Optimization: Optimization is a subject that is widely and increasingly used in science, engineering, economics, management, industry, and other areas.
It deals with selecting the best of many decisions in real-life problems, constructing computational methods to find the optimal solution of optimization problems, and studying the computational performances of numerical algorithms implemented best on the computational methods.
Optimization is used in
1. Data Science
2. Information science
3. Big data analysis
4. Mobile data analysis

· we consider the optimization problem

$$\begin{array}{ll} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \Omega. \end{array}$$

The function $f : \mathbb{R}^n \to \mathbb{R}$ that we wish to minimize is a real-valued function, and is called the *objective function*, or *cost function*. The vector $\boldsymbol{x}$ is an n-vector of independent variables, that is, $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^n$. The variables $x_1, \ldots, x_n$ are often referred to as *decision variables*. The set Ω is a subset of $\mathbb{R}^n$, called the *constraint set* or *feasible set*.

The above problem is a general form of a *constrained* optimization problem, because the decision variables are constrained to be in the constraint set Ω. If $\Omega = \mathbb{R}^n$, then we refer to the problem as an *unconstrained* optimization problem.

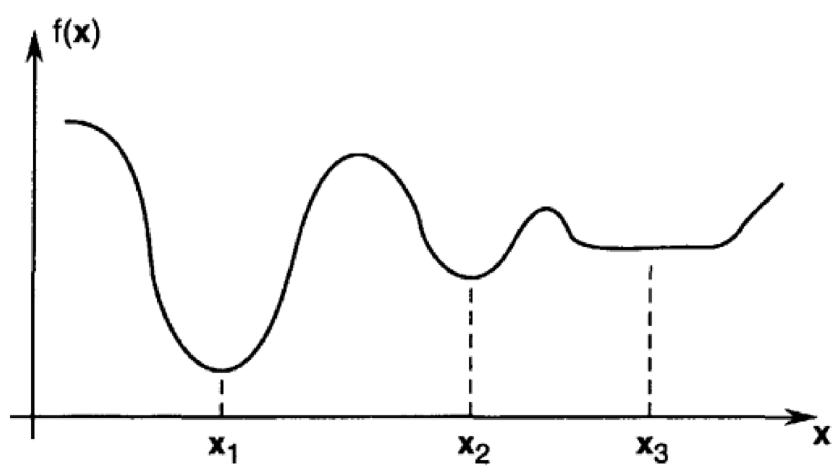

Figure 6.1 Examples of minimizers: $\boldsymbol{x}_1$: strict global minimizer; $\boldsymbol{x}_2$: strict local minimizer; $\boldsymbol{x}_3$: local (not strict) minimizer

Local minimizer. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a real-valued function defined on some set $\Omega \subset \mathbb{R}^n$. A point $\boldsymbol{x}^* \in \Omega$ is a *local minimizer* of f over Ω if there exists $\varepsilon > 0$ such that $f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$ for all $\boldsymbol{x} \in \Omega \setminus \{\boldsymbol{x}^*\}$ and $\|\boldsymbol{x} - \boldsymbol{x}^*\| < \varepsilon$.

Global minimizer. A point $\boldsymbol{x}^* \in \Omega$ is a *global minimizer* of f over Ω if $f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$ for all $\boldsymbol{x} \in \Omega \setminus \{\boldsymbol{x}^*\}$.

If, in the above definitions, we replace "$\geq$" with "$>$", then we have a *strict local minimizer* and a *strict global minimizer*, respectively.

In Figure 6.1, we graphically illustrate the above definitions for $n = 1$.

Given a real-valued function f, the notation $\arg\min f(\boldsymbol{x})$ denotes the argument that minimizes the function f (a point in the domain of f), assuming such a point is unique. For example, if $f : \mathbb{R} \to \mathbb{R}$ is given by $f(x) = (x+1)^2 + 3$, then $\arg\min f(x) = -1$. If we write $\arg\min_{\boldsymbol{x} \in \Omega}$, then we treat Ω as the domain of f. For example, for the function f above, $\arg\min_{x \geq 0} f(x) = 0$. In general, we can think of $\arg\min_{\boldsymbol{x} \in \Omega} f(\boldsymbol{x})$ as the global minimizer of f over Ω (assuming it exists and is unique).

Example 6.1 Let $f(x_1, x_2) = 5x_1 + 8x_2 + x_1x_2 - x_1^2 - 2x_2^2$. Then,

$$Df(\boldsymbol{x}) = (\nabla f(\boldsymbol{x}))^T = \left[\frac{\partial f}{\partial x_1}(\boldsymbol{x}), \frac{\partial f}{\partial x_2}(\boldsymbol{x})\right] = [5 + x_2 - 2x_1, 8 + x_1 - 4x_2],$$

and

$$\boldsymbol{F}(\boldsymbol{x}) = D^2 f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_2 \partial x_1}(\boldsymbol{x}) \\ \frac{\partial^2 f}{\partial x_1 \partial x_2}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_2^2}(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 1 & -4 \end{bmatrix}.$$

■

Definition 6.2 ***Feasible direction.*** A vector $\boldsymbol{d} \in \mathbb{R}^n$, $\boldsymbol{d} \neq \boldsymbol{0}$, is a *feasible direction* at $\boldsymbol{x} \in \Omega$ if there exists $\alpha_0 > 0$ such that $\boldsymbol{x} + \alpha\boldsymbol{d} \in \Omega$ for all $\alpha \in [0, \alpha_0]$. ■

Figure 6.2 illustrates the notion of feasible directions.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a real-valued function and let $\boldsymbol{d}$ be a feasible direction at $\boldsymbol{x} \in \Omega$. The *directional derivative of f in the direction* $\boldsymbol{d}$, denoted $\partial f / \partial \boldsymbol{d}$, is the real-valued function defined by

$$\frac{\partial f}{\partial \boldsymbol{d}}(\boldsymbol{x}) = \lim_{\alpha \to 0} \frac{f(\boldsymbol{x} + \alpha\boldsymbol{d}) - f(\boldsymbol{x})}{\alpha}.$$

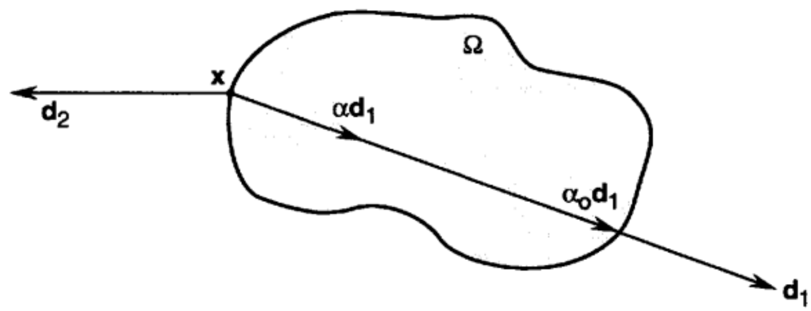

Figure 6.2 Two-dimensional illustration of feasible directions; $\boldsymbol{d}_1$ is a feasible direction, $\boldsymbol{d}_2$ is not a feasible direction

