ for all γ, $k = 1,..., n$, and (ii) if $\det \mathbf{B}_{\gamma,k}(\mathbf{x}) = 0$ for any γ, k, then $\det \mathbf{H}_{\gamma,k} \geq 0$ over some neighborhood of $\mathbf{x}$. Moreover, if f is quadratic, then these conditions are both necessary and sufficient. Also, in general, the condition $\det \mathbf{B}_k(\mathbf{x}) < 0$ for all $k = 1,..., n$, $\mathbf{x} \in S$, is sufficient for f to be pseudoconvex on S.

7. Let f: $S \subseteq R^n \rightarrow R$ be quadratic, where S is a convex subset of R^n. Then [f is pseudoconvex on S] $\Leftrightarrow$ [the bordered Hessian $\mathbf{B}(\mathbf{x})$ has exactly one simple negative eigenvalue for all $\mathbf{x} \in S$] $\Leftrightarrow$ [for each $\mathbf{y} \in R^n$ such that $\nabla f(\mathbf{x})^t \mathbf{y} = 0$, we have that $\mathbf{y}^t \mathbf{H}(\mathbf{x})\mathbf{y} \geq 0$ for all $\mathbf{x} \in S$]. Moreover, [f is strictly pseudoconvex on S] $\Leftrightarrow$ [for all $\mathbf{x} \in S$, and for all $k = 1,..., n$, we have (i) $\det \mathbf{B}_k(\mathbf{x}) \leq 0$, and (ii) if $\det \mathbf{B}_k(\mathbf{x}) = 0$, then $\det \mathbf{H}_k > 0$].

Quasiconvex Functions

1. The function f is quasiconvex over S if and only if $\{\mathbf{x} \in S : f(\mathbf{x}) \leq \alpha\}$ is convex for each real α.
2. A maximum of f over a nonempty compact polyhedral set $X \subseteq S$ is achieved at an extreme point of X.
3. A differentiable function f on S is quasiconvex over S if and only if $\mathbf{x}_1, \mathbf{x}_2 \in S$ with $f(\mathbf{x}_1) \leq f(\mathbf{x}_2)$ implies that $\nabla f(\mathbf{x}_2)^t(\mathbf{x}_1 - \mathbf{x}_2) \leq 0$.
4. Let f: $S \subseteq R^n \rightarrow R$, where f is twice differentiable and S is a *solid* (i.e., has a nonempty interior) convex subset of R^n. Define the bordered Hessian of f and its submatrices as in Property 6 of pseudoconvex functions. Then a sufficient condition for f to be quasiconvex on S is that for each $\mathbf{x} \in S$, $\det \mathbf{B}_k(\mathbf{x}) < 0$ for all $k = 1,..., n$. (Note that this condition actually implies that f is pseudoconvex.)

On the other hand, a necessary condition for f to be quasiconvex on S is that for each $\mathbf{x} \in S$, $\det \mathbf{B}_k(\mathbf{x}) \leq 0$ for all $k = 1,..., n$.

5. Let f: $S \subseteq R^n \to R$ be a quadratic function where $S \subseteq R^n$ is a solid (nonempty interior) convex subset of R^n. Then f is quasiconvex on S if and only if f is pseudoconvex on int(S).

A local minimum of a strictly quasiconvex function over a convex set $X \subseteq S$ is also a global minimum. Furthermore, if the function is strongly quasiconvex, the minimum is unique. If a function f is both strictly quasiconvex and lower semicontinuous, it is quasiconvex, so that the above properties for quasiconvexity hold true.

