Chapter 7
Basic Properties of Solutions and Algorithms

In this chapter we consider optimization problems of the form

$$\begin{aligned} &\text{minimize} && f(\mathbf{x}) \\ &\text{subject to} && \mathbf{x} \in \Omega, \end{aligned} \tag{7.1}$$

where f is a real-valued function and Ω, the feasible set, is a subset of E^n. Throughout most of the chapter attention is restricted to the case where $\Omega = E^n$, corresponding to the completely unconstrained case, but sometimes we consider cases where Ω is some particularly simple subset of E^n.

The first and third sections of the chapter characterize the first- and second-order conditions that must hold at a solution point of (7.1). These conditions are simply extensions to E^n of the well-known derivative conditions for a function of a single variable that hold at a maximum or a minimum point. The fourth and fifth sections of the chapter introduce the important classes of convex and concave functions that provide zeroth-order conditions as well as a natural formulation for a global theory of optimization and provide geometric interpretations of the derivative conditions derived in the first two sections.

The final sections of the chapter are devoted to basic convergence characteristics of algorithms. Although this material is not exclusively applicable to optimization problems but applies to general iterative algorithms for solving other problems as well, it can be regarded as a fundamental prerequisite for a modern treatment of optimization techniques. Two essential questions are addressed concerning iterative algorithms. The first question, which is qualitative in nature, is whether a given algorithm in some sense yields, at least in the limit, a solution to the original problem. This question is treated in Sect. 7.6, and conditions sufficient to guarantee appropriate convergence are established. The second question, the more quantitative one, is related to how fast the algorithm converges to a solution. This question is defined more precisely in Sect. 7.7. Several special types of convergence, which arise frequently in the development of algorithms for optimization, are explored.

D.G. Luenberger, Y. Ye, *Linear and Nonlinear Programming*, International Series in Operations Research & Management Science 228,
DOI 10.1007/978-3-319-18842-3_7

7.1 First-Order Necessary Conditions

Perhaps the first question that arises in the study of the minimization problem (7.1) is whether a solution exists. The main result that can be used to address this issue is the theorem of Weierstras, which states that if f is continuous and Ω is compact, a solution exists (see Appendix A.6). This is a valuable result that should be kept in mind throughout our development; however, our primary concern is with characterizing solution points and devising effective methods for finding them.

In an investigation of the general problem (7.1) we distinguish two kinds of solution points: *local minimum points*, and *global minimum points*.

Definition. A point $\mathbf{x}^* \in \Omega$ is said to be a *relative minimum point* or a *local minimum point* of f over Ω if there is an $\varepsilon > 0$ such that $f(\mathbf{x}) \geqslant f(\mathbf{x}^*)$ for all $\mathbf{x} \in \Omega$ within a distance ε of $\mathbf{x}^*$ (that is, $\mathbf{x} \in \Omega$ and $|\mathbf{x} - \mathbf{x}^*| < \varepsilon$). If $f(\mathbf{x}) > f(\mathbf{x}^*)$ for all $\mathbf{x} \in \Omega$, $\mathbf{x} \neq \mathbf{x}^*$, within a distance ε of $\mathbf{x}^*$, then $\mathbf{x}^*$ is said to be a *strict relative minimum point* of f over Ω.

Definition. A point $\mathbf{x}^* \in \Omega$ is said to be a *global minimum point* of f over Ω if $f(\mathbf{x}) \geqslant f(\mathbf{x}^*)$ for all $\mathbf{x} \in \Omega$. If $f(\mathbf{x}) > f(\mathbf{x}^*)$ for all $\mathbf{x} \in \Omega$, $\mathbf{x} \neq \mathbf{x}^*$, then $\mathbf{x}^*$ is said to be a *strict global minimum point* of f over Ω.

In formulating and attacking problem (7.1) we are, by definition, explicitly asking for a global minimum point of f over the set Ω. Practical reality, however, both from the theoretical and computational viewpoint, dictates that we must in many circumstances be content with a relative minimum point. In deriving necessary conditions based on the differential calculus, for instance, or when searching for the minimum point by a convergent stepwise procedure, comparisons of the values of nearby points is all that is possible and attention focuses on relative minimum points. Global conditions and global solutions can, as a rule, only be found if the problem possesses certain convexity properties that essentially guarantee that any relative minimum is a global minimum. Thus, in formulating and attacking problem (7.1) we shall, by the dictates of practicality, usually consider, implicitly, that we are asking for a relative minimum point. If appropriate conditions hold, this will also be a global minimum point.

Feasible Directions

To derive necessary conditions satisfied by a relative minimum point $\mathbf{x}^*$, the basic idea is to consider movement away from the point in some given direction. Along any given direction the objective function can be regarded as a function of a single variable, the parameter defining movement in this direction, and hence the ordinary calculus of a single variable is applicable. Thus given $\mathbf{x} \in \Omega$ we are motivated to say that a vector $\mathbf{d}$ is a *feasible direction at* $\mathbf{x}$ if there is an $\bar{\alpha} > 0$ such that $\mathbf{x} + \alpha\mathbf{d} \in \Omega$ for all α, $0 \leqslant \alpha \leqslant \bar{\alpha}$. With this simple concept we can state some simple conditions satisfied by relative minimum points.

Proposition 1 (First-Order Necessary Conditions). *Let Ω be a subset of E^n and let $f \in C^1$ be a function on Ω. If $\mathbf{x}^*$ is a relative minimum point of f over Ω, then for any $\mathbf{d} \in E^n$ that is a feasible direction at $\mathbf{x}^*$, we have $\nabla f(\mathbf{x}^*)\mathbf{d} \geqslant 0$.*

Proof. For any α, $0 \leqslant \alpha \leqslant \bar{\alpha}$, the point $\mathbf{x}(\alpha) = \mathbf{x}^* + \alpha\mathbf{d} \in \Omega$. For $0 \leqslant \alpha \leqslant \bar{\alpha}$ define the function $g(\alpha) = f(\mathbf{x}(\alpha))$. Then g has a relative minimum at $\alpha = 0$. A typical g is shown in Fig. 7.1. By the ordinary calculus we have

$$g(\alpha) - g(0) = g'(0)\alpha + o(\alpha), \tag{7.2}$$

where $o(\alpha)$ denotes terms that go to zero faster than α (see Appendix A). If $g'(0) < 0$ then, for sufficiently small values of $\alpha > 0$, the right side of (7.2) will be negative, and hence $g(\alpha) - g(0) < 0$, which contradicts the minimal nature of $g(0)$. Thus $g'(0) = \nabla f(\mathbf{x}^*)\mathbf{d} \geqslant 0$. ∎

A very important special case is where $\mathbf{x}^*$ is in the interior of Ω (as would be the case if $\Omega = E^n$). In this case there are feasible directions emanating in every direction from $\mathbf{x}^*$, and hence $\nabla f(\mathbf{x}^*)\mathbf{d} \geqslant 0$ for all $\mathbf{d} \in E^n$. This implies $\nabla f(\mathbf{x}^*) = 0$. We state this important result as a corollary.

Corollary (Unconstrained Case). *Let Ω be a subset of E^n, and let $f \in C^1$ be function' on Ω. If $\mathbf{x}^*$ is a relative minimum point of f over Ω and if $\mathbf{x}^*$ is an interior point of Ω, then $\nabla f(\mathbf{x}^*) = 0$.*

The necessary conditions in the pure unconstrained case lead to n equations (one for each component of ∇f) in n unknowns (the components of $\mathbf{x}^*$), which in many cases can be solved to determine the solution. In practice, however, as demonstrated in the following chapters, an optimization problem is solved directly without explicitly attempting to solve the equations arising from the necessary conditions. Nevertheless, these conditions form a foundation for the theory.

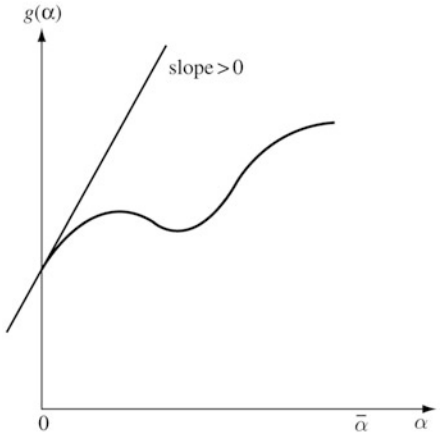

Fig. 7.1 Construction for proof

Example 1. Consider the problem

$$\text{minimize } f(x_1,\ x_2) = x_1^2 - x_1x_2 + x_2^2 - 3x_2.$$

There are no constraints, so $\Omega = E^2$. Setting the partial derivatives of f equal to zero yields the two equations

$$\begin{aligned} 2x_1 - x_2 &= 0 \\ -x_1 + 2x_2 &= 3. \end{aligned}$$

These have the unique solution $x_1 = 1,\ x_2 = 2$, which is a global minimum point of f.

Example 2. Consider the problem

$$\begin{aligned} \text{minimize} \quad & f(x_1,\ x_2) = x_1^2 - x_1 + x_2 + x_1x_2 \\ \text{subject to} \quad & x_1 \geqslant 0, \qquad x_2 \geqslant 0. \end{aligned}$$

This problem has a global minimum at $x_1 = \frac{1}{2},\ x_2 = 0$. At this point

$$\begin{aligned} \frac{\partial f}{\partial x_1} &= 2x_1 - 1 + x_2 = 0 \\ \frac{\partial f}{\partial x_2} &= 1 + x_1 = \frac{3}{2}. \end{aligned}$$

Thus, the partial derivatives do not both vanish at the solution, but since any feasible direction must have an x_2 component greater than or equal to zero, we have $\nabla f(\mathbf{x}^*)\mathbf{d} \geqslant 0$ for all $\mathbf{d} \in E^2$ such that $\mathbf{d}$ is a feasible direction at the point (1/2, 0).

7.2 Examples of Unconstrained Problems

Unconstrained optimization problems occur in a variety of contexts, but most frequently when the problem formulation is simple. More complex formulations often involve explicit functional constraints. However, many problems with constraints are frequently converted to unconstrained problems, such as using the barrier functions, e.g., the analytic center problem for (dual) linear programs. We present a few more examples here that should begin to indicate the wide scope to which the theory applies.

Example 1 (Logistic Regression). Recall the classification problem where we have vectors $\mathbf{a}_i \in E^d$ for $i = 1, 2, \ldots, n_1$ in a class, and vectors $\mathbf{b}_j \in E^d$ for $j = 1, 2, \ldots, n_2$ not. Then we wish to find $\mathbf{y} \in E^d$ and a number β such that

$$\frac{\exp(\mathbf{a}_i^T\mathbf{y} + \beta)}{1 + \exp(\mathbf{a}_i^T\mathbf{y} + \beta)}$$

is close to 1 for all i, and

$$\frac{\exp(\mathbf{b}_j^T\mathbf{y}+\beta)}{1+\exp(\mathbf{b}_j^T\mathbf{y}+\beta)}$$

is close to 0 for all j. The problem can be cast as a unconstrained optimization problem, called the max-likelihood,

$$\text{maximize}_{\mathbf{y},\,\beta}\left(\prod_i \frac{\exp(\mathbf{a}_i^T\mathbf{y}+\beta)}{1+\exp(\mathbf{a}_i^T\mathbf{y}+\beta)}\right)\left(\prod_j\left(1-\frac{\exp(\mathbf{b}_j^T\mathbf{y}+\beta)}{1+\exp(\mathbf{b}_j^T\mathbf{y}+\beta)}\right)\right),$$

which can be also equivalently, using a logarithmic transformation, written as

$$\text{minimize}_{\mathbf{y},\,\beta}\ \sum_i \log\left(1+\exp(-\mathbf{a}_i^T\mathbf{y}-\beta)\right)+\sum_j \log\left(1+\exp(\mathbf{b}_j^T\mathbf{y}+\beta)\right).$$

Example 2 (Utility Maximization). A common problem in economic theory is the determination of the best way to combine various inputs in order to maximize a utility function $f(x_1,\ x_2,\ \ldots,\ x_n)$ (in the monetary unit) of the amounts x_j of the inputs, $i = 1, 2,\ \ldots,\ n$. The unit prices of the inputs are $p_1,\ p_2,\ \ldots,\ p_n$. The producer wishing to maximize profit must solve the problem

$$\text{maximize } f(x_1,\ x_2,\ \ldots,\ x_n) - p_1x_1 - p_2x_2 \ldots - p_nx_n.$$

The first-order necessary conditions are that the partial derivatives with respect to the x_i's each vanish. This leads directly to the n equations

$$\frac{\partial f}{\partial x_i}(x_1,\ x_2,\ \ldots,\ x_n) = p_i,\ \ i = 1, 2,\ \ldots,\ n.$$

These equations can be interpreted as stating that, at the solution, the marginal value due to a small increase in the ith input must be equal to the price p_i.

