CHAPTER
TWELVE
OPTIMIZATION WITH EQUALITY CONSTRAINTS

The last chapter presented a general method for finding the relative extrema of an objective function of two or more choice variables. One important feature of that discussion is that all the choice variables are *independent* of one another, in the sense that the decision made regarding one variable does not impinge upon the choices of the remaining variables. For instance, a two-product firm can choose any value for Q_1 and any value for Q_2 it wishes, without the two choices limiting each other.

If the said firm is somehow required to observe a restriction (such as a production quota) in the form of $Q_1 + Q_2 = 950$, however, the independence between the choice variables will be lost. In that event, the firm's profit-maximizing output levels $\bar{Q}_1$ and $\bar{Q}_2$ will be not only simultaneous but also dependent, because the higher $\bar{Q}_1$ is, the lower $\bar{Q}_2$ must correspondingly be, in order to stay within the combined quota of 950. The new optimum satisfying the production quota constitutes a *constrained optimum*, which, in general, may be expected to differ from the *free optimum* discussed in the preceding chapter.

A restriction, such as the production quota mentioned above, establishes a relationship between the two variables in their roles as choice variables, but this should be distinguished from other types of relationships that may link the variables together. For instance, in Example 2 of Sec. 11.6, the two products of the firm are related in consumption (substitutes) as well as in production (as is reflected in the cost function), but that fact does not qualify the problem as one of constrained optimization, since the two output variables are still *independent as* *choice variables*. Only the dependence of the variables qua choice variables gives rise to a constrained optimum.

In the present chapter, we shall consider equality constraints only, such as $Q_1 + Q_2 = 950$. Our primary concern will be with *relative* constrained extrema, although *absolute* ones will also be discussed in Sec. 12.4.

12.1 EFFECTS OF A CONSTRAINT

The primary purpose of imposing a constraint is to give due cognizance to certain limiting factors present in the optimization problem under discussion.

We have already seen the limitation on output choices that result from a production quota. For further illustration, let us consider a consumer with the simple utility (index) function

$$U = x_1 x_2 + 2x_1 \tag{12.1}$$

Since the marginal utilities—the partial derivatives $U_1 \equiv \partial U/\partial x_1$ and $U_2 \equiv \partial U/\partial x_2$—are positive for all positive levels of x_1 and x_2 here, to have U maximized without any constraint, the consumer should purchase an *infinite* amount of both goods, a solution that obviously has little practical relevance. To render the optimization problem meaningful, the purchasing power of the consumer must also be taken into account; i.e., a *budget constraint* should be incorporated into the problem. If the consumer intends to spend a given sum, say, \$60, on the two goods and if the current prices are $P_{10} = 4$ and $P_{20} = 2$, then the budget constraint can be expressed by the linear equation

$$4x_1 + 2x_2 = 60 \tag{12.2}$$

Such a constraint, like the production quota referred to earlier, renders the choices of $\bar{x}_1$ and $\bar{x}_2$ mutually dependent.

The problem now is to maximize (12.1), subject to the constraint stated in (12.2). Mathematically, what the constraint (variously called *restraint*, *side relation*, or *subsidiary condition*) does is to narrow the domain, and hence the range of the objective function. The domain of (12.1) would normally be the set $\{(x_1, x_2) | x_1 \geq 0, x_2 \geq 0\}$. Graphically, the domain is represented by the nonnegative quadrant of the x_1x_2 plane in Fig. 12.1*a*. After the budget constraint (12.2) is added, however, we can admit only those values of the variables which satisfy this latter equation, so that the domain is immediately reduced to the set of points lying on the budget line. This will automatically affect the range of the objective function, too; only that subset of the utility surface lying directly above the budget-constraint line will now be relevant. The said subset (a cross section of the surface) may look like the curve in Fig. 12.1*b*, where U is plotted on the vertical axis, with the budget line of diagram *a* placed on the horizontal axis. Our interest, then, is only in locating the maximum on the curve in diagram *b*.

In general, for a function $z = f(x, y)$, the difference between a constrained extremum and a free extremum may be illustrated in the three-dimensional graph

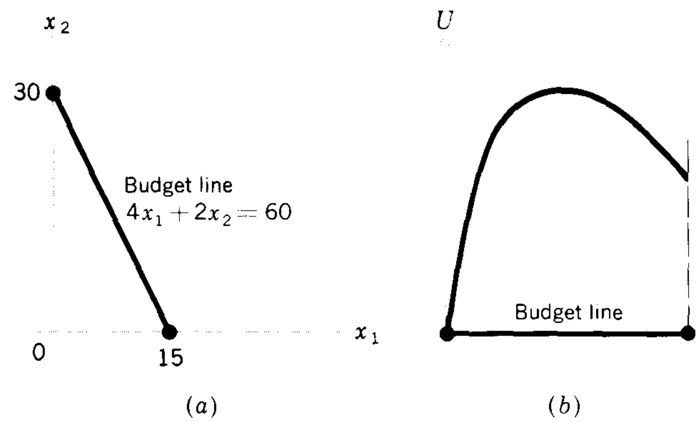

Figure 12.1

of Fig. 12.2. The free extremum in this particular graph is the peak point of the entire dome, but the constrained extremum is at the peak of the inverse U-shaped curve situated on top of (i.e., lying directly above) the constraint line. In general, a constrained maximum can be expected to have a lower value than the free maximum, although, by coincidence, the two maxima may happen to have the same value. But the constrained maximum can never exceed the free maximum.

It is interesting to note that, had we added another constraint intersecting the first constraint at a single point in the xy plane, the two constraints together would have restricted the domain to that single point. Then the locating of the

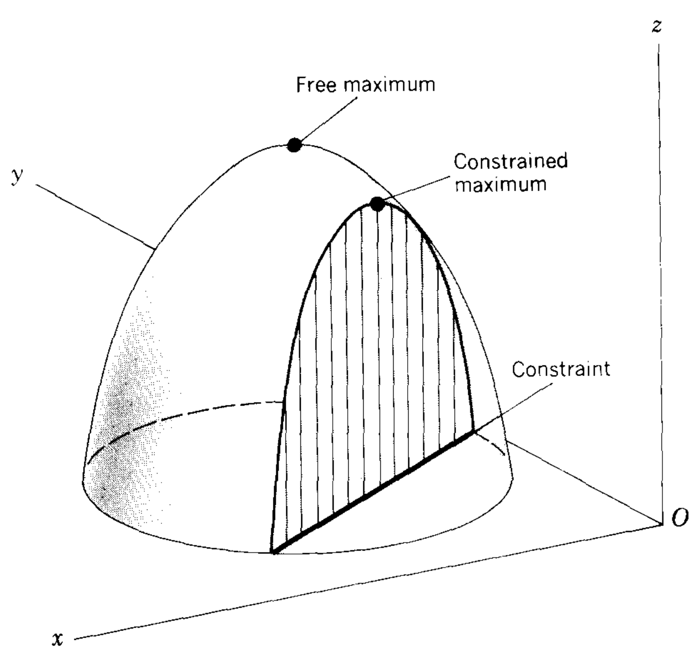

Figure 12.2

extremum would become a trivial matter. In a meaningful problem, the number and the nature of the constraints should be such as to restrict, but not eliminate, the possibility of choice. Generally, the number of constraints should be less than the number of choice variables.

12.2 FINDING THE STATIONARY VALUES

Even without any new technique of solution, the constrained maximum in the simple example defined by (12.1) and (12.2) can easily be found. Since the constraint (12.2) implies

$$x_2 = \frac{60 - 4x_1}{2} = 30 - 2x_1 \qquad (12.2')$$

we can combine the constraint with the objective function by substituting (12.2′) into (12.1). The result is an objective function in one variable only:

$$U = x_1(30 - 2x_1) + 2x_1 = 32x_1 - 2x_1^2$$

which can be handled with the method already learned. By setting $dU/dx_1 = 32 - 4x_1$ equal to zero, we get the solution $\bar{x}_1 = 8$, which by virtue of (12.2′) immediately leads to $\bar{x}_2 = 30 - 2(8) = 14$. From (12.1), we can then find the stationary value $\bar{U} = 128$; and since the second derivative is $d^2U/dx_1^2 = -4 < 0$, that stationary value constitutes a (constrained) maximum of U.*

When the constraint is itself a complicated function, or when there are several constraints to consider, however, the technique of substitution and elimination of variables could become a burdensome task. More importantly, when the constraint comes in a form such that we cannot solve it to express one variable (x_2) as an explicit function of the other (x_1), the elimination method would in fact be of no avail—even if x_2 were known to be an implicit function of x_1, that is, even if the conditions of the implicit-function theorem were satisfied. In such cases, we may resort to a method known as the *method of Lagrange (undetermined) multiplier*, which, as we shall see, has distinct analytical advantages.

Lagrange-Multiplier Method

The essence of the Lagrange-multiplier method is to convert a constrained-extremum problem into a form such that the first-order condition of the free-extremum problem can still be applied.

Given the problem of maximizing $U = x_1x_2 + 2x_1$, subject to the constraint $4x_1 + 2x_2 = 60$ [from (12.1) and (12.2)], let us write what is referred to as the *Lagrangian function*, which is a modified version of the objective function that

* You may recall that for the flower-bed problem of Exercise 9.4-2 the same technique of substitution was applied to find the maximum area, using a constraint (the available quantity of wire netting) to eliminate one of the two variables (the length or the width of the flower bed).

incorporates the constraint as follows:

$$Z = x_1x_2 + 2x_1 + \lambda(60 - 4x_1 - 2x_2) \tag{12.3}$$

The symbol λ (the Greek letter lambda), representing some as yet undetermined number, is called a *Lagrange (undetermined) multiplier*. If we can somehow be assured that $4x_1 + 2x_2 = 60$, so that the constraint will be satisfied, then the last term in (12.3) will vanish regardless of the value of λ. In that event, Z will be identical with U. Moreover, with the constraint out of the way, we only have to seek the *free* maximum of Z, in lieu of the *constrained* maximum of U, with respect to the two variables x_1 and x_2. The question is: How can we make the parenthetical expression in (12.3) vanish?

The tactic that will accomplish this is simply to treat λ as an additional variable in (12.3), i.e., to consider $Z = Z(\lambda, x_1, x_2)$. For then the first-order condition for free extremum will consist of the set of simultaneous equations

$$\begin{aligned} Z_\lambda(\equiv \partial Z/\partial\lambda) &= 60 - 4x_1 - 2x_2 = 0 \\ Z_1(\equiv \partial Z/\partial x_1) &= x_2 + 2 - 4\lambda = 0 \\ Z_2(\equiv \partial Z/\partial x_2) &= x_1 - 2\lambda = 0 \end{aligned} \tag{12.4}$$

and the first equation will automatically guarantee the satisfaction of the constraint. Thus, by incorporating the constraint into the Lagrangian function Z and by treating the Lagrange multiplier as an extra variable, we can obtain the constrained extremum $\bar{U}$ (two choice variables) simply by screening the stationary values of Z, taken as a *free* function of three choice variables.

Solving (12.4) for the critical values of the variables, we find $\bar{x}_1 = 8$, $\bar{x}_2 = 14$ (and $\bar{\lambda} = 4$). As expected, the values of $\bar{x}_1$ and $\bar{x}_2$ check with the answers already obtained by the substitution method. Furthermore, it is clear from (12.3) that $\bar{Z} = 128$; this is identical with the value of $\bar{U}$ found earlier, as it should be.

In general, given an objective function

$$z = f(x, y) \tag{12.5}$$

subject to the constraint

$$g(x, y) = c \tag{12.6}$$

where c is a constant,* we can write the Lagrangian function as

$$Z = f(x, y) + \lambda[c - g(x, y)] \tag{12.7}$$

For stationary values of Z, regarded as a function of the three variables λ, x, and

* It is also possible to subsume the constant c under the contraint function so that (12.6) appears instead as $G(x, y) = 0$, where $G(x, y) = g(x, y) - c$. In that case, (12.7) should be changed to $Z = f(x, y) + \lambda[0 - G(x, y)] = f(x, y) - \lambda G(x, y)$. The version in (12.6) is chosen because it facilitates the study of the comparative-static effect of a change in the constraint constant later [see (12.16)].

y, the necessary condition is

$$
\begin{aligned}
& Z_\lambda = c - g(x, y) = 0 \\
(12.8) \qquad & Z_x = f_x - \lambda g_x = 0 \\
& Z_y = f_y - \lambda g_y = 0
\end{aligned}
$$

Since the first equation in (12.8) is simply a restatement of (12.6), the stationary values of the Lagrangian function Z will automatically satisfy the constraint of the original function z. And since the expression $\lambda[c - g(x, y)]$ is now assuredly zero, the stationary values of Z in (12.7) must be identical with those of (12.5), subject to (12.6).

Let us illustrate the method with two more examples.

Example 1 Find the extremum of

$$z = xy \qquad \text{subject to} \qquad x + y = 6$$

The first step is to write the Lagrangian function

$$Z = xy + \lambda(6 - x - y)$$

For a stationary value of Z, it is necessary that

$$
\left.\begin{aligned}
Z_\lambda &= 6 - x - y = 0 \\
Z_x &= y - \lambda = 0 \\
Z_y &= x - \lambda = 0
\end{aligned}\right\}
\quad \text{or} \quad
\left\{\begin{aligned}
x + y &= 6 \\
-\lambda \qquad + y &= 0 \\
-\lambda + x \qquad &= 0
\end{aligned}\right.
$$

Thus, by Cramer's rule or some other method, we can find

$$\bar{\lambda} = 3 \qquad \bar{x} = 3 \qquad \bar{y} = 3$$

The stationary value is $\bar{Z} = \bar{z} = 9$, which needs to be tested against a second-order condition before we can tell whether it is a maximum or minimum (or neither). That will be taken up later.

Example 2 Find the extremum of

$$z = x_1^2 + x_2^2 \qquad \text{subject to} \qquad x_1 + 4x_2 = 2$$

The Lagrangian function is

$$Z = x_1^2 + x_2^2 + \lambda(2 - x_1 - 4x_2)$$

for which the necessary condition for a stationary value is

$$
\left.\begin{aligned}
Z_\lambda &= 2 - x_1 - 4x_2 = 0 \\
Z_1 &= 2x_1 - \lambda = 0 \\
Z_2 &= 2x_2 - 4\lambda = 0
\end{aligned}\right\}
\quad \text{or} \quad
\left\{\begin{aligned}
x_1 + 4x_2 &= 2 \\
-\lambda + 2x_1 \qquad &= 0 \\
-4\lambda \qquad + 2x_2 &= 0
\end{aligned}\right.
$$

The stationary value of Z, defined by the solution

$$\bar{\lambda} = \tfrac{4}{17} \qquad \bar{x}_1 = \tfrac{2}{17} \qquad \bar{x}_2 = \tfrac{8}{17}$$

is therefore $\bar{Z} = \bar{z} = \frac{4}{17}$. Again, a second-order condition should be consulted before we can tell whether $\bar{z}$ is a maximum or a minimum.

Total-Differential Approach

In the discussion of the free extremum of $z = f(x, y)$, it was learned that the first-order necessary condition may be stated in terms of the total differential dz as follows:

$$(12.9) \qquad dz = f_x\, dx + f_y\, dy = 0$$

This statement remains valid after a constraint $g(x, y) = c$ is added. However, with the constraint in the picture, we can no longer take dx and dy both as "arbitrary" changes as before. For if $g(x, y) = c$, then dg must be equal to dc, which is zero since c is a constant. Hence,

$$(12.10) \qquad (dg =)\, g_x\, dx + g_y\, dy = 0$$

and this relation makes dx and dy dependent on each other. The first-order necessary condition therefore becomes $dz = 0$ [(12.9)], subject to $g = c$, and hence also subject to $dg = 0$ [(12.10)]. By visual inspection of (12.9) and (12.10), it should be clear that, in order to satisfy this necessary condition, we must have

$$(12.11) \qquad \frac{f_x}{g_x} = \frac{f_y}{g_y}$$

This result can be verified by solving (12.10) for dy and substituting the result into (12.9). The condition (12.11), together with the constraint $g(x, y) = c$, will provide two equations from which to find the critical values of x and y.*

Does the total-differential approach yield the same first-order condition as the Lagrange-multiplier method? Let us compare (12.8) with the result just obtained. The first equation in (12.8) merely repeats the constraint; the new result requires its satisfaction also. The last two equations in (12.8) can be rewritten, respectively, as

$$(12.11') \qquad \frac{f_x}{g_x} = \lambda \qquad \text{and} \qquad \frac{f_y}{g_y} = \lambda$$

and these convey precisely the same information as (12.11). Note, however, that whereas the total-differential approach yields only the values of $\bar{x}$ and $\bar{y}$, the Lagrange-multiplier method also gives the value of $\bar{\lambda}$ as a direct by-product. As it turns out, $\bar{\lambda}$ provides a measure of the sensitivity of $\bar{Z}$ (and $\bar{z}$) to a shift of the constraint, as we shall presently demonstrate. Therefore, the Lagrange-multiplier

* Note that the constraint $g = c$ is still to be considered along with (12.11), even though we have utilized the equation $dg = 0$—that is, (12.10)—in deriving (12.11). While $g = c$ necessarily implies $dg = 0$, the converse is not true: $dg = 0$ merely implies $g =$ a constant (not necessarily c). Unless the constraint is explicitly considered, therefore, some information will be unwittingly left out of the problem.

method offers the advantage of containing certain built-in comparative-static information in the solution.

An Interpretation of the Lagrange Multiplier

To show that $\bar{\lambda}$ indeed measures the sensitivity of $\bar{Z}$ to changes in the constraint, let us perform a comparative-static analysis on the first-order condition (12.8). Since λ, x, and y are endogenous, the only available exogenous variable is the constraint parameter c. A change in c would cause a shift of the constraint curve in the xy plane and thereby alter the optimal solution. In particular, the effect of an *increase* in c (a larger budget, or a larger production quota) would indicate how the optimal solution is affected by a *relaxation* of the constraint.

To do the comparative-static analysis, we again resort to the implicit-function theorem. Taking the three equations in (12.8) to be in the form of $F^j(\lambda, x, y; c) = 0$ (with $j = 1, 2, 3$), and assuming them to have continuous partial derivatives, we must first check that the following endogenous-variable Jacobian (where $f_{xy} = f_{yx}$, and $g_{xy} = g_{yx}$)

$$(12.12)\qquad |J| = \begin{vmatrix} \frac{\partial F^1}{\partial \lambda} & \frac{\partial F^1}{\partial x} & \frac{\partial F^1}{\partial y} \\ \frac{\partial F^2}{\partial \lambda} & \frac{\partial F^2}{\partial x} & \frac{\partial F^2}{\partial y} \\ \frac{\partial F^3}{\partial \lambda} & \frac{\partial F^3}{\partial x} & \frac{\partial F^3}{\partial y} \end{vmatrix} = \begin{vmatrix} 0 & -g_x & -g_y \\ -g_x & f_{xx} - \lambda g_{xx} & f_{xy} - \lambda g_{xy} \\ -g_y & f_{xy} - \lambda g_{xy} & f_{yy} - \lambda g_{yy} \end{vmatrix}$$

does not vanish in the optimal state. At this moment, there is certainly no inkling that this would be the case. But our previous experience with the comparative statics of optimization problems [see the discussion of (11.42)] would suggest that this Jacobian is closely related to the second-order sufficient condition, and that if the sufficient condition is satisfied, then the Jacobian will be nonzero at the equilibrium (optimum). Leaving the full demonstration of this fact to the following section, let us proceed on the assumption that $|J| \neq 0$. If so, then we can express $\bar{\lambda}$, $\bar{x}$, and $\bar{y}$ all as implicit functions of the parameter c:

$$(12.13)\qquad \bar{\lambda} = \bar{\lambda}(c) \qquad \bar{x} = \bar{x}(c) \qquad \text{and} \qquad \bar{y} = \bar{y}(c)$$

all of which will have continuous derivatives. Also, we have the identities

$$(12.14)\qquad \begin{aligned} c - g(\bar{x}, \bar{y}) &\equiv 0 \\ f_x(\bar{x}, \bar{y}) - \bar{\lambda} g_x(\bar{x}, \bar{y}) &\equiv 0 \\ f_y(\bar{x}, \bar{y}) - \bar{\lambda} g_y(\bar{x}, \bar{y}) &\equiv 0 \end{aligned}$$

Now since the optimal value of Z depends on $\bar{\lambda}$, $\bar{x}$, and $\bar{y}$, that is,

$$(12.15)\qquad \bar{Z} = f(\bar{x}, \bar{y}) + \bar{\lambda}[c - g(\bar{x}, \bar{y})]$$

we may, in view of (12.13), consider $\bar{Z}$ to be a function of c alone. Differentiating

$\bar{Z}$ totally with respect to c, we find

$$\begin{aligned}\frac{d\bar{Z}}{dc} &= f_x\frac{d\bar{x}}{dc} + f_y\frac{d\bar{y}}{dc} + [c - g(\bar{x}, \bar{y})]\frac{d\bar{\lambda}}{dc} + \bar{\lambda}\left(1 - g_x\frac{d\bar{x}}{dc} - g_y\frac{d\bar{y}}{dc}\right)\\ &= (f_x - \bar{\lambda}g_x)\frac{d\bar{x}}{dc} + (f_y - \bar{\lambda}g_y)\frac{d\bar{y}}{dc} + [c - g(\bar{x}, \bar{y})]\frac{d\bar{\lambda}}{dc} + \bar{\lambda}\end{aligned}$$

where f_x, f_y, g_x, and g_y are all to be evaluated at the optimum. By (12.14), however, the first three terms on the right will all drop out. Thus we are left with the simple result

$$\frac{d\bar{Z}}{dc} = \bar{\lambda} \tag{12.16}$$

which validates our claim that the solution value of the Lagrange multiplier constitutes a measure of the effect of a change in the constraint via the parameter c on the optimal value of the objective function.