B.3 Optimality Conditions

Consider the following problem:

$$\begin{aligned} \text{P}: \text{ Minimize } & f(\mathbf{x}) \\ \text{subject to } & g_i(\mathbf{x}) \leq 0 \quad \text{for } i = 1,..., m \\ & h_i(\mathbf{x}) = 0 \quad \text{for } i = 1,..., \ell \\ & \mathbf{x} \in X, \end{aligned}$$

where f, g_i, h_i: $R^n \to R$ and X is a nonempty open set in R^n. We give below the *Fritz John necessary optimality conditions*. If a point $\bar{\mathbf{x}}$ is a local optimal solution to the above problem, then there must exist a nonzero vector $(u_0, \mathbf{u}, \mathbf{v})$ such that

$$u_0 \nabla f(\bar{\mathbf{x}}) + \sum_{i=1}^{m} u_i \nabla g_i(\bar{\mathbf{x}}) + \sum_{i=1}^{\ell} v_i \nabla h_i(\bar{\mathbf{x}}) = 0$$

$$u_i g_i(\bar{\mathbf{x}}) = 0 \qquad \text{for } i = 1,..., m$$

$$u_0 \geq 0,\ u_i \geq 0 \qquad \text{for } i = 1,..., m,$$

where $\mathbf{u}$ and $\mathbf{v}$ are m- and ℓ-vectors whose ith components are u_i and v_i, respectively. Here, u_0, u_i, and v_i are referred to as the *Lagrange* or *Lagrangian multipliers* associated, respectively, with the objective function, the ith inequality constraint $g_i(\mathbf{x}) \leq 0$, and the ith equality constraint $h_i(\mathbf{x}) = 0$. The condition $u_i g_i(\bar{\mathbf{x}}) = 0$ is called the *complementary slackness condition* and stipulates that either $u_i = 0$ or $g_i(\bar{\mathbf{x}}) = 0$. Thus, if $g_i(\bar{\mathbf{x}}) < 0$, then $u_i = 0$. By letting I be the set of binding inequality constraints at $\bar{\mathbf{x}}$, that is, $I = \{i : g_i(\bar{\mathbf{x}}) = 0\}$, then the Fritz John conditions can be written in the following equivalent form. If $\bar{\mathbf{x}}$ is a local optimal solution to Problem P above, then there exists a nonzero vector $(u_0, \mathbf{u}_I, \mathbf{v})$ satisfying the following, where $\mathbf{u}_I$ is the vector of Lagrange multipliers associated with $g_i(\mathbf{x}) \leq 0$ for $i \in I$:

$$u_0 \nabla f(\bar{\mathbf{x}}) + \sum_{i \in I} u_i \nabla g_i(\bar{\mathbf{x}}) + \sum_{i=1}^{\ell} v_i \nabla h_i(\bar{\mathbf{x}}) = 0$$

$$u_0 \geq 0,\ u_i \geq 0 \qquad \text{for } i \in I.$$

If $u_0 = 0$, the Fritz John conditions become less meaningful, since essentially, they simply state that the gradients of the binding inequality constraints and the gradients of the equality constraints are linearly dependent. Under suitable assumptions, referred to as *constraint qualifications*, u_0 is guaranteed to be positive, and the Fritz John conditions reduce to the Karush–Kuhn–Tucker (KKT) conditions. A typical constraint qualification is that the gradients of the inequality constraints for $i \in I$ and the gradients of the equality constraints at $\bar{\mathbf{x}}$ are linearly independent.

The KKT necessary optimality conditions can be stated as follows. If $\bar{\mathbf{x}}$ is a local optimal solution to Problem P, under a suitable constraint qualification, there exists a vector (u, v) such that

$$\nabla f(\bar{\mathbf{x}}) + \sum_{i=1}^{m} u_i \nabla g_i(\bar{\mathbf{x}}) + \sum_{i=1}^{\ell} v_i \nabla h_i(\bar{\mathbf{x}}) = 0$$

$$u_i g_i(\bar{\mathbf{x}}) = 0 \qquad \text{for } i = 1, \ldots, m$$

$$u_i \geq 0 \qquad \text{for } i = 1, \ldots, m.$$

Again, u_i and v_i are the *Lagrange* or *Lagrangian multipliers* associated with the constraints $g_i(\mathbf{x}) \leq 0$ and $h_i(\mathbf{x}) = 0$, respectively. Furthermore, $u_i g_i(\bar{\mathbf{x}}) = 0$ is referred to as a *complementary slackness condition*. If we let $I = \{i : g_i(\bar{\mathbf{x}}) = 0\}$, the above conditions can be rewritten as