Example 3 (Parametric Estimation). A common use of optimization is for the purpose of function approximation. Suppose, for example, that through an experiment the value of a function g is observed at m points, $x_1,\ x_2,\ \ldots,\ x_m$. Thus, values $g(x_1),\ g(x_2),\ \ldots,\ g(x_m)$ are known. We wish to approximate the function by a polynomial

$$h(x) = a_nx^n + a_{n-1}x^{n-1} + \ldots + a_0$$

of degree n (or less), where $n < m$. Corresponding to any choice of the approximating polynomial, there will be a set of errors $\varepsilon_k = g(x_k) - h(x_k)$. We define the best approximation as the polynomial that minimizes the sum of the squares of these errors; that is, minimizes

$$\sum_{k=1}^{m}(\varepsilon_k)^2.$$

This in turn means that we minimize

$$f(\mathbf{a}) = \sum_{k=1}^{m} \left[g(x_k) - \left(a_n x_k^n + a_{n-1} x_k^{n-1} + \ldots + a_0 \right) \right]^2$$

with respect to $\mathbf{a} = (a_0,\ a_1,\ \ldots,\ a_n)$ to find the best coefficients. This is a quadratic expression in the coefficients $\mathbf{a}$. To find a compact representation for this objective we define $q_{ij} = \sum_{k=1}^{m} (x_k)^{i+j}$, $b_j = \sum_{k=1}^{m} g(x_k)(x_k)^j$ and $c = \sum_{k=1}^{m} g(x_k)^2$. Then after a bit of algebra it can be shown that

$$f(\mathbf{a}) = \mathbf{a}^T \mathbf{Q} \mathbf{a} - 2\mathbf{b}^T \mathbf{a} + c$$

where $\mathbf{Q} = [q_{ij}]$, $\mathbf{b} = (b_1,\ b_2,\ \ldots,\ b_{n+1})$.

The first-order necessary conditions state that the gradient of f must vanish. This leads directly to the system of $n + 1$ equations

$$\mathbf{Q}\mathbf{a} = \mathbf{b}.$$

These can be solved to determine $\mathbf{a}$.

Example 4 (Selection Problem). It is often necessary to select an assortment of factors to meet a given set of requirements. An example is the problem faced by an electric utility when selecting its power-generating facilities. The level of power that the company must supply varies by time of the day, by day of the week, and by season. Its power-generating requirements are summarized by a curve, $h(x)$, as shown in Fig. 7.2a, which shows the total hours in a year that a power level of at least x is required for each x. For convenience the curve is normalized so that the upper limit is unity.

The power company may meet these requirements by installing generating equipment, such as (7.1) nuclear or (7.2) coal-fired, or by purchasing power from a central energy grid. Associated with type $i (i = 1,\ 2)$ of generating equipment is a yearly unit capital cost b_i and a unit operating cost c_i. The unit price of power purchased from the grid is c_3.

Nuclear plants have a high capital cost and low operating cost, so they are used to supply a base load. Coal-fired plants are used for the intermediate level, and power is purchased directly only for peak demand periods. The requirements are satisfied as shown in Fig. 7.2b, where x_1 and x_2 denote the capacities of the nuclear and coal-fired plants, respectively. (For example, the nuclear power plant can be visualized as consisting of x_1/Δ small generators of capacity Δ, where Δ is small. The first such generator is on for about $h(\Delta)$ hours, supplying $\Delta h(\Delta)$ units of energy; the next supplies $\Delta h(2\Delta)$ units, and so forth. The total energy supplied by the nuclear plant is thus the area shown.)

The total cost is

$$f(x_1,\ x_2) = b_1x_1 + b_2x_2 + c_1\int_0^{x_1} h(x)dx + c_2\int_{x_1}^{x_1+x_2} h(x)dx + c_3\int_{x_1+x_2}^{1} h(x)dx,$$

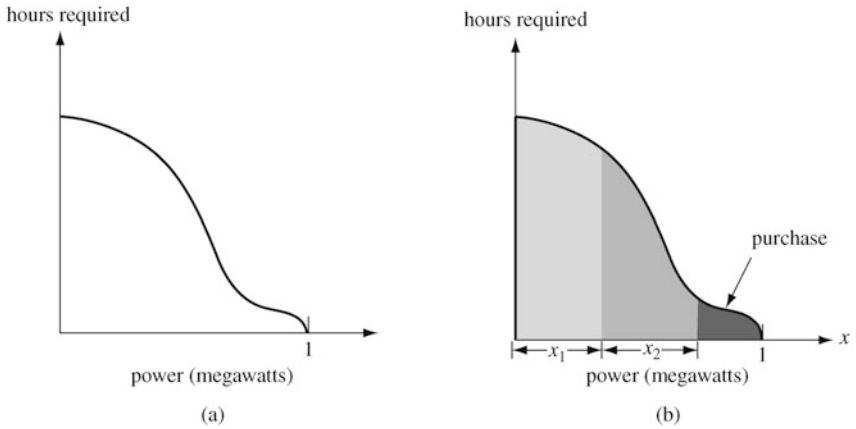

Fig. 7.2 (**a**) Power requirement curve; (**b**) x_1 and x_2 denote the capacities of the nuclear and coal-fired plants, respectively

and the company wishes to minimize this over the set defined by

$$x_1 \geqslant 0, \quad x_2 \geqslant 0, \quad x_1 + x_2 \leqslant 1.$$

Assuming that the solution is interior to the constraints, by setting the partial derivatives equal to zero, we obtain the two equations

$$b_1 + (c_1 - c_2)h(x_1) + (c_2 - c_3)h(x_1 + x_2) = 0$$
$$b_2 + (c_2 - c_3)h(x_1 + x_2) = 0,$$

which represent the necessary conditions.

If $x_1 = 0$, then the general necessary condition theorem shows that the first equality could relax to $\geqslant 0$. Likewise, if $x_2 = 0$, then the second equality could relax to $\geqslant 0$. The case $x_1 + x_2 = 1$ requires a bit more analysis (see Exercise 2).

7.3 Second-Order Conditions

The proof of Proposition 1 in Sect. 7.1 is based on making a first-order approximation to the function f in the neighborhood of the relative minimum point. Additional conditions can be obtained by considering higher-order approximations.

The second-order conditions, which are defined in terms of the Hessian matrix $\nabla^2 f$ of second partial derivatives of f (see Appendix A), are of extreme theoretical importance and dominate much of the analysis presented in later chapters.

Proposition 1 (Second-Order Necessary Conditions). *Let Ω be a subset of E^n and let $f \in C^2$ be a function on Ω. If $\mathbf{x}^*$ is a relative minimum point of f over Ω, then for any $\mathbf{d} \in E^n$ that is a feasible direction at $\mathbf{x}^*$ we have*

$$\text{i) } \nabla f(\mathbf{x}^*)\mathbf{d} \geqslant 0 \tag{7.3}$$

$$\text{ii) if } \nabla f(\mathbf{x}^*)\mathbf{d} = 0, \text{ then } \mathbf{d}^T \nabla^2 f(\mathbf{x}^*)\mathbf{d} \geqslant 0. \tag{7.4}$$

Proof. The first condition is just Proposition 1, and the second applies only if $\nabla f(\mathbf{x}^*)\mathbf{d} = 0$. In this case, introducing $\mathbf{x}(\alpha) = \mathbf{x}^* + \alpha\mathbf{d}$ and $g(\alpha) = f(\mathbf{x}(\alpha))$ as before, we have, in view of $g'(0) = 0$,

$$g(\alpha) - g(0) = \frac{1}{2} g''(0)\alpha^2 + o(\alpha^2).$$

If $g''(0) < 0$ the right side of the above equation is negative for sufficiently small α which contradicts the relative minimum nature of $g(0)$. Thus

$$g''(0) = \mathbf{d}^T \nabla^2 f(\mathbf{x}^*)\mathbf{d} \geqslant 0. \blacksquare$$

Example 1. For the same problem as Example 2 of Sect. 7.1, we have for $\mathbf{d} = (d_1, d_2)$

$$\nabla f(\mathbf{x}^*)\mathbf{d} = \frac{3}{2} d_2.$$

Thus condition (ii) of Proposition 1 applies only if $d_2 = 0$. In that case we have $\mathbf{d}^T \nabla^2 f(\mathbf{x}^*)\mathbf{d} = 2d_1^2 \geqslant 0$, so condition (ii) is satisfied.

Again of special interest is the case where the minimizing point is an interior point of Ω, as, for example, in the case of completely unconstrained problems. We then obtain the following classical result.

Proposition 2 (Second-Order Necessary Conditions—Unconstrained Case). *Let $\mathbf{x}^*$ be an interior point of the set Ω, and suppose $\mathbf{x}^*$ is a relative minimum point over Ω of the function $f \in C^2$. Then*

$$\text{i) } \nabla f(\mathbf{x}^*) = 0 \tag{7.5}$$

$$\text{ii) for all } \mathbf{d}, \ \mathbf{d}^T \nabla^2 f(\mathbf{x}^*)\mathbf{d} \geqslant 0. \tag{7.6}$$

For notational simplicity we often denote $\nabla^2 f(\mathbf{x})$, the $n \times n$ matrix of the second partial derivatives of f, the Hessian of f, by the alternative notation $\mathbf{F}(\mathbf{x})$. Condition (ii) is equivalent to stating that the matrix $\mathbf{F}(\mathbf{x}^*)$ is positive semidefinite. As we shall see, the matrix $\mathbf{F}(\mathbf{x}^*)$, which arises here quite naturally in a discussion of necessary conditions, plays a fundamental role in the analysis of iterative methods for solving unconstrained optimization problems. The structure of this matrix is the primary determinant of the rate of convergence of algorithms designed to minimize the function f.