If $\|\boldsymbol{d}\| = 1$, then $\partial f/\partial \boldsymbol{d}$ is the rate of increase of f at $\boldsymbol{x}$ in the direction $\boldsymbol{d}$. To compute the above directional derivative, suppose that $\boldsymbol{x}$ and $\boldsymbol{d}$ are given. Then, $f(\boldsymbol{x} + \alpha\boldsymbol{d})$ is a function of α, and

$$\frac{\partial f}{\partial \boldsymbol{d}}(\boldsymbol{x}) = \frac{d}{d\alpha} f(\boldsymbol{x} + \alpha\boldsymbol{d})\bigg|_{\alpha=0}.$$

Applying the chain rule yields

$$\frac{\partial f}{\partial \boldsymbol{d}}(\boldsymbol{x}) = \frac{d}{d\alpha} f(\boldsymbol{x} + \alpha\boldsymbol{d})\bigg|_{\alpha=0} = \nabla f(\boldsymbol{x})^T\boldsymbol{d} = \langle \nabla f(\boldsymbol{x}), \boldsymbol{d}\rangle = \boldsymbol{d}^T \nabla f(\boldsymbol{x}).$$

In summary, if $\boldsymbol{d}$ is a unit vector, that is, $\|\boldsymbol{d}\| = 1$, then $\langle \nabla f(\boldsymbol{x}), \boldsymbol{d}\rangle$ is the rate of increase of f at the point $\boldsymbol{x}$ in the direction $\boldsymbol{d}$.

Example 6.2 Define $f : \mathbb{R}^3 \to \mathbb{R}$ by $f(\boldsymbol{x}) = x_1 x_2 x_3$, and let

$$\boldsymbol{d} = \left[\frac{1}{2}, \frac{1}{2}, \frac{1}{\sqrt{2}}\right]^T.$$

The directional derivative of f in the direction $\boldsymbol{d}$ is

$$\frac{\partial f}{\partial \boldsymbol{d}}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})^T\boldsymbol{d} = [x_2x_3, x_1x_3, x_1x_2]\begin{bmatrix} 1/2 \\ 1/2 \\ 1/\sqrt{2} \end{bmatrix} = \frac{x_2x_3 + x_1x_3 + \sqrt{2}x_1x_2}{2}.$$

Note that because $\|\boldsymbol{d}\| = 1$, the above is also the rate of increase of f at $\boldsymbol{x}$ in the direction $\boldsymbol{d}$. ∎

We are now ready to state and prove the following theorem.

Theorem 6.1 ***First-Order Necessary Condition (FONC).*** *Let Ω be a subset of $\mathbb{R}^n$ and $f \in C^1$ a real-valued function on Ω. If $\boldsymbol{x}^*$ is a local minimizer of f over Ω, then for any feasible direction $\boldsymbol{d}$ at $\boldsymbol{x}^*$, we have*

$$\boldsymbol{d}^T \nabla f(\boldsymbol{x}^*) \geq 0.$$

Proof. Define

$$\boldsymbol{x}(\alpha) = \boldsymbol{x}^* + \alpha \boldsymbol{d} \in \Omega.$$

Note that $\boldsymbol{x}(0) = \boldsymbol{x}^*$. Define the composite function

$$\phi(\alpha) = f(\boldsymbol{x}(\alpha)).$$

Then, by Taylor's theorem,

$$f(\boldsymbol{x}^* + \alpha \boldsymbol{d}) - f(\boldsymbol{x}^*) = \phi(\alpha) - \phi(0) = \phi'(0)\alpha + o(\alpha) = \alpha \boldsymbol{d}^T \nabla f(\boldsymbol{x}(0)) + o(\alpha),$$

where $\alpha \geq 0$ (recall the definition of $o(\alpha)$ ("little-oh of α") in Part I). Thus, if $\phi(\alpha) \geq \phi(0)$, that is, $f(\boldsymbol{x}^* + \alpha \boldsymbol{d}) \geq f(\boldsymbol{x}^*)$ for sufficiently small values of $\alpha > 0$ ($\boldsymbol{x}^*$ is a local minimizer), then we have to have $\boldsymbol{d}^T \nabla f(\boldsymbol{x}^*) \geq 0$ (see Exercise 5.7). ∎

The above theorem is graphically illustrated in Figure 6.3.

An alternative way to express the FONC is:

$$\frac{\partial f}{\partial \boldsymbol{d}}(\boldsymbol{x}^*) \geq 0$$

for all feasible directions $\boldsymbol{d}$. In other words, if $\boldsymbol{x}^*$ is a local minimizer, then the rate of increase of f at $\boldsymbol{x}^*$ in any feasible direction $\boldsymbol{d}$ in Ω is nonnegative. Using directional derivatives, an alternative proof of Theorem 6.1 is as follows. Suppose that $\boldsymbol{x}^*$ is a local minimizer. Then, for any feasible direction $\boldsymbol{d}$, there exists $\bar{\alpha} > 0$ such that for all $\alpha \in (0, \bar{\alpha})$,

$$f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}^* + \alpha \boldsymbol{d}).$$

Hence, for all $\alpha \in (0, \bar{\alpha})$, we have

$$\frac{f(\boldsymbol{x}^* + \alpha \boldsymbol{d}) - f(\boldsymbol{x}^*)}{\alpha} \geq 0.$$

Taking the limit as $\alpha \to 0$, we conclude that

$$\frac{\partial f}{\partial \boldsymbol{d}}(\boldsymbol{x}^*) \geq 0.$$

A special case of interest is when $\boldsymbol{x}^*$ is an interior point of Ω (see Section 4.4). In this case, any direction is feasible, and we have the following result.

Corollary 6.1 ***Interior case.*** *Let Ω be a subset of $\mathbb{R}^n$ and $f \in \mathcal{C}^1$ a real-valued function on Ω. If $\boldsymbol{x}^*$ is a local minimizer of f over Ω and if $\boldsymbol{x}^*$ is an interior point of Ω, then*

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

Example 6.3 Consider the problem

$$\begin{array}{ll} \text{minimize} & x_1^2 + 0.5x_2^2 + 3x_2 + 4.5 \\ \text{subject to} & x_1, x_2 \geq 0. \end{array}$$

Questions:

a. Is the first-order necessary condition (FONC) for a local minimizer satisfied at $\boldsymbol{x} = [1,3]^T$?

b. Is the FONC for a local minimizer satisfied at $\boldsymbol{x} = [0,3]^T$?

c. Is the FONC for a local minimizer satisfied at $\boldsymbol{x} = [1,0]^T$?

d. Is the FONC for a local minimizer satisfied at $\boldsymbol{x} = [0,0]^T$?