A word of caution, however, is perhaps in order here. For this interpretation of $\bar{\lambda}$, you must express Z specifically as in (12.7). In particular, write the last term as $\lambda[c - g(x, y)]$, *not* $\lambda[g(x, y) - c]$.

n-Variable and Multiconstraint Cases

The generalization of the Lagrange-multiplier method to n variables can be easily carried out if we write the choice variables in subscript notation. The objective function will then be in the form

$$z = f(x_1, x_2, \ldots, x_n)$$

subject to the constraint

$$g(x_1, x_2, \ldots, x_n) = c$$

It follows that the Lagrangian function will be

$$Z = f(x_1, x_2, \ldots, x_n) + \lambda[c - g(x_1, x_2, \ldots, x_n)]$$

for which the first-order condition will consist of the following $(n + 1)$ simultaneous equations:

$$\begin{aligned}Z_\lambda &= c - g(x_1, x_2, \ldots, x_n) = 0\\ Z_1 &= f_1 - \lambda g_1 = 0\\ Z_2 &= f_2 - \lambda g_2 = 0\\ &\cdots\cdots\cdots\cdots\\ Z_n &= f_n - \lambda g_n = 0\end{aligned}$$

Again, the first of these equations will assure us that the constraint is met, even though we are to focus our attention on the *free* Lagrangian function.

When there is more than one constraint, the Lagrange-multiplier method is equally applicable, provided we introduce as many such multipliers as there are constraints in the Lagrangian function. Let an n-variable function be subject simultaneously to the two constraints

$$g(x_1, x_2, \ldots, x_n) = c \quad \text{and} \quad h(x_1, x_2, \ldots, x_n) = d$$

Then, adopting λ and μ (the Greek letter mu) as the two undetermined multipliers, we may construct a Lagrangian function as follows:

$$Z = f(x_1, x_2, \ldots, x_n) + \lambda[c - g(x_1, x_2, \ldots, x_n)] + \mu[d - h(x_1, x_2, \ldots, x_n)]$$

This function will have the same value as the original objective function f if both constraints are satisfied, i.e., if the last two terms in the Lagrangian function both vanish. Considering λ and μ as variables, we now count $(n + 2)$ variables altogether; thus the first-order condition will in this case consist of the following $(n + 2)$ simultaneous equations:

$$Z_\lambda = c - g(x_1, x_2, \ldots, x_n) = 0$$
$$Z_\mu = d - h(x_1, x_2, \ldots, x_n) = 0$$
$$Z_i = f_i - \lambda g_i - \mu h_i = 0 \qquad (i = 1, 2, \ldots, n)$$

These should normally enable us to solve for all the x_i as well as λ and μ. As before, the first two equations of the necessary condition represent essentially a mere restatement of the two constraints.

EXERCISE 12.2

1 Use the Lagrange-multiplier method to find the stationary values of z:
(a) $z = xy$, subject to $x + 2y = 2$
(b) $z = x(y + 4)$, subject to $x + y = 8$
(c) $z = x - 3y - xy$, subject to $x + y = 6$
(d) $z = 7 - y + x^2$, subject to $x + y = 0$

2 In the above problem, find whether a slight relaxation of the constraint will increase or decrease the optimal value of z. At what rate?

3 Write the Lagrangian function and the first-order condition for stationary values (without solving the equations) for each of the following:
(a) $z = x + 2y + 3w + xy - yw$, subject to $x + y + 2w = 10$
(b) $z = x^2 + 2xy + yw^2$, subject to $2x + y + w^2 = 24$ and $x + w = 8$

4 If, instead of $g(x, y) = c$, the constraint is written in the form of $G(x, y) = 0$, how should the Lagrangian function and the first-order condition be modified as a consequence?

5 In discussing the total-differential approach, it was pointed out that, given the constraint $g(x, y) = c$, we may deduce that $dg = 0$. By the same token, we can further deduce that $d^2g = d(dg) = d(0) = 0$. Yet, in our earlier discussion of the unconstrained extremum of a function $z = f(x, y)$, we had a situation where $dz = 0$ is accompanied by either a positive definite or a negative definite d^2z, rather than $d^2z = 0$. How would you account for this disparity of treatment in the two cases?

6 If the Lagrangian function is written as $Z = f(x, y) + \lambda[g(x, y) - c]$ rather than as in (12.7), can we still interpret the Lagrange multiplier as in (12.16)? Give the new interpretation, if any.

12.3 SECOND-ORDER CONDITIONS

The introduction of a Lagrange multiplier as an additional variable makes it possible to apply to the constrained-extremum problem the same first-order condition used in the free-extremum problem. It is tempting to go a step further and borrow the second-order necessary and sufficient conditions as well. This, however, should not be done. For even though $\bar{Z}$ is indeed a standard type of extremum with respect to the choice variables, it is *not* so with respect to the Lagrange multiplier. Specifically, we can see from (12.15) that, unlike $\bar{x}$ and $\bar{y}$, if $\bar{\lambda}$ is replaced by any other value of λ, no effect will be produced on $\bar{Z}$, since $[c - g(\bar{x}, \bar{y})]$ is identically zero. Thus the role played by λ in the optimal solution differs basically from that of x and y.* While it is harmless to treat λ as just another choice variable in the discussion of the first-order condition, we must be careful not to apply blindly the second-order conditions developed for the free-extremum problem to the present constrained case. Rather, we must derive a set of new ones. As we shall see, the new conditions can again be stated in terms of the second-order total differential d^2z. However, the presence of the constraint will entail certain significant modifications of the criterion.

Second-Order Total Differential

It has been mentioned that, inasmuch as the constraint $g(x, y) = c$ means $dg = g_x\,dx + g_y\,dy = 0$, as in (12.10), dx and dy no longer are both arbitrary. We may, of course, still take (say) dx as an arbitrary change, but then dy must be regarded as dependent on dx, always to be chosen so as to satisfy (12.10), i.e., to satisfy $dy = -(g_x/g_y)\,dx$. Viewed differently, once the value of dx is specified, dy will depend on g_x and g_y, but since the latter derivatives in turn depend on the variables x and y, dy will also depend on x and y. Obviously, then, the earlier formula for d^2z in (11.6), which is based on the arbitrariness of both dx and dy, can no longer apply.

To find an appropriate new expression for d^2z, we must treat dy as a variable dependent on x and y during differentiation (if dx is to be considered a constant).

* In a more general framework of constrained optimization known as "nonlinear programming," to be discussed in a later chapter, it will be shown that, with inequality constraints, if $\bar{Z}$ is a maximum (minimum) with respect to x and y, then it will in fact be a minimum (maximum) with respect to λ. In other words, the point $(\bar{\lambda}, \bar{x}, \bar{y})$ is a saddle point. The present case—where $\bar{Z}$ is a genuine extremum with respect to x and y, but is invariant with respect to λ—may be considered as a degenerate case of saddle point. The saddle-point nature of the solution $(\bar{\lambda}, \bar{x}, \bar{y})$ also leads to the important concept of "duality." But this subject is best to be pursued later.

Thus,

$$\begin{aligned} d^2z = d(dz) &= \frac{\partial(dz)}{\partial x}dx + \frac{\partial(dz)}{\partial y}dy \\ &= \frac{\partial}{\partial x}(f_x\,dx + f_y\,dy)\,dx + \frac{\partial}{\partial y}(f_x\,dx + f_y\,dy)\,dy \\ &= \left[f_{xx}\,dx + \left(f_{xy}\,dy + f_y\frac{\partial dy}{\partial x}\right)\right]dx + \left[f_{yx}\,dx + \left(f_{yy}\,dy + f_y\frac{\partial dy}{\partial y}\right)\right]dy \\ &= f_{xx}\,dx^2 + f_{xy}\,dy\,dx + f_y\frac{\partial(dy)}{\partial x}dx + f_{yx}\,dx\,dy + f_{yy}\,dy^2 + f_y\frac{\partial(dy)}{\partial y}dy \end{aligned}$$

Since the third and the sixth terms can be reduced to

$$f_y\left[\frac{\partial(dy)}{\partial x}dx + \frac{\partial(dy)}{\partial y}dy\right] = f_y\,d(dy) = f_y\,d^2y$$

the desired expression for d^2z is

$$d^2z = f_{xx}\,dx^2 + 2f_{xy}\,dx\,dy + f_{yy}\,dy^2 + f_y\,d^2y \tag{12.17}$$

which differs from (11.6) only by the last term, $f_y\,d^2y$.

It should be noted that this last term is in the *first* degree [d^2y is *not* the same as $(dy)^2$]; thus its presence in (12.17) disqualifies d^2z as a quadratic form. However, d^2z can be transformed into a quadratic form by virtue of the constraint $g(x, y) = c$. Since the constraint implies $dg = 0$ and also $d^2g = d(dg) = 0$, so by the procedure used in obtaining (12.17) we can get

$$(d^2g =)\,g_{xx}\,dx^2 + 2g_{xy}\,dx\,dy + g_{yy}\,dy^2 + g_y\,d^2y = 0$$

Solving this last equation for d^2y and substituting the result in (12.17), we are able to eliminate the first-degree expression d^2y and write d^2z as the following quadratic form:

$$d^2z = \left(f_{xx} - \frac{f_y}{g_y}g_{xx}\right)dx^2 + 2\left(f_{xy} - \frac{f_y}{g_y}g_{xy}\right)dx\,dy + \left(f_{yy} - \frac{f_y}{g_y}g_{yy}\right)dy^2$$

Because of (12.11′), the first parenthetical coefficient is reducible to $(f_{xx} - \lambda g_{xx})$, and similarly for the other terms. However, by partially differentiating the derivatives in (12.8), you will find that the following second derivatives

$$\begin{aligned} Z_{xx} &= f_{xx} - \lambda g_{xx} \\ Z_{xy} &= f_{xy} - \lambda g_{xy} = Z_{yx} \\ Z_{yy} &= f_{yy} - \lambda g_{yy} \end{aligned} \tag{12.18}$$

are precisely equal to these parenthetical coefficients. Hence, by making use of the Lagrangian function, we can finally express d^2z more neatly as follows:

$$\begin{aligned} d^2z = \; & Z_{xx}dx^2 + Z_{xy}dx\,dy \\ & + Z_{yx}\,dy\,dx + Z_{yy}\,dy^2 \end{aligned} \tag{12.17′}$$

The coefficients of (12.17′) are simply the second partial derivatives of Z with respect to the choice variables x and y; together, therefore, they can give rise to a Hessian determinant.

Second-Order Conditions

For a constrained extremum of $z = f(x, y)$, subject to $g(x, y) = c$, the second-order necessary and sufficient conditions still revolve around the algebraic sign of the second-order total differential d^2z, evaluated at a stationary point. However, there is one important change. In the present context, we are concerned with the sign definiteness or semidefiniteness of d^2z, *not* for all possible values of dx and dy (not both zero), but *only* for those dx and dy values (not both zero) satisfying the linear constraint (12.10), $g_x\,dx + g_y\,dy = 0$. Thus the second-order *necessary* conditions are:

For maximum of z: d^2z negative semidefinite, subject to $dg = 0$

For minimum of z: d^2z positive semidefinite, subject to $dg = 0$

and the second-order *sufficient* conditions are:

For maximum of z: d^2z negative definite, subject to $dg = 0$

For minimum of z: d^2z positive definite, subject to $dg = 0$

In the following, we shall concentrate on the second-order sufficient conditions.

Inasmuch as the (dx, dy) pairs satisfying the constraint $g_x\,dx + g_y\,dy = 0$ constitute merely a subset of the set of all possible dx and dy, the constrained sign definiteness is less stringent—that is, easier to satisfy—than the unconstrained sign definiteness discussed in the preceding chapter. In other words, the second-order sufficient condition for a constrained extremum problem is a weaker condition than that for a free extremum problem. This is welcome news because, unlike necessary conditions which must be stringent in order to serve as effective screening devices, sufficient conditions should be weak to be truly serviceable.*

The Bordered Hessian

As in the case of free extremum, it is possible to express the second-order sufficient condition in determinantal form. In place of the Hessian determinant $|H|$, however, in the constrained-extremum case we shall encounter what is known as a *bordered Hessian*.

In preparation for the development of this idea, let us first analyze the conditions for the sign definiteness of a two-variable quadratic form, subject to a

* "A million-dollar bank deposit" is clearly a sufficient condition for "being able to afford a steak dinner." But the extremely limited applicability of that condition renders it practically useless. A more meaningful sufficient condition might be something like "twenty dollars in one's wallet," which is a much less stringent financial requirement.

linear constraint, say,

$$q = au^2 + 2huv + bv^2 \qquad \text{subject to} \qquad \alpha u + \beta v = 0$$

Since the constraint implies $v = -(\alpha/\beta)u$, we can rewrite q as a function of one variable only:

$$q = au^2 - 2h\frac{\alpha}{\beta}u^2 + b\frac{\alpha^2}{\beta^2}u^2 = (a\beta^2 - 2h\alpha\beta + b\alpha^2)\frac{u^2}{\beta^2}$$

It is obvious that q is positive (negative) definite if and only if the expression in parentheses is positive (negative). Now, it so happens that the following symmetric determinant

$$\begin{vmatrix} 0 & \alpha & \beta \\ \alpha & a & h \\ \beta & h & b \end{vmatrix} = 2h\alpha\beta - a\beta^2 - b\alpha^2$$

is exactly the *negative* of the said parenthetical expression. Consequently, we can state that

$$q \text{ is } \begin{Bmatrix} \text{positive definite} \\ \text{negative definite} \end{Bmatrix} \text{ subject to } \alpha u + \beta v = 0$$

$$\text{iff} \qquad \begin{vmatrix} 0 & \alpha & \beta \\ \alpha & a & h \\ \beta & h & b \end{vmatrix} \begin{Bmatrix} < 0 \\ > 0 \end{Bmatrix}$$

It is noteworthy that the determinant used in this criterion is nothing but the discriminant of the original quadratic form $\begin{vmatrix} a & h \\ h & b \end{vmatrix}$, with a border placed on top and a similar border on the left. Furthermore, the border is merely composed of the two coefficients α and β from the constraint, plus a zero in the principal diagonal. This bordered discriminant is symmetric.

Example 1 Determine whether $q = 4u^2 + 4uv + 3v^2$, subject to $u - 2v = 0$, is either positive or negative definite. We first form the bordered discriminant $\begin{vmatrix} 0 & 1 & -2 \\ 1 & 4 & 2 \\ -2 & 2 & 3 \end{vmatrix}$, which is made symmetric by splitting the coefficient of uv into two equal parts for insertion into the determinant. Inasmuch as the determinant has a negative value (-27), q must be positive definite.

When applied to the quadratic form d^2z in (12.17′), the variables u and v become dx and dy, respectively, and the (plain) discriminant consists of the Hessian $\begin{vmatrix} Z_{xx} & Z_{xy} \\ Z_{yx} & Z_{yy} \end{vmatrix}$. Moreover, the constraint to the quadratic form being $g_x\,dx + g_y\,dy = 0$, we have $\alpha = g_x$ and $\beta = g_y$. Thus, for values of dx and dy that satisfy the said constraint, we now have the following determinantal criterion for the sign definiteness of d^2z:

$$d^2z \text{ is } \begin{Bmatrix} \text{positive definite} \\ \text{negative definite} \end{Bmatrix} \text{ subject to } dg = 0$$

$$\text{iff} \quad \begin{vmatrix} 0 & g_x & g_y \\ g_x & Z_{xx} & Z_{xy} \\ g_y & Z_{yx} & Z_{yy} \end{vmatrix} \begin{Bmatrix} < 0 \\ > 0 \end{Bmatrix}$$

The determinant to the right, often referred to as a bordered Hessian, shall be denoted by $|\bar{H}|$, where the bar on top symbolizes the border. On the basis of this, we may conclude that, given a stationary value of $z = f(x, y)$ or of $Z = f(x, y) + \lambda[c - g(x, y)]$, a positive $|\bar{H}|$ is sufficient to establish it as a relative maximum of z; similarly, a negative $|\bar{H}|$ is sufficient to establish it as a minimum—all the derivatives involved in $|\bar{H}|$ being evaluated at the critical values of x and y.

Now that we have derived the second-order sufficient condition, it is an easy matter to verify that, as earlier claimed, the satisfaction of this condition will guarantee that the endogenous-variable Jacobian (12.12) does not vanish in the optimal state. Substituting (12.18) into (12.12), and multiplying both the first column and the first row of the Jacobian by -1 (which will leave the value of the determinant unaltered), we see that

$$(12.19) \qquad |J| = \begin{vmatrix} 0 & g_x & g_y \\ g_x & Z_{xx} & Z_{xy} \\ g_y & Z_{yx} & Z_{yy} \end{vmatrix} = |\bar{H}|$$

That is, the endogenous-variable Jacobian is identical with the bordered Hessian —a result similar to (11.42) where it was shown that, in the free-extremum context, the endogenous-variable Jacobian is identical with the plain Hessian. If, in fulfillment of the sufficient condition, we have $|\bar{H}| \neq 0$ at the optimum, then $|J|$ must also be nonzero. Consequently, in applying the implicit-function theorem to the present context, it would not be amiss to substitute the condition $|\bar{H}| \neq 0$ for the usual condition $|J| \neq 0$. This practice will be followed when we analyze the comparative statics of constrained-optimization problems below.

Example 2 Let us now return to Example 1 of Sec. 12.2 and ascertain whether the stationary value found there gives a maximum or a minimum. Since $Z_x = y - \lambda$ and $Z_y = x - \lambda$, the second-order partial derivatives are $Z_{xx} = 0$, $Z_{xy} = Z_{yx} = 1$, and $Z_{yy} = 0$. The border elements we need are $g_x = 1$ and $g_y = 1$. Thus we find that

$$|\bar{H}| = \begin{vmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{vmatrix} = 2 > 0$$

which establishes the value $\bar{z} = 9$ as a maximum.

Example 3 Continuing on to Example 2 of Sec. 12.2, we see that $Z_1 = 2x_1 - \lambda$ and $Z_2 = 2x_2 - 4\lambda$. These yield $Z_{11} = 2$, $Z_{12} = Z_{21} = 0$, and $Z_{22} = 2$. From the constraint $x_1 + 4x_2 = 2$, we obtain $g_1 = 1$ and $g_2 = 4$. It follows that the bordered Hessian is

$$|\bar{H}| = \begin{vmatrix} 0 & 1 & 4 \\ 1 & 2 & 0 \\ 4 & 0 & 2 \end{vmatrix} = -34 < 0$$

and the value $\bar{z} = \frac{4}{17}$ is a minimum.

n-Variable Case

When the objective function takes the form

$$z = f(x_1, x_2, \ldots, x_n) \qquad \text{subject to} \qquad g(x_1, x_2, \ldots, x_n) = c$$

the second-order condition still hinges on the sign of d^2z. Since the latter is a constrained quadratic form in the variables $dx_1, dx_2, \ldots, dx_n$, subject to the relation

$$(dg =) g_1\,dx_1 + g_2\,dx_2 + \cdots + g_n\,dx_n = 0$$

the conditions for the positive or negative definiteness of d^2z again involve a bordered Hessian. But this time these conditions must be expressed in terms of the bordered principal minors of the Hessian.

Given a bordered Hessian

$$|\bar{H}| = \begin{vmatrix} 0 & g_1 & g_2 & \cdots & g_n \\ g_1 & Z_{11} & Z_{12} & \cdots & Z_{1n} \\ g_2 & Z_{21} & Z_{22} & \cdots & Z_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ g_n & Z_{n1} & Z_{n2} & \cdots & Z_{nn} \end{vmatrix}$$

its successive bordered principal minors can be defined as

$$|\bar{H}_2| \equiv \begin{vmatrix} 0 & g_1 & g_2 \\ g_1 & Z_{11} & Z_{12} \\ g_2 & Z_{21} & Z_{22} \end{vmatrix} \qquad |\bar{H}_3| \equiv \begin{vmatrix} 0 & g_1 & g_2 & g_3 \\ g_1 & Z_{11} & Z_{12} & Z_{13} \\ g_2 & Z_{21} & Z_{22} & Z_{23} \\ g_3 & Z_{31} & Z_{32} & Z_{33} \end{vmatrix} \qquad \text{(etc.)}$$

with the last one being $|\bar{H}_n| = |\bar{H}|$. In the newly introduced symbols, the horizontal bar above H again means bordered, and the subscript indicates the order of the principal minor being bordered. For instance, $|\bar{H}_2|$ involves the second principal minor of the (plain) Hessian, bordered with 0, g_1, and g_2; and similarly for the others. The conditions for positive and negative definiteness of

Table 12.1 Determinantal test for relative constrained extremum:
$z = f(x_1, x_2, \ldots, x_n)$, **subject to** $g(x_1, x_2, \ldots, x_n) = c$; **with**
$Z = f(x_1, x_2, \ldots, x_n) + \lambda[c - g(x_1, x_2, \ldots, x_n)]$

Condition	Maximum	Minimum
First-order necessary condition	$Z_\lambda = Z_1 = Z_2 = \cdots = Z_n = 0$	$Z_\lambda = Z_1 = Z_2 = \cdots = Z_n = 0$
Second-order sufficient condition*	$\lvert\bar{H}_2\rvert > 0$; $\lvert\bar{H}_3\rvert < 0$; $\lvert\bar{H}_4\rvert > 0; \ldots; (-1)^n\lvert\bar{H}_n\rvert > 0$	$\lvert\bar{H}_2\rvert, \lvert\bar{H}_3\rvert, \ldots, \lvert\bar{H}_n\rvert < 0$

*Applicable only after the first-order necessary condition has been satisfied.

d^2z are then

$$d^2z \text{ is } \begin{Bmatrix} \text{positive definite} \\ \text{negative definite} \end{Bmatrix} \text{ subject to } dg = 0$$

$$\text{iff} \quad \begin{cases} |\bar{H}_2|, |\bar{H}_3|, \ldots, |\bar{H}_n| < 0 \\ |\bar{H}_2| > 0; |\bar{H}_3| < 0; |\bar{H}_4| > 0; \text{etc.} \end{cases}$$

In the former, all the bordered principal minors, starting with $|\bar{H}_2|$, must be negative; in the latter, they must alternate in sign. As previously, a positive definite d^2z is sufficient to establish a stationary value of z as its minimum, whereas a negative definite d^2z is sufficient to establish it as a maximum.

