Recursive Methods in Economic Dynamics

NANCY L. STOKEY
AND
ROBERT E. LUCAS, JR.

with Edward C. Prescott

Harvard University Press
Cambridge, Massachusetts, and London, England 1989

Preface

This book was motivated by our conviction that recursive methods should be part of every economist's set of analytical tools. Applications of these methods appear in almost every substantive area of economics—the theory of investment, the theory of the consumer, search theory, public finance, growth theory, and so on—but neither the methods nor the applications have ever been drawn together and presented in a systematic way. Our goal has been to do precisely this. We have attempted to develop the basic tools of recursive analysis in a systematic, rigorous way, while at the same time stressing the wide applicability of recursive methods and suggesting new areas where they might usefully be exploited.

Our first outlines for the book included a few chapters devoted to mathematical preliminaries, followed by numerous chapters treating the various substantive areas of economics to which recursive methods have been applied. We hoped to keep the technical material to a minimum, by simply citing the existing literature for most of the required mathematical results, and to focus on substantive issues. This plan failed rather quickly, as it soon became apparent that the reader would be required either to take most of the important results on faith or else to keep a dozen mathematics books close at hand and refer to them constantly. Neither approach seemed reasonable, and we were led to make major alterations in the overall structure of the book.

The methods became the organizing principle, and we began to focus on providing a fairly comprehensive, rigorous, and self-contained treatment of the tools and techniques used in recursive analysis. We then found it natural to group applications by the nature of the technical tools involved rather than by their economic substance. Thus Parts II–IV of the book deal with deterministic models, stochastic models, and equilibrium theory, respectively, with substantive applications appearing in all three places. Indeed, many of the applications appear more than once, with different aspects of the same problem treated as the appropriate tools are developed.

Once we had decided to write a book focused on analytical tools rather than on economic substance, the choice of technical level became more important than ever. We wanted the book to be rigorous enough to be useful to researchers and at the same time to be accessible to as wide an audience as possible. In pursuing these twin goals we have aimed for a rigorous and fairly general treatment of the analytical tools, but one that requires relatively little by way of mathematical background. The reader should have had a course in advanced calculus or real analysis and should be comfortable with delta-epsilon arguments. A little background in probability theory is also useful, although not at all essential. The other mathematical topics that arise—and there are a wide variety—are treated in a largely self-contained way.

The most difficult decision we faced was choosing the appropriate level at which to treat probability theory. Our first inclination was to restrict attention to discrete probabilities and continuous densities, but in the end we found that this approach caused more trouble than it saved. We were pleased to find that a relatively small investment in measure theory produced enormous returns. We provide a modest number of definitions and basic results from the abstract theory of measure and integration in Chapter 7, and then draw on them repeatedly in our treatment of stochastic models. The reader will find that this investment yields returns elsewhere as well: measure theory is rapidly becoming the standard language of the economics of uncertainty.

The term *recursive methods* is broad enough to include a variety of interesting topics that might have been included in the book but are not. There is a large literature on linear-quadratic formulations of dynamic problems that, except for examples discussed briefly in Chapters 4 and 9, we ignore. There is also a growing body of expertise on methods for the numerical solution of recursive models that we have not attempted to incorporate in this volume. Although a wide variety of dynamic games can be analyzed by recursive methods, our examples of equilibrium are almost exclusively competitive. We have included a large collection of applications, but we certainly have not exhausted the many applied literatures where recursive methods are being used. Yet these omissions are

not, we feel, cause for apology. The book is long enough as it is, and we will certainly not be disappointed if one of the functions it serves is to stimulate the reader to a more serious exploration of some of these closely related areas.

We have tried to write this book in a way that will make it useful for several different types of readers. Those who are familiar with dynamic economic models and have specific questions in mind are invited simply to consult the table of contents and proceed to the particular topics that interest them. We have tried to make chapters and sections sufficiently self-contained so that the book can be used in this way. Primarily, however, the book is directed at the reader with little or no background in dynamic models. The manuscript has, at a variety of stages, been used for graduate-level courses at Chicago, Minnesota, Northwestern, and elsewhere, and we have been gratified with the response from students. The book is about the right length and level for a year-long course for second-year students but can easily be adapted for shorter courses as well. After the introductory material in Chapters 1 and 2, it is probably advisable to cover Chapter 3 in detail, skim Section 4.1, cover Section 4.2 in detail, and then choose a few applications from Chapter 5. For a one-quarter course, there are then several possibilities. One could skip to Chapters 15 and 16, and if time permits, go on to 17 and 18. Alternatively, with measure theory as a prerequisite, one could proceed to Section 8.1, then to Section 9.2, and then to applications from Chapter 10. Covering the required measure theory, Sections 7.0–7.5, takes about three weeks and could be done in a one-semester course.