Example 2. Consider the problem

$$\begin{aligned} &\text{minimize} \quad f(x_1,\ x_2) = x_1^3 - x_1^2 x_2 + 2x_2^2 \\ &\text{subject to} \quad x_1 \geqslant 0, \quad x_2 \geqslant 0. \end{aligned}$$

If we assume that the solution is in the interior of the feasible set, that is, if $x_1 > 0,\ x_2 > 0$, then the first-order necessary conditions are

$$3x_1^2 - 2x_1 x_2 = 0, \quad -x_1^2 + 4x_2 = 0.$$

There is a solution to these at $x_1 = x_2 = 0$ which is a boundary point, but there is also a solution at $x_1 = 6,\ x_2 = 9$. We note that for x_1 fixed at $x_1 = 6$, the objective attains a relative minimum with respect to x_2 at $x_2 = 9$. Conversely, with x_2 fixed at $x_2 = 9$, the objective attains a relative minimum with respect to x_1 at $x_1 = 6$. Despite this fact, the point $x_1 = 6,\ x_2 = 9$ is not a relative minimum point, because the Hessian matrix is

$$\mathbf{F} = \begin{bmatrix} 6x_1 - 2x_2 & -2x_1 \\ -2x_1 & 4 \end{bmatrix},$$

which, evaluated at the proposed solution $x_1 = 6,\ x_2 = 9$, is

$$\mathbf{F} = \begin{bmatrix} 18 & -12 \\ -12 & 4 \end{bmatrix}.$$

This matrix is not positive semidefinite, since its determinant is negative. Thus the proposed solution is not a relative minimum point.

Sufficient Conditions for a Relative Minimum

By slightly strengthening the second condition of Proposition 2 above, we obtain a set of conditions that imply that the point $\mathbf{x}^*$ is a relative minimum. We give here the conditions that apply only to unconstrained problems, or to problems where the minimum point is interior to the feasible region, since the corresponding conditions for problems where the minimum is achieved on a boundary point of the feasible set are a good deal more difficult and of marginal practical or theoretical value. A more general result, applicable to problems with functional constraints, is given in Chap. 11.

Proposition 3 (Second-Order Sufficient Conditions—Unconstrained Case). *Let $f \in C^2$ be function defined on a region in which the point $\mathbf{x}^*$ is an interior point. Suppose in addition that*

$$\text{i) } \nabla f(\mathbf{x}^*) = \mathbf{0} \tag{7.7}$$

$$\text{ii) } \mathbf{F}(\mathbf{x}^*) \text{ is positive definite} \tag{7.8}$$

Then $\mathbf{x}^*$ is a strict relative minimum point of f.

Proof. Since $\mathbf{F}(\mathbf{x}^*)$ is positive definite, there is an $a > 0$ such that for all $\mathbf{d}$, $\mathbf{d}^T\mathbf{F}(\mathbf{x}^*)\mathbf{d} \geqslant a|\mathbf{d}|^2$. Thus by the Taylor's Theorem (with remainder)

$$\begin{aligned} f(\mathbf{x}^* + \mathbf{d}) - f(\mathbf{x}^*) &= \frac{1}{2}\mathbf{d}^T\mathbf{F}(\mathbf{x}^*)\mathbf{d} + o(|\mathbf{d}|^2) \\ &\geqslant (a/2)|\mathbf{d}|^2 + o(|\mathbf{d}|^2) \end{aligned}$$

For small $|\mathbf{d}|$ the first term on the right dominates the second, implying that both sides are positive for small $\mathbf{d}$. ∎

7.4 Convex and Concave Functions

In order to develop a theory directed toward characterizing global, rather than local, minimum points, it is necessary to introduce some sort of convexity assumptions. This results not only in a more potent, although more restrictive, theory but also provides an interesting geometric interpretation of the second-order sufficiency result derived above.

Definition. A function f defined on a convex set Ω is said to be *convex* if, for every $\mathbf{x}_1, \mathbf{x}_2 \in \Omega$ and every α, $0 \leqslant \alpha \leqslant 1$, there holds

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

If, for every α, $0 < \alpha < 1$, and $\mathbf{x}_1 \neq \mathbf{x}_2$, there holds

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) < \alpha f(\mathbf{x}_1) + (1-\alpha)f(\mathbf{x}_2),$$

then f is said to be *strictly convex*.

Several examples of convex or nonconvex functions are shown in Fig. 7.3. Geometrically, a function is convex if the line joining two points on its graph lies nowhere below the graph, as shown in Fig. 7.3a, or, thinking of a function in two dimensions, it is convex if its graph is bowl shaped.

Next we turn to the definition of a concave function.

Definition. A function g defined on a convex set Ω is said to be *concave* if the function $f = -g$ is convex. The function g is *strictly concave* if $-g$ is strictly convex.

Combinations of Convex Functions

We show that convex functions can be combined to yield new convex functions and that convex functions when used as constraints yield convex constraint sets.

Proposition 1. *Let f_1 and f_2 be convex functions on the convex set Ω. Then the function $f_1 + f_2$ is convex on Ω.*

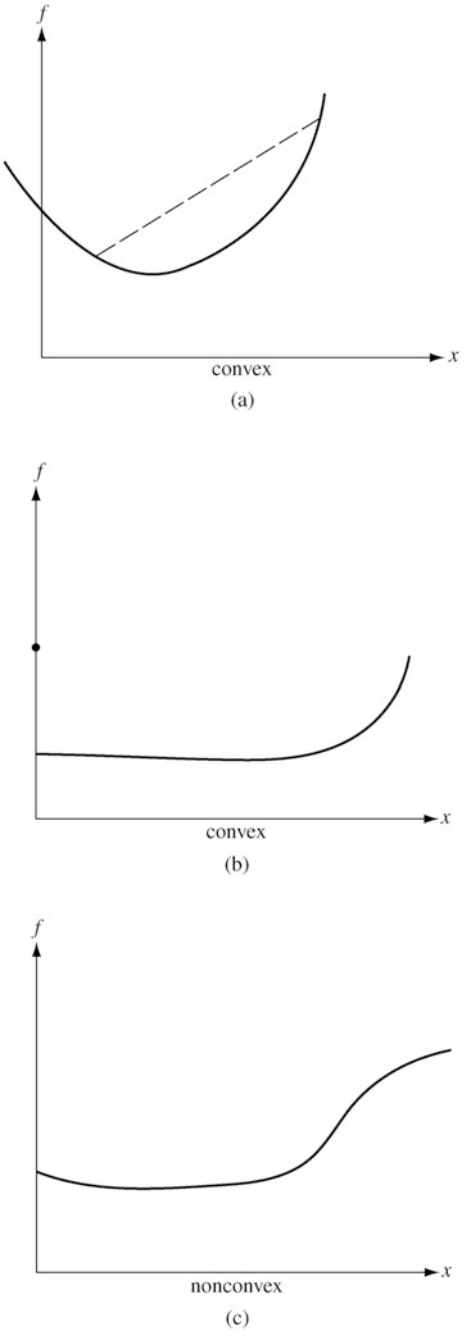

Fig. 7.3 Convex and nonconvex functions

Proof. Let $\mathbf{x}_1,\ \mathbf{x}_2 \in \Omega$, and $0 < \alpha < 1$. Then

$$f_1(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) + f_2(\alpha\mathbf{x}_1) + (1-\alpha)\mathbf{x}_2) \leqslant \alpha[f_1(\mathbf{x}_1) + f_2(\mathbf{x}_1)] + (1-\alpha)[f_1(\mathbf{x}_2) + f_2(\mathbf{x}_2)].$$ ∎

Proposition 2. *Let f be a convex function over the convex set Ω. Then the function af is convex for any $a \geqslant 0$.*

Proof. Immediate. ∎

Note that through repeated application of the above two propositions it follows that a positive combination $a_1 f_1 + a_2 f_2 + \ldots + a_m f_m$ of convex functions is again convex.

Finally, we consider sets defined by convex inequality constraints.

Proposition 3. *Let f be a convex function on a convex set Ω. The set $\Gamma_c = \{\mathbf{x} : \mathbf{x} \in \Omega,\ f(\mathbf{x}) \leqslant c\}$ is convex for every real number c.*

Proof. Let $\mathbf{x}_1,\ \mathbf{x}_2 \in \Gamma_c$. Then $f(\mathbf{x}_1) \leqslant c,\ f(\mathbf{x}_2) \leqslant c$ and for $0 < \alpha < 1$,

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leqslant \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2) \leqslant c.$$

Thus $\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in \Gamma_c$. ∎

We note that, since the intersection of convex sets is also convex, the set of points simultaneously satisfying

$$f_1(\mathbf{x}) \leqslant c_1,\ f_2(\mathbf{x}) \leqslant c_2,\ \ldots,\ f_m(\mathbf{x}) \leqslant c_m,$$

where each f_i is a convex function, defines a convex set. This is important in mathematical programming, since the constraint set is often defined this way.

Properties of Differentiable Convex Functions

If a function f is differentiable, then there are alternative characterizations of convexity.

Proposition 4. *Let $f \in C^1$. Then f is convex over a convex set Ω if and only if*

$$f(\mathbf{y}) \geqslant f(\mathbf{x}) + \nabla f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \tag{7.9}$$

for all $\mathbf{x},\ \mathbf{y} \in \Omega$.

Proof. First suppose f is convex. Then for all $\alpha,\ 0 \leqslant \alpha \leqslant 1$,

$$f(\alpha\mathbf{y} + (1-\alpha)\mathbf{x}) \leqslant \alpha f(\mathbf{y}) + (1-\alpha) f(\mathbf{x}).$$

Thus for $0 < \alpha \leqslant 1$

$$\frac{f(\mathbf{x} + \alpha(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{\alpha} \leqslant f(\mathbf{y}) - f(\mathbf{x}).$$

Letting $\alpha \to 0$ we obtain

$$\nabla f(\mathbf{x})\,(\mathbf{y} - \mathbf{x}) \leqslant f(\mathbf{y}) - f(\mathbf{x}).$$

This proves the "only if" part.

Now assume

$$f(\mathbf{y}) \geqslant f(\mathbf{x}) + \nabla f(\mathbf{x})\,(\mathbf{y} - \mathbf{x})$$

for all $\mathbf{x},\ \mathbf{y} \in \Omega$. Fix $\mathbf{x}_1,\ \mathbf{x}_2 \in \Omega$ and $\alpha,\ 0 \leqslant \alpha \leqslant 1$. Setting $\mathbf{x} = \alpha\mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2$ and alternatively $\mathbf{y} = \mathbf{x}_1$ or $\mathbf{y} = \mathbf{x}_2$, we have

$$f(\mathbf{x}_1) \geqslant f(\mathbf{x}) + \nabla f(\mathbf{x})(\mathbf{x}_1 - \mathbf{x}) \tag{7.10}$$

$$f(\mathbf{x}_2) \geqslant f(\mathbf{x}) + \nabla f(\mathbf{x})(\mathbf{x}_2 - \mathbf{x}). \tag{7.11}$$

Multiplying (7.10) by α and (7.11) by $(1 - \alpha)$ and adding, we obtain

$$\alpha f(\mathbf{x}_1) + (1 - \alpha) f(\mathbf{x}_2) \geqslant f(\mathbf{x}) + \nabla f(\mathbf{x})[\alpha\mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2 - \mathbf{x}].$$

But substituting $\mathbf{x} = \alpha\mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2$, we obtain

$$\alpha f(\mathbf{x}_1) + (1 - \alpha) f(\mathbf{x}_2) \geqslant f(\alpha\mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2). \blacksquare$$

The statement of the above proposition is illustrated in Fig. 7.4. It can be regarded as a sort of dual characterization of the original definition illustrated in Fig. 7.3. The original definition essentially states that linear interpolation between two points overestimates the function, while the above proposition states that linear approximation based on the local derivative underestimates the function.

For twice continuously differentiable functions, there is another characterization of convexity.