Drawing the threads of the discussion together, we may summarize the conditions for a constrained relative extremum in Table 12.1. You will recognize, however, that the criterion stated in the table is not complete. Because the second-order sufficient condition is *not* necessary, failure to satisfy the criteria stated does not preclude the possibility that the stationary value is nonetheless a maximum or a minimum as the case may be. In many economic applications, however, this (relatively less stringent) second-order sufficient condition is either satisfied, or assumed to be satisfied, so that the information in the table is adequate. It should prove instructive for you to compare the results contained in Table 12.1 with those in Table 11.2 for the free-extremum case.

Multiconstraint Case

When more than one constraint appears in the problem, the second-order condition involves a Hessian with more than one border. Suppose that there are n choice variables and m constraints ($m < n$) of the form $g^j(x_1, \ldots, x_n) = c_j$. Then the Lagrangian function will be

$$Z = f(x_1, \ldots, x_n) + \sum_{j=1}^{m} \lambda_j \left[c_j - g^j(x_1, \ldots, x_n)\right]$$

and the bordered Hessian will appear as

$$|\bar{H}| \equiv \left|\begin{array}{cccc|cccc} 0 & 0 & \cdots & 0 & g_1^1 & g_2^1 & \cdots & g_n^1 \\ 0 & 0 & \cdots & 0 & g_1^2 & g_2^2 & \cdots & g_n^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & g_1^m & g_2^m & \cdots & g_n^m \\ \hline g_1^1 & g_1^2 & \cdots & g_1^m & Z_{11} & Z_{12} & \cdots & Z_{1n} \\ g_2^1 & g_2^2 & \cdots & g_2^m & Z_{21} & Z_{22} & \cdots & Z_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ g_n^1 & g_n^2 & \cdots & g_n^m & Z_{n1} & Z_{n2} & \cdots & Z_{nn} \end{array}\right|$$

where $g_i^j \equiv \partial g^j/\partial x_i$ are the partial derivatives of the constraint functions, and the double-subscripted Z symbols denote, as before, the second-order partial derivatives of the Lagrangian function. Note that we have partitioned the bordered Hessian into four *areas* for visual clarity. The upper-left area consists of zeros only, and the lower-right area is simply the plain Hessian. The other two areas, containing the g_i^j derivatives, bear a mirror-image relationship to each other with reference to the principal diagonal, thereby resulting in a symmetric array of elements in the entire bordered Hessian.

Various bordered principal minors can be formed from $|\bar{H}|$. The one that contains Z_{22} as the last element of its principal diagonal may be denoted by $|\bar{H}_2|$, as before. By including one more row and one more column, so that Z_{33} enters into the scene, we will have $|\bar{H}_3|$, and so forth. With this symbolism, we can state the second-order sufficient condition in terms of the signs of the following $(n-m)$ bordered principal minors:

$$|\bar{H}_{m+1}|, |\bar{H}_{m+2}|, \ldots, |\bar{H}_n| (= |\bar{H}|)$$

For a maximum of z, a sufficient condition is that these bordered principal minors alternate in sign, the sign of $|\bar{H}_{m+1}|$ being that of $(-1)^{m+1}$. For a minimum of z, a sufficient condition is that these bordered principal minors all take the same sign, namely, that of $(-1)^m$.

Note that it makes an important difference whether we have an odd or even number of constraints, because (-1) raised to an odd power will yield the opposite sign to the case of an even power. Note, also, that when $m = 1$, the condition just stated reduces to that presented in Table 12.1.

EXERCISE 12.3

1 Use the bordered Hessian to determine whether the stationary value of z obtained in each part of Exercise 12.2-1 is a maximum or a minimum.

2 In stating the second-order sufficient conditions for constrained maximum and minimum, we specified the algebraic signs of $|\bar{H}_2|$, $|\bar{H}_3|$, $|\bar{H}_4|$, etc., but not of $|\bar{H}_1|$. Write out an appropriate expression for $|\bar{H}_1|$, and verify that it invariably takes the negative sign.

3 Recalling Property II of determinants (Sec. 5.3), show that:

(*a*) By appropriately interchanging two rows and/or two columns of $|\bar{H}_2|$ and duly altering the sign of the determinant after each interchange, it can be transformed into

$$\begin{vmatrix} Z_{11} & Z_{12} & g_1 \\ Z_{21} & Z_{22} & g_2 \\ g_1 & g_2 & 0 \end{vmatrix}$$

(*b*) By a similar procedure, $|\bar{H}_3|$ can be transformed into

$$\begin{vmatrix} Z_{11} & Z_{12} & Z_{13} & g_1 \\ Z_{21} & Z_{22} & Z_{23} & g_2 \\ Z_{31} & Z_{32} & Z_{33} & g_3 \\ g_1 & g_2 & g_3 & 0 \end{vmatrix}$$

What alternative way of "bordering" the principal minors of the Hessian do these results suggest?

4 Write out the bordered Hessian for a constrained optimization problem with four choice variables and two constraints. Then state specifically the second-order sufficient condition for a maximum and for a minimum of z, respectively.

12.4 QUASICONCAVITY AND QUASICONVEXITY

In Sec. 11.5 it was shown that, for a problem of free extremum, a knowledge of the concavity or convexity of the objective function obviates the need to check the second-order condition. In the context of constrained optimization, it is again possible to dispense with the second-order condition if the surface or hypersurface has the appropriate type of configuration. But this time the desired configuration is quasiconcavity (rather than concavity) for a maximum, and quasiconvexity (rather than convexity) for a minimum. As we shall demonstrate, quasiconcavity (quasiconvexity) is a weaker condition than concavity (convexity). This is only to be expected, since the second-order sufficient condition to be dispensed with is also weaker for the constrained optimization problem (d^2z definite in sign only for those dx_i satisfying $dg = 0$) than for the free one (d^2z definite in sign for *all* dx_i).

Geometric Characterization

Quasiconcavity and quasiconvexity, like concavity and convexity, can be either strict or nonstrict. We shall first present the geometric characterization of these concepts:

> Let u and v be any two distinct points in the domain (a convex set) of a function f, and let line segment uv in the domain give rise to arc MN on the graph of the function, such that point N is higher than or equal in height to

point M. Then function f is said to be *quasiconcave* (*quasiconvex*) if all points on arc MN other than M and N are higher than or equal in height to point M (lower than or equal in height to point N). The function f is said to be *strictly quasiconcave* (*strictly quasiconvex*) if all the points on arc MN other than M and N are strictly higher than point M (strictly lower than point N).

It should be clear from this that any strictly quasiconcave (strictly quasiconvex) function is quasiconcave (quasiconvex), but the converse is not true.

For a better grasp, let us examine the illustrations in Fig. 12.3, all drawn for the one-variable case. In diagram a, line segment uv in the domain gives rise to arc MN on the curve such that N is higher than M. Since all the points between M and N on the said arc are strictly higher than M, this particular arc satisfies the condition for strict quasiconcavity. For the curve to qualify as strictly quasiconcave, however, *all* possible (u, v) pairs must have arcs that satisfy the same condition. This is indeed the case for the function in diagram a. Note that this function also satisfies the condition for (nonstrict) quasiconcavity. But it fails the condition for quasiconvexity, because some points on arc MN are higher than N, which is forbidden for a quasiconvex function. The function in diagram b has the opposite configuration. All the points on arc $M'N'$ are lower than N', the higher of the two ends, and the same is true of all arcs that can be drawn. Thus the function in diagram b is strictly quasiconvex. As you can verify, it also satisfies the condition for (nonstrict) quasiconvexity, but fails the condition for quasiconcavity. What distinguishes diagram c is the presence of a horizontal line segment $M''N''$, where all the points have the same height. As a result, that line segment—and hence the entire curve—can only meet the condition for quasiconcavity, but not strict quasiconcavity.

Generally speaking, a quasiconcave function that is not also concave has a graph roughly shaped like a bell, or a portion thereof, and a quasiconvex function has a graph shaped like an inverted bell, or a portion thereof. On the bell, it is admissible (though not required) to have both concave and convex segments. This more permissive nature of the characterization makes quasiconcavity (quasicon-

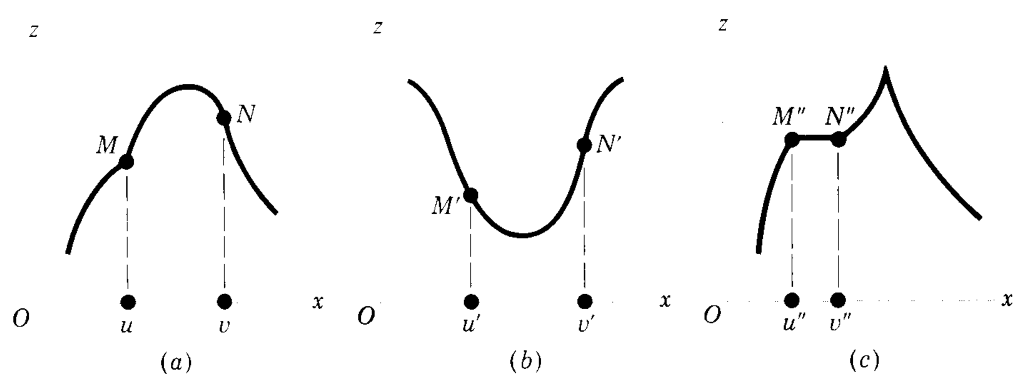

Figure 12.3

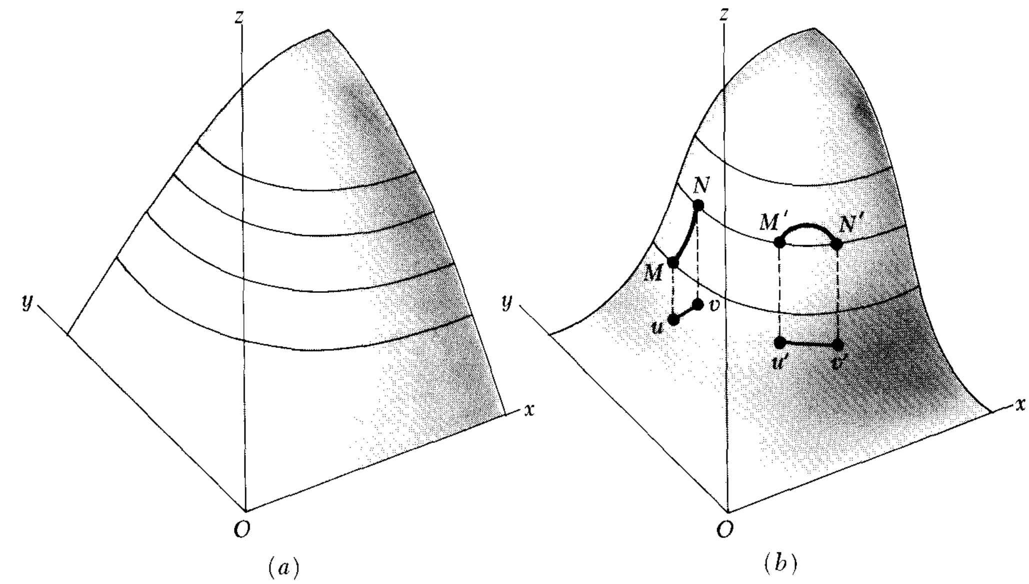

Figure 12.4

vexity) a weaker condition than concavity (convexity). In Fig. 12.4, we contrast strict concavity against strict quasiconcavity for the two-variable case. As drawn, both surfaces depict increasing functions, as they contain only the ascending portions of a dome and a bell, respectively. The surface in diagram a is strictly concave, but the one in diagram b is certainly not, since it contains convex portions near the base of the bell. Yet it is strictly quasiconcave; all the arcs on the surface, exemplified by MN and $M'N'$, satisfy the condition that all the points on each arc between the two end points are higher than the lower end point. Returning to diagram a, we should note that the surface therein is also strictly quasiconcave. Although we have not drawn any illustrative arcs MN and $M'N'$ in diagram a, it is not difficult to check that all possible arcs do indeed satisfy the condition for strict quasiconcavity. In general, a strictly concave function must be strictly quasiconcave, although the converse is not true. We shall demonstrate this more formally in the paragraphs that follow.

Algebraic Definition

The geometric characterization above can be translated into an algebraic definition for easier generalization to higher-dimensional cases:

A function f is $\left\{ \begin{matrix} \text{quasiconcave} \\ \text{quasiconvex} \end{matrix} \right\}$ iff, for any pair of distinct points u and v in the (convex) domain of f, and for $0 < \theta < 1$,

$$(12.20) \qquad f(v) \geq f(u) \quad \Rightarrow \quad f[\theta u + (1-\theta)v] \left\{ \begin{matrix} \geq f(u) \\ \leq f(v) \end{matrix} \right\}$$

To adapt this definition to *strict* quasiconcavity and quasiconvexity, the two weak inequalities on the right should be changed into strict inequalities $\begin{Bmatrix} > f(u) \\ < f(v) \end{Bmatrix}$. You may find it instructive to compare (12.20) with (11.20).

From this definition, the following three theorems readily follow. These will be stated in terms of a function $f(x)$, where x can be interpreted as a vector of variables, $x = (x_1, \dots, x_n)$.

Theorem I (negative of a function) If $f(x)$ is quasiconcave (strictly quasiconcave), then $-f(x)$ is quasiconvex (strictly quasiconvex).

Theorem II (concavity versus quasiconcavity) Any concave (convex) function is quasiconcave (quasiconvex), but the converse is not true. Similarly, any strictly concave (strictly convex) function is strictly quasiconcave (strictly quasiconvex), but the converse is not true.

Theorem III (linear function) If $f(x)$ is a linear function, then it is quasiconcave as well as quasiconvex.

Theorem I follows from the fact that multiplying an inequality by -1 reverses the sense of inequality. Let $f(x)$ be quasiconcave, with $f(v) \geq f(u)$. Then, by (12.20), $f[\theta u + (1 - \theta)v] \geq f(u)$. As far as the function $-f(x)$ is concerned, however, we have (after multiplying the two inequalities through by -1) $-f(u) \geq -f(v)$ and $-f[\theta u + (1 - \theta)v] \leq -f(u)$. Interpreting $-f(u)$ as the height of point N, and $-f(v)$ the height of M, we see that the function $-f(x)$ satisfies the condition for quasiconvexity in (12.20). This proves one of the four cases cited in Theorem I; the proofs for the other three are similar.

For Theorem II, we shall only prove that concavity implies quasiconcavity. Let $f(x)$ be concave. Then, by (11.20),

$$f[\theta u + (1 - \theta)v] \geq \theta f(u) + (1 - \theta)f(v)$$

Now assume that $f(v) \geq f(u)$; then any weighted average of $f(v)$ and $f(u)$ cannot possibly be less than $f(u)$, i.e.,

$$\theta f(u) + (1 - \theta)f(v) \geq f(u)$$

Combining the above two results, we find that

$$f[\theta u + (1 - \theta)v] \geq f(u) \qquad \text{for } f(v) \geq f(u)$$

which satisfies the definition of quasiconcavity in (12.20). Note, however, that the condition for quasiconcavity cannot guarantee concavity.

Once Theorem II is established, Theorem III follows immediately. We already know that a linear function is both concave and convex, though not strictly so. In view of Theorem II, a linear function must also be both quasiconcave and quasiconvex, though not strictly so.

In the case of concave and convex functions, there is a useful theorem to the effect that the sum of concave (convex) functions is also concave (convex). Unfortunately, this theorem cannot be generalized to quasiconcave and quasiconvex functions. For instance, a sum of two quasiconcave functions is *not necessarily* quasiconcave (see Exercise 12.4-3).

Sometimes it may prove easier to check quasiconcavity and quasiconvexity by the following alternative definition:

A function $f(x)$, where x is a vector of variables, is $\left\{\begin{matrix}\text{quasiconcave}\\ \text{quasiconvex}\end{matrix}\right\}$ iff, for any constant k, the set

$$\left\{\begin{matrix} S^{\geq} \equiv \{x \mid f(x) \geq k\} \\ S^{\leq} \equiv \{x \mid f(x) \leq k\} \end{matrix}\right\} \text{ is a convex set} \tag{12.21}$$

The sets $S^{\geq}$ and $S^{\leq}$ were introduced earlier (Fig. 11.10) to show that a convex function (or even a concave function) can give rise to a convex set. Here we are employing these two sets as tests for quasiconcavity and quasiconvexity. The three functions in Fig. 12.5 all contain concave as well as convex segments and hence are neither convex nor concave. But the function in diagram *a* is quasiconcave, because for any value of k (only one of which has been illustrated), the set $S^{\geq}$ is convex. The function in diagram *b* is, on the other hand, quasiconvex since the set $S^{\leq}$ is convex. The function in diagram *c*—a monotonic function—differs from the other two in that both $S^{\geq}$ and $S^{\leq}$ are convex sets. Hence that function is quasiconcave as well as quasiconvex.

Note that while (12.21) can be used to check quasiconcavity and quasiconvexity, it is incapable of distinguishing between strict and nonstrict varieties of these properties. Note, also, that the defining properties in (12.21) are in themselves not sufficient for concavity and convexity, respectively. In particular, given a concave function which must perforce be quasiconcave, we can conclude that $S^{\geq}$ is a

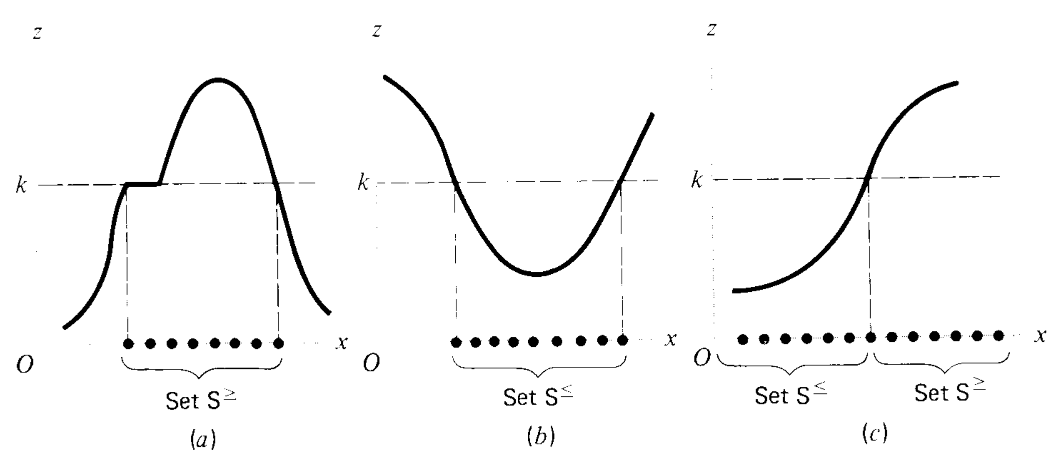

Figure 12.5

convex set; but given that $S^{\geq}$ is a convex set, we can conclude only that the function f is *quasi*concave (but not necessarily concave).

Example 1 Check $z = x^2$ $(x \geq 0)$ for quasiconcavity and quasiconvexity. This function is easily verified geometrically to be convex, in fact strictly so. Hence it is quasiconvex. Interestingly, it is also quasiconcave. For its graph—the right half of a U-shaped curve, initiating from the point of origin and increasing at an increasing rate—is, similarly to Fig. 12.5*c*, capable of generating a convex $S^{\geq}$ as well as a convex $S^{\leq}$.

If we wish to apply (12.20) instead, we first let u and v be any two distinct nonnegative values of x. Then

$$f(u) = u^2 \quad f(v) = v^2 \quad \text{and} \quad f[\theta u + (1-\theta)v] = [\theta u + (1-\theta)v]^2$$

Suppose that $f(v) \geq f(u)$, that is, $v^2 \geq u^2$; then $v \geq u$, or more specifically, $v > u$ (since u and v are distinct). Inasmuch as the weighted average $[\theta u + (1-\theta)v]$ must lie between u and v, we may write the continuous inequality

$$v^2 > [\theta u + (1-\theta)v]^2 > u^2 \qquad \text{for } 0 < \theta < 1$$

$$\text{or} \quad f(v) > f[\theta u + (1-\theta)v] > f(u) \qquad \text{for } 0 < \theta < 1$$

By (12.20), this result makes the function f *both* quasiconcave and quasiconvex—indeed strictly so.

Example 2 Show that $z = f(x, y) = xy$ $(x, y \geq 0)$ is quasiconcave. We shall use the criterion in (12.21) and establish that the set $S^{\geq} = \{(x, y) \mid xy \geq k\}$ is convex for any k. For this purpose, we set $xy = k$ to obtain an isovalue curve for each value of k. Like x and y, k should be nonnegative. In case $k > 0$, the isovalue curve is a rectangular hyperbola in the first quadrant of the xy plane. The set $S^{\geq}$, consisting of all the points on or above a rectangular hyperbola, is a convex set. In the other case, with $k = 0$, the isovalue curve as defined by $xy = 0$ is L-shaped, with the L coinciding with the nonnegative segments of the x and y axes. The set $S^{\geq}$, consisting this time of the entire nonnegative quadrant, is again a convex set. Thus, by (12.21), the function $z = xy$ $(x, y \geq 0)$ is quasiconcave.

You should be careful not to confuse the shape of the isovalue curves $xy = k$ (which is defined in the xy plane) with the shape of the surface $z = xy$ (which is defined in the xyz space). The characteristic of the z surface (quasiconcave in 3-space) is what we wish to ascertain; the shape of the isovalue curves (convex in 2-space for positive k) is of interest here only as a means to delineate the sets $S^{\geq}$ in order to apply the criterion in (12.21).