A consequence of our decision to make the book technically self-contained is that completing it involved a much higher ratio of exposition to new results than any of us had anticipated. Ed Prescott found he did not wish to spend so much of his time away from the research frontier, and so proposed the reduced level of involvement reflected in the phrase "with the collaboration of." However, there is no part of the book that has not benefited from his ideas and contributions.

We are grateful also to many friends and colleagues for their comments and criticism. In particular we thank Andrew Caplin, V. V. Chari, Lars Hansen, Hugo Hopenhayn, Larry Jones, Lars Ljungquist, Rodolfo Manuelli, Masao Ogaki, José Victor Rios-Rull, and José Scheinkman for fruitful discussions. Arthur Kupferman read large portions of the manuscript at an early stage, and his detailed comments enhanced both the content and the style of the final product. We are also indebted to Ricard Torres, whose comments on the entire manuscript led to many improvements and, in several places, to major revisions along lines he proposed.

We owe special thanks to Michael Aronson, whose patience and enthusiasm have supported this project from its beginning—more years ago than any of us cares to remember. We are grateful too to Jodi Simpson, whose editing led to many refinements of style and logic; her skillful work is much valued. June Nason began typing our early drafts on an IBM Selectric and stayed to finish the job on a LaserJet printer. We appreciate her cheerful assistance, and the tact she showed by never asking how a job could remain urgent for six years. Finally, we would like to thank Mary Ellen Geer for helping us see the book through to its completion.

Contents

Symbols Used

$x \in X$	element
$A \subseteq B, A \subset B$	subset, strict subset
$A \supseteq B, A \supset B$	superset, strict superset
$\cup, \cap$	union, intersection
$\emptyset$	empty set
$A \backslash B$	difference, defined only if $A \supseteq B$
A^c	complement
$\mathring{A}, \bar{A}$	interior, closure
∂A	boundary
χ_A	indicator function
$X \times Y$	Cartesian product
$\mathbf{R}, \bar{\mathbf{R}}$	real numbers, extended real numbers
$\mathbf{R}^l$	l-dimensional Euclidean space
$\mathbf{R}^l_+, \mathbf{R}^l_{++}$	subspace of $\mathbf{R}^l$ containing nonnegative vectors, strictly positive vectors
$(a, b), [a, b]$	open interval, closed interval
$(a, b], [a, b)$	half-open intervals
$\mathcal{B}, \mathcal{B}^l$	Borel subsets of $\mathbf{R}$, of $\mathbf{R}^l$
$\mathcal{B}_X$	Borel subsets of X, defined for $X \in \mathcal{B}^l$
$\rho(x, y)$	distance
$\|x\|$	norm
$C(X)$	space of bounded continuous functions on X
f^+, f^-	positive and negative parts of the function f
$\{x_i\}_{i=1}^n$	finite sequence
$\{x_i\}_{i=1}^\infty$	infinite sequence

$x_i \to x$	converges
$x_i \uparrow x$	converges from below
$x_i \downarrow x$	converges from above
$(X, \mathscr{X})$	measurable space
$M(X, \mathscr{X})$	space of measurable real-valued functions on $(X, \mathscr{X})$
$M^+(X, \mathscr{X})$	subset of $M(X, \mathscr{X})$ containing nonnegative functions
$B(X, \mathscr{X})$	space of bounded measurable real-valued functions on $(X, \mathscr{X})$
$(X, \mathscr{X}, \mu)$	measure space
$L(X, \mathscr{X}, \mu)$	space of μ-integrable functions on $(X, \mathscr{X})$
$\Lambda(X, \mathscr{X})$	space of probability measures on $(X, \mathscr{X})$
μ-a.e.	except on a set A with $\mu(A) = 0$
$\mu \perp \lambda$	mutually singular
$\mu \ll \lambda$	absolutely continuous with respect to
$\lambda_n \to \lambda$	converges in the total variation norm
$\lambda_n \Rightarrow \lambda$	converges weakly
$\mathscr{A} \times \mathscr{B}$	product σ-algebra

PART I

The Recursive Approach

1 Introduction

Research in economic dynamics has undergone a remarkable transformation in recent decades. A generation ago, empirical researchers were typically obliged to add dynamic and stochastic elements as afterthoughts to predictions about behavior derived from static, deterministic economic models. Today, in every field of application, we have theories that deal explicitly with rational economic agents operating through time in stochastic environments. The idea of an economic equilibrium has undergone a similar evolution: it no longer carries the connotation of a system at rest. Powerful methods are now available for analyzing theoretical models with equilibrium outcomes described by the same kinds of complicated stochastic processes that we use to describe observed economic behavior.