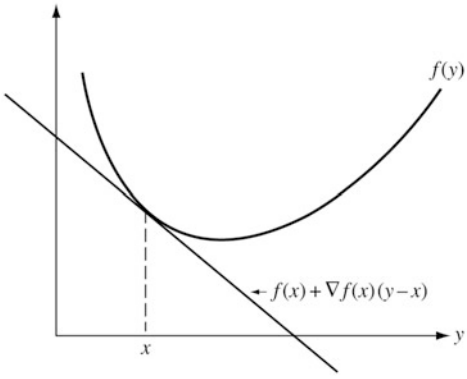

Fig. 7.4 Illustration of Proposition 4

Proposition 5. *Let $f \in C^2$. Then f is convex over a convex set Ω containing an interior point if and only if the Hessian matrix $\mathbf{F}$ of f is positive semidefinite throughout Ω.*

Proof. By Taylor's theorem we have

$$f(\mathbf{y}) = f(\mathbf{x}) = \nabla f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^T \mathbf{F}(\mathbf{x} + \alpha(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}) \tag{7.12}$$

for some $\alpha,\ 0 \leqslant \alpha \leqslant 1$. Clearly, if the Hessian is everywhere positive semidefinite, we have

$$f(\mathbf{y}) \geqslant f(\mathbf{x}) + \nabla f(\mathbf{x})(\mathbf{y} - \mathbf{x}), \tag{7.13}$$

which in view of Proposition 4 implies that f is convex.

Now suppose the Hessian is not positive semidefinite at some point $\mathbf{x} \in \Omega$. By continuity of the Hessian it can be assumed, without loss of generality, that $\mathbf{x}$ is an interior point of Ω. There is a $\mathbf{y} \in \Omega$ such that $(\mathbf{y} - \mathbf{x})^T \mathbf{F}(\mathbf{x})(\mathbf{y} - \mathbf{x}) < 0$. Again by the continuity of the Hessian, $\mathbf{y}$ may be selected so that for all $\alpha,\ 0 \leqslant \alpha \leqslant 1$,

$$(\mathbf{y} - \mathbf{x})^T \mathbf{F}(\mathbf{x} + \alpha(\mathbf{y} - \mathbf{x}))\,(\mathbf{y} - \mathbf{x}) < 0.$$

This in view of (7.12) implies that (7.13) does not hold; which in view of Proposition 4 implies that f is not convex. ∎

The Hessian matrix is the generalization to E^n of the concept of the curvature of a function, and correspondingly, positive definiteness of the Hessian is the generalization of positive curvature. Convex functions have positive (or at least nonnegative) curvature in every direction. Motivated by these observations, we sometimes refer to a function as being *locally convex* if its Hessian matrix is positive semidefinite in a small region, and *locally strictly convex* if the Hessian is positive definite in the region. In these terms we see that the second-order sufficiency result of the last section requires that the function be locally strictly convex at the point $\mathbf{x}^*$. Thus, even the local theory, derived solely in terms of the elementary calculus, is actually intimately related to convexity—at least locally. For this reason we can view the two theories, local and global, not as disjoint parallel developments but as complementary and interactive. Results that are based on convexity apply even to nonconvex problems in a region near the solution, and conversely, local results apply to a global minimum point.

7.5 Minimization and Maximization of Convex Functions

We turn now to the three classic results concerning minimization or maximization of convex functions.

Theorem 1. *Let f be a convex function defined on the convex set Ω. Then the set Γ where f achieves its minimum is convex, and any relative minimum of f is a global minimum.*

Proof. If f has no relative minima the theorem is valid by default. Assume now that c_0 is the minimum of f. Then clearly $\Gamma = \{\mathbf{x} : f(\mathbf{x}) \leqslant c_0,\ \mathbf{x} \in \Omega\}$ and this is convex by Proposition 3 of the last section.

Suppose now that $\mathbf{x}^* \in \Omega$ is a relative minimum point of f, but that there is another point $\mathbf{y} \in \Omega$ with $f(\mathbf{y}) < f(\mathbf{x}^*)$. On the line $\alpha\mathbf{y} + (1-\alpha)\mathbf{x}^*$, $0 < \alpha < 1$ we have

$$f(\alpha\mathbf{y} + (1-\alpha)\mathbf{x}^*) \leqslant \alpha f(\mathbf{y}) + (1-\alpha)f(\mathbf{x}^*) < f(\mathbf{x}^*),$$

contradicting the fact that $\mathbf{x}^*$ is a relative minimum point. ∎

We might paraphrase the above theorem as saying that for convex functions, all minimum points are located together (in a convex set) and all relative minima are global minima. The next theorem says that if f is continuously differentiable and convex, *then* satisfaction of the first-order necessary conditions are both necessary and sufficient for a point to be a global minimizing point.

Theorem 2. *Let $f \in C^1$ be convex on the convex set Ω. If there is a point $\mathbf{x}^* \in \Omega$ such that, for all $\mathbf{y} \in \Omega$, $\nabla f(\mathbf{x}^*)(\mathbf{y} - \mathbf{x}^*) \geqslant 0$, then $\mathbf{x}^*$ is a global minimum point of f over Ω.*

Proof. We note parenthetically that since $\mathbf{y} - \mathbf{x}^*$ is a feasible direction at $\mathbf{x}^*$, the given condition is equivalent to the first-order necessary condition stated in Sect. 7.1. The proof of the proposition is immediate, since by Proposition 4 of the last section

$$f(\mathbf{y}) \geqslant f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)(\mathbf{y} - \mathbf{x}^*) \geqslant f(\mathbf{x}^*). \blacksquare$$

Next we turn to the question of maximizing a convex function over a convex set. There is, however, no analog of Theorem 1 for maximization; indeed, the tendency is for the occurrence of numerous nonglobal relative maximum points. Nevertheless, it is possible to prove one important result. It is not used in subsequent chapters, but it is useful for some areas of optimization.

Theorem 3. *Let f be a convex function defined on the bounded, closed convex set Ω. If f has a maximum over Ω it is achieved at an extreme point of Ω.*

Proof. Suppose f achieves a global maximum at $\mathbf{x}^* \in \Omega$. We show first that this maximum is achieved at some boundary point of Ω. If $\mathbf{x}^*$ is itself a boundary point, then there is nothing to prove, so assume $\mathbf{x}^*$ is not a boundary point. Let L be any line passing through the point $\mathbf{x}^*$. The intersection of this line with Ω is an interval of the line L having end points $\mathbf{y}_1$, $\mathbf{y}_2$ which are boundary points of Ω, and we have $\mathbf{x}^* = \alpha\mathbf{y}_1 + (1-\alpha)\mathbf{y}_2$ for some α, $0 < \alpha < 1$. By convexity of f

$$f(\mathbf{x}^*) \leqslant \alpha f(\mathbf{y}_1) + (1-\alpha)f(\mathbf{y}_2) \leqslant \max\{f(\mathbf{y}_1),\ f(\mathbf{y}_2)\}.$$

Thus either $f(\mathbf{y}_1)$ or $f(\mathbf{y}_2)$ must be at least as great as $f(\mathbf{x}^*)$. Since $\mathbf{x}^*$ is a maximum point, so is either $\mathbf{y}_1$ or $\mathbf{y}_2$.

We have shown that the maximum, if achieved, must be achieved at a boundary point of Ω. If this boundary point, $\mathbf{x}^*$, is an extreme point of Ω there is nothing more to prove. If it is not an extreme point, consider the intersection of Ω with a

supporting hyperplane H at $\mathbf{x}^*$. This intersection, T_1, is of dimension $n-1$ or less and the global maximum of f over T_1 is equal to $f(\mathbf{x}^*)$ and must be achieved at a boundary point $\mathbf{x}_1$ of T_1. If this boundary point is an extreme point of T_1, it is also an extreme point of Ω by Lemma 1, Sect. B.4, and hence the theorem is proved. If $\mathbf{x}_1$ is not an extreme point of T_1, we form T_2, the intersection of T_1 with a hyperplane in E^{n-1} supporting T_1 at $\mathbf{x}_1$. This process can continue at most a total of n times when a set T_n of dimension zero, consisting of a single point, is obtained. This single point is an extreme point of T_n and also, by repeated application of Lemma 1, Sect. B.4, an extreme point of Ω. ∎

*7.6 *Zero-Order Conditions

We have considered the problem

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) \\ &\text{subject to} \quad \mathbf{x} \in \Omega \end{aligned} \tag{7.14}$$

to be unconstrained because there are no functional constraints of the form $g(\mathbf{x}) \leqslant b$ or $h(\mathbf{x}) = c$. However, the problem is of course constrained by the set Ω. This constraint influences the first- and second-order necessary and sufficient conditions through the relation between feasible directions and derivatives of the function f. Nevertheless, there is a way to treat this constraint without reference to derivatives. The resulting conditions are then of zero order. These necessary conditions require that the problem be convex is a certain way, while the sufficient conditions require no assumptions at all. The simplest assumptions for the necessary conditions are that Ω is a convex set and that f is a convex function on all of E^n.

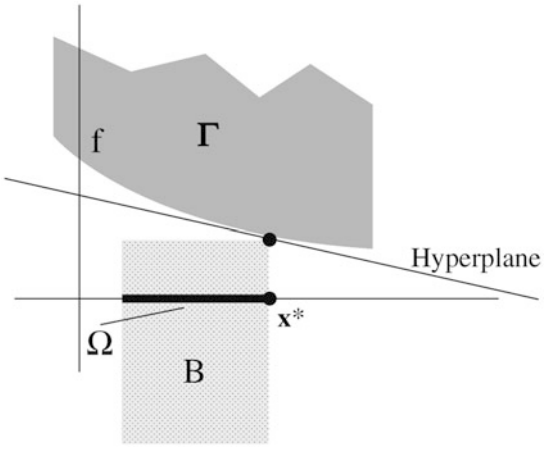

Fig. 7.5 The epigraph, the tubular region, and the hyperplane

To derive the necessary conditions under these assumptions consider the set $\Gamma \subset E^{n+1} = \{(r,\ \mathbf{x}) : r \geqslant f(\mathbf{x}),\ \mathbf{x} \in E^n\}$. In a figure of the graph of f, the set Γ is the region above the graph, shown in the upper part of Fig. 7.5. This set is called the *epigraph* of f. It is easy to verify that the set Γ is convex if f is a convex function.

Suppose that $\mathbf{x}^* \in \Omega$ is the minimizing point with value $f^* = f(\mathbf{x}^*)$. We construct a tubular region with cross section Ω and extending vertically from $-\infty$ up to f^*, shown as B in the upper part of Fig. 7.5. This is also a convex set, and it overlaps the set Γ only at the boundary point $(f^*,\ \mathbf{b}^*)$ above $\mathbf{x}^*$(or possibly many boundary points if f is flat near $\mathbf{x}^*$).

According to the separating hyperplane theorem (Appendix B), there is a hyperplane separating these two sets. This hyperplane can be represented by a nonzero vector of the form $(s,\ \boldsymbol{\lambda}) \in E^{n+1}$ with s a scalar and $\boldsymbol{\lambda} \in E^n$, and a separation constant c. The separation conditions are

$$sr + \boldsymbol{\lambda}^T\mathbf{x} \geqslant c \text{ for all } \mathbf{x} \in E^n \text{ and } r \geqslant f(\mathbf{x}) \tag{7.15}$$

$$sr + \boldsymbol{\lambda}^T\mathbf{x} \leqslant c \text{ for all } \mathbf{x} \in \Omega \text{ and } r \leqslant f^*. \tag{7.16}$$

It follows that $s \neq 0$; for otherwise $\boldsymbol{\lambda} \neq \mathbf{0}$ and then (7.15) would be violated for some $\mathrm{x} \in E^n$. It also follows that $s \geqslant 0$ since otherwise (7.16) would be violated by very negative values of r. Hence, together we find $s > 0$ and by appropriate scaling we may take $s = 1$.