Example 3 Show that $z = f(x, y) = (x - a)^2 + (y - b)^2$ is quasiconvex. Let us again apply (12.21). Setting $(x - a)^2 + (y - b)^2 = k$, we see that k must be nonnegative. For each k, the isovalue curve is a circle in the xy plane with its center at (a, b) and with radius $\sqrt{k}$. Since $S^{\leq} = \{(x, y) \mid (x - a)^2 + (y - b)^2 \leq k\}$ is the set of all points on or inside a circle, it constitutes a convex set. This is true even when $k = 0$—when the circle degenerates into a single point, (a, b)—since by convention a single point is considered as a convex set. Thus the given function is quasiconvex.

Differentiable Functions

The definitions (12.20) and (12.21) do not require differentiability of the function f. If f is differentiable, however, quasiconcavity and quasiconvexity can alternatively be defined in terms of its first derivatives:

> A differentiable function of one variable, $f(x)$, is $\begin{Bmatrix} \text{quasiconcave} \\ \text{quasiconvex} \end{Bmatrix}$ iff, for any pair of distinct points u and v in the domain,

$$(12.22) \qquad f(v) \geq f(u) \quad \Rightarrow \quad \begin{Bmatrix} f'(u)(v-u) \\ f'(v)(v-u) \end{Bmatrix} \geq 0$$

Quasiconcavity and quasiconvexity will be *strict*, if the weak inequality on the right is changed to the strict inequality > 0. When there are two or more independent variables, the definition is to be modified as follows:

> A differentiable function $f(x_1, \ldots, x_n)$ is $\begin{Bmatrix} \text{quasiconcave} \\ \text{quasiconvex} \end{Bmatrix}$ iff, for any two distinct points $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$ in the domain,

$$(12.22') \qquad f(v) \geq f(u) \quad \Rightarrow \quad \begin{Bmatrix} \sum_{j=1}^{n} f_j(u)(v_j - u_j) \\ \sum_{j=1}^{n} f_j(v)(v_j - u_j) \end{Bmatrix} \geq 0$$

> where $f_j \equiv \partial f / \partial x_j$, to be evaluated at u or v as the case may be.

Again, for *strict* quasiconcavity and quasiconvexity, the weak inequality on the right should be changed to the strict inequality > 0.

Finally, if a function $z = f(x_1, \ldots, x_n)$ is twice continuously differentiable, quasiconcavity and quasiconvexity can be checked by means of the first and second partial derivatives of the function, arranged into the bordered determinant

$$(12.23) \qquad |B| = \begin{vmatrix} 0 & f_1 & f_2 & \cdots & f_n \\ f_1 & f_{11} & f_{12} & \cdots & f_{1n} \\ f_2 & f_{21} & f_{22} & \cdots & f_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ f_n & f_{n1} & f_{n2} & \cdots & f_{nn} \end{vmatrix}$$

This bordered determinant resembles the bordered Hessian $|\bar{H}|$ introduced in the preceding section. But unlike the latter, the border in $|B|$ is composed of the first derivatives of the function f rather than an extraneous constraint function g. It is because $|B|$ depends exclusively on the derivatives of function f itself that we can use $|B|$, along with its successive principal minors

$$(12.24) \qquad |B_1| = \begin{vmatrix} 0 & f_1 \\ f_1 & f_{11} \end{vmatrix} \qquad |B_2| = \begin{vmatrix} 0 & f_1 & f_2 \\ f_1 & f_{11} & f_{12} \\ f_2 & f_{21} & f_{22} \end{vmatrix} \qquad \cdots \qquad |B_n| = |B|$$

to characterize the configuration of that function.

We shall state here two conditions; one is necessary, and the other is sufficient. Both relate to quasiconcavity and quasiconvexity on a domain consisting only of the *nonnegative orthant* (the n-dimensional analog of the nonnegative quadrant), that is, with $x_1, \ldots, x_n \geq 0$.*

For $z = f(x_1, \ldots, x_n)$ to be quasiconcave on the nonnegative orthant, it is necessary that

$$(12.25) \qquad |B_1| \leq 0, \quad |B_2| \geq 0, \quad \ldots, \quad |B_n| \begin{Bmatrix} \leq \\ \geq \end{Bmatrix} 0 \text{ if } n \text{ is} \begin{Bmatrix} \text{odd} \\ \text{even} \end{Bmatrix}$$

wherever the partial derivatives are evaluated in the nonnegative orthant. For quasiconvexity, it is necessary that

$$(12.25') \qquad |B_1| \leq 0, \quad |B_2| \leq 0, \quad \ldots, \quad |B_n| \leq 0$$

A sufficient condition for f to be quasiconcave on the nonnegative orthant is that

$$(12.26) \qquad |B_1| < 0, \quad |B_2| > 0, \quad \ldots, \quad |B_n| \begin{Bmatrix} < \\ > \end{Bmatrix} 0 \text{ if } n \text{ is} \begin{Bmatrix} \text{odd} \\ \text{even} \end{Bmatrix}$$

wherever the partial derivatives are evaluated in the nonnegative orthant. For quasiconvexity, the corresponding sufficient condition is that

$$(12.26') \qquad |B_1| < 0, \quad |B_2| < 0, \quad \ldots, \quad |B_n| < 0$$

Note that the condition $|B_1| \leq 0$ in (12.25) and (12.25′) is automatically satisfied

* Whereas concavity (convexity) of a function on a convex domain can always be extended to concavity (convexity) over the entire space, quasiconcavity and quasiconvexity cannot. For instance, our conclusions in Examples 1 and 2 above will not hold if the variables are allowed to take negative values. The two conditions given here are based on Kenneth J. Arrow and Alain C. Enthoven, "Quasi-Concave Programming," *Econometrica*, October 1961, p. 797, (Theorem 5). Their attention is confined to quasiconcave functions; our extension to quasiconvex functions makes use of the fact that the negative of a quasiconcave function is quasiconvex.

because $|B_1| = -f_1^2$; it is listed here only for the sake of symmetry. In contrast, the condition $|B_1| < 0$ in (12.26) and (12.26′) is not automatically satisfied.

Example 4 The function $z = f(x_1, x_2) = x_1 x_2$ $(x_1, x_2 \geq 0)$ is quasiconcave (cf. Example 2 above). We shall now check this by (12.22′). Let $u = (u_1, u_2)$ and $v = (v_1, v_2)$ be any two points in the domain. Then $f(u) = u_1 u_2$ and $f(v) = v_1 v_2$. Assume that

$$(12.27) \qquad f(v) \geq f(u) \qquad \text{or} \qquad v_1 v_2 \geq u_1 u_2 \qquad (v_1, v_2, u_1, u_2 \geq 0)$$

Since the partial derivatives of f are $f_1 = x_2$ and $f_2 = x_1$, (12.22′) amounts to the condition that

$$f_1(u)(v_1 - u_1) + f_2(u)(v_2 - u_2) = u_2(v_1 - u_1) + u_1(v_2 - u_2) \geq 0$$

or, upon rearrangement,

$$(12.28) \qquad u_2(v_1 - u_1) \geq u_1(u_2 - v_2)$$

We need to consider four possibilities regarding the values of u_1 and u_2. First, if $u_1 = u_2 = 0$, then (12.28) is trivially satisfied. Second, if $u_1 = 0$ but $u_2 > 0$, then (12.28) reduces to the condition $u_2 v_1 \geq 0$, which is again satisfied since u_2 and v_1 are both nonnegative. Third, if $u_1 > 0$ and $u_2 = 0$, then (12.28) reduces to the condition $0 \geq -u_1 v_2$, which is still satisfied. Fourth and last, suppose that u_1 and u_2 are both positive, so that v_1 and v_2 are also positive. Subtracting $v_2 u_1$ from both sides of (12.27), we obtain

$$(12.29) \qquad v_2(v_1 - u_1) \geq u_1(u_2 - v_2)$$

Three subpossibilities now present themselves:

1. If $u_2 = v_2$, then $v_1 \geq u_1$. In fact, we should have $v_1 > u_1$ since (u_1, u_2) and (v_1, v_2) are distinct points. The fact that $u_2 = v_2$ and $v_1 > u_1$ implies that condition (12.28) is satisfied.
2. If $u_2 > v_2$, then we must also have $v_1 > u_1$ by (12.29). Multiplying both sides of (12.29) by u_2/v_2, we get

$$(12.30) \qquad u_2(v_1 - u_1) \geq \frac{u_2}{v_2} u_1(u_2 - v_2)$$
$$> u_1(u_2 - v_2) \qquad \left[\text{since } \frac{u_2}{v_2} > 1\right]$$

   Thus (12.28) is again satisfied.
3. The final subpossibility is that $u_2 < v_2$, implying that u_2/v_2 is a positive fraction. In this case, the first line of (12.30) still holds. The second line also holds, but now for a different reason: a fraction (u_2/v_2) of a negative number $(u_2 - v_2)$ is greater than the latter number itself.

Inasmuch (12.28) is satisfied in every possible situation that can arise, the function $z = x_1 x_2$ $(x_1, x_2 \geq 0)$ is quasiconcave. Therefore, the necessary condition (12.25) should hold. Because the partial derivatives of f are

$$f_1 = x_2 \qquad f_2 = x_1 \qquad f_{11} = f_{22} = 0 \qquad f_{12} = f_{21} = 1$$

the relevant principal minors turn out to be

$$|B_1| = \begin{vmatrix} 0 & x_2 \\ x_2 & 0 \end{vmatrix} = -x_2^2 \le 0 \qquad |B_2| = \begin{vmatrix} 0 & x_2 & x_1 \\ x_2 & 0 & 1 \\ x_1 & 1 & 0 \end{vmatrix} = 2x_1x_2 \ge 0$$

Thus (12.25) is indeed satisfied. Note, however, that the sufficient condition (12.26) is satisfied only over the positive orthant.

Example 5 Show that $z = f(x, y) = x^a y^b$ ($x, y > 0$; $0 < a, b < 1$) is quasiconcave. The partial derivatives of this function are

$$f_x = ax^{a-1}y^b \qquad f_y = bx^a y^{b-1}$$

$$f_{xx} = a(a-1)x^{a-2}y^b \qquad f_{xy} = f_{yx} = abx^{a-1}y^{b-1} \qquad f_{yy} = b(b-1)x^a y^{b-2}$$

Thus the principal minors of $|B|$ have the following signs:

$$|B_1| = \begin{vmatrix} 0 & f_x \\ f_x & f_{xx} \end{vmatrix} = -\left(ax^{a-1}y^b\right)^2 < 0$$

$$|B_2| = \begin{vmatrix} 0 & f_x & f_y \\ f_x & f_{xx} & f_{xy} \\ f_y & f_{yx} & f_{yy} \end{vmatrix} = \left[2a^2b^2 - a(a-1)b^2 - a^2b(b-1)\right]x^{3a-2}y^{3b-2} > 0$$

This satisfies the sufficient condition for quasiconcavity in (12.26). The given function is in fact strictly quasiconcave, although the criterion in (12.26) is incapable of confirming that.

A Further Look at the Bordered Hessian

The bordered determinant $|B|$, as defined in (12.23), differs from the bordered Hessian

$$|\overline{H}| = \begin{vmatrix} 0 & g_1 & g_2 & \cdots & g_n \\ g_1 & Z_{11} & Z_{12} & \cdots & Z_{1n} \\ g_2 & Z_{21} & Z_{22} & \cdots & Z_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ g_n & Z_{n1} & Z_{n2} & \cdots & Z_{nn} \end{vmatrix}$$

in two ways: (1) the border elements in $|B|$ are the first-order partial derivatives of function f rather than g; and (2) the remaining elements in $|B|$ are the second-order partial derivatives of f rather than the Lagrangian function Z. However, in the special case of a linear constraint equation, $g(x_1, \ldots, x_n) = a_1 x_1 + \cdots + a_n x_n = c$—a case frequently encountered in economics (see Secs. 12.5 and 12.7)—Z_{ij} reduces to f_{ij}. For then the Lagrangian function is

$$Z = f(x_1, \ldots, x_n) + \lambda[c - a_1 x_1 - \cdots - a_n x_n]$$

so that

$$Z_j = f_j - \lambda a_j \qquad \text{and} \qquad Z_{ij} = f_{ij}$$

Turning to the borders, we note that the linear constraint function yields the first derivative $g_j = a_j$. Moreover, when the first-order condition is satisfied, we have $Z_j = f_j - \lambda a_j = 0$, so that $f_j = \lambda a_j$, or $f_j = \lambda g_j$. Thus the border in $|B|$ is simply that of $|\bar{H}|$ multiplied by a positive scalar λ. By factoring out λ successively from the horizontal and vertical borders of $|\bar{H}|$ (see Sec. 5.3, Example 5), we have

$$|B| = \lambda^2 |\bar{H}|$$

Consequently, in the linear-constraint case, the two bordered determinants always possess the same sign at the stationary point of Z. By the same token, the principal minors $|B_i|$ and $|\bar{H}_i|$ $(i = 1, \ldots, n)$ must also share the same sign at that point. It then follows that if the bordered determinant $|B|$ satisfies the sufficient condition for quasiconcavity in (12.26), the bordered Hessian $|\bar{H}|$ must then satisfy the second-order sufficient condition for constrained maximization in Table 12.1. A similar link exists between quasiconvexity and the second-order condition for minimization subject to a linear constraint.

Absolute versus Relative Extrema

A more comprehensive picture of the relationship between quasiconcavity and second-order conditions is presented in Fig. 12.6. (A suitable modification will adapt the figure for quasiconvexity.) Constructed in the same spirit—and to be read in the same manner—as Fig. 11.5, this figure relates quasiconcavity to *absolute* as well as *relative* constrained maxima. The three ovals in the upper part summarize the first- and second-order conditions for a relative constrained maximum. And the rectangles in the middle column, like those in Fig. 11.5, tie the concepts of relative maximum, absolute maximum, and unique absolute maximum to one another.

But the really interesting information can be found in the two diamonds and the elongated $\Rightarrow$ symbols passing through them. The one on the left tells us that, once the first-order condition is satisfied, and if the two provisos listed in the diamond are also satisfied, we have a sufficient condition for an absolute constrained maximum. The first proviso is that the function f be explicitly quasiconcave—a new term which we must hasten to define.

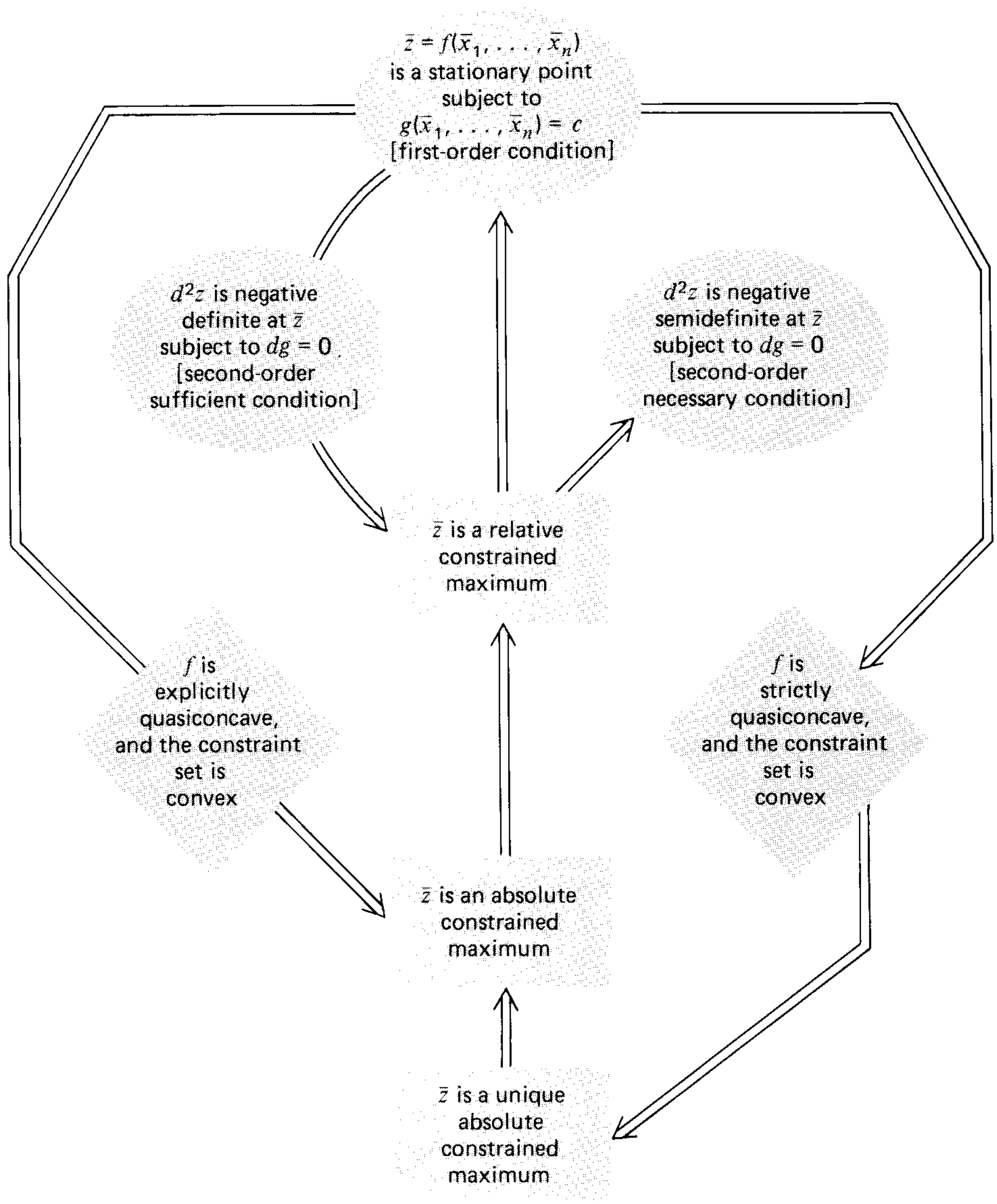

Figure 12.6

A function f is *explicitly quasiconcave* iff

$$f(v) > f(u) \quad \Rightarrow \quad f[\theta u + (1 - \theta)v] > f(u)$$

This defining property means that whenever a point on the surface, $f(v)$, is higher than another, $f(u)$, then all the intermediate points—the points on the surface lying directly above line segment uv in the domain—must also be higher than

$f(u)$. What such a stipulation does is to rule out any *horizontal* plane segments on the surface.* Note that the condition for *explicit* quasiconcavity is not as strong as the condition for *strict* quasiconcavity, since the latter requires $f[\theta u + (1-\theta)v] > f(u)$ even for $f(v) = f(u)$, implying that *non*horizontal plane segments are ruled out, too.† The other proviso in the left-side diamond is that the set $\{(x_1, \ldots, x_n) \mid g(x_1, \ldots, x_n) = c\}$ be convex. If an *equality* constraint is specified, as in our present context, however, the said set can be convex if and only if the constraint function g is linear (e.g., a straight line in a two-dimensional domain). Thus, in the present context, the second proviso simply means a linear constraint equation. When both provisos are met, we shall be dealing with that portion of a bell-shaped, horizontal-segment-free surface (or hypersurface) lying directly above a line (or plane or hyperplane) in the domain. A local maximum found on such a subset of the surface must be an absolute constrained maximum.

The diamond on the right in Fig. 12.6 involves the stronger condition of *strict* quasiconcavity. A strictly quasiconcave function must be explicitly quasiconcave, although the converse is not true. Hence, when strict quasiconcavity replaces explicit quasiconcavity, an absolute constrained maximum is still ensured. But this time that absolute constrained maximum must also be unique, since the absence of any plane segment anywhere on the surface decidedly precludes the possibility of multiple constrained maxima.

EXERCISE 12.4

1 Draw a strictly quasiconcave curve $z = f(x)$ which is
- (*a*) also quasiconvex
- (*b*) not quasiconvex
- (*c*) not convex
- (*d*) not concave
- (*e*) neither concave nor convex
- (*f*) both concave and convex

2 Are the following functions quasiconcave? Strictly so? First check graphically, then algebraically by (12.20). Assume that $x \geq 0$.
(*a*) $f(x) = a$ (*b*) $f(x) = a + bx \; (b > 0)$ (*c*) $f(x) = a + cx^2 \; (c < 0)$

3 (*a*) Let $z = f(x)$ plot as a negatively sloped curve shaped like the right half of a bell in the first quadrant, passing through the points (0, 5), (2, 4), (3, 2), and (5, 1). Let $z = g(x)$ plot as a positively sloped 45° line. Are $f(x)$ and $g(x)$ quasiconcave?
(*b*) Now plot the sum $f(x) + g(x)$. Is the sum function quasiconcave?

4 By examining their graphs, and using (12.21), check whether the following functions are quasiconcave, quasiconvex, both, or neither:
(*a*) $f(x) = x^3 - 2x$ (*b*) $f(x_1, x_2) = 6x_1 - 9x_2$ (*c*) $f(x_1, x_2) = x_2 - \ln x_1$

* Let the surface contain a horizontal plane segment P such that $f(u) \in P$ and $f(v) \notin P$. Then those intermediate points that are located on P will be of equal height to $f(u)$, thereby violating the first proviso.

† Let the surface contain a slanted plane segment P' such that $f(u) = f(v)$ are both located on P'. Then all the intermediate points will also be on P' and be of equal height to $f(u)$, thereby violating the cited requirement for strict quasiconcavity.

5 (*a*) Verify that a cubic function $z = ax^3 + bx^2 + cx + d$ is in general neither quasiconcave nor quasiconvex.

(*b*) Is it possible to impose restrictions on the parameters such that the function becomes both quasiconcave and quasiconvex for $x \geq 0$?