$$\nabla f(\bar{\mathbf{x}}) + \sum_{i=1}^{m} u_i \nabla g_i(\bar{\mathbf{x}}) + \sum_{i=1}^{\ell} v_i \nabla h_i(\bar{\mathbf{x}}) = 0$$

$$u_i \geq 0 \qquad \text{for } i \in I.$$

Under suitable convexity assumptions, the KKT conditions are also *sufficient* for optimality. In particular, suppose that $\bar{\mathbf{x}}$ is a feasible solution to Problem P and that the KKT conditions stated below hold true:

$$\nabla f(\bar{\mathbf{x}}) + \sum_{i \in I} u_i \nabla g_i(\bar{\mathbf{x}}) + \sum_{i=1}^{\ell} v_i \nabla h_i(\bar{\mathbf{x}}) = 0$$

$$u_i \geq 0 \qquad \text{for } i \in I,$$

where $I = \{i : g_i(\bar{\mathbf{x}}) = 0\}$. If f is pseudoconvex, g_i is quasiconvex for $i \in I$; and if h_i is quasiconvex if $v_i > 0$ and quasiconcave if $v_i < 0$, then $\bar{\mathbf{x}}$ is an optimal solution to Problem P.

To illustrate the KKT conditions, consider the following problem:

$$\begin{aligned} \text{Minimize} \quad & (x_1 - 3)^2 + (x_2 - 2)^2 \\ \text{subject to} \quad & x_1^2 + x_2^2 \le 5 \\ & x_1 + 2x_2 \le 4 \\ & -x_1 \le 0 \\ & -x_2 \le 0. \end{aligned}$$

The problem is illustrated in Figure B.6. Note that the optimal solution is $\bar{\mathbf{x}} = (2, 1)^t$. We first verify that the KKT conditions hold true at $\bar{\mathbf{x}}$. Here, the set of binding inequality constraints is $I = \{1, 2\}$, so that we must have $u_3 = u_4 = 0$ to satisfy the complementary slackness conditions. Note that

$$\nabla f(\bar{\mathbf{x}}) = (-2, -2)^t, \quad \nabla g_i(\bar{\mathbf{x}}) = (4, 2)^t, \quad \text{and} \quad \nabla g_2(\bar{\mathbf{x}}) = (1, 2)^t.$$

Thus, $\nabla f(\bar{\mathbf{x}}) + u_1 \nabla g_1(\bar{\mathbf{x}}) + u_2 \nabla g_2(\bar{\mathbf{x}}) = 0$ holds true by letting $u_1 = 1/3$ and $u_2 = 2/3$, so that the KKT conditions are satisfied at $\bar{\mathbf{x}}$. Noting that f, g_1, and g_2 are convex, we have that $\bar{\mathbf{x}}$ is indeed an optimal solution by the consequent sufficiency of the KKT conditions.

Now, let us check whether the KKT conditions hold true at the point $\hat{\mathbf{x}} = (0, 0)^t$. Here, $I = \{3, 4\}$, so that we must have $u_1 = u_2 = 0$ to satisfy complementary slackness. Note that

$$\nabla f(\hat{\mathbf{x}}) = (-6, -4)^t, \quad \nabla g_3(\hat{\mathbf{x}}) = (-1, 0)^t, \quad \text{and} \quad \nabla g_4(\hat{\mathbf{x}}) = (0, -1)^t.$$