It is easy to see that the above conditions can be expressed alternatively as two optimization problems, as stated in the following proposition.

Proposition 1 (Zero-Order Necessary Conditions). *If $\mathbf{x}^*$ solves (7.14) under the stated convexity conditions, then there is a nonzero vector $\boldsymbol{\lambda} \in E^n$ such that $\mathbf{x}^*$ is a solution to the two problems:*

$$\begin{aligned} &\text{minimize} && f(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{x} \\ &\text{subject to} && \mathbf{x} \in E^n \end{aligned} \tag{7.17}$$

and

$$\begin{aligned} &\text{maximize} && \boldsymbol{\lambda}^T\mathbf{x} \\ &\text{subject to} && \mathbf{x} \in \Omega. \end{aligned} \tag{7.18}$$

Proof. Problem (7.17) follows from (7.15) (with $s = 1$) and the fact that $f(\mathbf{x}) \leqslant r$ for $r \geqslant f(\mathbf{x})$. The value c is attained from above at $(f^*,\ \mathbf{x}^*)$. Likewise (7.18) follows from (7.16) and the fact that $\mathbf{x}^*$ and the appropriate r attain c from below. ∎

Notice that problem (7.17) is completely unconstrained, since $\mathbf{x}$ may range over all of E^n. The second problem (7.18) is constrained by Ω but has a linear objective function. It is clear from Fig. 7.5 that the slope of the hyperplane is equal to the slope of the function f when f is continuously differentiable at the solution $\mathbf{x}^*$.

If the optimal solution $\mathbf{x}^*$ is in the interior of Ω, then the second problem (7.18) implies that $\boldsymbol{\lambda} = \mathbf{0}$, for otherwise there would be a direction of movement from $\mathbf{x}^*$ that increases the product $\boldsymbol{\lambda}^T\mathbf{x}$ above $\boldsymbol{\lambda}^T\mathbf{x}^*$. The hyperplane is horizontal in that case.

The zeroth-order conditions provide no new information in this situation. However, when the solution is on a boundary point of Ω the conditions give very useful information.

Example 1 (Minimization Over an Interval). Consider a continuously differentiable function f of a single variable $x \in E^1$ defined on the unit interval [0,1] which plays the role of Ω here. The first problem (7.17) implies $f'(x^*) = -\lambda$. If the solution is at the left end of the interval (at $x = 0$) then the second problem (7.18) implies that $\lambda \leq 0$ which means that $f'(x^*) \geq 0$. The reverse holds if x^* is at the right end. These together are identical to the first-order conditions of Sect. 7.1.

Example 2. As a generalization of the above example, let $f \in C^1$ on E^n, and let f have a minimum with respect to Ω at $\mathbf{x}^*$. Let $\mathbf{d} \in E^n$ be a feasible direction at x^*. Then it follows again from (7.17) that $\nabla f(\mathbf{x}^*)\mathbf{d} \geq 0$.

Sufficient Conditions Theorem. The conditions of Proposition 1 are sufficient for $\mathbf{x}^*$ to be a minimum even without the convexity assumptions.

Proposition 2 (Zero-Order Sufficiency Conditions). *If there is a $\boldsymbol{\lambda}$ such that $\mathbf{x}^* \in \Omega$ solves the problems (7.17) and (7.18), then $\mathbf{x}^*$ solves (7.14).*

Proof. Suppose $\mathbf{x}_1$ is any other point in $\boldsymbol{\Omega}$. Then from (7.17)

$$f(\mathbf{x}_1) + \boldsymbol{\lambda}^T \mathbf{x}_1 \geqslant f(\mathbf{x}^*) + \boldsymbol{\lambda}^T \mathbf{x}^*.$$

This can be rewritten as

$$f(\mathbf{x}_1) - f(\mathbf{x}^*) \geqslant \boldsymbol{\lambda}^T \mathbf{x}^* - \boldsymbol{\lambda}^T \mathbf{x}_1.$$

By problem (7.18) the right hand side of this is greater than or equal to zero. Hence $f(\mathbf{x}_1) - f(\mathbf{x}^*) \geqslant 0$ which establishes the result. ∎

7.7 Global Convergence of Descent Algorithms

A good portion of the remainder of this book is devoted to presentation and analysis of various algorithms designed to solve nonlinear programming problems. Although these algorithms vary substantially in their motivation, application, and detailed analysis, ranging from the simple to the highly complex, they have the common heritage of all being iterative descent algorithms. By *iterative*, we mean, roughly, that the algorithm generates a series of points, each point being calculated on the basis of the points preceding it. By *descent*, we mean that as each new point is generated by the algorithm the corresponding value of some function (evaluated at the most recent point) decreases in value. Ideally, the sequence of points generated by the algorithm in this way converges in a finite or infinite number of steps to a solution of the original problem.

An iterative algorithm is initiated by specifying a starting point. If for arbitrary starting points the algorithm is guaranteed to generate a sequence of points converging to a solution, then the algorithm is said to be *globally convergent*. Quite definitely, not all algorithms have this obviously desirable property. Indeed, many of the most important algorithms for solving nonlinear programming problems are not globally convergent in their purest form and thus occasionally generate sequences that either do not converge at all or converge to points that are not solutions. It is often possible, however, to modify such algorithms, by appending special devices, so as to guarantee global convergence.

Fortunately, the subject of global convergence can be treated in a unified manner through the analysis of a general theory of algorithms developed mainly by Zangwill. From this analysis, which is presented in this section, we derive the Global Convergence Theorem that is applicable to the study of any iterative descent algorithm. Frequent reference to this important result is made in subsequent chapters.

Iterative Algorithms

We think of an algorithm as a mapping. Given a point $\mathbf{x}$ in some space X, the output of an algorithm applied to $\mathbf{x}$ is a new point. Operated iteratively, an algorithm is repeatedly reapplied to the new points it generates so as to produce a whole sequence of points. Thus, as a preliminary definition, we might formally define an algorithm A as a mapping taking points in a space X into (other) points in X. Operated iteratively, the algorithm A initiated at $\mathbf{x}_0 \in X$ would generate the sequence $\{\mathbf{x}_k\}$ defined by

$$\mathbf{x}_{k+1} = \mathbf{A}(\mathbf{x}_k).$$

In practice, the mapping $\mathbf{A}$ might be defined explicitly by a simple mathematical expression or it might be defined implicitly by, say, a lengthy complex computer program. Given an input vector, both define a corresponding output.

With this intuitive idea of an algorithm in mind, we now generalize the concept somewhat so as to provide greater flexibility in our analyses.

Definition. An *algorithm* $\mathbf{A}$ is a mapping defined on a space X that assigns to every point $\mathbf{x} \in X$ a subset of X.

In this definition the term "space" can be interpreted loosely. Usually X is the vector space E^n but it may be only a subset of E^n or even a more general metric space. The most important aspect of the definition, however, is that the mapping $\mathbf{A}$, rather than being a point-to-point mapping of X, is a *point-to-set mapping* of X.

An algorithm $\mathbf{A}$ generates a sequence of points in the following way. Given $\mathbf{x}_k \in X$ the algorithm yields $\mathbf{A}(\mathbf{x}_k)$ which is a subset of X. From this subset an arbitrary element $\mathbf{x}_{k+1}$ is selected. In this way, given an initial point $\mathbf{x}_0$, the algorithm generates sequences through the iteration

$$\mathbf{x}_{k+1} \in \mathbf{A}(\mathbf{x}_k).$$

It is clear that, unlike the case where **A** is a point-to-point mapping, the sequence generated by the algorithm **A** cannot, in general, be predicted solely from knowledge of the initial point $\mathbf{x}_0$. This degree of uncertainty is designed to reflect uncertainty that we may have in practice as to specific details of an algorithm.

Example 1. Suppose for x on the real line we define

$$A(x) = [-|x|/2,\ |x|/2]$$

so that $A(x)$ is an interval of the real line. Starting at $x_0 = 100$, each of the sequences below might be generated from iterative application of this algorithm.

$$100,\ 50,\ 25,\ 12,\ -6,\ -2,\ 1,\ 1/2,\ldots$$
$$100,\ -40,\ 20,\ -5,\ -2,\ 1,\ 1/4,\ 1/8,\ldots$$
$$100,\ 10,\ -1, 1/16, 1/100,\ -1/1000,\ 1/10,\ 100,\ldots$$

The apparent ambiguity that is built into this definition of an algorithm is not meant to imply that actual algorithms are random in character. In actual implementation algorithms are not defined ambiguously. Indeed, a particular computer program executed twice from the same starting point will generate two copies of the same sequence. In other words, in practice algorithms are point-to-point mappings. The utility of the more general definition is that it allows one to analyze, in a single step, the convergence of an infinite family of similar algorithms. Thus, two computer programs, designed from the same basic idea, may differ slightly in some details, and therefore perhaps may not produce identical results when given the same starting point. Both programs may, however, be regarded as implementations of the same point-to-set mappings. In the example above, for instance, it is not necessary to know exactly how x_{k+1} is determined from x_k so long as it is known that its absolute value is no greater than one-half x_k's absolute value. The result will always tend toward zero. In this manner, the generalized concept of an algorithm sometimes leads to simpler analysis.

Descent

In order to describe the idea of a descent algorithm we first must agree on a subset Γ of the space X, referred to as the *solution set*. The basic idea of a *descent function*, which is defined below, is that for points outside the solution set, a single step of the algorithm yields a decrease in the value of the descent function.

Definition. Let $\Gamma \subset X$ be a given solution set and let **A** be an algorithm on X. A continuous real-valued function Z on X is said to be a *descent function* for Γ and **A** if it satisfies

i) if $\mathbf{x} \notin \Gamma$ and $\mathbf{y} \in \mathbf{A}(\mathbf{x})$, then $Z(\mathbf{y}) < Z(\mathbf{x})$
ii) if $\mathbf{x} \in \Gamma$ and $\mathbf{y} \in \mathbf{A}(\mathbf{x})$, then $Z(\mathbf{y}) \leqslant Z(\mathbf{x})$.

There are a number of ways a solution set, algorithm, and descent function can be defined. A natural set-up for the problem

$$\begin{aligned} &\text{minimize } f(\mathbf{x}) \\ &\text{subject to } \mathbf{x} \in \Omega \end{aligned} \tag{7.19}$$

is to let Γ be the set of minimizing points, and define an algorithm $\mathbf{A}$ on Ω in such a way that f decreases at each step and thereby serves as a descent function. Indeed, this is the procedure followed in a majority of cases. Another possibility for unconstrained problems is to let Γ be the set of points $\mathbf{x}$ satisfying $\nabla f(\mathbf{x}) = 0$. In this case we might design an algorithm for which $|\nabla f(\mathbf{x})|$ serves as a descent function or for which $f(\mathbf{x})$ serves as a descent function.

*Closed Mappings

An important property possessed by some algorithms is that they are closed. This property, which is a generalization for point-to-set mappings of the concept of continuity for point-to-point mappings, turns out to be the key to establishing a general global convergence theorem. In defining this property we allow the point-to-set mapping to map points in one space X into subsets of another space Y.