6 Use (12.22) to check $z = x^2$ $(x \geq 0)$ for quasiconcavity and quasiconvexity.

7 Show that $z = xy$ $(x, y \geq 0)$ is not quasiconvex.

8 Use bordered determinants to check the following functions for quasiconcavity and quasiconvexity:

(*a*) $z = -x^2 - y^2 \quad (x, y > 0)$ (*b*) $z = -(x+1)^2 - (y+2)^2 \quad (x, y > 0)$

12.5 UTILITY MAXIMIZATION AND CONSUMER DEMAND

The maximization of a utility function was cited earlier as an example of constrained optimization. Let us now reexamine this problem in more detail. For simplicity, we shall still allow our hypothetical consumer the choice of only two goods, both of which have continuous, positive marginal-utility functions. The prices of both goods are market-determined, hence exogenous, although in this section we shall omit the zero subscript from the price symbols. If the purchasing power of the consumer is a given amount B (for budget), the problem posed will be that of maximizing a smooth utility (index) function

$$U = U(x, y) \qquad (U_x, U_y > 0)$$

subject to

$$xP_x + yP_y = B$$

First-Order Condition

The Lagrangian function of this optimization model is

$$Z = U(x, y) + \lambda(B - xP_x - yP_y)$$

As the first-order condition, we have the following set of simultaneous equations:

$$\begin{aligned} Z_\lambda &= B - xP_x - yP_y = 0 \\ Z_x &= U_x - \lambda P_x = 0 \\ Z_y &= U_y - \lambda P_y = 0 \end{aligned} \tag{12.31}$$

Since the last two equations are equivalent to

$$\frac{U_x}{P_x} = \frac{U_y}{P_y} = \lambda \tag{12.31'}$$

the first-order condition in effect calls for the satisfaction of (12.31′), subject to the budget constraint—the first equation in (12.31). What (12.31′) states is merely the familiar proposition in classical consumer theory that, in order to maximize utility, consumers must allocate their budgets so as to equalize the ratio of marginal utility to price for every commodity. Specifically, in the equilibrium or optimum, these ratios should have the common value $\bar{\lambda}$. As we learned earlier, $\bar{\lambda}$ measures the comparative-static effect of the constraint constant on the optimal value of the objective function. Hence, we have in the present context $\bar{\lambda} = (\partial \bar{U}/\partial B)$; that is, the optimal value of the Lagrange multiplier can be interpreted as the *marginal utility of money* (budget money) when the consumer's utility is maximized.

If we restate the condition in (12.31′) in the form

$$\frac{U_x}{U_y} = \frac{P_x}{P_y} \tag{12.31''}$$

the first-order condition can be given an alternative interpretation, in terms of indifference curves.

An *indifference curve* is defined as the locus of the combinations of x and y that will yield a constant level of U. This means that on an indifference curve we must find

$$dU = U_x\,dx + U_y\,dy = 0$$

with the implication that $dy/dx = -U_x/U_y$. Accordingly, if we plot an indifference curve in the xy plane, as in Fig. 12.7, its slope, dy/dx, must be equal to the negative of the marginal-utility ratio U_x/U_y. (Since we assume $U_x, U_y > 0$, the slope of the indifference curve must be negative.) Conversely, since U_x/U_y is the negative of the indifference-curve slope, it must represent the *marginal rate of substitution* between the two goods.

What about the meaning of P_x/P_y? As we shall presently see, this ratio represents the negative of the slope of the graph of the budget constraint. The

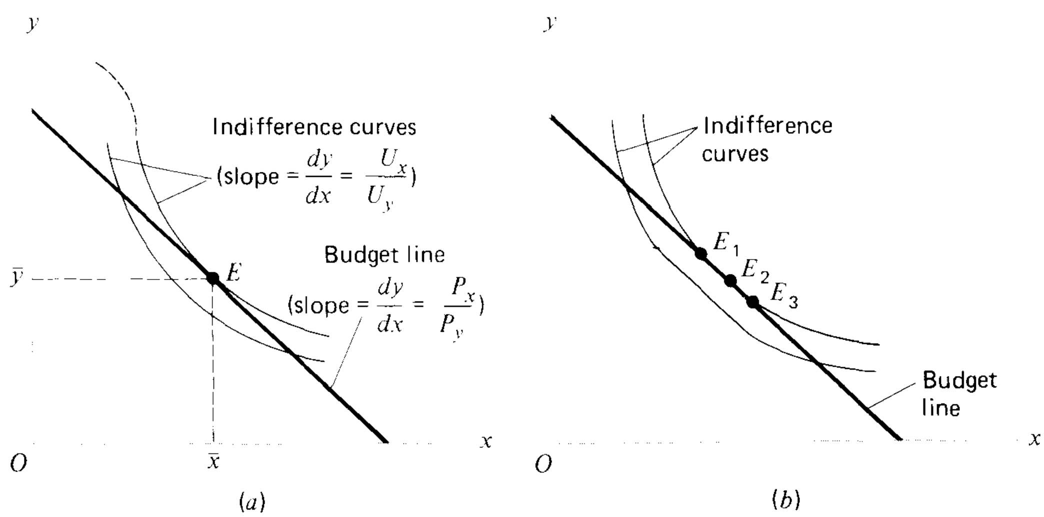

Figure 12.7

budget constraint, $xP_x + yP_y = B$, can be written alternatively as

$$y = \frac{B}{P_y} - \frac{P_x}{P_y}x$$

so that, when plotted in the xy plane as in Fig. 12.7, it emerges as a straight line with slope $-P_x/P_y$ (and vertical intercept B/P_y).

In this light, the new version of the first-order condition—(12.31″) plus the budget constraint—discloses that, to maximize utility, a consumer must allocate the budget such that the slope of the budget line (on which the consumer must remain) is equal to the slope of some indifference curve. This condition is met at point E in Fig. 12.7*a*, where the budget line is tangent to an indifference curve.

Second-Order Condition

If the bordered Hessian in the present problem is positive, i.e., if

$$(12.32) \qquad |\overline{H}| = \begin{vmatrix} 0 & P_x & P_y \\ P_x & U_{xx} & U_{xy} \\ P_y & U_{yx} & U_{yy} \end{vmatrix} = 2P_xP_yU_{xy} - P_y^2U_{xx} - P_x^2U_{yy} > 0$$

(with all the derivatives evaluated at the critical values $\bar{x}$ and $\bar{y}$), then the stationary value of U will assuredly be a maximum. The presence of the derivatives U_{xx}, U_{yy}, and U_{xy} in (12.32) clearly suggests that meeting this condition would entail certain restrictions on the utility function and, hence, on the shape of the indifference curves. What are these restrictions?

Considering first the shape of the indifference curves, we can show that a positive $|\overline{H}|$ means the strict convexity of the (downward-sloping) indifference curve at the point of tangency E. Just as the downward slope of an indifference curve is guaranteed by a negative $dy/dx(= -U_x/U_y)$, its strict convexity would be ensured by a positive d^2y/dx^2. To get the expression for d^2y/dx^2, we can differentiate $-U_x/U_y$ with respect to x; but in doing so, we should bear in mind not only that both U_x and U_y (being derivatives) are functions of x and y but also that, along a given indifference curve, y is itself a function of x. Accordingly, U_x and U_y can both be considered as functions of x alone; therefore, we can get a total derivative

$$(12.33) \qquad \frac{d^2y}{dx^2} = \frac{d}{dx}\left(-\frac{U_x}{U_y}\right) = -\frac{1}{U_y^2}\left(U_y\frac{dU_x}{dx} - U_x\frac{dU_y}{dx}\right)$$

Since x can affect U_x and U_y not only directly but also indirectly, via the intermediary of y, we have

$$(12.34) \qquad \frac{dU_x}{dx} = U_{xx} + U_{yx}\frac{dy}{dx} \qquad \frac{dU_y}{dx} = U_{xy} + U_{yy}\frac{dy}{dx}$$

where dy/dx refers to the slope of the indifference curve. Now, at the point of tangency E—the only point relevant to the discussion of the second-order condition—this slope is identical with that of the budget constraint; that is, $dy/dx = -P_x/P_y$. Thus we can rewrite (12.34) as

$$(12.34') \qquad \frac{dU_x}{dx} = U_{xx} - U_{yx}\frac{P_x}{P_y} \qquad \frac{dU_y}{dx} = U_{xy} - U_{yy}\frac{P_x}{P_y}$$

Substituting (12.34′) into (12.33) and utilizing the information that

$$U_x = \frac{U_y P_x}{P_y} \qquad \text{[from (12.31″)]}$$

and then factoring out U_y/P_y^2, we can finally transform (12.33) into

$$(12.33') \qquad \frac{d^2y}{dx^2} = \frac{2P_xP_yU_{xy} - P_y^2U_{xx} - P_x^2U_{yy}}{U_yP_y^2} = \frac{|\overline{H}|}{U_yP_y^2}$$

It is clear that when the second-order sufficient condition (12.32) is satisfied, the second derivative in (12.33′) is positive, and the relevant indifference curve is strictly convex at the point of tangency. In the present context, it is also true that the strict convexity of the indifference curve at the tangency implies the satisfaction of the sufficient condition (12.32). This is because, given that the indifference curves are negatively sloped, with no stationary points anywhere, the possibility of a zero d^2y/dx^2 value on a strictly convex curve is ruled out. Thus strict convexity can now result only in a positive d^2y/dx^2, and hence a positive $|\overline{H}|$, by (12.33′).

Recall, however, that the derivatives in $|\overline{H}|$ are to be evaluated at the critical values $\bar{x}$ and $\bar{y}$ only. Thus the strict convexity of the indifference curve, as a sufficient condition, pertains only to the point of tangency, and it is not inconceivable for the curve to contain a concave segment away from point E, as illustrated by the broken curve segment in Fig. 12.7*a*. On the other hand, if the utility function is known to be a smooth, increasing, strictly quasiconcave function, then every indifference curve will be everywhere strictly convex. Such a utility function has a surface like the one in Fig. 12.4*b*. When such a surface is cut with a plane parallel to the xy plane, we obtain for each of such cuts a cross section which, when projected onto the xy plane, becomes a strictly convex, downward-sloping indifference curve. In that event, no matter where the point of tangency may occur, the second-order sufficient condition will always be satisfied. Besides, there can exist only one point of tangency, one that yields the unique absolute maximum level of utility attainable on the given linear budget. This result, of course, conforms perfectly to what the diamond on the right of Fig. 12.6 states.

You have been repeatedly reminded that the second-order sufficient condition is not necessary. Let us illustrate here the maximization of utility while (12.32) fails to hold. Suppose that, as illustrated in Fig. 12.7*b*, the relevant indifference curve contains a linear segment that coincides with a portion of the budget line. Then clearly we have multiple maxima, since the first-order condition $U_x/U_y =$

P_x/P_y is now satisfied at every point on the linear segment of the indifference curve, including E_1, E_2, and E_3. In fact, these are absolute constrained maxima. But since on a line segment d^2y/dx^2 is zero, we have $|\overline{H}| = 0$ by (12.33′). Thus maximization is achieved in this case even though the second-order sufficient condition (12.32) is violated.

The fact that a linear segment appears on the indifference curve suggests the presence of a slanted plane segment on the utility surface. This occurs when the utility function is explicitly quasiconcave rather than strictly quasiconcave. As Fig. 12.7*b* shows, points E_1, E_2, and E_3, all of which are located on the same (highest attainable) indifference curve, yield the same absolute maximum utility under the given linear budget constraint. Referring to Fig. 12.6 again, we note that this result is perfectly consistent with the message conveyed by the diamond on the left.

Comparative-Static Analysis

In our consumer model, the prices P_x and P_y are exogenous, as is the amount of the budget, B. If we assume the satisfaction of the second-order sufficient condition, we can analyze the comparative-static properties of the model on the basis of the first-order condition (12.31), viewed as a set of equations $F^j = 0$ $(j = 1, 2, 3)$, where each F^j function has continuous partial derivatives. As pointed out in (12.19), the endogenous-variable Jacobian of this set of equations must have the same value as the bordered Hessian; that is, $|J| = |\overline{H}|$. Thus, when the second-order condition (12.32) is met, $|J|$ must be positive and it does not vanish at the initial optimum. Consequently, the implicit-function theorem is applicable, and we may express the optimal values of the endogenous variables as implicit functions of the exogenous variables:

$$\begin{aligned} \bar{\lambda} &= \bar{\lambda}(P_x, P_y, B) \\ \bar{x} &= \bar{x}(P_x, P_y, B) \\ \bar{y} &= \bar{y}(P_x, P_y, B) \end{aligned} \tag{12.35}$$

These are known to possess continuous derivatives that give comparative-static information. In particular, the derivatives of the last two functions $\bar{x}$ and $\bar{y}$, which are descriptive of the consumer's demand behavior, can tell us how the consumer will react to changes in prices and in the budget. To find these derivatives, however, we must first convert (12.31) into a set of equilibrium identities as follows:

$$\begin{aligned} B - \bar{x}P_x - \bar{y}P_y &\equiv 0 \\ U_x(\bar{x}, \bar{y}) - \bar{\lambda}P_x &\equiv 0 \\ U_y(\bar{x}, \bar{y}) - \bar{\lambda}P_y &\equiv 0 \end{aligned} \tag{12.36}$$

By taking the total differential of each identity in turn (allowing every variable to change), and noting that $U_{xy} = U_{yx}$, we then arrive at the linear system

$$
\begin{aligned}
-P_x\, d\bar{x} - P_y\, d\bar{y} &= \bar{x}\, dP_x + \bar{y}\, dP_y - dB \\
-P_x\, d\bar{\lambda} + U_{xx}\, d\bar{x} + U_{xy}\, d\bar{y} &= \bar{\lambda}\, dP_x \\
-P_y\, d\bar{\lambda} + U_{yx}\, d\bar{x} + U_{yy}\, d\bar{y} &= \qquad\qquad \bar{\lambda}\, dP_y
\end{aligned}
\tag{12.37}
$$

To study the effect of a change in the budget size (also referred to as the *income* of the consumer), let $dP_x = dP_y = 0$, but keep $dB \neq 0$. Then, after dividing (12.37) through by dB, and interpreting each ratio of differentials as a partial derivative, we can write the matrix equation*

$$
\begin{bmatrix} 0 & -P_x & -P_y \\ -P_x & U_{xx} & U_{xy} \\ -P_y & U_{yx} & U_{yy} \end{bmatrix}
\begin{bmatrix} (\partial\bar{\lambda}/\partial B) \\ (\partial\bar{x}/\partial B) \\ (\partial\bar{y}/\partial B) \end{bmatrix}
=
\begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}
\tag{12.38}
$$

As you can verify, the array of elements in the coefficient matrix is exactly the same as what would appear in the Jacobian $|J|$, which has the same value as the bordered Hessian $|\bar{H}|$ although the latter has P_x and P_y (rather than $-P_x$ and $-P_y$) in the first row and the first column. By Cramer's rule, we can solve for all three comparative-static derivatives, but we shall confine our attention to the following two:

$$
\left(\frac{\partial\bar{x}}{\partial B}\right) = \frac{1}{|J|}
\begin{vmatrix} 0 & -1 & -P_y \\ -P_x & 0 & U_{xy} \\ -P_y & 0 & U_{yy} \end{vmatrix}
= \frac{1}{|J|}
\begin{vmatrix} -P_x & U_{xy} \\ -P_y & U_{yy} \end{vmatrix}
\tag{12.39}
$$

$$
\left(\frac{\partial\bar{y}}{\partial B}\right) = \frac{1}{|J|}
\begin{vmatrix} 0 & -P_x & -1 \\ -P_x & U_{xx} & 0 \\ -P_y & U_{yx} & 0 \end{vmatrix}
= \frac{-1}{|J|}
\begin{vmatrix} -P_x & U_{xx} \\ -P_y & U_{yx} \end{vmatrix}
\tag{12.40}
$$

By the second-order condition, $|J| = |\bar{H}|$ is positive, as are P_x and P_y. Unfortunately, in the absence of additional information about the relative magnitudes of P_x, P_y, and the U_{ij}, we are still unable to ascertain the signs of these two comparative-static derivatives. This means that, as the consumer's budget (or income) increases, his optimal purchases $\bar{x}$ and $\bar{y}$ may *either* increase *or* decrease. In case, say, $\bar{x}$ decreases as B increases, product x is referred to as an *inferior good* as against a *normal good*.

* The matrix equation (12.38) can also be obtained by totally differentiating (12.36) with respect to B, while bearing in mind the implicit solutions in (12.35).

Next, we may analyze the effect of a change in P_x. Letting $dP_y = dB = 0$ this time, but keeping $dP_x \neq 0$, and then dividing (12.37) through by dP_x, we obtain another matrix equation

$$(12.41) \qquad \begin{bmatrix} 0 & -P_x & -P_y \\ -P_x & U_{xx} & U_{xy} \\ -P_y & U_{yx} & U_{yy} \end{bmatrix} \begin{bmatrix} (\partial\bar{\lambda}/\partial P_x) \\ (\partial\bar{x}/\partial P_x) \\ (\partial\bar{y}/\partial P_x) \end{bmatrix} = \begin{bmatrix} \bar{x} \\ \bar{\lambda} \\ 0 \end{bmatrix}$$

From this, the following comparative-static derivatives emerge:

$$(12.42) \qquad \left(\frac{\partial\bar{x}}{\partial P_x}\right) = \frac{1}{|J|}\begin{vmatrix} 0 & \bar{x} & -P_y \\ -P_x & \bar{\lambda} & U_{xy} \\ -P_y & 0 & U_{yy} \end{vmatrix}$$

$$= \frac{-\bar{x}}{|J|}\begin{vmatrix} -P_x & U_{xy} \\ -P_y & U_{yy} \end{vmatrix} + \frac{\bar{\lambda}}{|J|}\begin{vmatrix} 0 & -P_y \\ -P_y & U_{yy} \end{vmatrix}$$

$$\equiv T_1 + T_2 \qquad [T_i \text{ means the } i\text{th term}]$$

$$(12.43) \qquad \left(\frac{\partial\bar{y}}{\partial P_x}\right) = \frac{1}{|J|}\begin{vmatrix} 0 & -P_x & \bar{x} \\ -P_x & U_{xx} & \bar{\lambda} \\ -P_y & U_{yx} & 0 \end{vmatrix}$$

$$= \frac{\bar{x}}{|J|}\begin{vmatrix} -P_x & U_{xx} \\ -P_y & U_{yx} \end{vmatrix} - \frac{\bar{\lambda}}{|J|}\begin{vmatrix} 0 & -P_x \\ -P_y & U_{yx} \end{vmatrix}$$

$$\equiv T_3 + T_4$$

How do we interpret these two results? The first one, $(\partial\bar{x}/\partial P_x)$, tells how a change in P_x affects the optimal purchase of x; it thus provides the basis for the study of our consumer's demand function for x. There are two component terms in this effect. The first term, T_1, can be rewritten, by using (12.39), as $-(\partial\bar{x}/\partial B)\bar{x}$. In this light, T_1 seems to be a measure of the effect of a change in B (budget, or income) upon the optimal purchase $\bar{x}$, with $\bar{x}$ itself serving as a weighting factor. However, since this derivative obviously is concerned with a price change, T_1 must be interpreted as the *income effect* of a *price change*. As P_x rises, the decline in the consumer's real income will produce an effect on $\bar{x}$ similar to that of an actual decrease in B; hence the use of the term $-(\partial\bar{x}/\partial B)$. Understandably, the more prominent the place of commodity x in the total budget, the greater this income effect will be—and hence the appearance of the weighting factor $\bar{x}$ in T_1. This interpretation can be demonstrated more formally by expressing the consumer's effectual income loss by the differential $dB = -\bar{x}\, dP_x$. Then we have

(12.44) $$\bar{x} = -\frac{dB}{dP_x}$$

and $$T_1 = -\left(\frac{\partial \bar{x}}{\partial B}\right)\bar{x} = \left(\frac{\partial \bar{x}}{\partial B}\right)\frac{dB}{dP_x}$$

which shows T_1 to be the measure of the effect of dP_x on $\bar{x}$ via B, that is, the income effect.

If we now compensate the consumer for the effectual income loss by a cash payment numerically equal to dB, then, because of the neutralization of the income effect, the remaining component in the comparative-static derivative $(\partial \bar{x}/\partial P_x)$, namely, T_2, will measure the change in $\bar{x}$ due entirely to price-induced substitution of one commodity for another, i.e., the *substitution effect* of the change in P_x. To see this more clearly, let us return to (12.37), and see how the income compensation will modify the situation. When studying the effect of dP_x only (with $dP_y = dB = 0$), the first equation in (12.37) can be written as $-P_x\, d\bar{x} - P_y\, d\bar{y} = \bar{x}\, dP_x$. Since the indication of the effectual income loss to the consumer lies in the expression $\bar{x}\, dP_x$ (which, incidentally, appears only in the first equation), to compensate the consumer means to set this term equal to zero. If so, the vector of constants in (12.41) must be changed from $\begin{bmatrix} \bar{x} \\ \bar{\lambda} \\ 0 \end{bmatrix}$ to $\begin{bmatrix} 0 \\ \bar{\lambda} \\ 0 \end{bmatrix}$, and the income-compensated version of the derivative $(\partial \bar{x}/\partial P_x)$ will be

$$\left(\frac{\partial \bar{x}}{\partial P_x}\right)_{\text{compensated}} = \frac{1}{|J|}\begin{vmatrix} 0 & 0 & -P_y \\ -P_x & \bar{\lambda} & U_{xy} \\ -P_y & 0 & U_{yy} \end{vmatrix} = \frac{\bar{\lambda}}{|J|}\begin{vmatrix} 0 & -P_y \\ -P_y & U_{yy} \end{vmatrix} = T_2$$

Hence, we may express (12.42) in the form

(12.42′) $$\left(\frac{\partial \bar{x}}{\partial P_x}\right) = T_1 + T_2 = \underbrace{-\left(\frac{\partial \bar{x}}{\partial B}\right)\bar{x}}_{\text{income effect}} + \underbrace{\left(\frac{\partial \bar{x}}{\partial P_x}\right)_{\text{compensated}}}_{\text{substitution effect}}$$

This result, which decomposes the comparative-static derivative $(\partial \bar{x}/\partial P_x)$ into two components, an income effect and a substitution effect, is the two-good version of the so-called "Slutsky equation."