Answers: First, let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(\boldsymbol{x}) = x_1^2 + 0.5x_2^2 + 3x_2 + 4.5$, where $\boldsymbol{x} = [x_1, x_2]^T$. A plot of the level sets of f is shown in Figure 6.4.

a. At $\boldsymbol{x} = [1,3]^T$, we have $\nabla f(\boldsymbol{x}) = [2x_1, x_2 + 3]^T = [2,6]^T$. The point $\boldsymbol{x} = [1,3]^T$ is an interior point of $\Omega = \{\boldsymbol{x} : x_1 \geq 0, x_2 \geq 0\}$. Hence, the FONC requires $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$. The point $\boldsymbol{x} = [1,3]^T$ does not satisfy the FONC for a local minimizer.

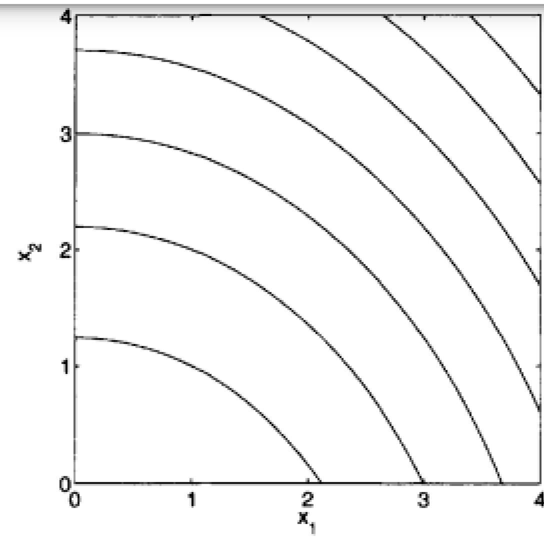

Figure 6.4 Level sets of the function in Example 6.3

b. At $\boldsymbol{x} = [0, 3]^T$, we have $\nabla f(\boldsymbol{x}) = [0, 6]^T$, and hence $\boldsymbol{d}^T \nabla f(\boldsymbol{x}) = 6d_2$, where $\boldsymbol{d} = [d_1, d_2]^T$. For $\boldsymbol{d}$ to be feasible at $\boldsymbol{x}$, we need $d_1 \geq 0$, and d_2 can take an arbitrary value in $\mathbb{R}$. The point $\boldsymbol{x} = [0, 3]^T$ does not satisfy the FONC for a minimizer because d_2 is allowed to be less than zero. For example, $\boldsymbol{d} = [1, -1]^T$ is a feasible direction, but $\boldsymbol{d}^T \nabla f(\boldsymbol{x}) = -6 < 0$.

c. At $\boldsymbol{x} = [1, 0]^T$, we have $\nabla f(\boldsymbol{x}) = [2, 3]^T$, and hence $\boldsymbol{d}^T \nabla f(\boldsymbol{x}) = 2d_1 + 3d_2$. For $\boldsymbol{d}$ to be feasible, we need $d_2 \geq 0$, and d_1 can take an arbitrary value in $\mathbb{R}$. For example, $\boldsymbol{d} = [-5, 1]^T$ is a feasible direction. But $\boldsymbol{d}^T \nabla f(\boldsymbol{x}) = -7 < 0$. Thus, $\boldsymbol{x} = [1, 0]^T$ does not satisfy the FONC for a local minimizer.

d. At $\boldsymbol{x} = [0, 0]^T$, we have $\nabla f(\boldsymbol{x}) = [0, 3]^T$, and hence $\boldsymbol{d}^T \nabla f(\boldsymbol{x}) = 3d_2$. For $\boldsymbol{d}$ to be feasible, we need $d_2 \geq 0$ and $d_1 \geq 0$. Hence, $\boldsymbol{x} = [0, 0]^T$ satisfies the FONC for a local minimizer.

∎

Example 6.4 Figure 6.5 shows a simplified model of a cellular wireless system (the distances shown have been scaled down to make the calculations simpler). A mobile user (also called a "mobile") is located at position x (see Figure 6.5).

There are two basestation antennas, one for the primary basestation and another for the neighboring basestation. Both antennas are transmitting signals to the mobile user, at equal power. However, the power of the received signal as measured by the mobile is the reciprocal of the squared distance from the associated antenna (primary or neighboring basestation). We are interested in finding the position of the mobile that maximizes the *signal-to-interference ratio*, which is the ratio of the received signal power from the primary basestation to the received signal power from the neighboring basestation.

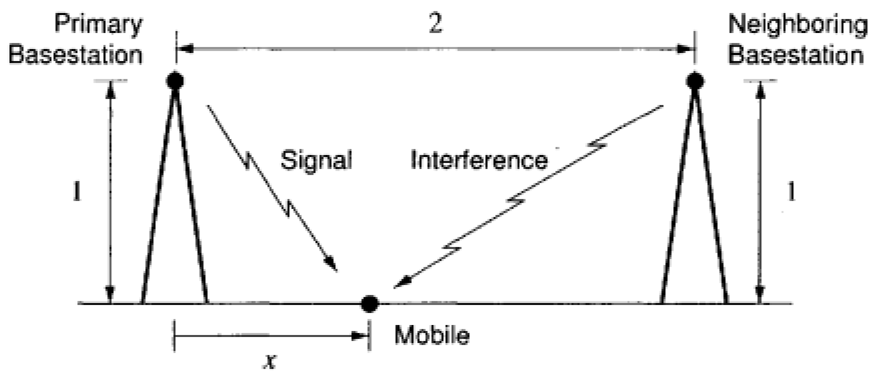

Figure 6.5 Simplified cellular wireless system in Example 6.4

We use the FONC to solve this problem. The squared distance from the mobile to the primary antenna is $1 + x^2$, while the squared distance from the mobile to the neighboring antenna is $1 + (2 - x)^2$. Therefore, the signal-to-interference ratio is

$$f(x) = \frac{1 + x^2}{1 + (2 - x)^2}.$$

We have

$$\begin{aligned} f'(x) &= \frac{-2x(1 + (2 - x)^2) - 2(2 - x)(1 + x^2)}{1 + (2 - x)^2} \\ &= \frac{4(x^2 - 2x - 1)}{1 + (2 - x)^2}. \end{aligned}$$

By the FONC, at the optimal position x^*, we have $f'(x^*) = 0$. Hence, either $x^* = 1 - \sqrt{2}$ or $x^* = 1 + \sqrt{2}$. Evaluating the objective function at these two candidate points, it easy to see that $x^* = 1 - \sqrt{2}$ is the optimal position. ∎

We now derive a second-order necessary condition that is satisfied by a local minimizer.

7.1 GOLDEN SECTION SEARCH

The search methods we discuss in this and the next section allow us to determine the minimizer of a function $f : \mathbb{R} \to \mathbb{R}$ over a closed interval, say $[a_0, b_0]$. The only property that we assume of the objective function f is that it is *unimodal*, which means that f has only one local minimizer. An example of such a function is depicted in Figure 7.1.

The methods we discuss are based on evaluating the objective function at different points in the interval $[a_0, b_0]$. We choose these points in such a way that an approximation to the minimizer of f may be achieved in as few evaluations as possible. Our goal is to progressively narrow the range until the minimizer is "boxed in" with sufficient accuracy.