These theoretical developments are based on a wide variety of results in economics, mathematics, and statistics: the contingent-claim view of economic equilibria introduced by Arrow (1953) and Debreu (1959), the economic applications of the calculus of variations pioneered long ago by Ramsey (1928) and Hotelling (1931), the theory of dynamic programming of Bellman (1957) and Blackwell (1965). Our goal in this book is to provide self-contained treatments of these theoretical ideas that form the basis of modern economic dynamics. Our approach is distinguished by its systematic use of recursive methods, methods that make it possible to treat a wide variety of dynamic economic problems—both deterministic and stochastic—from a fairly unified point of view.

To illustrate what we mean by a recursive approach to economic dynamics, we begin with a list of concrete examples, drawn from the much longer list of applications to be treated in detail in later chapters. These examples also serve to illustrate the kinds of substantive economic questions that can be studied by the analytical methods in this book.

First consider an economy that produces a single good that can be either consumed or invested. The quantity consumed yields immediate utility to the single decision-maker, a "social planner." The quantity invested augments the capital stock, thereby making increased production possible in the future. What is the consumption–investment policy that maximizes the sum of utilities over an infinite planning horizon?

Next consider an economy that is otherwise similar to the one just described, but that is subject to random shocks affecting the amount of output that can be produced with a given stock of capital. How should the consumption-investment decision be made if the objective is to maximize the expected sum of utilities?

Suppose a worker wishes to maximize the present value of his earnings. In any period he is presented with a wage offer at which he can work one unit of time or zero. If he works, he takes the earnings and retains the same job next period. If he does not work, he searches, an activity that yields him a new wage offer from a known probability distribution. What decision rule should he adopt if his goal is to maximize the expected present discounted value of his lifetime earnings?

A store manager has in stock a given number of items of a specific type. Demand is stochastic, so in any period he may either stock out and forgo the sales he would have made with a larger inventory or incur the costs of carrying over unsold items. At the beginning of each period he can place an order for more items. The cost of this action includes a fixed delivery charge plus a charge per item ordered. The order must be placed before the manager knows the current period demand. If his goal is to maximize the expected discounted present value of profits, when should he place an order; and when an order is placed, how large should it be?

An economy is endowed with a fixed number of productive assets that have exogenously given yields described by a stochastic process. These assets are privately owned, and claims to all of them are traded on a competitive equities market. How are the competitive equilibrium prices in this market related to consumer preferences over consumption of goods and to the current state of the yield process? How is the answer to this question altered if assets can be produced?

A monopolist faces a stochastically shifting demand curve for his product. His current production capacity is determined by his past investments, but he has the option to invest in additions to capacity, additions that will be available for production in the future. What investment strategy maximizes the expected discounted present value of profits? Alternatively, suppose there are many firms in this industry. In competitive equilibrium what are the investment strategies for all of these firms, and what do they imply for the behavior of industry production and prices?

These problems evidently have much in common. In each case a decision-maker—a social planner, a worker, a manager, an entire market, a firm, or collection of firms—must choose a sequence of actions through time. In the first example there is no uncertainty, so the entire sequence may as well be chosen at the outset. In the other five examples the environment is subject to unpredictable outside shocks, and it is clear that the best future actions depend on the magnitudes of these shocks. Consider how we might formulate each of these problems mathematically and what we might mean by a recursive approach to each.

The first example is the problem of optimal savings that Frank Ramsey formulated and solved in 1928. Ramsey viewed the problem as one of maximizing a function (total utility) of an infinity of variables (consumption and capital stock at each date) subject to the constraints imposed by the technology. He set up the problem in continuous time and applied the calculus of variations to obtain a very sharp characterization of the utility-maximizing dynamics: the capital stock should converge monotonically to the level that, if sustained, maximizes consumption per unit of time.

In the Ramsey problem the feature of the production possibility set that changes over time is the current stock of capital. This observation suggests that an alternative way to describe the optimal policy is in terms of a function that gives the society's optimal current investment as a function of its current capital stock; and, in fact, Ramsey's solution can be expressed in this way. Thus an alternative mathematical strategy is to seek the optimal savings function directly and then to use this function to compute the optimal sequence of investments from any initial stock. This way of looking at the problem—decide on the immediate action to take as a function of the current situation—is called a recursive formulation because it exploits the observation that a decision problem of the same general structure recurs each period.