Definition. A point-to-set mapping $\mathbf{A}$ from X to Y is said to be *closed* at $\mathbf{x} \in X$ if the assumptions

i) $\mathbf{x}_k \to \mathbf{x}$, $\mathbf{x}_k \in X$,
ii) $\mathbf{y}_k \to \mathbf{y}$, $\mathbf{y}_k \in \mathbf{A}(\mathbf{x}_k)$
imply
iii) $\mathbf{y} \in \mathbf{A}(\mathbf{x})$.

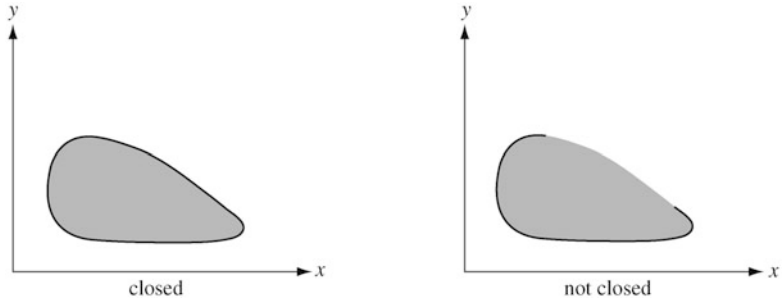

Fig. 7.6 Graphs of mappings

The point-to-set map $\mathbf{A}$ is said to be *closed* on X if it is closed at each point of X.

Example 2. As a special case, suppose that the mapping $\mathbf{A}$ is a point-to-point mapping; that is, for each $\mathbf{x} \in X$ the set $\mathbf{A}(\mathbf{x})$ consists of a single point in Y. Suppose also that $\mathbf{A}$ is continuous at $\mathbf{x} \in X$. This means that if $\mathbf{x}_k \to \mathbf{x}$ then $\mathbf{A}(\mathbf{x}_k) \to \mathbf{A}(\mathbf{x})$, and it follows that $\mathbf{A}$ is closed at $\mathbf{x}$. Thus for point-to-point mappings continuity implies closedness. The converse is, however, not true in general.

The definition of a closed mapping can be visualized in terms of the *graph* of the mapping, which is the set $\{(\mathbf{x}, \mathbf{y}) : \mathbf{x} \in X, \mathbf{y} \in \mathbf{A}(\mathbf{x})\}$. If X is closed, then $\mathbf{A}$ is closed throughout X if and only if this graph is a closed set. This is illustrated in Fig. 7.6. However, this equivalence is valid only when considering closedness everywhere. In general a mapping may be closed at some points and not at others.

Example 3. The reader should verify that the point-to-set mapping defined in Example 1 is closed.

Many complex algorithms that we analyze are most conveniently regarded as the composition of two or more simple point-to-set mappings. It is therefore natural to ask whether closedness of the individual maps implies closedness of the composite. The answer is a qualified "yes." The technical details of composition are described in the remainder of this subsection. They can safely be omitted at first reading while proceeding to the Global Convergence Theorem.

Definition. Let $\mathbf{A} : X \to Y$ and $\mathbf{B} : Y \to Z$ be point-to-set mappings. The composite mapping $\mathbf{C} = \mathbf{BA}$ is defined as the point-to-set mapping $\mathbf{C} : X \to Z$ with

$$\mathbf{C}(\mathbf{x}) = \bigcup_{\mathbf{y} \in \mathbf{A}(\mathbf{x})} \mathbf{B}(\mathbf{y}).$$

This definition is illustrated in Fig. 7.7.

Proposition. *Let $\mathbf{A} : X \to Y$ and $\mathbf{B} : Y \to Z$ be point-to-set mappings. Suppose $\mathbf{A}$ is closed at $\mathbf{x}$ and $\mathbf{B}$ is closed on $\mathbf{A}(\mathbf{x})$. Suppose also that if $\mathbf{x}_k \to \mathbf{x}$ and $\mathbf{y}_k \in \mathbf{A}(\mathbf{x}_k)$, there is a $\mathbf{y}$ such that, for some subsequence $\{\mathbf{y}_{ki}\}$, $\mathbf{y}_{ki} \to \mathbf{y}$. Then the composite mapping $\mathbf{C} = \mathbf{BA}$ is closed at $\mathbf{x}$.*

Proof. Let $\mathbf{x}_k \to \mathbf{x}$ and $\mathbf{z}_k \to \mathbf{z}$ with $\mathbf{z}_k \in \mathbf{C}(\mathbf{x}_k)$. It must be shown that $\mathbf{z} \in \mathbf{C}(\mathbf{x})$.

Select $\mathbf{y}_k \in \mathbf{A}(\mathbf{x}_k)$ such that $\mathbf{z}_k \in \mathbf{B}(\mathbf{y}_k)$ and according to the hypothesis let $\mathbf{y}$ and $\{\mathbf{y}_{ki}\}$ be such that $\mathbf{y}_{ki} \to \mathbf{y}$. Since $\mathbf{A}$ is closed at $\mathbf{x}$ it follows that $\mathbf{y} \in \mathbf{A}(\mathbf{x})$.

Likewise, since $\mathbf{y}_{ki} \to \mathbf{y}$, $\mathbf{z}_{ki} \to \mathbf{z}$ and $\mathbf{B}$ is closed at $\mathbf{y}$, it follows that $\mathbf{z} \in \mathbf{B}(\mathbf{y}) \subset \mathbf{BA}(\mathbf{x}) = \mathbf{C}(\mathbf{x})$. ∎

Two important corollaries follow immediately.

Corollary 1. *Let $\mathbf{A} : X \to Y$ and $\mathbf{B} : Y \to Z$ be point-to-set mappings. If $\mathbf{A}$ is closed at $\mathbf{x}$, $\mathbf{B}$ is closed on $\mathbf{A}(\mathbf{x})$ and Y is compact, then the composite map $\mathbf{C} = \mathbf{BA}$ is closed at $\mathbf{x}$.*

Corollary 2. *Let $\mathbf{A} : X \to Y$ be a point-to-point mapping and $\mathbf{B} : Y \to Z$ a point-to-set mapping. If $\mathbf{A}$ is continuous at $\mathbf{x}$ and $\mathbf{B}$ is closed at $\mathbf{A}(\mathbf{x})$, then the composite mapping $\mathbf{C} = \mathbf{BA}$ is closed at $\mathbf{x}$.*

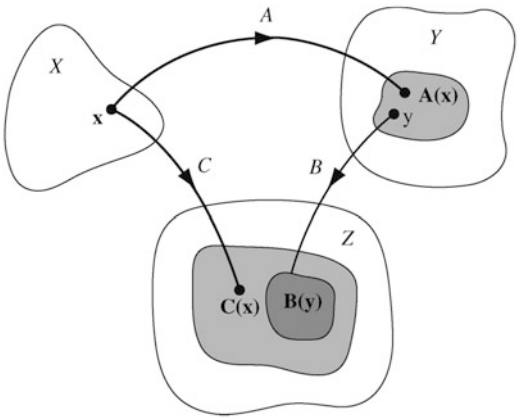

Fig. 7.7 Composition of mappings

Global Convergence Theorem

The Global Convergence Theorem is used to establish convergence for the following general situation. There is a solution set Γ. Points are generated according to the algorithm $\mathbf{x}_{k+1} \in \mathbf{A}(\mathbf{x}_k)$, and each new point always strictly decreases a descent function Z unless the solution set Γ is reached. For example, in nonlinear programming, the solution set may be the set of minimum points (perhaps only one point), and the descent function may be the objective function itself. A suitable algorithm is found that generates points such that each new point strictly reduces the value of the objective. Then, under appropriate conditions, it follows that the sequence converges to the solution set. The Global Convergence Theorem establishes technical conditions for which convergence is guaranteed.

Global Convergence Theorem. *Let* $\mathbf{A}$ *be an algorithm on* X, *and suppose that, given* $\mathbf{x}_0$ *the sequence* $\{\mathbf{x}_k\}_{k=0}^{\infty}$ *is generated satisfying*

$$\mathbf{x}_{k+1} \in \mathbf{A}(\mathbf{x}_k).$$

Let a solution set $\Gamma \subset X$ *be given, and suppose*

i) *all points* $\mathbf{x}_k$ *are contained in a compact set* $S \subset X$

ii) *there is a continuous function* Z *on* X *such that*

(a) *if* $\mathbf{x} \notin \Gamma$, *then* $Z(\mathbf{y}) < Z(\mathbf{x})$ *for all* $\mathbf{y} \in \mathbf{A}(\mathbf{x})$

(b) *if* $\mathbf{x} \in \Gamma$, *then* $Z(\mathbf{y}) \leqslant Z(\mathbf{x})$ *for all* $\mathbf{y} \in \mathbf{A}(\mathbf{x})$

iii) *the mapping* $\mathbf{A}$ *is closed at points outside* Γ.

Then the limit of any convergent subsequence of $\{\mathbf{x}_k\}$ *is a solution.*

Proof. Suppose the convergent subsequence $\{\mathbf{x}_k\}$, $k \in \mathcal{K}$ converges to the limit $\mathbf{x}$. Since Z is continuous, it follows that for $k \in \mathcal{K}$, $Z(\mathbf{x}_k) \to Z(\mathbf{x})$. This means that Z is convergent with respect to the subsequence, and we shall show that it is convergent

with respect to the entire sequence. By the monotonicity of Z on the sequence $\{\mathbf{x}_k\}$ we have $Z(\mathbf{x}_k) - Z(\mathbf{x}) \geqslant 0$ for all k. By the convergence of Z on the subsequence, there is, for a given $\varepsilon > 0$, a $K \in \mathcal{K}$ such that $Z(\mathbf{x}_k) - Z(\mathbf{x}) < \varepsilon$ for all $k > K,\ k \in \mathcal{K}$.

Thus for all $k > K$

$$Z(\mathbf{x}_k) - Z(\mathbf{x}) = Z(\mathbf{x}_k) - Z(\mathbf{x}_K) + Z(\mathbf{x}_K) - Z(\mathbf{x}) < \varepsilon,$$

which shows that $Z(\mathbf{x}_k) \to Z(\mathbf{x})$.

To complete the proof it is only necessary to show that $\mathbf{x}$ is a solution. Suppose $\mathbf{x}$ is not a solution. Consider the subsequence $\{\mathbf{x}_{k+1}\}_{\mathcal{K}}$. Since all members of this sequence are contained in a compact set, there is a $\bar{\mathcal{K}} \subset \mathcal{K}$ such that $\{\mathbf{x}_{k+1}\}_{\bar{\mathcal{K}}}$ converges to some limit $\bar{\mathbf{x}}$. We thus have $\mathbf{x}_k \to \mathbf{x},\ k \in \bar{\mathcal{K}}$, and $\mathbf{x}_{k+1} \in \mathbf{A}(\mathbf{x}_k)$ with $\mathbf{x}_{k+1} \to \bar{\mathbf{x}},\ k \in \bar{\mathcal{K}}$. Thus since $\mathbf{A}$ is closed at $\mathbf{x}$ it follows that $\bar{\mathbf{x}} \in \mathbf{A}(\mathbf{x})$. But from above, $Z(\bar{\mathbf{x}}) = Z(\mathbf{x})$ which contradicts the fact that Z is a descent function. ∎

Corollary. *If under the conditions of the Global Convergence Theorem* Γ *consists of a single point* $\bar{\mathbf{x}}$*, then the sequence* $\{\mathbf{x}_k\}$ *converges to* $\bar{\mathbf{x}}$*.*

Proof. Suppose to the contrary that there is a subsequence $\{\mathbf{x}_k\}_{\mathcal{K}}$ and an $\varepsilon > 0$ such that $|\mathbf{x}_k - \bar{\mathbf{x}}| > \varepsilon$ for all $k \in \mathcal{K}$. By compactness there must be $\mathcal{K}' \subset \mathcal{K}$ such that $\{\mathbf{x}_k\}_{\mathcal{K}'}$, converges, say to $\mathbf{x}'$. Clearly, $|\mathbf{x}' - \bar{\mathbf{x}}| \geqslant \varepsilon$, but by the Global Convergence Theorem $\mathbf{x}' \in \Gamma$, which is a contradiction. ∎

In later chapters the Global Convergence Theorem is used to establish the convergence of several standard algorithms. Here we consider some simple examples designed to illustrate the roles of the various conditions of the theorem.