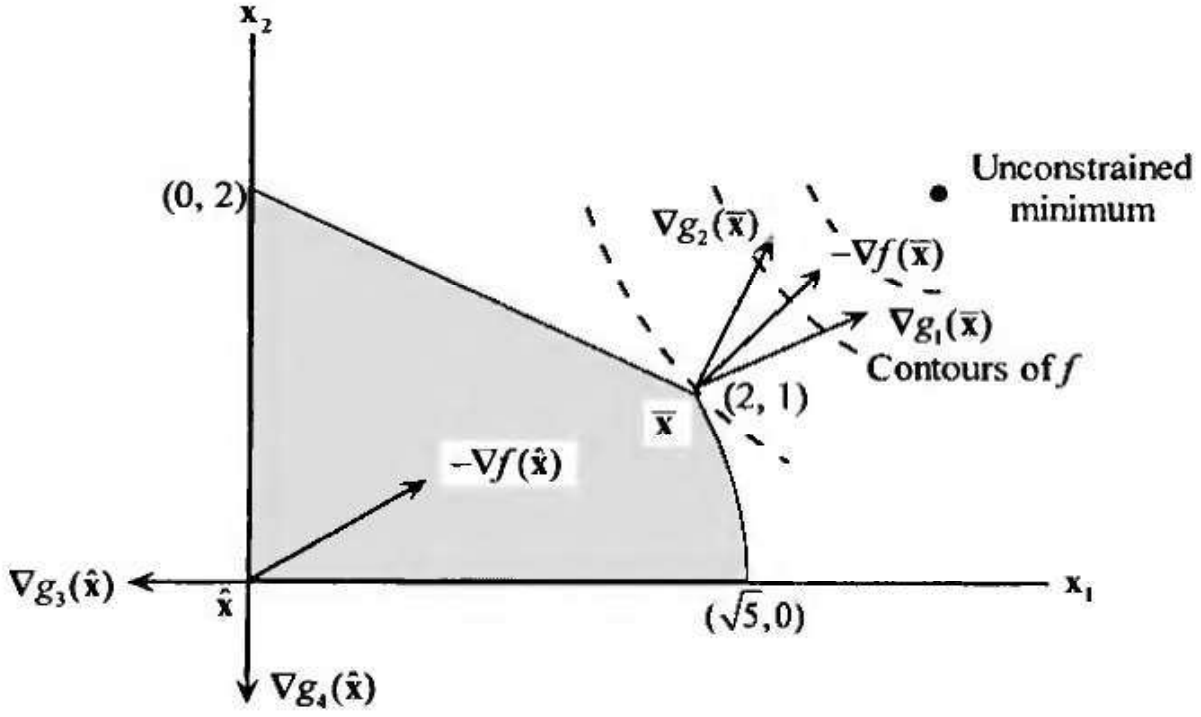


Figure B.6 The KKT conditions.

Thus, $\nabla f(\hat{\mathbf{x}}) + u_3 \nabla g_3(\hat{\mathbf{x}}) + u_4 \nabla g_4(\hat{\mathbf{x}}) = \mathbf{0}$ holds true only by letting $u_3 = -6$ and $u_4 = -4$, violating the nonnegativity of the Lagrange multipliers. This shows that $\hat{\mathbf{x}}$ is not a KKT point and hence could not be a candidate for an optimal solution.

In Figure B.6, the gradients of the objective function and the binding constraints are illustrated for both $\bar{\mathbf{x}}$ and $\hat{\mathbf{x}}$. Note that $-\nabla f(\bar{\mathbf{x}})$ lies in the cone spanned by the gradients of the binding constraints at $\bar{\mathbf{x}}$, whereas $-\nabla f(\hat{\mathbf{x}})$ does not lie in the corresponding cone. Indeed, the KKT conditions for a problem having inequality constraints could be interpreted geometrically as follows. A vector $\bar{\mathbf{x}}$ is a KKT point if and only if $-\nabla f(\bar{\mathbf{x}})$ lies in the cone spanned by the gradients of the binding constraints at $\bar{\mathbf{x}}$.