Since Ramsey completely solved his problem using variational methods, this example is better suited to defining a recursive approach than to motivating it. Consider next the stochastic variation on this problem. In this case it obviously makes no sense to choose a deterministic plan of investments for all future dates: the best future choices will depend on how much output is available at the time, and that in turn will depend on as-yet-unrealized shocks to the productivity of capital. To carry out the analogue to Ramsey's strategy, one must follow the contingent-claim formulation introduced by Kenneth Arrow (1953) and Gerard Debreu (1959) and view an investment plan as a sequence of investments, each of which is made contingent on the history of shocks that have been realized up to the time the decision is actually implemented.

This contingent-claim formulation is an enormously useful point of view for many purposes (and is, indeed, essential to the analysis in parts of this book), but in the current context it leads to a maximization problem in a space that is much more difficult to work in than the space of sequences Ramsey used. Yet a recursive formulation of the stochastic Ramsey problem is hardly more complicated than the one for the deterministic case. With random shocks the current state of the system is described by two variables: capital and the current shock. We search, in this case, for a savings function that expresses the optimal investment decision as a function of these two variables only.

The optimal job-search problem can also be set up as one of choosing a sequence of contingent actions, but it is awkward to do so. In this problem the shocks the decision-maker observes depend on his actions: if he takes a job, he never learns what wage offers he would have received if he had kept looking. Formulated recursively, the problem becomes one of choosing a single number, the reservation wage. The worker should then accept any job offering a wage above this level and reject any offer below it. The inventory problem discussed next has a similar structure, although in this case two numbers must be chosen: the inventory level that triggers an order and the order size.

The asset-pricing example does not have a single decision-maker, as did the first four examples. The issue here is the determination of market equilibrium prices. We can, for this economy, calculate the Arrow-Debreu prices for dated claims to goods, contingent on the histories of shocks up to the date at which the exchange is to occur. Alternatively (and, one can show, equivalently), we can think of prices as being set in a sequence of spot equity markets. From this second, recursive viewpoint we seek an expression for equilibrium prices as functions of the system's state that is exactly analogous to expressing agents' decisions as functions of the state.

Our final example was a microeconomic problem involving investment in a single industry with given consumer demand behavior. When the industry is a monopoly, this problem has a single decision-maker and thus is similar in structure to the social planner's problem of choosing optimal savings in a stochastic economy. If the industry is competitive, we would like to solve simultaneously for equilibrium prices and investment levels, both as functions of the industry state. As Harold Hotelling conjectured in his 1931 paper on exhaustible resources, in this case the industry as a whole solves a consumer surplus maximization problem that has exactly the same mathematical structure as does the monopoly problem.

As we hope these examples illustrate, a great variety of economic and other decision problems are quite naturally cast in a recursive framework. A first purpose of this book, then, is to present in a unified way the theory of recursive decision-making—dynamic programming (in Richard Bellman's terminology)—and to illustrate the application of this theory to a wide variety of economic problems.

A second purpose is to show how the methods of dynamic programming can be combined with those of modern general equilibrium theory to yield tractable models of dynamic economic systems. This possibility is easiest to see when the system as a whole itself solves a maximum problem, and some of our applications take this form. We will also consider systems the behavior of which cannot be mimicked by any individual decision problem but to which recursive methods can still fruitfully be applied.

These examples give a sense of the kinds of economic issues we will address and of the general point of view from which we intend to study them. Because at this informal level it is not possible to discuss the technical questions such problems raise, we cannot at this juncture provide a useful overview of the remainder of the book. Accordingly, in the next chapter we study a concrete economic example that illustrates the range of analytical methods we will be dealing with. At that point we will be in a position to outline the rest of the book.

2 An Overview

In this chapter we preview the recursive methods of analysis to be developed in detail in the rest of the book. This material falls into three broad parts, and the remainder of the book is structured accordingly. Part II deals with methods for solving deterministic optimization problems, Part III with the extension of these methods to problems that include stochastic shocks, and Part IV with ways of using solutions of either type within a competitive equilibrium framework.

To make this preview as concrete as possible, we examine these three sets of issues by looking at a specific example, a one-sector model of economic growth. Our goal is not to provide a substantive treatment of growth theory but to illustrate the types of arguments and results that are developed in the later chapters of the book—arguments that can be applied to a wide variety of problems. A few of these problems were mentioned in Chapter 1, and many more will be discussed in detail in Chapters 5, 10, 13, 16, 17, and 18, all of which are devoted exclusively to substantive applications. With that said, in this chapter we focus exclusively on the example of economic growth.