Example 4. In many respects condition (iii) of the theorem, the closedness of $\mathbf{A}$ outside the solution set, is the most important condition. The failure of many popular algorithms can be traced to nonsatisfaction of this condition. On the real line consider the point-to-point algorithm

$$A(x) = \begin{cases} \frac{1}{2}(x-1) + 1 & x > 1 \\ \frac{1}{2}x & x \leqslant 1 \end{cases}$$

and the solution set $\Gamma = \{0\}$. It is easily verified that a descent function for this solution set and this algorithm is $Z(x) = |x|$. However, starting from $x > 1$, the algorithm generates a sequence converging to $x = 1$ which is not a solution. The difficulty is that A is not closed at $x = 1$.

Example 5. On the real line X consider the solution set to be empty, the descent function $Z(x) = e^{-x}$, and the algorithm $A(x) = x + 1$. All conditions of the convergence theorem except (i) hold. The sequence generated from any starting condition diverges to infinity. This is not strictly a violation of the conclusion of the theorem but simply an example illustrating that if no compactness assumption is introduced, the generated sequence may have no convergent subsequence.

Example 6. Consider the point-to-set algorithm A defined by the graph in Fig. 7.8 and given explicitly on $X = [0,\ 1]$ by

$$A(x) = \begin{cases} [0,\ x) & 1 \geqslant x > 0 \\ 0 & x = 0, \end{cases}$$

where $[0,\ x)$ denotes a half-open interval (see Appendix A). Letting $\Gamma = \{0\}$, the function $Z(x) = x$ serves as a descent function, because for $x \neq 0$ all points in $A(x)$ are less than x.

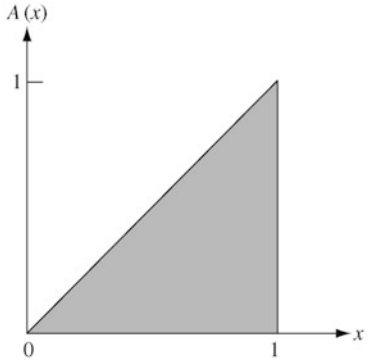

Fig. 7.8 Graph for Example 6

The sequence defined by

$$x_0 = 1$$
$$x_{k+1} = x_k - \frac{1}{2^{k+2}}$$

satisfies $x_{k+1} \in A(x_k)$ but it can easily be seen that $x_k \to \frac{1}{2} \notin \Gamma$. The difficulty here, of course, is that the algorithm A is not closed outside the solution set.

*Spacer Steps

In some of the more complex algorithms presented in later chapters, the rule used to determine a succeeding point in an iteration may depend on several previous points rather than just the current point, or it may depend on the iteration index k. Such features are generally introduced in order to obtain a rapid rate of convergence but they can grossly complicate the analysis of global convergence.

If in such a complex sequence of steps there is inserted, perhaps irregularly but infinitely often, a step of an algorithm such as steepest descent that is known to converge, then it is not difficult to insure that the entire complex process converges. The step which is repeated infinitely often and guarantees convergence is called a *spacer step*, since it separates disjoint portions of the complex sequence. Essentially the only requirement imposed on the other steps of the process is that they do not increase the value of the descent function.

This type of situation can be analyzed easily from the following viewpoint. Suppose **B** is an algorithm which together with the descent function Z and solution set Γ, satisfies all the requirements of the Global Convergence Theorem. Define the algorithm **C** by $\mathbf{C}(\mathbf{x}) = \{\mathbf{y} : Z(\mathbf{y}) \leqslant Z(\mathbf{x})\}$. In other words, **C** applied to **x** can give any point so long as it does not increase the value of Z. It is easy to verify that **C** is closed. We imagine that **B** represents the spacer step and the complex process between spacer steps is just some realization of C. Thus the overall process amounts merely to repeated applications of the composite algorithm **CB**. With this viewpoint we may state the Spacer Step Theorem.

Spacer Step Theorem. *Suppose* **B** *is an algorithm on X which is closed outside the solution set* Γ. *Let Z be a descent function corresponding to* **B** *and* Γ.
Suppose that the sequence $\{\mathbf{x}_k\}_{k=0}^{\infty}$ *is generated satisfying*

$$\mathbf{x}_{k+1} \in \mathbf{B}(\mathbf{x}_k)$$

for k in an infinite index set $\mathcal{K}$, *and that*

$$Z(\mathbf{x}_{k+1}) \leqslant Z(\mathbf{x}_k)$$

for all k. Suppose also that the set $S = \{\mathbf{x} : Z(\mathbf{x}) \leqslant Z(\mathbf{x}_0)\}$ *is compact. Then the limit of any convergent subsequence of* $\{\mathbf{x}_k\}_{\mathcal{K}}$ *is a solution.*

Proof. We first define for any $\mathbf{x} \in X$, $\bar{\mathbf{B}}(\mathbf{x}) = S \cap \mathbf{B}(\mathbf{x})$ and then observe that $\mathbf{A} = \mathbf{C}\bar{\mathbf{B}}$ is closed outside the solution set by Corollary 1. The Global Convergence Theorem can then be applied to A. Since S is compact, there is a subsequence of $\{\mathbf{x}_k\}_{k\in\mathcal{K}}$ converging to a limit **x**. In view of the above we conclude that $\mathbf{x} \in \Gamma$. ∎

7.8 Speed of Convergence

The study of speed of convergence is an important but sometimes complex subject. Nevertheless, there is a rich and yet elementary theory of convergence rates that enables one to predict with confidence the relative effectiveness of a wide class of algorithms. In this section we introduce various concepts designed to measure speed of convergence, and prepare for a study of this most important aspect of nonlinear programming.

Order of Convergence

Consider a sequence of real numbers $\{r_k\}_{k=0}^{\infty}$ converging to the limit r^*. We define several notions related to the speed of convergence of such a sequence.

Definition. Let the sequence $\{r_k\}$ converge to r^*. The *order* of convergence of $\{r_k\}$ is defined as the supremum of the nonnegative numbers p satisfying

$$0 \leqslant \overline{\lim_{k\to\infty}} \frac{|r_{k+1} - r^*|}{|r_k - r^*|^p} < \infty.$$

To ensure that the definition is applicable to any sequence, it is stated in terms of limit superior rather than just limit and 0/0 (which occurs if $r_k = r^*$ for all k) is regarded as finite. But these technicalities are rarely necessary in actual analysis, since the sequences generated by algorithms are generally quite well behaved.

It should be noted that the order of convergence, as with all other notions related to speed of convergence that are introduced, is determined only by the properties of the sequence that hold as $k \to \infty$. Somewhat loosely but picturesquely, we are therefore led to refer to the *tail* of a sequence—that part of the sequence that is arbitrarily far out. In this language we might say that the order of convergence is a measure of how good the worst part of the tail is. Larger values of the order p imply, in a sense, faster convergence, since the distance from the limit r^* is reduced, at least in the tail, by the pth power in a single step. Indeed, if the sequence has order p and (as is the usual case) the limit

$$\beta = \lim_{k\to\infty} \frac{|r_{k+1} - r^*|}{|r_k - r^*|^p}$$

exists, then asymptotically we have

$$|r_{k+1} - r^*| = \beta |r_k - r^*|^p.$$

Example 1. The sequence with $r_k = a^k$ where $0 < a < 1$ converges to zero with order unity, since $r_{k+1}/r_k = a$.

Example 2. The sequence with $r_k = a^{(2^k)}$ for $0 < a < 1$ converges to zero with order two, since $r_{k+1}/r_k^2 = 1$.

Linear Convergence

Most algorithms discussed in this book have an order of convergence equal to unity. It is therefore appropriate to consider this class in greater detail and distinguish certain cases within it.

Definition. If the sequence $\{r_k\}$ converges to r^* in such a way that

$$\lim_{k\to\infty} \frac{|r_{k+1} - r^*|}{|r_k - r^*|} = \beta < 1,$$

the sequence is said to converge *linearly* to r^* with *convergence ratio* (*or rate*) β.

Linear convergence is, for our purposes, without doubt the most important type of convergence behavior. A linearly convergent sequence, with convergence ratio β, can be said to have a tail that converges at least as fast as the geometric sequence $c\beta^k$ for some constant c. Thus linear convergence is sometimes referred to as *geometric convergence*, although in this book we reserve that phrase for the case when a sequence is exactly geometric.

As a rule, when comparing the relative effectiveness of two competing algorithms both of which produce linearly convergent sequences, the comparison is based on their corresponding convergence ratios—the smaller the ratio the faster the rate. The ultimate case where $\beta = 0$ is referred to as *superlinear convergence*. We note immediately that convergence of any order greater than unity is superlinear, but it is also possible for superlinear convergence to correspond to unity order.

Example 3. The sequence $r_k = (1/k)^k$ is of order unity, since $r_{k+1}/r_k^p \to \infty$ for $p > 1$. However, $r_{k+1}/r_k \to 0$ as $k \to \infty$ and hence this is superlinear convergence.

Arithmetic Convergence

Linear convergence is also called geometric convergence. There is another (slower) type of convergence:

Definition. If the sequence $\{r_k\}$ converges to r^* in such a way that

$$|r_k - r^*| \le C\frac{|r_0 - r^*|}{k^p},\ k \ge 1,\ 0 < p < \infty$$

where C is a fixed positive number, the sequence is said to converge *arithmetically* to r^* with order p.

When $p = 1$, it is referred as arithmetic convergence. The greater of p the faster of the convergence.

Example 4. The sequence $r_k = 1/k$ converges to zero arithmetically. The convergence is of order one but it is not linear, since $\lim_{k\to\infty}(r_{k+1}/r_k) = 1$, that is, β is not strictly less than one.

**Average Rates*

All the definitions given above can be referred to as *step-wise* concepts of convergence, since they define bounds on the progress made by going a single step: from k to $k + 1$. Another approach is to define concepts related to the average progress per step over a large number of steps. We briefly illustrate how this can be done.

Definition. Let the sequence $\{r_k\}$ converge to r^*. The *average order* of convergence is the infimum of the numbers $p > 1$ such that

$$\overline{\lim_{k\to\infty}} |r_k - r^*|^{1/p^k} = 1.$$

The order is infinity if the equality holds for no $p > 1$.

Example 5. For the sequence $r_k = a^{(2^k)}$, $0 < a < 1$, given in Example 2, we have

$$|r_k|^{1/2^k} = a,$$

while

$$|r_k|^{1/p^k} = a^{(2/p)^k} \to 1$$

for $p > 2$. Thus the average order is two.

Example 6. For $r_k = a^k$ with $0 < a < 1$ we have

$$(r_k)^{1/p^k} = a^{k(1/p)^k} \to 1$$

for any $p > 1$. Thus the average order is unity.