Consider a unimodal function f of one variable and the interval $[a_0, b_0]$. If we evaluate f at only one intermediate point of the interval, we cannot narrow the range within which we know the minimizer is located. We have to evaluate f at two intermediate points, as illustrated in Figure 7.2. We choose the intermediate points in such a way that the reduction in the range is symmetric, in the sense that

$$a_1 - a_0 = b_0 - b_1 = \rho(b_0 - a_0),$$

where

$$\rho < \frac{1}{2}.$$

We then evaluate f at the intermediate points. If $f(a_1) < f(b_1)$, then the minimizer must lie in the range $[a_0, b_1]$. If, on the other hand, $f(a_1) \geq f(b_1)$, then the minimizer is located in the range $[a_1, b_0]$ (see Figure 7.3).

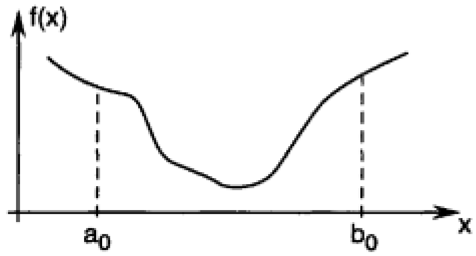

Figure 7.1 A unimodal function

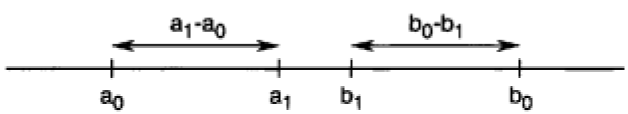

Figure 7.2 Evaluating the objective function at two intermediate points

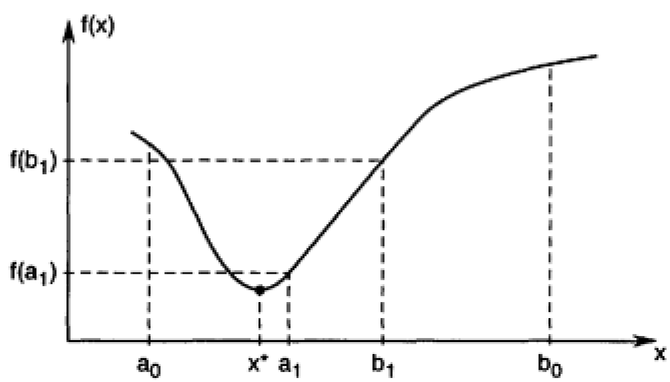

Figure 7.3 The case where $f(a_1) < f(b_1)$; the minimizer $x^* \in [a_0, b_1]$

Starting with the reduced range of uncertainty we can repeat the process and similarly find two new points, say a_2 and b_2, using the same value of $\rho < \frac{1}{2}$ as before. However, we would like to minimize the number of the objective function evaluations while reducing the width of the uncertainty interval. Suppose, for example, that $f(a_1) < f(b_1)$, as in Figure 7.3. Then, we know that $x^* \in [a_0, b_1]$. Because a_1 is already in the uncertainty interval and $f(a_1)$ is already known, we can make a_1 coincide with b_2. Thus, only one new evaluation of f at a_2 would be necessary. To find the value of ρ that results in only one new evaluation of f, see Figure 7.4. Without loss of generality, imagine that the original range $[a_0, b_0]$ is of unit length. Then, to have only one new evaluation of f it is enough to choose ρ so that

$$\rho(b_1 - a_0) = b_1 - b_2.$$

Because $b_1 - a_0 = 1 - \rho$ and $b_1 - b_2 = 1 - 2\rho$, we have

$$\rho(1 - \rho) = 1 - 2\rho.$$

We write the above quadratic function of ρ as

$$\rho^2 - 3\rho + 1 = 0.$$

The solutions are

$$\rho_1 = \frac{3 + \sqrt{5}}{2}, \qquad \rho_2 = \frac{3 - \sqrt{5}}{2}.$$

Because we require $\rho < \frac{1}{2}$, we take

$$\rho = \frac{3 - \sqrt{5}}{2} \approx 0.382.$$

Observe that

$$1 - \rho = \frac{\sqrt{5} - 1}{2},$$

and

$$\frac{\rho}{1 - \rho} = \frac{3 - \sqrt{5}}{\sqrt{5} - 1} = \frac{\sqrt{5} - 1}{2} = \frac{1 - \rho}{1},$$

that is,

$$\frac{\rho}{1 - \rho} = \frac{1 - \rho}{1}.$$

Thus, dividing a range in the ratio of ρ to $1 - \rho$ has the effect that the ratio of the shorter segment to the longer equals the ratio of the longer to the sum of the two. This rule was referred to by ancient Greek geometers as the *Golden Section*.

Using this Golden Section rule means that at every stage of the uncertainty range reduction (except the first one), the objective function f need only be evaluated at one new point. The uncertainty range is reduced by the ratio $1 - \rho \approx 0.61803$ at every stage. Hence, N steps of reduction using the Golden Section method reduces the range by the factor

$$(1 - \rho)^N \approx (0.61803)^N.$$

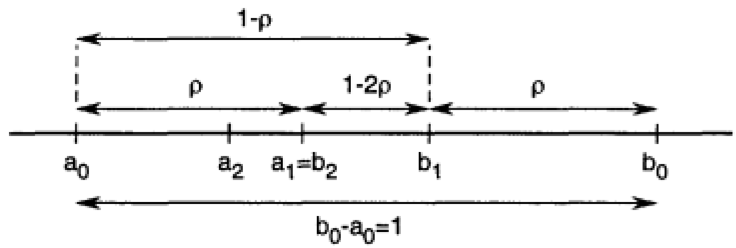

Figure 7.4 Finding value of ρ resulting in only one new evaluation of f

Example 7.1 Use the Golden Section search to find the value of x that minimizes

$$f(x) = x^4 - 14x^3 + 60x^2 - 70x$$

in the range $[0, 2]$ (this function comes from an example in [16]). Locate this value of x to within a range of 0.3.

After N stages the range $[0, 2]$ is reduced by $(0.61803)^N$. So, we choose N so that

$$(0.61803)^N \leq 0.3/2.$$

Four stages of reduction will do; that is, $N = 4$.