What can we say about the sign of $(\partial \bar{x}/\partial P_x)$? The substitution effect T_2 is clearly negative, because $|J| > 0$ and $\bar{\lambda} > 0$ [see (12.31′)]. The income effect T_1, on the other hand, is indeterminate in sign according to (12.39). Should it be negative, it would reinforce T_2; in that event, an increase in P_x must decrease the purchase of x, and the demand curve of the utility-maximizing consumer would be negatively sloped. Should it be positive, but relatively small in magnitude, it would dilute the substitution effect, though the overall result would still be a downward-sloping demand curve. But in case T_1 is positive and dominates T_2 (such as when $\bar{x}$ is a significant item in the consumer budget, thus providing an overwhelming weighting factor), then a rise in P_x will actually lead to a *larger* purchase of x, a special demand situation characteristic of what are called *Giffen goods*. Normally, of course, we would expect $(\partial\bar{x}/\partial P_x)$ to be negative.

Finally, let us examine the comparative-static derivative in (12.43), $(\partial\bar{y}/\partial P_x) = T_3 + T_4$, which has to do with the *cross effect* of a change in the price of x on the optimal purchase of y. The term T_3 bears a striking resemblance to term T_1 and again has the interpretation of an income effect.* Note that the weighting factor here is again $\bar{x}$ (rather than $\bar{y}$); this is because we are studying the effect of a change in P_x on effectual income, which depends for its magnitude upon the relative importance of $\bar{x}$ (not $\bar{y}$) in the consumer budget. Naturally, the remaining term, T_4, is again a measure of the substitution effect.

The sign of T_3 is, according to (12.40), dependent on such factors as U_{xx}, U_{yx}, etc., and is indeterminate without further restrictions on the model. However, the substitution effect T_4 will surely be positive in our model, since $\bar{\lambda}$, P_x, P_y and $|J|$ are all positive. This means that, unless more than offset by a negative income effect, an increase in the price of x will always increase the purchase of y in our two-commodity model. In other words, in the context of the present model, where the consumer can choose only between two goods, these goods must bear a relationship to each other as substitutes.

Even though the above analysis relates to the effects of a change in P_x, our results are readily adaptable to the case of a change in P_y. Our model happens to be such that the positions occupied by the variables x and y are perfectly symmetrical. Thus, to infer the effects of a change in P_y, all that it takes is to interchange the roles of x and y in the results already obtained above.

Proportionate Changes in Prices and Income

It is also of interest to ask how $\bar{x}$ and $\bar{y}$ will be affected when all three parameters P_x, P_y, and B are changed in the same proportion. Such a question still lies within the realm of comparative statics, but unlike the preceding analysis, the present inquiry now involves the simultaneous change of all the parameters.

When both prices are raised, along with income, by the same multiple j, every term in the budget constraint will increase j-fold, to become

$$jB - jxP_x - jyP_y = 0$$

* If you need a stronger dose of assurance that T_3 represents the income effect, you can use (12.40) and (12.44) to write

$$T_3 = -\left(\frac{\partial\bar{y}}{\partial B}\right)\bar{x} = \left(\frac{\partial\bar{y}}{\partial B}\right)\frac{dB}{dP_x}$$

Thus T_3 is the effect of a change in P_x on $\bar{y}$ via the income factor B.

Inasmuch as the common factor j can be canceled out, however, this new constraint is in fact identical with the old. The utility function, moreover, is independent of these parameters. Consequently, the old equilibrium levels of x and y will continue to prevail; that is, the consumer equilibrium position in our model is invariant to *equal* proportionate changes in all the prices and in the income. Thus, in the present model, the consumer is seen to be free from any "money illusion."

Symbolically, this situation can be described by the equations

$$\bar{x}(P_x, P_y, B) = \bar{x}(jP_x, jP_y, jB)$$

$$\bar{y}(P_x, P_y, B) = \bar{y}(jP_x, jP_y, jB)$$

The functions $\bar{x}$ and $\bar{y}$, with the *invariance* property just cited, are no ordinary functions; they are examples of a special class of function known as *homogeneous functions*, which have interesting economic applications. We shall therefore examine these in the next section.

EXERCISE 12.5

1 Given $U = (x + 2)(y + 1)$ and $P_x = 4$, $P_y = 6$, and $B = 130$:

(*a*) Write the Lagrangian function.

(*b*) Find the optimal levels of purchase $\bar{x}$ and $\bar{y}$.

(*c*) Is the second-order sufficient condition for maximum satisfied?

(*d*) Does the answer in (*b*) give any comparative-static information?

2 Assume that $U = (x + 2)(y + 1)$, but this time assign no specific numerical values to the price and income parameters.

(*a*) Write the Lagrangian function.

(*b*) Find $\bar{x}$, $\bar{y}$, and $\bar{\lambda}$ in terms of the parameters P_x, P_y, and B.

(*c*) Check the second-order sufficient condition for maximum.

(*d*) By setting $P_x = 4$, $P_y = 6$, and $B = 130$, check the validity of your answer to the preceding problem.

3 Can your solution ($\bar{x}$ and $\bar{y}$) in Exercise 12.5-2 yield any comparative-static information? Find all the comparative-static derivatives you can, evaluate their signs, and interpret their economic meanings.

4 From the utility function $U = (x + 2)(y + 1)$ and the constraint $xP_x + yP_y = B$ of Exercise 12.5-2, we have already found the U_{ij} and $|\bar{H}|$, as well as $\bar{x}$ and $\bar{\lambda}$. Moreover, we recall that $|J| = |\bar{H}|$.

(*a*) Substitute these into (12.39) and (12.40) to find $(\partial\bar{x}/\partial B)$ and $(\partial\bar{y}/\partial B)$.

(*b*) Substitute into (12.42) and (12.43) to find $(\partial\bar{x}/\partial P_x)$ and $(\partial\bar{y}/\partial P_x)$.

Do these results check with those obtained in Exercise 12.5-3?

5 Comment on the validity of the statement "If the derivate $(\partial\bar{x}/\partial P_x)$ is negative, then x cannot possibly represent an inferior good."

6 When studying the effect of dP_x alone, the first equation in (12.37) reduces to $-P_x\,d\bar{x} - P_y\,d\bar{y} = \bar{x}\,dP_x$, and when we compensate for the consumer's effectual income loss by dropping the term $\bar{x}\,dP_x$, the equation becomes $-P_x\,d\bar{x} - P_y\,d\bar{y} = 0$. Show that this last result can be obtained alternatively from a compensation procedure whereby we try to keep the consumer's optimal utility level $\bar{U}$ (rather than effectual income) unchanged, so that the term T_2 can alternatively be interpreted as $(\partial\bar{x}/\partial P_x)_{\bar{U}=\text{constant}}$. [*Hint*: Make use of (12.31″).]

7 (*a*) Does the assumption of diminishing marginal utility to goods x and y imply strictly convex indifference curves?

(*b*) Does the assumption of strict convexity in the indifference curves imply diminishing marginal utility to goods x and y?

12.6 HOMOGENEOUS FUNCTIONS

A function is said to be homogeneous of degree r, if multiplication of each of its independent variables by a constant j will alter the value of the function by the proportion j^r, that is, if

$$f(jx_1, \ldots, jx_n) = j^r f(x_1, \ldots, x_n)$$

In general, j can take any value. However, in order for the above equation to make sense, $(jx_1, \ldots, jx_n)$ must not lie outside the domain of the function f. For this reason, in economic applications the constant j is usually taken to be positive, as most economic variables do not admit negative values.

Example 1 Given the function $f(x, y, w) = x/y + 2w/3x$, if we multiply each variable by j, we get

$$f(jx, jy, jw) = \frac{(jx)}{(jy)} + \frac{2(jw)}{3(jx)} = \frac{x}{y} + \frac{2w}{3x} = f(x, y, w) = j^0 f(x, y, w)$$

In this particular example, the value of the function will *not* be affected at all by equal proportionate changes in all the independent variables; or, one might say, the value of the function is changed by a multiple of j^0 $(= 1)$. This makes the function f a homogeneous function of degree zero.

You will observe that the functions $\bar{x}$ and $\bar{y}$ cited at the end of the preceding section are both homogeneous of degree zero.

Example 2 When we multiply each variable in the function

$$g(x, y, w) = \frac{x^2}{y} + \frac{2w^2}{x}$$

by j, we get

$$g(jx, jy, jw) = \frac{(jx)^2}{(jy)} + \frac{2(jw)^2}{(jx)} = j\left(\frac{x^2}{y} + \frac{2w^2}{x}\right) = jg(x, y, w)$$

The function g is homogeneous of degree one (or, of the first degree); multiplication of each variable by j will alter the value of the function exactly j-fold as well.

Example 3 Now, consider the function $h(x, y, w) = 2x^2 + 3yw - w^2$. A similar multiplication this time will give us

$$h(jx, jy, jw) = 2(jx)^2 + 3(jy)(jw) - (jw)^2 = j^2h(x, y, w)$$

Thus the function h is homogeneous of degree two; in this case, a doubling of all variables, for example, will quadruple the value of the function.

Linear Homogeneity

In the discussion of production functions, wide use is made of homogeneous functions of the first degree. These are often referred to as *linearly homogeneous* functions, the adverb "linearly" modifying the adjective "homogeneous." Some writers, however, seem to prefer the somewhat misleading terminology *linear* homogeneous functions, or even *linear and* homogeneous functions, which tends to convey, wrongly, the impression that the functions themselves are linear. On the basis of the function g in Example 2 above, we know that a function which is homogeneous of the first degree is *not necessarily* linear in itself. Hence you should avoid using the terms "linear homogeneous functions" and "linear and homogeneous functions" unless, of course, the functions in question are indeed linear. Note, however, that it is not incorrect to speak of "linear homogeneity," meaning homogeneity of degree one, because to modify a noun (homogeneity) does call for the use of an adjective (linear).

Since the primary field of application of linearly homogeneous functions is in the theory of production, let us adopt as the framework of our discussion a production function in the form, say,

$$Q = f(K, L) \tag{12.45}$$

Whether applied at the *micro* or the *macro* level, the mathematical assumption of linear homogeneity would amount to the economic assumption of constant returns to scale, because linear homogeneity means that raising all inputs (independent variables) j-fold will always raise the output (value of the function) exactly j-fold also.

What unique properties characterize this linearly homogeneous production function?

Property I Given the linearly homogeneous production function $Q = f(K, L)$, the average physical product of labor (APP_L) and of capital (APP_K) can be expressed as functions of the capital-labor ratio, $k \equiv K/L$, alone.

To prove this, we multiply each independent variable in (12.45) by a factor $j = 1/L$. By virtue of linear homogeneity, this will change the output from Q to $jQ = Q/L$. The right side of (12.45) will correspondingly become

$$f\left(\frac{K}{L}, \frac{L}{L}\right) = f\left(\frac{K}{L}, 1\right) = f(k, 1)$$

Since the variables K and L in the original function are to be replaced (whenever they appear) by k and 1, respectively, the right side in effect becomes a function of the capital-labor ratio k alone, say, $\phi(k)$, which is a function with a single argument, k, even though two independent variables K and L are actually involved in that argument. Equating the two sides, we have

$$(12.46) \qquad \mathrm{APP}_L \equiv \frac{Q}{L} = \phi(k)$$

The expression for APP_K is then found to be

$$(12.47) \qquad \mathrm{APP}_K \equiv \frac{Q}{K} = \frac{Q}{L}\frac{L}{K} = \frac{\phi(k)}{k}$$

Since both average products depend on k alone, linear homogeneity implies that, as long as the K/L ratio is kept constant (whatever the absolute levels of K and L), the average products will be constant, too. Therefore, while the production function is homogeneous of degree one, both APP_L and APP_K are homogeneous of degree *zero* in the variables K and L, since equal proportionate changes in K and L (maintaining a constant k) will not alter the magnitudes of the average products.

Property II Given the linearly homogeneous production function $Q = f(K, L)$, the marginal physical products MPP_L and MPP_K can be expressed as functions of k alone.

To find the marginal products, we first write the total product as

$$(12.45') \qquad Q = L\phi(k) \qquad [\text{by } (12.46)]$$

and then differentiate Q with respect to K and L. For this purpose, we shall find the following two preliminary results to be of service:

$$(12.48) \qquad \frac{\partial k}{\partial K} = \frac{\partial}{\partial K}\left(\frac{K}{L}\right) = \frac{1}{L} \qquad \frac{\partial k}{\partial L} = \frac{\partial}{\partial L}\left(\frac{K}{L}\right) = \frac{-K}{L^2}$$

The results of differentiation are

$$(12.49) \quad \text{MPP}_K \equiv \frac{\partial Q}{\partial K} = \frac{\partial}{\partial K}[L\phi(k)]$$

$$= L\frac{\partial \phi(k)}{\partial K} = L\frac{d\phi(k)}{dk}\frac{\partial k}{\partial K} \qquad \text{[chain rule]}$$

$$= L\phi'(k)\left(\frac{1}{L}\right) = \phi'(k) \qquad \text{[by (12.48)]}$$

$$(12.50) \quad \text{MPP}_L \equiv \frac{\partial Q}{\partial L} = \frac{\partial}{\partial L}[L\phi(k)]$$

$$= \phi(k) + L\frac{\partial \phi(k)}{\partial L} \qquad \text{[product rule]}$$

$$= \phi(k) + L\phi'(k)\frac{\partial k}{\partial L} \qquad \text{[chain rule]}$$

$$= \phi(k) + L\phi'(k)\frac{-K}{L^2} \qquad \text{[by (12.48)]}$$

$$= \phi(k) - k\phi'(k)$$

which indeed show that MPP_K and MPP_L are functions of k alone.

Like average products, the marginal products will remain the same as long as the capital-labor ratio is held constant; they are homogeneous of degree zero in the variables K and L.

Property III (Euler's theorem) If $Q = f(K, L)$ is linearly homogeneous, then

$$K\frac{\partial Q}{\partial K} + L\frac{\partial Q}{\partial L} \equiv Q$$

PROOF

$$K\frac{\partial Q}{\partial K} + L\frac{\partial Q}{\partial L} = K\phi'(k) + L[\phi(k) - k\phi'(k)] \qquad \text{[by (12.49), (12.50)]}$$

$$= K\phi'(k) + L\phi(k) - K\phi'(k) \qquad [k \equiv K/L]$$

$$= L\phi(k) = Q \qquad \text{[by (12.45')]}$$

Note that this result is valid for *any* values of K and L; this is why the property can be written as an identical equality. What this property says is that the value of a linearly homogeneous function can always be expressed as a sum of terms, each of which is the product of one of the independent variables and the first-order partial derivative with respect to that variable, regardless of the levels of the two inputs actually employed. Be careful, however, to distinguish between the identity $K\frac{\partial Q}{\partial K} + L\frac{\partial Q}{\partial L} \equiv Q$ [Euler's theorem, which applies only to the constant-returns-to-scale case of $Q = f(K, L)$] and the equation $dQ = \frac{\partial Q}{\partial K}dK + \frac{\partial Q}{\partial L}dL$ [total differential of Q, for *any* function $Q = f(K, L)$].

Economically, this property means that under conditions of constant returns to scale, if each input factor is paid the amount of its marginal product, the total product will be exactly exhausted by the distributive shares fcr all the input factors, or, equivalently, the pure economic profit will be zero. Since this situation is descriptive of the long-run equilibrium under pure competition, it was once thought that only linearly homogeneous production functions would make sense in economics. This, of course, is not the case. The zero economic profit in the long-run equilibrium is brought about by the forces of competition through the entry and exit of firms, regardless of the specific nature of the production functions actually prevailing. Thus it is not mandatory to have a production function that ensures product exhaustion for any and all (K, L) pairs. Moreover, when imperfect competition exists in the factor markets, the remuneration to the factors may not be equal to the marginal products, and, consequently, Euler's theorem becomes irrelevant to the distribution picture. However, linearly homogeneous production functions are often convenient to work with because of the various nice mathematical properties they are known to possess.

Cobb-Douglas Production Function

One specific production function widely used in economic analysis is the *Cobb-Douglas production function*:

$$(12.51) \qquad Q = AK^{\alpha}L^{1-\alpha}$$

where A is a positive constant, and α is a positive fraction. What we shall consider here first is a generalized version of this function, namely,

$$(12.52) \qquad Q = AK^{\alpha}L^{\beta}$$

where β is another positive fraction which may or may not be equal to $1 - \alpha$. Some of the major features of this function are: (1) it is homogeneous of degree $(\alpha + \beta)$; (2) in the special case of $\alpha + \beta = 1$, it is linearly homogeneous; (3) its isoquants are negatively sloped throughout and strictly convex for positive values of K and L; and (4) it is strictly quasiconcave for positive K and L.

Its homogeneity is easily seen from the fact that, by changing K and L to jK and jL, respectively, the output will be changed to

$$A(jK)^{\alpha}(jL)^{\beta} = j^{\alpha+\beta}(AK^{\alpha}L^{\beta}) = j^{\alpha+\beta}Q$$

That is, the function is homogeneous of degree $(\alpha + \beta)$. In case $\alpha + \beta = 1$, there wili be constant returns to scale, because the function will be linearly homogeneous. (Note, however, that this function is *not* linear! It would thus be confusing to refer to it as a "*linear* homogeneous" or "*linear* and homogeneous" function.) That its isoquants have negative slopes and strict convexity can be verified from the signs of the derivatives dK/dL and d^2K/dL^2 (or the signs of dL/dK and d^2L/dK^2). For any positive output Q_0, (12.52) can be written as

$$AK^{\alpha}L^{\beta} = Q_0 \qquad (A, K, L, Q_0 > 0)$$

Taking the natural log of both sides and transposing, we find that

$$\ln A + \alpha \ln K + \beta \ln L - \ln Q_0 = 0$$

which implicitly defines K as a function of L.* By the implicit-function rule and the log rule, therefore, we have

$$\frac{dK}{dL} = -\frac{\partial F/\partial L}{\partial F/\partial K} = -\frac{(\beta/L)}{(\alpha/K)} = -\frac{\beta K}{\alpha L} < 0$$

Then it follows that

$$\frac{d^2K}{dL^2} = \frac{d}{dL}\left(-\frac{\beta K}{\alpha L}\right) = -\frac{\beta}{\alpha}\frac{d}{dL}\left(\frac{K}{L}\right) = -\frac{\beta}{\alpha}\frac{1}{L^2}\left(L\frac{dK}{dL} - K\right) > 0$$

The signs of these derivatives establish the isoquant (any isoquant) to be downward-sloping throughout and strictly convex in the LK plane for positive values of K and L. This, of course, is only to be expected from a function that is strictly quasiconcave for positive K and L. For the strict quasiconcavity feature of this function, see Example 5 of Sec. 12.4, where a similar function was discussed.

Let us now examine the $\alpha + \beta = 1$ case (the Cobb-Douglas function proper), to verify the three properties of linear homogeneity cited earlier. First of all, the total product in this special case is expressible as

$$(12.51') \qquad Q = AK^{\alpha}L^{1-\alpha} = A\left(\frac{K}{L}\right)^{\alpha} L = LAk^{\alpha}$$

where the expression Ak^{α} is a specific version of the general expression $\phi(k)$ used before. Therefore, the average products are

$$(12.53) \qquad \begin{aligned} \text{APP}_L &= \frac{Q}{L} = Ak^{\alpha} \\ \text{APP}_K &= \frac{Q}{K} = \frac{Q}{L}\frac{L}{K} = \frac{Ak^{\alpha}}{k} = Ak^{\alpha-1} \end{aligned}$$

both of which are now functions of k alone.

* The conditions of the implicit-function theorem are satisfied, because F (the left-side expression) has continuous partial derivatives, and because $\partial F/\partial K = \alpha/K \neq 0$ for positive values of K.

Second, differentiation of $Q = AK^{\alpha}L^{1-\alpha}$ yields the marginal products:

$$\text{(12.54)}\qquad \begin{aligned}\frac{\partial Q}{\partial K} &= A\alpha K^{\alpha-1}L^{-(\alpha-1)} = A\alpha\left(\frac{K}{L}\right)^{\alpha-1} = A\alpha k^{\alpha-1}\\ \frac{\partial Q}{\partial L} &= AK^{\alpha}(1-\alpha)L^{-\alpha} = A(1-\alpha)\left(\frac{K}{L}\right)^{\alpha} = A(1-\alpha)k^{\alpha}\end{aligned}$$

and these are also functions of k alone.

Last, we can verify Euler's theorem by using (12.54) as follows:

$$\begin{aligned}K\frac{\partial Q}{\partial K} + L\frac{\partial Q}{\partial L} &= KA\alpha k^{\alpha-1} + LA(1-\alpha)k^{\alpha}\\ &= LAk^{\alpha}\left[\frac{K\alpha}{Lk} + 1 - \alpha\right]\\ &= LAk^{\alpha}[\alpha + 1 - \alpha] = LAk^{\alpha} = Q \qquad [\text{by } (12.51')]\end{aligned}$$

Interesting economic meanings can be assigned to the exponents α and $(1-\alpha)$ in the linearly homogeneous Cobb-Douglas production function. If each input is assumed to be paid by the amount of its marginal product, the relative share of total product accruing to capital will be

$$\frac{K(\partial Q/\partial K)}{Q} = \frac{KA\alpha k^{\alpha-1}}{LAk^{\alpha}} = \alpha$$

Similarly, labor's relative share will be

$$\frac{L(\partial Q/\partial L)}{Q} = \frac{LA(1-\alpha)k^{\alpha}}{LAk^{\alpha}} = 1 - \alpha$$

Thus the exponent of each input variable indicates the relative share of that input in the total product. Looking at it another way, we can also interpret the exponent of each input variable as the partial elasticity of output with respect to that input. This is because the capital-share expression given above is equivalent to the expression $\dfrac{\partial Q/\partial K}{Q/K} \equiv \varepsilon_{QK}$ and, similarly, the labor-share expression above is precisely that of ε_{QL}.