As before, the most important case is that of unity order, and in this case we define the *average convergence ratio* as $\overline{\lim_{k\to\infty}} |r_k - r^*|^{1/k}$. Thus for the geometric sequence $r_k = ca^k$, $0 < a < 1$, the average convergence ratio is a. Paralleling the earlier definitions, the reader can then in a similar manner define corresponding notions of average linear and average superlinear convergence.

Although the above array of definitions can be further embellished and expanded, it is quite adequate for our purposes. For the most part we work with the step-wise definitions, since in analyzing iterative algorithms it is natural to compare one step with the next. In most situations, moreover, when the sequences are well behaved and the limits exist in the definitions, then the step-wise and average concepts of convergence rates coincide.

*Convergence of Vectors

Suppose $\{\mathbf{x}_k\}_{k=0}^{\infty}$ is a sequence of vectors in E^n converging to a vector $\mathbf{x}^*$. The convergence properties of such a sequence are defined with respect to some particular function that converts the sequence of vectors into a sequence of numbers. Thus, if f is a given continuous function on E^n, the convergence properties of $\{\mathbf{x}_k\}$ can be defined with respect to f by analyzing the convergence of $f(\mathbf{x}_k)$ to $f(\mathbf{x}^*)$. The function f used in this way to measure convergence is called the *error function*.

In optimization theory it is common to choose the error function by which to measure convergence as the same function that defines the objective function of the original optimization problem. This means we measure convergence by how fast the

objective converges to its minimum. alternatively, we sometimes use the function $|\mathbf{x} - \mathbf{x}^*|^2$ and thereby measure convergence by how fast the (squared) distance from the solution point decreases to zero.

Generally, the order of convergence of a sequence is insensitive to the particular error function used; but for step-wise linear convergence the associated convergence ratio is not. Nevertheless, the average convergence ratio is not too sensitive, as the following proposition demonstrates, and hence the particular error function used to measure convergence is not really very important.

Proposition. *Let f and g be two error functions satisfying $f(\mathbf{x}^*) = g(\mathbf{x}^*) = 0$ and, for all $\mathbf{x}$, a relation of the form*

$$0 \leqslant a_1 g(\mathbf{x}) \leqslant f(\mathbf{x}) \leqslant a_2 g(\mathbf{x})$$

for some fixed $a_1 > 0$, $a_2 > 0$. If the sequence $\{\mathbf{x}_k\}_{k=0}^{\infty}$ converges to $\mathbf{x}^*$ linearly with average ratio β with respect to one of these functions, it also does so with respect to the other.

Proof. The statement is easily seen to be symmetric in f and g. Thus we assume $\{\mathbf{x}_k\}$ is linearly convergent with average convergence ratio β with respect to f, and will prove that the same is true with respect to g. We have

$$\beta = \overline{\lim_{k\to\infty}} f(\mathbf{x}_k)^{1/k} \leqslant \overline{\lim_{k\to\infty}} a_2^{1/k} g(\mathbf{x}_k)^{1/k} = \overline{\lim_{k\to\infty}} g(\mathbf{x}_k)^{1/k}$$

and

$$\beta = \overline{\lim_{k\to\infty}} f(\mathbf{x}_k)^{1/k} \geqslant \overline{\lim_{k\to\infty}} a_1^{1/k} g(\mathbf{x}_k)^{1/k} = \overline{\lim_{k\to\infty}} g(\mathbf{x}_k)^{1/k}.$$

Thus

$$\beta = \overline{\lim_{k\to\infty}} g(\mathbf{x}_k)^{1/k}. \blacksquare$$

As an example of an application of the above proposition, consider the case where $g(\mathbf{x}) = |\mathbf{x} - \mathbf{x}^*|^2$ and $f(\mathbf{x}) = (\mathbf{x} - \mathbf{x}^*)^T \mathbf{Q}(\mathbf{x} - \mathbf{x}^*)$, where $\mathbf{Q}$ is a positive definite symmetric matrix. Then a_1 and a_2 correspond, respectively, to the smallest and largest eigenvalues of Q. Thus average linear convergence is identical with respect to any error function constructed from a positive definite quadratic form.

Complexity

Complexity theory as outlined in Sect. 5.1 is an important aspect of convergence theory. This theory can be used in conjunction with the theory of local convergence. If an algorithm converges according to any order greater than zero, then for a fixed problem, the sequence generated by the algorithm will converge in a time that is a function of the convergence order (and rate, if convergence is linear). For example, if the order is one with rate $0 < c < 1$ and the process begins with an error of R, a final error of r can be achieved by a number of steps n satisfying $c^n R \leqslant r$. Thus it requires approximately $n = \log(R/r)/\log(1/c)$ steps. In this form the number of steps is not affected by the size of the problem. However, problem size enters in two possible ways. First, the rate c may depend on the size-say going toward 1 as

the size increases so that the speed is slower for large problems. The second way that size may enter, and this is the more important way, is that the time to execute a single step almost always increases with problem size. For instance if, for a problem seeking an optimal vector of dimension m, each step requires a Gaussian elimination inversion of an $m \times m$ matrix, the solution time will increase by a factor proportional to m^3. Overall the algorithm is therefore a polynomial time algorithm. Essentially all algorithms in this book employ steps, such as matrix multiplications or inversion or other algebraic operations, which are polynomial-time in character. Convergence analysis, therefore, focuses on whether an algorithm is globally convergent, on its local convergence properties, and also on the order of the algebraic operations required to execute the steps required. The last of these is usually easily deduced by listing the number and size of the required vector and matrix operations.

7.9 Summary

There are two different but complementary ways to characterize the solution to unconstrained optimization problems. In the local approach, one examines the relation of a given point to its neighbors. This leads to the conclusion that, at an unconstrained relative minimum point of a smooth function, the gradient of the function vanishes and the Hessian is positive semidefinite; and conversely, if at a point the gradient vanishes and the Hessian is positive definite, that point is a relative minimum point. This characterization has a natural extension to the global approach where convexity ensures that if the gradient vanishes at a point, that point is a global minimum point.

In considering iterative algorithms for finding either local or global minimum points, there are two distinct issues: global convergence properties and local convergence properties. The first is concerned with whether starting at an arbitrary point the sequence generated will converge to a solution. This is ensured if the algorithm is closed, has a descent function, and generates a bounded sequence. It is also explained that global convergence is guaranteed simply by the inclusion, in a complex algorithm, of spacer steps. This result is called upon frequently in what follows. Local convergence properties are a measure of the ultimate speed of convergence and generally determine the relative advantage of one algorithm to another.

7.10 Exercises

1. To approximate a function g over the interval $[0, 1]$ by a polynomial p of degree n (or less), we minimize the criterion

$$f(\mathbf{a}) = \int_0^1 [g(x) - p(x)]^2 dx,$$

where $p(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_0$. Find the equations satisfied by the optimal coefficients $\mathbf{a} = (a_0, a_1, \ldots, a_n)$.

2. In Example 4 of Sect. 7.2 show that if the solution has $x_1 > 0,\ x_1 + x_2 = 1$, then it is necessary that

$$b_1 - b_2 + (c_1 - c_2)h(x_1) = 0$$
$$b_2 + (c_2 - c_3)h(x_1 + x_2) \leqslant 0.$$

Hint: One way is to reformulate the problem in terms of the variables x_1 and $y = x_1 + x_2$.

3. (a) Using the first-order necessary conditions, find a minimum point of the function

$$f(x,\ y,\ z) = 2x^2 + xy + y^2 + yz + z^2 - 6x\ - 7y - 8z + 9.$$

(b) Verify that the point is a relative minimum point by verifying that the second-order sufficiency conditions hold.

(c) Prove that the point is a global minimum point.

4. In this exercise and the next we develop a method for determining whether a given symmetric matrix is positive definite. Given an $n \times n$ matrix $\mathbf{A}$ let $\mathbf{A}_k$ denote the principal submatrix made up of the first k rows and columns. Show (by induction) that if the first $n - 1$ principal submatrices are nonsingular, then there is a unique lower triangular matrix $\mathbf{L}$ with unit diagonal and a unique upper triangular matrix $\mathbf{U}$ such that $\mathbf{A} = \mathbf{LU}$. (See Appendix C.)

5. A symmetric matrix is positive definite if and only if the determinant of each of its principal submatrices is positive. Using this fact and the considerations of Exercise 4, show that an $n \times n$ symmetric matrix $\mathbf{A}$ is positive definite if and only if it has an $\mathbf{LU}$ decomposition (without interchange of rows) and the diagonal elements of $\mathbf{U}$ are all positive.

6. Using Exercise 5 show that an $n \times n$ matrix $\mathbf{A}$ is symmetric and positive definite if and only if it can be written as $\mathbf{A} = \mathbf{G}\mathbf{G}^T$ where $\mathbf{G}$ is a lower triangular matrix with positive diagonal elements. This representation is known as the *Cholesky factorization* of $\mathbf{A}$.

7. Let $f_j,\ i \in I$ be a collection of convex functions defined on a convex set Ω. Show that the function f defined by $f(\mathbf{x}) = \sup_{i \in I} f_i(\mathbf{x})$ is convex on the region where it is finite.

8. Let γ be a monotone nondecreasing function of a single variable (that is, $\gamma(r) \leqslant \gamma(r')$ for $r' > r$) which is also convex; and let f be a convex function defined on a convex set Ω. Show that the function $\gamma(f)$ defined by $\gamma(f)(\mathbf{x}) = \gamma[f(\mathbf{x})]$ is convex on Ω.

9. Let f be twice continuously differentiable on a region $\Omega \subset E^n$. Show that a sufficient condition for a point $\mathbf{x}^*$ in the interior of Ω to be a relative minimum point of f is that $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and that f be locally convex at $\mathbf{x}^*$.

10. Define the point-to-set mapping on E^n by

$$\mathbf{A}(\mathbf{x}) = \{\mathbf{y} : \mathbf{y}^T\mathbf{x} \leqslant b\},$$

where b is a fixed constant. Is $\mathbf{A}$ closed?
11. Prove the two corollaries in Sect. 7.6 on the closedness of composite mappings.
12. Show that if $\mathbf{A}$ is a continuous point-to-point mapping, the Global Convergence Theorem is valid even without assumption (i). Compare with Example 2, Sect. 7.7.
13. Let $\{r_k\}_{k=0}^{\infty}$ and $\{c_k\}_{k=0}^{\infty}$ be sequences of real numbers. Suppose $r_k \to 0$ average linearly and that there are constants $c > 0$ and C such that $c \leqslant c_k \leqslant C$ for all k. Show that $c_k r_k \to 0$ average linearly.
14. Prove a proposition, similar to the one in Sect. 7.8, showing that the order of convergence is insensitive to the error function.
15. Show that if $r_k \to r^*$ (step-wise) linearly with convergence ratio β, then $r_k \to r^*$(average) linearly with average convergence ratio no greater than β.

References

7.1–7.5 For alternative discussions of the material in these sections, see Hadley [H2], Fiacco and McCormick [F4], Zangwill [Z2] and Luenberger [L8].

7.6 Although the general concepts of this section are well known, the formulation as zero-order conditions appears to be new.

7.7 The idea of using a descent function (usually the objective itself) in order to guarantee convergence of minimization algorithms is an old one that runs through most literature on optimization, and has long been used to establish global convergence. Formulation of the general Global Convergence Theorem, which captures the essence of many previously diverse arguments, and the idea of representing an algorithm as a point-to-set mapping are both due to Zangwill [Z2]. A version of the Spacer Step Theorem can be found in Zangwill [Z2] as well.

7.8 Most of the definitions given in this section have been standard for quite some time. A thorough discussion which contributes substantially to the unification of these concepts is contained in Ortega and Rheinboldt [O7].