Let Problem P be as defined above, where all objective and constraint functions are continuously twice differentiable, and let $\bar{\mathbf{x}}$ be a KKT solution having associated Lagrange multipliers $(\bar{\mathbf{u}}, \bar{\mathbf{v}})$. Define the (restricted) Lagrangian function $L(\mathbf{x}) = f(\mathbf{x}) + \bar{\mathbf{u}}^t \mathbf{g}(\mathbf{x}) + \bar{\mathbf{v}}^t \mathbf{h}(\mathbf{x})$, and let $\nabla^2 L(\bar{\mathbf{x}})$ denote its Hessian at $\bar{\mathbf{x}}$. Let C denote the cone $\{\mathbf{d} : \nabla g_i(\bar{\mathbf{x}})^t \mathbf{d} = 0$ for all $i \in I^+$, $\nabla g_i(\bar{\mathbf{x}})^t \mathbf{d} \leq 0$ for all $i \in I^0$, and $\nabla h_i(\bar{\mathbf{x}})^t \mathbf{d} = 0$ for all $i = 1,...,\ell\}$, where $I^+ = \{i \in \{1,..., m\} : \bar{u}_i > 0\}$ and $I^0 = \{1,..., m\} - I^+$. Then we have the following *second-order sufficient conditions* holding true: If $\nabla^2 L(\bar{\mathbf{x}})$ is positive definite on C, that is, $\mathbf{d}^t \nabla^2 L(\bar{\mathbf{x}}) \mathbf{d} > 0$ for all $\mathbf{d} \in C$, $\mathbf{d} \neq \mathbf{0}$, then $\bar{\mathbf{x}}$ is a strict local minimum for Problem P. We also remark that if $\nabla^2 L(\mathbf{x})$ is positive semidefinite for all feasible $\mathbf{x}$ [respectively, for all feasible $\mathbf{x}$ in $N_\varepsilon(\bar{\mathbf{x}})$ for some $\varepsilon > 0$], then $\bar{\mathbf{x}}$ is a global (respectively, local) minimum for P.

Conversely, suppose that $\bar{\mathbf{x}}$ is a local minimum for P, and let the gradients $\nabla g_i(\bar{\mathbf{x}})$, $i \in I$, $\nabla h_i(\bar{\mathbf{x}})$, $i = 1,..., \ell$ be linearly independent, where $I = \{i \in \{1,.., m\} : g_i(\bar{\mathbf{x}}) = 0\}$. Define the cone C as stated above for the second-order sufficiency conditions. Then $\bar{\mathbf{x}}$ is a KKT point having associated Lagrange multipliers $(\bar{\mathbf{u}}, \bar{\mathbf{v}})$. Moreover, defining the (restricted) Lagrangian function $L(\mathbf{x}) = f(\mathbf{x}) + \bar{\mathbf{u}}^t \mathbf{g}(\mathbf{x}) + \bar{\mathbf{v}}^t \mathbf{h}(\mathbf{x})$, the *second-order necessary condition* is that $\nabla^2 L(\bar{\mathbf{x}})$ is positive semidefinite on C.

B.4 Lagrangian Duality

Given a nonlinear programming problem, called the *primal problem*, there exists a problem that is closely associated with it, called the *Lagrangian dual problem*. These two problems are given below.

$$\begin{aligned} \text{Primal Problem P: Minimize } & f(\mathbf{x}) \\ \text{subject to } & g_i(\mathbf{x}) \le 0 \qquad \text{for } i = 1,\ldots,m \\ & h_i(\mathbf{x}) = 0 \qquad \text{for } i = 1,\ldots,\ell \\ & \mathbf{x} \in X, \end{aligned}$$

where f, g_i, and h_i: $R^n \to R$ and X is a nonempty set in R^n. Let $\mathbf{g}$ and $\mathbf{h}$ be the m- and ℓ-vector functions whose ith components are, respectively, g_i and h_i.