What about the meaning of the constant A? For given values of K and L, the magnitude of A will proportionately affect the level of Q. Hence A may be considered as an *efficiency parameter*, i.e., as an indicator of the state of technology.

Extensions of the Results

We have discussed linear homogeneity in the specific context of production functions, but the properties cited are equally valid in other contexts, provided the variables K, L, and Q are properly reinterpreted.

Furthermore it is possible to extend our results to the case of more than two variables. With a linearly homogeneous function

$$y = f(x_1, x_2, \ldots, x_n)$$

we can again divide each variable by x_1 (that is, multiply by $1/x_1$) and get the result

$$y = x_1 \phi\left(\frac{x_2}{x_1}, \frac{x_3}{x_1}, \ldots, \frac{x_n}{x_1}\right) \qquad \text{[homogeneity of degree 1]}$$

which is comparable to (12.45′). Moreover, Euler's theorem is easily extended to the form

$$\sum_{i=1}^{n} x_i f_i \equiv y \qquad \text{[Euler's theorem]}$$

where the partial derivatives of the original function f (namely, f_i) are again homogeneous of degree zero in the variables x_i, as in the two-variable case.

The above extensions can, in fact, also be generalized with relative ease to a homogeneous function of degree r. In the first place, by definition of homogeneity, we can in the present case write

$$y = x_1^r \phi\left(\frac{x_2}{x_1}, \frac{x_3}{x_1}, \ldots, \frac{x_n}{x_1}\right) \qquad \text{[homogeneity of degree } r\text{]}$$

The modified version of Euler's theorem will now appear in the form

$$\sum_{i=1}^{n} x_i f_i \equiv ry \qquad \text{[Euler's theorem]}$$

where a multiplicative constant r has been attached to the dependent variable y on the right. And, finally, the partial derivatives of the original function f, the f_i, will all be homogeneous of degree $(r - 1)$ in the variables x_i. You can thus see that the linear-homogeneity case is merely a special case thereof, in which $r = 1$.

EXERCISE 12.6

1 Determine whether the following functions are homogeneous. If so, of what degree?

(*a*) $f(x, y) = \sqrt{xy}$ (*d*) $f(x, y) = 2x + y + 3\sqrt{xy}$

(*b*) $f(x, y) = (x^2 - y^2)^{1/2}$ (*e*) $f(x, y, w) = \dfrac{xy^2}{w} + 2xw$

(*c*) $f(x, y) = x^3 - xy + y^3$ (*f*) $f(x, y, w) = x^4 - 5yw^3$

2 Show that the function (12.45) can be expressed alternatively as $Q = K\psi\left(\dfrac{L}{K}\right)$ instead of $Q = L\phi\left(\dfrac{K}{L}\right)$.

3 Deduce from Euler's theorem that, with constant returns to scale:
(*a*) When $\mathrm{MPP}_K = 0$, APP_L is equal to MPP_L.
(*b*) When $\mathrm{MPP}_L = 0$, APP_K is equal to MPP_K.

4 On the basis of (12.46) through (12.50), check whether the following are true under conditions of constant returns to scale:
(*a*) An APP_L curve can be plotted against k $(= K/L)$ as the independent variable (on the horizontal axis).
(*b*) MPP_K is measured by the slope of that APP_L curve.
(*c*) APP_K is measured by the slope of the radius vector to the APP_L curve.
(*d*) $\mathrm{MPP}_L = \mathrm{APP}_L - k(\mathrm{MPP}_K) = \mathrm{APP}_L - k$ (slope of APP_L).

5 Use (12.53) and (12.54) to verify that the relations described in parts *b*, *c*, and *d* of the preceding problem are obeyed by the Cobb-Douglas production function.

6 Given the production function $Q = AK^{\alpha}L^{\beta}$, show that:
(*a*) $\alpha + \beta > 1$ implies increasing returns to scale.
(*b*) $\alpha + \beta < 1$ implies decreasing returns to scale.
(*c*) α and β are, respectively, the partial elasticities of output with respect to the capital and labor inputs.

7 Let output be a function of three inputs: $Q = AK^{a}L^{b}N^{c}$.
(*a*) Is this function homogeneous? If so, of what degree?
(*b*) Under what condition would there be constant returns to scale? Increasing returns to scale?
(*c*) Find the share of product for input N, if it is paid by the amount of its marginal product.

8 Let the production function $Q = g(K, L)$ be homogeneous of degree 2.
(*a*) Write an equation to express the second-degree homogeneity property of this function.
(*b*) Find an expression for Q in terms of $\phi(k)$, in the vein of (12.45′).
(*c*) Find the MPP_K function. Is MPP_K still a function of k alone, as in the linear-homogeneity case?
(*d*) Is the MPP_K function homogeneous in K and L? If so, of what degree?

12.7 LEAST-COST COMBINATION OF INPUTS

As another example of constrained optimization, let us discuss the problem of finding the least-cost input combination for the production of a specified level of output Q_0 representing, say, a customer's special order. Here we shall work with a general production function; later on, however, reference will be made to homogeneous production functions.

First-Order Condition

Assuming a smooth production function with two variable inputs, $Q = Q(a, b)$, where $Q_a, Q_b > 0$, and assuming both input prices to be exogenous (though again omitting the zero subscript), we may formulate the problem as one of minimizing the cost

$$C = aP_a + bP_b$$

subject to the output constraint

$$Q(a, b) = Q_0$$

Hence, the Lagrangian function is

$$Z = aP_a + bP_b + \mu[Q_0 - Q(a, b)]$$

To satisfy the first-order condition for a minimum C, the input levels (the choice variables) must satisfy the following simultaneous equations:

$$Z_\mu = Q_0 - Q(a, b) = 0$$
$$Z_a = P_a - \mu Q_a = 0$$
$$Z_b = P_b - \mu Q_b = 0$$

The first equation in this set is merely the constraint restated, and the last two imply the condition

$$\frac{P_a}{Q_a} = \frac{P_b}{Q_b} = \mu \tag{12.55}$$

At the point of optimal input combination, the input-price–marginal-product ratio must be the same for each input. Since this ratio measures the amount of outlay per unit of marginal product of the input in question, the Lagrange multiplier can be given the interpretation of the marginal cost of production in the optimum state. This interpretation is, of course, entirely consistent with our earlier discovery in (12.16) that the optimal value of the Lagrange multiplier measures the comparative-static effect of the constraint constant on the optimal value of the objective function, that is, $\bar{\mu} = (\S\bar{C}/\S Q_0)$, where the § symbol indicates that this is a partial total derivative.

Equation (12.55) can be alternatively written in the form

$$\frac{P_a}{P_b} = \frac{Q_a}{Q_b} \tag{12.55'}$$

which you should compare with (12.31″). Presented in this form, the first-order condition can be explained in terms of isoquants and isocosts. As we learned in (11.36), the Q_a/Q_b ratio is the negative of the slope of an isoquant; that is, it is a measure of the *marginal rate of technical substitution of a for b* (MRTS_{ab}). In the present model, the output level is specified at Q_0; thus only one isoquant is involved, as shown in Fig. 12.8, with a negative slope.

The P_a/P_b ratio, on the other hand, represents the negative of the slope of *isocosts* (a notion comparable with the budget line in consumer theory). An isocost, defined as the locus of the input combinations that entail the same total cost, is expressible by the equation

$$C_0 = aP_a + bP_b \qquad \text{or} \qquad b = \frac{C_0}{P_b} - \frac{P_a}{P_b}a$$

where C_0 stands for a (parametric) cost figure. When plotted in the *ab* plane, as in Fig. 12.8, therefore, it yields a family of straight lines with (negative) slope $-P_a/P_b$ (and vertical intercept C_0/P_b). The equality of the two ratios therefore amounts to the equality of the slopes of the isoquant and a selected isocost. Since we are compelled to stay on the given isoquant, this condition leads us to the point of tangency E and the input combination $(\bar{a}, \bar{b})$.

Second-Order Condition

To ensure a *minimum* cost, it is sufficient (after the first-order condition is met) to have a negative bordered Hessian, i.e., to have

$$|\bar{H}| = \begin{vmatrix} 0 & Q_a & Q_b \\ Q_a & -\mu Q_{aa} & -\mu Q_{ab} \\ Q_b & -\mu Q_{ba} & -\mu Q_{bb} \end{vmatrix} = \mu\left(Q_{aa}Q_b^2 - 2Q_{ab}Q_aQ_b + Q_{bb}Q_a^2\right) < 0$$

Since the optimal value of μ (marginal cost) is positive, this reduces to the condition that the expression in parentheses is negative when evaluated at E.

From (11.40), we recall that the curvature of an isoquant is represented by the second derivative

$$\frac{d^2b}{da^2} = \frac{-1}{Q_b^3}\left(Q_{aa}Q_b^2 - 2Q_{ab}Q_aQ_b + Q_{bb}Q_a^2\right)$$

in which the same parenthetical expression appears. Inasmuch as Q_b is positive, the satisfaction of the second-order sufficient condition would imply that d^2b/da^2

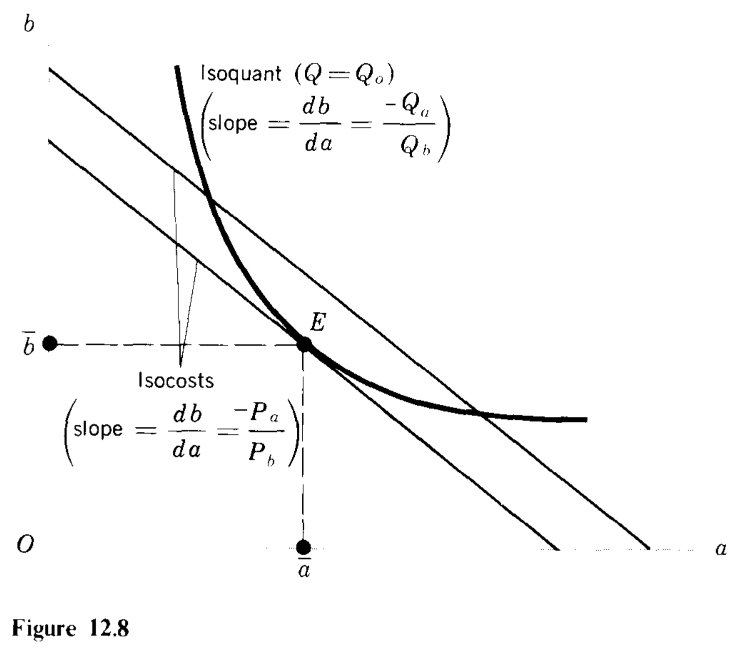

Figure 12.8

is positive—that is, the isoquant is strictly convex—at the point of tangency. In the present context, the strict convexity of the isoquant would also imply the satisfaction of the second-order sufficient condition. For, since the isoquant is negatively sloped, strict convexity can mean only a positive d^2b/da^2 (zero d^2b/da^2 is possible only at a stationary point on the isoquant), which would in turn ensure that $|\overline{H}| < 0$. However, it should again be borne in mind that the sufficient condition $|\overline{H}| < 0$ (and hence the strict convexity of the isoquant) at the tangency is, per se, not necessary for the minimization of C. Specifically, C can be minimized even when the isoquant is (nonstrictly) convex, in a multiple-minimum situation analogous to Fig. 12.7b, with $d^2b/da^2 = 0$ and $|\overline{H}| = 0$ at each minimum.

In discussing the utility-maximization model (Sec. 12.5), it was pointed out that a smooth, increasing, strictly quasiconcave utility function $U = U(x, y)$ gives rise to everywhere strictly convex, downward-sloping indifference curves in the xy plane. Since the notion of isoquants is almost identical with that of indifference curves,* we can reason by analogy that a smooth, increasing, strictly quasiconcave production function $Q = Q(a, b)$ can generate everywhere strictly convex, downward-sloping isoquants in the ab plane. If such a production function is assumed, then obviously the second-order sufficient condition will always be satisfied. Moreover, it should be clear that the resulting $\overline{C}$ will be a unique absolute constrained minimum.

The Expansion Path

Let us now turn to one of the comparative-static aspects of this model. Assuming *fixed* input prices, let us postulate successive increases of Q_0 (ascent to higher and higher isoquants) and trace the effect on the least-cost combination $\bar{b}/\bar{a}$. Each shift of the isoquant, of course, will result in a new point of tangency, with a higher isocost. The locus of such points of tangency, known as the *expansion path* of the firm, serves to describe the least-cost combinations required to produce varying levels of Q_0. Two possible shapes of the expansion path are shown in Fig. 12.9.

If we assume the strict convexity of the isoquants (hence, satisfaction of the second-order condition), the expansion path will be derivable directly from the first-order condition (12.55′). Let us illustrate this for the generalized version of the Cobb-Douglas production function.

The condition (12.55′) requires the equality of the input-price ratio and the marginal-product ratio. For the function $Q = Aa^{\alpha}b^{\beta}$, this means that each point on the expansion path must satisfy

$$\text{(12.56)} \qquad \frac{P_a}{P_b} = \frac{Q_a}{Q_b} = \frac{A\alpha a^{\alpha-1}b^{\beta}}{Aa^{\alpha}\beta b^{\beta-1}} = \frac{\alpha b}{\beta a}$$

* Both are in the nature of "isovalue" curves. They differ only in the field of application: indifference curves are used in models of consumption and isoquants, in models of production.

implying that the optimal input ratio should be

$$\frac{\bar{b}}{\bar{a}} = \frac{\beta P_a}{\alpha P_b} = \text{a constant} \tag{12.57}$$

since α, β, and the input prices are all constant. As a result, all points on the expansion path must show the same *fixed* input ratio; i.e., the expansion path must be a straight line emanating from the point of origin. This is illustrated in Fig. 12.9*b*, where the input ratios at the various points of tangency (AE/OA, $A'E'/OA'$ and $A''E''/OA''$) are all equal.

The linearity of expansion path is characteristic of the generalized Cobb-Douglas function whether or not $\alpha + \beta = 1$, because the derivation of the result in (12.57) does not rely on the assumption $\alpha + \beta = 1$. As a matter of fact, any homogeneous production function (not necessarily the Cobb-Douglas) will give rise to a linear expansion path for each set of input prices, because of the following reason: if it is homogeneous of (say) degree r, the marginal-product functions Q_a and Q_b must both be homogeneous of degree $(r - 1)$ in the inputs a and b; thus a j-fold increase in both inputs will produce a j^{r-1}-fold change in the values of *both* Q_a and Q_b, which will leave the Q_a/Q_b ratio intact. Therefore, if the first-order condition $P_a/P_b = Q_a/Q_b$ is satisfied at given input prices by a particular input combination (a_0, b_0), it must also be satisfied by a combination (ja_0, jb_0)—precisely as is depicted by the linear expansion path in Fig. 12.9*b*.

Although *any* homogeneous production function can give rise to a linear expansion path, the specific degree of homogeneity does make a significant difference in the interpretation of the expansion path. In Fig. 12.9*b*, we have drawn the distance OE equal to that of EE', so that point E' involves a doubling of the scale of point E. Now if the production function is homogeneous of degree *one*, the output at E' must be twice ($2^1 = 2$) that of E. But if the degree of

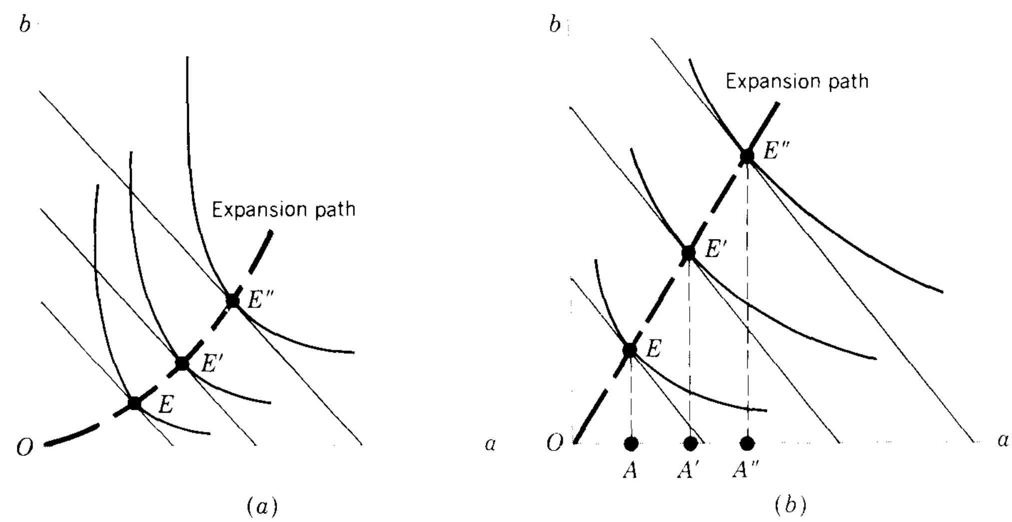

Figure 12.9

homogeneity is *two*, the output at E' will be four times ($2^2 = 4$) that of E. Thus, the spacing of the isoquants for $Q = 1$, $Q = 2, \ldots,$ will be widely different for different degrees of homogeneity.

Homothetic Functions

We have explained that, given a set of input prices, homogeneity (of any degree) of the production function produces a linear expansion path. But linear expansion paths are not unique to homogeneous production functions; for a more general class of functions, known as *homothetic functions*, can produce them, too.

A homothetic function is a composite function in the form

$$H = h[Q(a, b)] \qquad [h'(Q) \neq 0] \tag{12.58}$$

where $Q(a, b)$ is homogeneous of degree r. Although derived from a homogeneous function, the function $H = H(a, b)$ is in general *not* homogeneous in the variables a and b. Nonetheless, the expansion paths of $H(a, b)$, like those of $Q(a, b)$, are linear. The key to this result is that, at any given point in the *ab* plane, the H isoquant shares the same slope as the Q isoquant:

$$\begin{aligned} \text{Slope of } H \text{ isoquant} &= -\frac{H_a}{H_b} = -\frac{h'(Q)Q_a}{h'(Q)Q_b} \\ &= -\frac{Q_a}{Q_b} = \text{slope of } Q \text{ isoquant} \end{aligned} \tag{12.59}$$

Now the linearity of the expansion paths of $Q(a, b)$ implies, and is implied by, the condition

$$-\frac{Q_a}{Q_b} = \text{constant for any given } \frac{b}{a}$$

In view of (12.59), however, we immediately have

$$-\frac{H_a}{H_b} = \text{constant for any given } \frac{b}{a} \tag{12.60}$$

as well. And this establishes that $H(a, b)$ also produces linear expansion paths.

The concept of homotheticity is more general than that of homogeneity. In fact, every homogeneous function is automatically a member of the homothetic family, but a homothetic function may be a function outside the homogeneous family. The fact that a homogeneous function is always homothetic can be seen from (12.58), where if we let the function $H = h(Q)$ take the specific form $H = Q$ —with $h'(Q) = dH/dQ = 1$—then the function Q, being identical with the function H itself, is obviously homothetic. That a homothetic function may not be homogeneous will be illustrated in Example 2 below.

In defining the homothetic function H, we specified in (12.58) that $h'(Q) \neq 0$. This enables us to avoid division by zero in (12.59). While the specification $h'(Q) \neq 0$ is the only requirement from the mathematical standpoint, economic considerations would suggest the stronger restriction $h'(Q) > 0$. For if $H(a, b)$ is, like $Q(a, b)$, to serve as a production function, that is, if H is to denote output, then H_a and H_b should, respectively, be made to go in the same direction as Q_a and Q_b in the $Q(a, b)$ function. Thus $H(a, b)$ needs to be restricted to be a monotonically increasing transformation of $Q(a, b)$.

Homothetic production functions (including the special case of homogeneous ones) possess the interesting property that the (partial) elasticity of optimal input level with respect to the output level is uniform for all inputs. To see this, recall that the linearity of expansion paths of homothetic functions means that the optimal input ratio $\bar{b}/\bar{a}$ is unaffected by a change in the exogenous output level H_0. Thus $\partial(\bar{b}/\bar{a})/\partial H_0 = 0$ or

$$\frac{1}{\bar{a}^2}\left(\bar{a}\frac{\partial \bar{b}}{\partial H_0} - \bar{b}\frac{\partial \bar{a}}{\partial H_0}\right) = 0 \qquad \text{[quotient rule]}$$

Multiplying through by $\bar{a}^2 H_0$, and rearranging, we then get

$$\frac{\partial \bar{a}}{\partial H_0}\frac{H_0}{\bar{a}} = \frac{\partial \bar{b}}{\partial H_0}\frac{H_0}{\bar{b}} \qquad \text{or} \qquad \varepsilon_{\bar{a}H_0} = \varepsilon_{\bar{b}H_0}$$

which is what we asserted above.

Example 1 Let $H = Q^2$, where $Q = Aa^\alpha b^\beta$. Since $Q(a, b)$ is homogeneous and $h'(Q) = 2Q$ is positive for positive output, $H(a, b)$ is homothetic for $Q > 0$. We shall verify that it satisfies (12.60). First, by substitution, we have

$$H = Q^2 = (Aa^\alpha b^\beta)^2 = A^2 a^{2\alpha} b^{2\beta}$$

Thus the slope of the isoquants of H is expressed by

$$(12.61) \qquad -\frac{H_a}{H_b} = -\frac{A^2 2\alpha a^{2\alpha-1} b^{2\beta}}{A^2 a^{2\alpha} 2\beta b^{2\beta-1}} = -\frac{\alpha b}{\beta a}$$

This result satisfies (12.60) and implies linear expansion paths. A comparison of (12.61) with (12.56) also shows that the function H satisfies (12.59).

In this example, $Q(a, b)$ is homogeneous of degree $(\alpha + \beta)$. As it turns out, $H(a, b)$ is also homogeneous, but of degree $2(\alpha + \beta)$. As a rule, however, a homothetic function is not necessarily homogeneous.