Iteration 1. We evaluate f at two intermediate points a_1 and b_1. We have

$$\begin{aligned} a_1 &= a_0 + \rho(b_0 - a_0) = 0.7639, \\ b_1 &= a_0 + (1 - \rho)(b_0 - a_0) = 1.236, \end{aligned}$$

where $\rho = (3 - \sqrt{5})/2$. We compute

$$\begin{aligned} f(a_1) &= -24.36, \\ f(b_1) &= -18.96. \end{aligned}$$

Thus, $f(a_1) < f(b_1)$, and so the uncertainty interval is reduced to

$$[a_0, b_1] = [0, 1.236].$$

Iteration 2. We choose b_2 to coincide with a_1, and f need only be evaluated at one new point

$$a_2 = a_0 + \rho(b_1 - a_0) = 0.4721.$$

We have

$$\begin{aligned} f(a_2) &= -21.10, \\ f(b_2) &= f(a_1) = -24.36. \end{aligned}$$

Now, $f(b_2) < f(a_2)$, so the uncertainty interval is reduced to

$$[a_2, b_1] = [0.4721, 1.236].$$

Iteration 3. We set $a_3 = b_2$, and compute b_3:

$$b_3 = a_2 + (1-\rho)(b_1 - a_2) = 0.9443.$$

We have

$$\begin{aligned} f(a_3) &= f(b_2) = -24.36, \\ f(b_3) &= -23.59. \end{aligned}$$

So $f(b_3) > f(a_3)$. Hence, the uncertainty interval is further reduced to

$$[a_2, b_3] = [0.4721, 0.9443].$$

Iteration 4. We set $b_4 = a_3$, and

$$a_4 = a_2 + \rho(b_3 - a_2) = 0.6525.$$

We have

$$\begin{aligned} f(a_4) &= -23.84, \\ f(b_4) &= f(a_3) = -24.36. \end{aligned}$$

Hence, $f(a_4) > f(b_4)$. Thus, the value of x that minimizes f is located in the interval

$$[a_4, b_3] = [0.6525, 0.9443].$$

Note that $b_3 - a_4 = 0.292 < 0.3$. ∎

8.1 INTRODUCTION

In this chapter, we consider a class of search methods for real-valued functions on $\mathbb{R}^n$. These methods use the gradient of the given function. In our discussion, we use terms like level sets, normal vectors, tangent vectors, and so on. These notions were discussed in some detail in Part I.

Recall that a level set of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the set of points $\boldsymbol{x}$ satisfying $f(\boldsymbol{x}) = c$ for some constant c. Thus, a point $\boldsymbol{x}_0 \in \mathbb{R}^n$ is on the level set corresponding to level c if $f(\boldsymbol{x}_0) = c$. In the case of functions of two real variables, $f : \mathbb{R}^2 \to \mathbb{R}$, the notion of the level set is illustrated in Figure 8.1.

The gradient of f at $\boldsymbol{x}_0$, denoted $\nabla f(\boldsymbol{x}_0)$, if it is not a zero vector, is orthogonal to the tangent vector to an arbitrary smooth curve passing through $\boldsymbol{x}_0$ on the level set $f(\boldsymbol{x}) = c$. Thus, the direction of maximum rate of increase of a real-valued differentiable function at a point is orthogonal to the level set of the function through that point. In other words, the gradient acts in such a direction that for a given small displacement, the function f increases more in the direction of the gradient than in any other direction. To prove this statement, recall that $\langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle$, $\|\boldsymbol{d}\| = 1$, is the rate of increase of f in the direction $\boldsymbol{d}$ at the point $\boldsymbol{x}$. By the Cauchy-Schwarz inequality,

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle \leq \|\nabla f(\boldsymbol{x})\|$$

because $\|\boldsymbol{d}\| = 1$. But if $\boldsymbol{d} = \nabla f(\boldsymbol{x}) / \|\nabla f(\boldsymbol{x})\|$, then

$$\langle \nabla f(\boldsymbol{x}), \frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|} \rangle = \|\nabla f(\boldsymbol{x})\|.$$

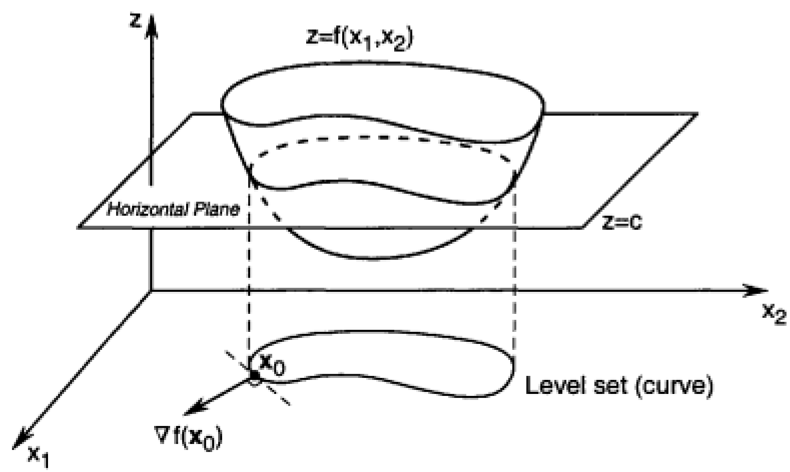

Figure 8.1 Constructing a level set corresponding to level c for f

Thus, the direction in which $\nabla f(\boldsymbol{x})$ points is the direction of maximum rate of increase of f at $\boldsymbol{x}$. The direction in which $-\nabla f(\boldsymbol{x})$ points is the direction of maximum rate of decrease of f at $\boldsymbol{x}$. Hence, the direction of negative gradient is a good direction to search if we want to find a function minimizer.

We proceed as follows. Let $\boldsymbol{x}^{(0)}$ be a starting point, and consider the point $\boldsymbol{x}^{(0)} - \alpha\nabla f(\boldsymbol{x}^{(0)})$. Then, by Taylor's theorem we obtain

$$f(\boldsymbol{x}^{(0)} - \alpha\nabla f(\boldsymbol{x}^{(0)})) = f(\boldsymbol{x}^{(0)}) - \alpha\|\nabla f(\boldsymbol{x}^{(0)})\|^2 + o(\alpha).$$

Thus, if $\nabla f(\boldsymbol{x}^{(0)}) \neq \mathbf{0}$, then for sufficiently small $\alpha > 0$, we have

$$f(\boldsymbol{x}^{(0)} - \alpha\nabla f(\boldsymbol{x}^{(0)})) < f(\boldsymbol{x}^{(0)}).$$

This means that the point $\boldsymbol{x}^{(0)} - \alpha\nabla f(\boldsymbol{x}^{(0)})$ is an improvement over the point $\boldsymbol{x}^{(0)}$ if we are searching for a minimizer.

To formulate an algorithm that implements the above idea, suppose that we are given a point $\boldsymbol{x}^{(k)}$. To find the next point $\boldsymbol{x}^{(k+1)}$, we start at $\boldsymbol{x}^{(k)}$ and move by an amount $-\alpha_k\nabla f(\boldsymbol{x}^{(k)})$, where α_k is a positive scalar called the *step size*. The above procedure leads to the following iterative algorithm:

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k\nabla f(\boldsymbol{x}^{(k)}).$$

We refer to the above as a *gradient descent algorithm* (or simply a *gradient algorithm*). The gradient varies as the search proceeds, tending to zero as we approach the minimizer. We have the option of either taking very small steps and re-evaluating the gradient at every step, or we can take large steps each time. The first approach results in a laborious method of reaching the minimizer, whereas the second approach may result in a more zigzag path to the minimizer. The advantage of the second approach is

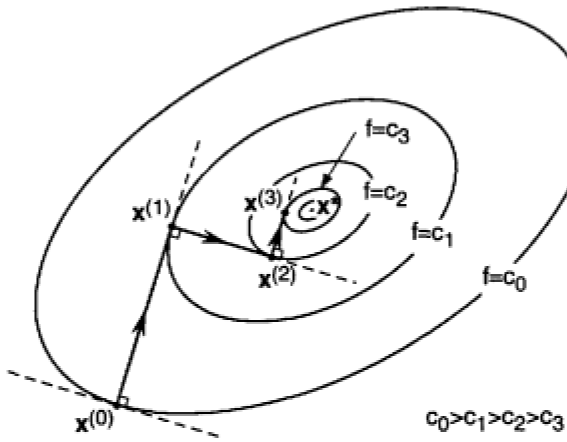

Figure 8.2 Typical sequence resulting from the method of steepest descent

a possibly fewer number of the gradient evaluations. Among many different methods that use the above philosophy the most popular is the method of *steepest descent*, which we discuss next.