$$\begin{aligned} \text{Lagrangian Dual Problem D: Maximize } & \theta(\mathbf{u}, \mathbf{v}) \\ \text{subject to } & \mathbf{u} \ge \mathbf{0}, \end{aligned}$$

where $\theta(\mathbf{u}, \mathbf{v}) = \inf\{f(\mathbf{x}) + \sum_{i=1}^{m} u_i g_i(\mathbf{x}) + \sum_{i=1}^{\ell} v_i h_i(\mathbf{x}) : \mathbf{x} \in X\}$. Here the vectors $\mathbf{u}$ and $\mathbf{v}$ belong to R^m and R^ℓ, respectively. The ith component u_i of $\mathbf{u}$ is referred to as the dual variable or Lagrange/Lagrangian multiplier associated with the constraint $g_i(\mathbf{x}) \le 0$, and the ith component v_i of $\mathbf{v}$ is referred to as the dual variable or Lagrange/Lagrangian multiplier associated with the constraint $h_i(\mathbf{x}) = 0$. It may be noted that θ is a *concave* function, even in the absence of any convexity or concavity assumptions on f, g_i, or h_i, or convexity of the set X.

We summarize below some important relationships between the primal and dual problems:

1. If $\mathbf{x}$ is feasible to Problem P and if $(\mathbf{u}, \mathbf{v})$ is feasible to Problem D, then $f(\mathbf{x}) \ge \theta(\mathbf{u}, \mathbf{v})$. Thus,

$$\inf\{f(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \le \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}, \mathbf{x} \in X\} \ge \sup\{\theta(\mathbf{u}, \mathbf{v}) : \mathbf{u} \ge \mathbf{0}\}.$$

   This result is called the *weak duality theorem.*
2. If $\sup\{\theta(\mathbf{u}, \mathbf{v}) : \mathbf{u} \ge \mathbf{0}\} = \infty$, then there exists no point $\mathbf{x} \in X$ such that $\mathbf{g}(\mathbf{x}) \le \mathbf{0}$ and $\mathbf{h}(\mathbf{x}) = \mathbf{0}$, so that the primal problem is infeasible.
3. If $\inf\{f(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \le \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}, \mathbf{x} \in X\} = -\infty$, then $\theta(\mathbf{u}, \mathbf{v}) = -\infty$ for each $(\mathbf{u}, \mathbf{v})$ with $\mathbf{u} \ge \mathbf{0}$.
4. If there exists a feasible $\mathbf{x}$ to the primal problem and a feasible $(\mathbf{u}, \mathbf{v})$ to the dual problem such that $f(\mathbf{x}) = \theta(\mathbf{u}, \mathbf{v})$, then $\mathbf{x}$ is an optimal solution to Problem P and $(\mathbf{u}, \mathbf{v})$ is an optimal solution to Problem D. Furthermore, the complementary slackness condition $u_i g_i(\mathbf{x}) = 0$ for $i = 1,\ldots, m$ holds true.
5. Suppose that X is convex, that f, g_i: $R^n \to R$ for $i = 1,\ldots, m$ are convex, and that $\mathbf{h}$ is of the form $\mathbf{h}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, where $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{b}$ is an m-vector. Under a suitable constraint qualification, the optimal objective values of Problems P and D are equal; that is,

$$\inf\{f(\mathbf{x}) : \mathbf{x} \in X, \mathbf{g}(\mathbf{x}) \le \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}\} = \sup\{\theta(\mathbf{u}, \mathbf{v}) : \mathbf{u} \ge \mathbf{0}\}.$$

Furthermore, if the inf is finite, then the sup is achieved at $(\bar{\mathbf{u}}, \bar{\mathbf{v}})$ with $\bar{\mathbf{u}} \geq \mathbf{0}$. Also, if the inf is achieved at $\bar{\mathbf{x}}$, then $u_i g_i(\bar{\mathbf{x}}) = 0$ for $i = 1,\ldots, m$. This result is referred to as the *strong duality theorem*.