Example 2 Let $H = e^Q$, where $Q = Aa^\alpha b^\beta$. Since $Q(a, b)$ is homogeneous and $h'(Q) = e^Q$ is positive, $H(a, b)$ is homothetic. From this function,

$$H(a, b) = \exp(Aa^\alpha b^\beta)$$

it is easily found that

$$-\frac{H_a}{H_b} = -\frac{A\alpha a^{\alpha-1} b^\beta \exp(Aa^\alpha b^\beta)}{Aa^\alpha \beta b^{\beta-1} \exp(Aa^\alpha b^\beta)} = -\frac{\alpha b}{\beta a}$$

This result is, of course, identical with (12.61) in Example 1.

This time, however, the homothetic function is *not* homogeneous, because

$$H(ja,\, jb) = \exp\left[A(ja)^{\alpha}(jb)^{\beta}\right] = \exp(Aa^{\alpha}b^{\beta}j^{\alpha+\beta})$$
$$= \left[\exp(Aa^{\alpha}b^{\beta})\right]^{j^{\alpha+\beta}} = \left[H(a,b)\right]^{j^{\alpha+\beta}} \neq j^{r}H(a,b)$$

Elasticity of Substitution

Another aspect of comparative statics has to do with the effect of a change in the P_a/P_b ratio upon the least-cost input combination $\bar{b}/\bar{a}$ for producing the same given output Q_0 (that is, while we stay on the same isoquant).

When the (exogenous) input-price ratio P_a/P_b rises, we can normally expect the optimal input ratio $\bar{b}/\bar{a}$ also to rise, because input b (now relatively cheaper) will tend to be substituted for input a. The *direction* of substitution is clear, but what about its *extent*? The extent of input substitution can be measured by the following point-elasticity expression, called the *elasticity of substitution* and denoted by σ (lower-case Greek letter sigma, for "substitution"):

$$(12.62) \qquad \sigma \equiv \frac{\text{relative change in } (\bar{b}/\bar{a})}{\text{relative change in } (P_a/P_b)} = \frac{\dfrac{d(\bar{b}/\bar{a})}{\bar{b}/\bar{a}}}{\dfrac{d(P_a/P_b)}{P_a/P_b}} = \frac{\dfrac{d(\bar{b}/\bar{a})}{d(P_a/P_b)}}{\dfrac{\bar{b}/\bar{a}}{P_a/P_b}}$$

The value of σ can be anywhere between 0 and ∞; the larger the σ, the greater the substitutability between the two inputs. The limiting case of $\sigma = 0$ is where the two inputs must be used in a fixed proportion as complements to each other. The other limiting case, with σ infinite, is where the two inputs are perfect substitutes for each other. Note that, if $(\bar{b}/\bar{a})$ is considered as a function of (P_a/P_b), then the elasticity σ will again be the ratio of a *marginal* function to an *average* function.*

For illustration, let us calculate the elasticity of substitution for the generalized Cobb-Douglas production function. We learned earlier that, for this case,

*There is an alternative way of expressing σ. Since, at the point of tangency, we always have

$$\frac{P_a}{P_b} = \frac{Q_a}{Q_b} = \text{MRTS}_{ab}$$

the elasticity of substitution can be defined equivalently as

$$(12.62') \qquad \sigma = \frac{\text{relative change in } (\bar{b}/\bar{a})}{\text{relative change in } \text{MRTS}_{ab}} = \frac{\dfrac{d(\bar{b}/\bar{a})}{\bar{b}/\bar{a}}}{\dfrac{d(Q_a/Q_b)}{Q_a/Q_b}} = \frac{\dfrac{d(\bar{b}/\bar{a})}{d(Q_a/Q_b)}}{\dfrac{\bar{b}/\bar{a}}{Q_a/Q_b}}$$

the least-cost input combination is specified by

$$\left(\frac{\bar{b}}{\bar{a}}\right) = \frac{\beta}{\alpha}\left(\frac{P_a}{P_b}\right) \qquad \text{[from (12.57)]}$$

This equation is in the form $y = ax$, for which dy/dx (the marginal) and y/x (the average) are both equal to the constant a. That is,

$$\frac{d(\bar{b}/\bar{a})}{d(P_a/P_b)} = \frac{\beta}{\alpha} \qquad \text{and} \qquad \frac{\bar{b}/\bar{a}}{P_a/P_b} = \frac{\beta}{\alpha}$$

Substituting these values into (12.62), we immediately find that $\sigma = 1$; that is, the generalized Cobb-Douglas production function is characterized by a *constant, unitary* elasticity of substitution. Note that the derivation of this result in no way relies upon the assumption that $\alpha + \beta = 1$. Thus the elasticity of substitution of the production function $Q = Aa^{\alpha}b^{\beta}$ will be unitary even if $\alpha + \beta \neq 1$.

CES Production Function

More recently, there has come into common use another form of production function which, while still characterized by a constant elasticity of substitution (CES), can yield a σ with a (constant) value other than 1.* The equation of this function, known as the *CES production function*, is

$$\text{(12.63)} \qquad Q = A[\delta K^{-\rho} + (1-\delta)L^{-\rho}]^{-1/\rho}$$
$$(A > 0;\ 0 < \delta < 1;\ -1 < \rho \neq 0)$$

where K and L represent two factors of production, and A, δ, and ρ (lowercase Greek letter rho) are three parameters. The parameter A (the *efficiency parameter*) plays the same role as the coefficient A in the Cobb-Douglas function; it serves as an indicator of the state of technology. The parameter δ (the *distribution parameter*), like the α in the Cobb-Douglas function, has to do with the relative factor shares in the product. And the parameter ρ (the *substitution parameter*)—which has no counterpart in the Cobb-Douglas function—is what determines the value of the (constant) elasticity of substitution, as will be shown later.

First, however, let us observe that this function is homogeneous of degree one. If we replace K and L by jK and jL, respectively, the output will change from Q to

$$A[\delta(jK)^{-\rho} + (1-\delta)(jL)^{-\rho}]^{-1/\rho} = A\{j^{-\rho}[\delta K^{-\rho} + (1-\delta)L^{-\rho}]\}^{-1/\rho}$$
$$= (j^{-\rho})^{-1/\rho}Q = jQ$$

Consequently, the CES function, like all linearly homogeneous production functions, displays constant returns to scale, qualifies for the application of Euler's theorem, and possesses average products and marginal products that are homogeneous of degree zero in the variables K and L.

* K. J. Arrow, H. B. Chenery, B. S. Minhas, and R. M. Solow, "Capital-Labor Substitution and Economic Efficiency," *Review of Economics and Statistics*, August 1961, pp. 225–250.

We may also note that the isoquants generated by the CES production function are always negatively sloped and strictly convex for positive values of K and L. To show this, let us first find the expressions for the marginal products Q_L and Q_K. Using the notation $[\cdots]$ as a shorthand for $[\delta K^{-\rho} + (1-\delta)L^{-\rho}]$, we have

$$(12.64) \qquad Q_L \equiv \frac{\partial Q}{\partial L} = A\left(-\frac{1}{\rho}\right)[\cdots]^{-(1/\rho)-1}(1-\delta)(-\rho)L^{-\rho-1}$$

$$= (1-\delta)A[\cdots]^{-(1+\rho)/\rho}L^{-(1+\rho)}$$

$$= (1-\delta)\frac{A^{1+\rho}}{A^\rho}[\cdots]^{-(1+\rho)/\rho}L^{-(1+\rho)}$$

$$= \frac{(1-\delta)}{A^\rho}\left(\frac{Q}{L}\right)^{1+\rho} > 0 \qquad [\text{by } (12.63)]$$

and similarly,

$$(12.65) \qquad Q_K \equiv \frac{\partial Q}{\partial K} = \frac{\delta}{A^\rho}\left(\frac{Q}{K}\right)^{1+\rho} > 0$$

which are defined for positive values of K and L. Thus the slope of isoquants (with K plotted vertically and L horizontally) is

$$(12.66) \qquad \frac{dK}{dL} = -\frac{Q_L}{Q_K} = -\frac{(1-\delta)}{\delta}\left(\frac{K}{L}\right)^{1+\rho} < 0 \qquad [\text{see } (11.36)]$$

It can then be easily checked that $d^2K/dL^2 > 0$ (which we leave to you as an exercise), implying that the isoquants are strictly convex for positive K and L.

It can also be shown that the CES production function is quasiconcave for positive K and L. Further differentiation of (12.64) and (12.65) shows that the second derivatives of the function have the following signs:

$$Q_{LL} = \frac{\partial}{\partial L}Q_L = \frac{(1-\delta)(1+\rho)}{A^\rho}\left(\frac{Q}{L}\right)^\rho \frac{Q_L L - Q}{L^2} < 0$$

$$[Q_L L - Q < 0, \text{ by Euler's theorem}]$$

$$Q_{KK} = \frac{\partial}{\partial K}Q_K = \frac{\delta(1+\rho)}{A^\rho}\left(\frac{Q}{K}\right)^\rho \frac{Q_K K - Q}{K^2} < 0$$

$$[Q_K K - Q < 0, \text{ by Euler's theorem}]$$

$$Q_{KL} = Q_{LK} = \frac{(1-\delta)(1+\rho)}{A^\rho}\left(\frac{Q}{L}\right)^\rho \frac{Q_K}{L} > 0$$

These derivative signs, valid for positive K and L, enable us to check the sufficient condition for quasiconcavity (12.26). As you can verify,

$$|B_1| = -Q_K^2 < 0$$

and $$|B_2| = 2Q_K Q_L Q_{KL} - Q_K^2 Q_{LL} - Q_L^2 Q_{KK} > 0$$

Thus the CES function is quasiconcave for positive K and L.

Last, we shall use the marginal products in (12.64) and (12.65) to find the elasticity of substitution of the CES function. To satisfy the least-cost combination condition $Q_L/Q_K = P_L/P_K$, where P_L and P_K denote the prices of labor service (wage rate) and capital service (rental charge for capital goods), respectively, we must have

$$\frac{1-\delta}{\delta}\left(\frac{K}{L}\right)^{1+\rho} = \frac{P_L}{P_K} \qquad [\text{see } (12.66)]$$

Thus the optimal input ratio is (introducing a shorthand symbol c)

$$(12.67) \qquad \left(\frac{\bar{K}}{\bar{L}}\right) = \left(\frac{\delta}{1-\delta}\right)^{1/(1+\rho)}\left(\frac{P_L}{P_K}\right)^{1/(1+\rho)} \equiv c\left(\frac{P_L}{P_K}\right)^{1/(1+\rho)}$$

Taking $(\bar{K}/\bar{L})$ to be a function of (P_L/P_K), we find the associated marginal and average functions to be

$$\text{Marginal function} = \frac{d(\bar{K}/\bar{L})}{d(P_L/P_K)} = \frac{c}{1+\rho}\left(\frac{P_L}{P_K}\right)^{1/(1+\rho)-1}$$

$$\text{Average function} = \frac{\bar{K}/\bar{L}}{P_L/P_K} = c\left(\frac{P_L}{P_K}\right)^{1/(1+\rho)-1}$$

Therefore the elasticity of substitution is*

$$(12.68) \qquad \sigma = \frac{\text{Marginal function}}{\text{Average function}} = \frac{1}{1+\rho}$$

What this shows is that σ is a constant whose magnitude depends on the value of the parameter ρ as follows:

$$\left.\begin{array}{r} -1 < \rho < 0 \\ \rho = 0 \\ 0 < \rho < \infty \end{array}\right\} \Rightarrow \left\{\begin{array}{l} \sigma > 1 \\ \sigma = 1 \\ \sigma < 1 \end{array}\right.$$

Cobb-Douglas Function as a Special Case of the CES Function

In this last result, the middle case of $\rho = 0$ leads to a unitary elasticity of substitution which, as we know, is characteristic of the Cobb-Douglas function. This suggests that the (linearly homogeneous) Cobb-Douglas function is a special

* Of course, we could also have obtained the same result by first taking the logarithms of both sides of (12.67):

$$\ln\left(\frac{\bar{K}}{\bar{L}}\right) = \ln c + \frac{1}{1+\rho}\ln\left(\frac{P_L}{P_K}\right)$$

and then applying the formula for elasticity in (10.28), to get

$$\sigma = \frac{d(\ln \bar{K}/\bar{L})}{d(\ln P_L/P_K)} = \frac{1}{1+\rho}$$

case of the (linearly homogeneous) CES function. The difficulty is that the CES function, as given in (12.63), is undefined when $\rho = 0$, because division by zero is not possible. Nevertheless, we can demonstrate that, as $\rho \to 0$, the CES function approaches the Cobb-Douglas function.

For this demonstration, we shall rely on a technique known as *L'Hôpital's rule*. This rule has to do with the evaluation of the limit of a function $f(x) = \dfrac{m(x)}{n(x)}$ as $x \to a$ (where a can be either finite or infinite), when the numerator $m(x)$ and the denominator $n(x)$ either (1) both tend to zero as $x \to a$, thus resulting in an expression of the 0/0 form, or (2) both tend to $\pm\infty$ as $x \to a$, thus resulting in an expression in the form of ∞/∞ (or $\infty/-\infty$, or $-\infty/\infty$, or $-\infty/-\infty$). Even though the limit of $f(x)$ cannot be evaluated as the expression stands under these two circumstances, its value can nevertheless be found by using the formula

$$(12.69) \qquad \lim_{x \to a} \frac{m(x)}{n(x)} = \lim_{x \to a} \frac{m'(x)}{n'(x)} \qquad \text{[L'Hôpital's rule]}$$

Example 3 Find the limit of $(1 - x^2)/(1 - x)$ as $x \to 1$. Here, both $m(x)$ and $n(x)$ approach zero as x approaches unity, thus exemplifying circumstance (1). Since $m'(x) = -2x$ and $n'(x) = -1$, we can write

$$\lim_{x \to 1} \frac{1 - x^2}{1 - x} = \lim_{x \to 1} \frac{-2x}{-1} = \lim_{x \to 1} 2x = 2$$

This answer is identical with that obtained by another method in Example 2 of Sec. 6.4.

Example 4 Find the limit of $(2x + 5)/(x + 1)$ as $x \to \infty$. When x becomes infinite, both $m(x)$ and $n(x)$ become infinite in the present case; thus we have here an example of circumstance (2). Since $m'(x) = 2$ and $n'(x) = 1$, we can write

$$\lim_{x \to \infty} \frac{2x + 5}{x + 1} = \lim_{x \to \infty} \frac{2}{1} = 2$$

Again, this answer is identical with that obtained by another method in Example 3 of Sec. 6.4.

It may turn out that the right-side expression in (12.69) again falls into the 0/0 or the ∞/∞ format, same as the left-side expression. In such an event, we may reapply L'Hôpital's rule, i.e., we may look for the limit of $m''(x)/n''(x)$ as $x \to a$, and take that limit as our answer. It may also turn out that even though the given function $f(x)$, whose limit we wish to evaluate, is originally not in the form of $m(x)/n(x)$ that falls into the 0/0 or the ∞/∞ format upon limit-taking, a suitable transformation will make $f(x)$ amenable to the application of the rule in (12.69). This latter possibility can be illustrated by the problem of finding the limit of the CES function (12.63)—now viewed as a function $Q(\rho)$—as $\rho \to 0$.

As given, $Q(\rho)$ is not in the form of $m(\rho)/n(\rho)$. Dividing both sides of (12.63) by A, and taking the natural log, however, we do get an expression in that form, namely,

$$\ln \frac{Q}{A} = \frac{-\ln[\delta K^{-\rho} + (1-\delta)L^{-\rho}]}{\rho} \equiv \frac{m(\rho)}{n(\rho)} \tag{12.70}$$

Moreover, as $\rho \to 0$, we find that $m(\rho) \to -\ln[\delta + 1 - \delta] = -\ln 1 = 0$, and $n(\rho) \to 0$, too. Thus L'Hôpital's rule can be used to find the limit of $\ln(Q/A)$. Once that is done, the limit of Q can also be found: since $Q/A = e^{\ln(Q/A)}$, so that $Q = Ae^{\ln(Q/A)}$, it follows that

$$\lim Q = \lim Ae^{\ln(Q/A)} = Ae^{\lim \ln(Q/A)} \tag{12.71}$$

From (12.70), let us first find $m'(\rho)$ and $n'(\rho)$, as required by L'Hôpital's rule. The latter is simply $n'(\rho) = 1$. The former is

$$m'(\rho) = \frac{-1}{[\delta K^{-\rho} + (1-\delta)L^{-\rho}]} \frac{d}{d\rho}[\delta K^{-\rho} + (1-\delta)L^{-\rho}] \qquad \text{[chain rule]}$$

$$= \frac{-[-\delta K^{-\rho}\ln K - (1-\delta)L^{-\rho}\ln L]}{[\delta K^{-\rho} + (1-\delta)L^{-\rho}]} \qquad \text{[by (10.21')]}$$

By L'Hôpital's rule, therefore, we have

$$\lim_{\rho \to 0} \ln \frac{Q}{A} = \lim_{\rho \to 0} \frac{m'(\rho)}{n'(\rho)} = \frac{\delta \ln K + (1-\delta)\ln L}{1} = \ln(K^{\delta} L^{1-\delta})$$

In view of this result, when e is raised to the power of $\lim_{\rho \to 0} \ln(Q/A)$, the outcome is simply $K^{\delta}L^{1-\delta}$. Hence, by (12.71), we finally arrive at the result

$$\lim_{\rho \to 0} Q = AK^{\delta}L^{1-\delta}$$

showing that, as $\rho \to 0$, the CES function indeed tends to the Cobb-Douglas function.

EXERCISE 12.7

1 Suppose that the isoquants in Fig. 12.9*b* are derived from a particular homogeneous production function $Q = Q(a, b)$. Noting that $OE = EE' = E'E''$, what must be the ratios between the output levels represented by the three isoquants if the function Q is homogeneous

(*a*) of degree one? (*b*) of degree two?

2 For the generalized Cobb-Douglas case, if we plot the ratio $\bar{b}/\bar{a}$ against the ratio P_a/P_b, what type of curve will result? Does this result depend on the assumption that $\alpha + \beta = 1$? Read the elasticity of substitution graphically from this curve.

3 Is the CES production function characterized by diminishing returns to each input for all positive levels of input?

4 Show that, on an isoquant of the CES function, $d^2K/dL^2 > 0$.

5 (*a*) For the CES function, if each factor of production is paid according to its marginal product, what is the ratio of labor's share of product to capital's share of product? Would a larger value of δ mean a larger relative share for capital?

(*b*) For the Cobb-Douglas function, is the ratio of labor's share to capital's share dependent on the K/L ratio? Does the same answer apply to the CES function?

6 (*a*) The CES production function rules out $\rho = -1$. If $\rho = -1$, however, what would be the general shape of the isoquants for positive K and L?

(*b*) Is σ defined for $\rho = -1$? What is the limit of σ as $\rho \to -1$?

(*c*) Interpret economically the above results.

7 Show that by writing the CES function as $Q = A[\delta K^{-\rho} + (1-\delta)L^{-\rho}]^{-r/\rho}$, where $r > 0$ is a new parameter, we can introduce increasing returns to scale and decreasing returns to scale.

8 Evaluate the following:

(*a*) $\lim\limits_{x \to 4} \dfrac{x^2 - x - 12}{x - 4}$ (*c*) $\lim\limits_{x \to 0} \dfrac{5^x - e^x}{x}$

(*b*) $\lim\limits_{x \to 0} \dfrac{e^x - 1}{x}$ (d) $\lim\limits_{x \to \infty} \dfrac{\ln x}{x}$

9 By use of L'Hôpital's rule, show that

(*a*) $\lim\limits_{x \to \infty} \dfrac{x^n}{e^x} = 0$ (*b*) $\lim\limits_{x \to 0^+} x \ln x = 0$ (*c*) $\lim\limits_{x \to 0^+} x^x = 1$

12.8 SOME CONCLUDING REMARKS

In the present part of the book, we have covered the basic techniques of optimization. The somewhat arduous journey has taken us (1) from the case of a single choice variable to the more general n-variable case, (2) from the polynomial objective function to the exponential and logarithmic, and (3) from the unconstrained to the constrained variety of extremum.

Most of this discussion consists of the "classical" methods of optimization, with differential calculus as the mainstay, and derivatives of various orders as the primary tools. One weakness of the calculus approach to optimization is its essentially myopic nature. While the first- and second-order conditions in terms of derivatives or differentials can normally locate relative or local extrema without difficulty, additional information or further investigation is often required for identification of absolute or global extrema. Our detailed discussion of concavity, convexity, quasiconcavity, and quasiconvexity is intended as a useful stepping stone from the realm of relative extrema to that of absolute ones.

A more serious limitation of the calculus approach is its inability to cope with constraints in the inequality form. For this reason, the budget constraint in the utility-maximization model, for instance, is stated in the form that the total expenditure be exactly *equal to* (and not "less than or equal to") a specified sum. In other words, the limitation of the calculus approach makes it necessary to deny the consumer the option of saving part of the available funds. And it is for precisely the same reason that we have not explicitly constrained the choice variables to be nonnegative, as economic common sense may dictate. In fact, we have only been able to use inequalities as model specifications (such as $Q_a > 0$ and $Q_{aa} < 0$). These play a role in evaluating the signs of mathematical solutions but are not objects of mathematical operations themselves.

We shall deal with the matter of inequality constraints when we study mathematical programming (linear and nonlinear programming), which represents the "nonclassical" approach to optimization. That topic, however, is reserved for Part 6 of the book. Meanwhile, so that you can develop an appreciation of the full sweep of categories of economic analysis—statics → comparative statics → dynamics—we shall introduce in Part 5 methods of dynamic analysis. This is also pedagogically preferable, since the mathematical techniques of dynamic analysis are closely related to methods of differential calculus which we have just learned.

For those of you who are anxious to turn to mathematical programming, however, it is perfectly feasible to skip Part 5 and proceed directly to Part 6. No methodological difficulties should arise.