Gradient methods are simple to implement and often perform well. For this reason, they are widely used in practical applications. For a discussion of applications of the steepest descent method to the computation of optimal controllers, we recommend [62, pp. 481–515]. In Chapter 13, we apply a gradient method to the training of a class of neural networks.

8.2 THE METHOD OF STEEPEST DESCENT

The method of steepest descent is a gradient algorithm where the step size α_k is chosen to achieve the maximum amount of decrease of the objective function at each individual step. Specifically, α_k is chosen to minimize $\phi_k(\alpha) \triangleq f(\boldsymbol{x}^{(k)} - \alpha \nabla f(\boldsymbol{x}^{(k)}))$. In other words,

$$\alpha_k = \underset{\alpha \geq 0}{\arg\min}\, f(\boldsymbol{x}^{(k)} - \alpha \nabla f(\boldsymbol{x}^{(k)})).$$

To summarize, the steepest descent algorithm proceeds as follows: at each step, starting from the point $\boldsymbol{x}^{(k)}$, we conduct a line search in the direction $-\nabla f(\boldsymbol{x}^{(k)})$ until a minimizer, $\boldsymbol{x}^{(k+1)}$, is found. A typical sequence resulting from the method of steepest descent is depicted in Figure 8.2.

Observe that the method of steepest descent moves in orthogonal steps, as stated in the following proposition.

Proposition 8.1 *If $\{x^{(k)}\}_{k=0}^{\infty}$ is a steepest descent sequence for a given function $f : \mathbb{R}^n \to \mathbb{R}$, then for each k the vector $x^{(k+1)} - x^{(k)}$ is orthogonal to the vector $x^{(k+2)} - x^{(k+1)}$.* □

Proof. From the iterative formula of the method of steepest descent it follows that

$$\langle x^{(k+1)} - x^{(k)}, x^{(k+2)} - x^{(k+1)} \rangle = \alpha_k \alpha_{k+1} \langle \nabla f(x^{(k)}), \nabla f(x^{(k+1)}) \rangle.$$

To complete the proof it is enough to show that

$$\langle \nabla f(x^{(k)}), \nabla f(x^{(k+1)}) \rangle = 0.$$

To this end, observe that α_k is a nonnegative scalar that minimizes $\phi_k(\alpha) \triangleq f(x^{(k)} - \alpha \nabla f(x^{(k)}))$. Hence, using the FONC and the chain rule,

$$\begin{aligned} 0 &= \phi_k'(\alpha_k) \\ &= \frac{d\phi_k}{d\alpha}(\alpha_k) \\ &= \nabla f(x^{(k)} - \alpha_k \nabla f(x^{(k)}))^T (-\nabla f(x^{(k)})) \\ &= -\langle \nabla f(x^{(k+1)}), \nabla f(x^{(k)}) \rangle \end{aligned}$$

and the proof is completed. ■

The above proposition implies that $\nabla f(x^{(k)})$ is parallel to the tangent plane to the level set $\{f(x) = f(x^{(k+1)})\}$ at $x^{(k+1)}$. Note that as each new point is generated by the steepest descent algorithm, the corresponding value of the function f decreases in value, as stated below.

Consider the problem

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} x^T Q x \\ \text{subject to} \quad & Ax = b, \end{aligned}$$

where $Q > 0$, $A \in \mathbb{R}^{m \times n}$, $m < n$, $\operatorname{rank} A = m$. This problem is a special case of what is called a *quadratic programming* problem (the general form of a quadratic programming problem includes the constraint $x \geq 0$). Note that the constraint set contains an infinite number of points (see Section 2.3). We now show,

21.2 CONVEX FUNCTIONS

We begin with a definition of the graph of a real-valued function.

Definition 21.1 The *graph* of $f : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}^n$, is the set of points in $\Omega \times \mathbb{R} \subset \mathbb{R}^{n+1}$ given by

$$\{[\boldsymbol{x}, f(\boldsymbol{x})]^T : \boldsymbol{x} \in \Omega\}.$$

∎

We can visualize the graph of f as simply the set of points on a "plot" of $f(\boldsymbol{x})$ versus $\boldsymbol{x}$ (see Figure 21.4). We next define the "epigraph" of a real-valued function.

Definition 21.2 The *epigraph* of a function $f : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}^n$, denoted $\text{epi}(f)$, is the set of points in $\Omega \times \mathbb{R}$ given by

$$\text{epi}(f) = \{[\boldsymbol{x}, \beta]^T : \boldsymbol{x} \in \Omega, \beta \in \mathbb{R}, \beta \geq f(\boldsymbol{x})\}.$$

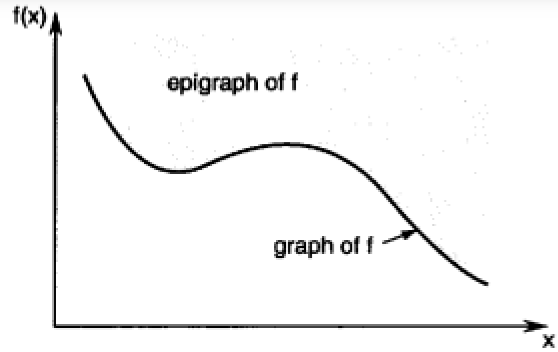

Figure 21.4 The graph and epigraph of a function $f : \mathbb{R} \to \mathbb{R}$

The epigraph $\text{epi}(f)$ of a function f is simply the set of points in $\Omega \times \mathbb{R}$ on or above the graph of f (see Figure 21.4). We can also think of $\text{epi}(f)$ as a subset of $\mathbb{R}^{n+1}$.

Recall that a set $\Omega \subset \mathbb{R}^n$ is convex if for every $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \Omega$ and $\alpha \in (0,1)$, $\alpha\boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2 \in \Omega$ (see Section 4.3). We now introduce the notion of a "convex function."

Definition 21.3 A function $f : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}^n$, is *convex* on Ω if its epigraph is a convex set. ∎

Theorem 21.1 *If a function $f : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}^n$, is convex on Ω, then Ω is a convex set.* □

Proof. We prove this theorem by contraposition. Suppose that Ω is not a convex set. Then, there exist two points $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ such that for some $\alpha \in (0,1)$,

$$\boldsymbol{z} = \alpha\boldsymbol{y}_1 + (1-\alpha)\boldsymbol{y}_2 \notin \Omega.$$

Let

$$\beta_1 = f(\boldsymbol{y}_1), \quad \beta_2 = f(\boldsymbol{y}_2).$$

Then, the pairs

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \beta_1 \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{y}_2 \\ \beta_2 \end{bmatrix}$$

belong to the graph of f, and hence also the epigraph of f. Let

$$\boldsymbol{w} = \alpha \begin{bmatrix} \boldsymbol{y}_1 \\ \beta_1 \end{bmatrix} + (1-\alpha) \begin{bmatrix} \boldsymbol{y}_2 \\ \beta_2 \end{bmatrix}.$$

We have

$$\boldsymbol{w} = \begin{bmatrix} \boldsymbol{z} \\ \alpha\beta_1 + (1-\alpha)\beta_2 \end{bmatrix}.$$

But note that $\boldsymbol{w} \notin \text{epi}(f)$, because $\boldsymbol{z} \notin \Omega$. Therefore, $\text{epi}(f)$ is not convex, and hence f is not a convex function. ∎

The next theorem gives a very useful characterization of convex functions. This characterization is often used as a definition for a convex function.

Proposition 21.1 *A quadratic form $f : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}^n$, given by $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}$, $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$, $\boldsymbol{Q} = \boldsymbol{Q}^T$, is convex on Ω if and only if for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$, $(\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{Q} (\boldsymbol{x} - \boldsymbol{y}) \geq 0$.* □

Proof. The result follows from Theorem 21.2. Indeed, the function $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}$ is convex if and only if for every $\alpha \in (0, 1)$, and every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ we have

$$f(\alpha \boldsymbol{x} + (1 - \alpha) \boldsymbol{y}) \leq \alpha f(\boldsymbol{x}) + (1 - \alpha) f(\boldsymbol{y}),$$

or equivalently

$$\alpha f(\boldsymbol{x}) + (1 - \alpha) f(\boldsymbol{y}) - f(\alpha \boldsymbol{x} + (1 - \alpha) \boldsymbol{y}) \geq 0.$$

Substituting for f into the left-hand side of the above equation yields

$$\begin{aligned}
\alpha \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} &+ (1 - \alpha) \boldsymbol{y}^T \boldsymbol{Q} \boldsymbol{y} - (\alpha \boldsymbol{x} + (1 - \alpha) \boldsymbol{y})^T \boldsymbol{Q} (\alpha \boldsymbol{x} + (1 - \alpha) \boldsymbol{y}) \\
&= \alpha \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{y}^T \boldsymbol{Q} \boldsymbol{y} - \alpha \boldsymbol{y}^T \boldsymbol{Q} \boldsymbol{y} - \alpha^2 \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} \\
&\quad - (2\alpha - 2\alpha^2) \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{y} - (1 - 2\alpha + \alpha^2) \boldsymbol{y}^T \boldsymbol{Q} \boldsymbol{y} \\
&= \alpha (1 - \alpha) \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} - 2\alpha (1 - \alpha) \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{y} + \alpha (1 - \alpha) \boldsymbol{y}^T \boldsymbol{Q} \boldsymbol{y} \\
&= \alpha (1 - \alpha) (\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{Q} (\boldsymbol{x} - \boldsymbol{y}).
\end{aligned}$$

Therefore, f is convex if and only if

$$\alpha (1 - \alpha) (\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{Q} (\boldsymbol{x} - \boldsymbol{y}) \geq 0,$$

which proves the result. ■

Example 21.5 In the previous example, $f(\boldsymbol{x}) = x_1 x_2$, which can be written as $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}$, where

$$\boldsymbol{Q} = \frac{1}{2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Let $\Omega = \{\boldsymbol{x} : \boldsymbol{x} \geq \boldsymbol{0}\}$, and $\boldsymbol{x} = [2, 2]^T \in \Omega$, $\boldsymbol{y} = [1, 3]^T \in \Omega$. We have

$$\boldsymbol{y} - \boldsymbol{x} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

and

$$(\boldsymbol{y} - \boldsymbol{x})^T \boldsymbol{Q} (\boldsymbol{y} - \boldsymbol{x}) = \frac{1}{2} [-1, 1] \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = -1 < 0.$$

Hence, by the above theorem, f is not convex on Ω. ■

Differentiable convex functions can be characterized using the following theorem.

19.4 LAGRANGE CONDITION

In this section, we present a first-order necessary condition for extremum problems with constraints. The result is the well-known *Lagrange's theorem*. To better understand the idea underlying this theorem, we first consider functions of two variables and only one equality constraint. Let $h : \mathbb{R}^2 \to \mathbb{R}$ be the constraint function. Recall that at each point $\boldsymbol{x}$ of the domain, the gradient vector $\nabla h(\boldsymbol{x})$ is orthogonal to the level set that passes through that point. Indeed, let us choose a point $\boldsymbol{x}^* = [x_1^*, x_2^*]^T$ such that $h(\boldsymbol{x}^*) = 0$, and assume $\nabla h(\boldsymbol{x}^*) \neq \mathbf{0}$. The level set through the point $\boldsymbol{x}^*$ is the set $\{\boldsymbol{x} : h(\boldsymbol{x}) = 0\}$. We then parameterize this level set in a neighborhood of $\boldsymbol{x}^*$ by a curve $\{\boldsymbol{x}(t)\}$, that is, a continuously differentiable vector function $\boldsymbol{x} : \mathbb{R} \to \mathbb{R}^2$ such that

$$\boldsymbol{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}, \quad t \in (a,b), \quad \boldsymbol{x}^* = \boldsymbol{x}(t^*), \quad \dot{\boldsymbol{x}}(t^*) \neq \boldsymbol{0}, \quad t^* \in (a,b).$$

We can now show that $\nabla h(\boldsymbol{x}^*)$ is orthogonal to $\dot{\boldsymbol{x}}(t^*)$. Indeed, because h is constant on the curve $\{\boldsymbol{x}(t) : t \in (a,b)\}$, we have that for all $t \in (a,b)$,

$$h(\boldsymbol{x}(t)) = 0.$$

Hence, for all $t \in (a,b)$,

$$\frac{d}{dt} h(\boldsymbol{x}(t)) = 0.$$

Applying the chain rule, we get

$$\frac{d}{dt} h(\boldsymbol{x}(t)) = \nabla h(\boldsymbol{x}(t))^T \dot{\boldsymbol{x}}(t) = 0.$$

Therefore, $\nabla h(\boldsymbol{x}^*)$ is orthogonal to $\dot{\boldsymbol{x}}(t^*)$.

Now suppose that $\boldsymbol{x}^*$ is a minimizer of $f : \mathbb{R}^2 \to \mathbb{R}$ on the set $\{\boldsymbol{x} : h(\boldsymbol{x}) = 0\}$. We claim that $\nabla f(\boldsymbol{x}^*)$ is orthogonal to $\dot{\boldsymbol{x}}(t^*)$. To see this, it is enough to observe that the composite function of t given by

$$\phi(t) = f(\boldsymbol{x}(t))$$

achieves a minimum at t^*. Consequently, the first-order necessary condition for the unconstrained extremum problem implies

$$\frac{d\phi}{dt}(t^*) = 0.$$

Applying the chain rule yields

$$0 = \frac{d}{dt}\phi(t^*) = \nabla f(\boldsymbol{x}(t^*))^T \dot{\boldsymbol{x}}(t^*) = \nabla f(\boldsymbol{x}^*)^T \dot{\boldsymbol{x}}(t^*).$$

Thus, $\nabla f(\boldsymbol{x}^*)$ is orthogonal to $\dot{\boldsymbol{x}}(t^*)$. The fact that $\dot{\boldsymbol{x}}(t^*)$ is tangent to the curve $\{\boldsymbol{x}(t)\}$ at $\boldsymbol{x}^*$ means that $\nabla f(\boldsymbol{x}^*)$ is orthogonal to the curve at $\boldsymbol{x}^*$ (see Figure 19.10).

Recall that $\nabla h(\boldsymbol{x}^*)$ is also orthogonal to $\dot{\boldsymbol{x}}(t^*)$. Therefore, the vectors $\nabla h(\boldsymbol{x}^*)$ and $\nabla f(\boldsymbol{x}^*)$ are "parallel", that is, $\nabla f(\boldsymbol{x}^*)$ is a scalar multiple of $\nabla h(\boldsymbol{x}^*)$. The above observations allow us now to formulate *Lagrange's theorem* for functions of two variables with one constraint.

Theorem 19.2 ***Lagrange's Theorem for*** $n = 2$, $m = 1$. *Let the point $\boldsymbol{x}^*$ be a minimizer of $f : \mathbb{R}^2 \to \mathbb{R}$ subject to the constraint $h(\boldsymbol{x}) = 0$, $h : \mathbb{R}^2 \to \mathbb{R}$. Then, $\nabla f(\boldsymbol{x}^*)$ and $\nabla h(\boldsymbol{x}^*)$ are parallel. That is, if $\nabla h(\boldsymbol{x}^*) \neq \boldsymbol{0}$, then there exists a scalar λ^* such that*

$$\nabla f(\boldsymbol{x}^*) + \lambda^* \nabla h(\boldsymbol{x}^*) = \boldsymbol{0}.$$

Theorem 19.3 ***Lagrange's Theorem.*** *Let $\boldsymbol{x}^*$ be a local minimizer (or maximizer) of $f : \mathbb{R}^n \to \mathbb{R}$, subject to $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$, $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^m$, $m \leq n$. Assume that $\boldsymbol{x}^*$ is a regular point. Then, there exists $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ such that*

$$Df(\boldsymbol{x}^*) + \boldsymbol{\lambda}^{*T} D\boldsymbol{h}(\boldsymbol{x}^*) = \boldsymbol{0}^T.$$

Example 19.6 Consider the problem of extremizing the objective function

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

on the ellipse

$$\{[x_1, x_2]^T : h(\boldsymbol{x}) = x_1^2 + 2x_2^2 - 1 = 0\}.$$

We have

$$\begin{aligned} \nabla f(\boldsymbol{x}) &= [2x_1, 2x_2]^T, \\ \nabla h(\boldsymbol{x}) &= [2x_1, 4x_2]^T. \end{aligned}$$

Thus,

$$D_x l(\boldsymbol{x}, \lambda) = D_x[f(\boldsymbol{x}) + \lambda h(\boldsymbol{x})] = [2x_1 + 2\lambda x_1, 2x_2 + 4\lambda x_2],$$

and

$$D_\lambda l(\boldsymbol{x}, \lambda) = h(\boldsymbol{x}) = x_1^2 + 2x_2^2 - 1.$$

Setting $D_x l(\boldsymbol{x}, \lambda) = \boldsymbol{0}^T$ and $D_\lambda l(\boldsymbol{x}, \lambda) = 0$ we obtain three equations in three unknowns

$$\begin{aligned} 2x_1 + 2\lambda x_1 &= 0 \\ 2x_2 + 4\lambda x_2 &= 0 \\ x_1^2 + 2x_2^2 &= 1. \end{aligned}$$

All feasible points in this problem are regular. From the first of the above equations, we get either $x_1 = 0$ or $\lambda = -1$. For the case where $x_1 = 0$, the second and third equations imply that $\lambda = -1/2$ and $x_2 = \pm 1/\sqrt{2}$. For the case where $\lambda = -1$, the second and third equations imply that $x_1 = \pm 1$ and $x_2 = 0$. Thus, the points that satisfy the Lagrange condition for extrema are

$$\boldsymbol{x}^{(1)} = \begin{bmatrix} 0 \\ 1/\sqrt{2} \end{bmatrix}, \quad \boldsymbol{x}^{(2)} = \begin{bmatrix} 0 \\ -1/\sqrt{2} \end{bmatrix}, \quad \boldsymbol{x}^{(3)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{x}^{(4)} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}.$$

Because

$$f(\boldsymbol{x}^{(1)}) = f(\boldsymbol{x}^{(2)}) = \frac{1}{2}$$

and

$$f(\boldsymbol{x}^{(3)}) = f(\boldsymbol{x}^{(4)}) = 1$$

we conclude that if there are minimizers, then they are located at $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$, and if there are maximizers, then they are located at $\boldsymbol{x}^{(3)}$ and $\boldsymbol{x}^{(4)}$. It turns out that, indeed, $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ are minimizers and $\boldsymbol{x}^{(3)}$ and $\boldsymbol{x}^{(4)}$ are maximizers. This problem can be solved graphically, as illustrated in Figure 19.14. ∎

In the above example both the objective function f and the constraint function h are quadratic functions. In the next example, we take a closer look at a class

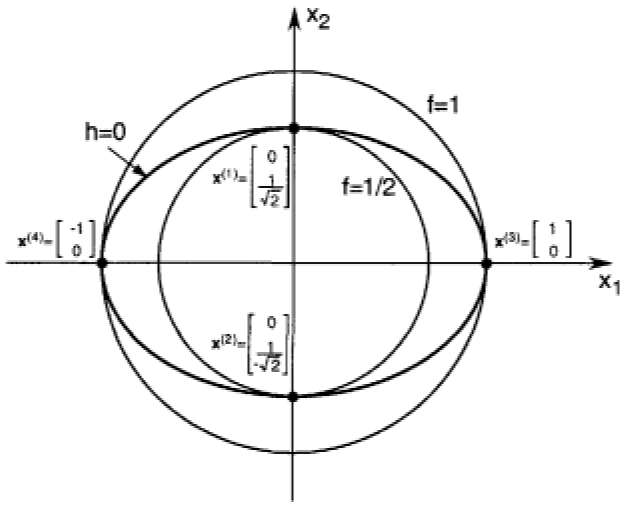

Figure 19.14 Graphical solution of the problem in Example 19.6

of problems where both the objective function f and the constraint h are quadratic functions of n variables.