In the next three sections we consider resource allocation in an economy composed of many identical, infinitely lived households. In each period t there is a single good, y_t, that is produced using two inputs: capital, k_t, in place at the beginning of the period, and labor, n_t. A production function relates output to inputs, $y_t = F(k_t, n_t)$. In each period current output must be divided between current consumption, c_t, and gross investment, i_t,

$$(1) \qquad c_t + i_t \leq y_t = F(k_t, n_t).$$

This consumption-savings decision is the only allocation decision the economy must make. Capital is assumed to depreciate at a constant rate

$0 < \delta < 1$, so capital is related to gross investment by

$$(2) \qquad k_{t+1} = (1 - \delta)k_t + i_t.$$

Labor is taken to be supplied inelastically, so $n_t = 1$, all t. Finally, prefer ences over consumption, common to all households, are taken to be o the form

$$(3) \qquad \sum_{t=0}^{\infty} \beta^t U(c_t),$$

where $0 < \beta < 1$ is a discount factor.

In Sections 2.1 and 2.2 we study the problem of optimal growth Specifically, in Section 2.1 we examine the problem of maximizing (3 subject to (1) and (2), given an initial capital stock k_0. In Section 2.2 w modify this planning problem to include exogenous random shocks t the technology in (1), in this case taking the preferences of household over random consumption sequences to be the expected value of th function in (3). In Section 2.3 we return to the deterministic model. W begin by characterizing the paths for consumption and capital accumula tion that would arise in a competitive market economy composed o many households, each with the preferences in (3), and many firms, eac with the technology in (1) and (2). We then consider the relationshi between the competitive equilibrium allocation and the solution to th planning problem found earlier. We conclude in Section 2.4 with a mor detailed overview of the remainder of the book, discussing briefly th content of each of the later chapters.

2.1 A Deterministic Model of Optimal Growth

In this section we study the problem of optimal growth when there is n uncertainty. Assume that the production function is $y_t = F(k_t, n_t)$, wher $F: \mathbf{R}^2_+ \to \mathbf{R}_+$ is continuously differentiable, strictly increasing, homoge neous of degree one, and strictly quasi-concave, with

$$F(0, n) = 0, \quad F_k(k, n) > 0, \quad F_n(k, n) > 0, \quad \text{all } k, n > 0;$$

$$\lim_{k \to 0} F_k(k, 1) = \infty, \quad \lim_{k \to \infty} F_k(k, 1) = 0.$$

Assume that the size of the population is constant over time and normalize the size of the available labor force to unity. Then actual labor supply must satisfy

(1a) $0 \le n_t \le 1$, all t.

Assume that capital decays at the fixed rate $0 < \delta \le 1$. Then consumption c_t, gross investment $i_t = k_{t+1} - (1-\delta)k_t$, and output $y_t = F(k_t, n_t)$ must satisfy the feasibility constraint

(1b) $c_t + k_{t+1} - (1-\delta)k_t \le F(k_t, n_t)$, all t.

Assume that all of the households in this economy have identical preferences over intertemporal consumption sequences. These common preferences take the additively separable form

(2) $$u(c_0, c_1, \ldots) = \sum_{t=0}^{\infty} \beta^t U(c_t),$$

where the discount factor is $0 < \beta < 1$, and where the current-period utility function $U: \mathbf{R}_+ \to \mathbf{R}$ is bounded, continuously differentiable, strictly increasing, and strictly concave, with $\lim_{c \to 0} U'(c) = \infty$. Households do not value leisure.

Now consider the problem faced by a benevolent social planner, one whose objective is to maximize (2) by choosing sequences $\{(c_t, k_{t+1}, n_t)\}_{t=0}^{\infty}$, subject to the feasibility constraints in (1), given $k_0 > 0$. Two features of any optimum are apparent. First, it is clear that output will not be wasted. That is, (1b) will hold with equality for all t, and we can use it to eliminate c_t from (2). Second, since leisure is not valued and the marginal product of labor is always positive, it is clear that an optimum requires $n_t = 1$, all t. Hence k_t and y_t represent both capital and output per worker and capital and output in total. It is therefore convenient to define $f(k) = F(k, 1) + (1-\delta)k$ to be the total supply of goods available per worker, including undepreciated capital, when beginning-of-period capital is k.

Exercise 2.1 Show that the assumptions on F above imply that $f: \mathbf{R}_+ \to \mathbf{R}_+$ is continuously differentiable, strictly increasing, and strictly concave, with

$$f(0) = 0, \quad f'(k) > 0, \quad \lim_{k \to 0} f'(k) = \infty, \quad \lim_{k \to \infty} f'(k) = 1 - \delta.$$

The planning problem can then be written as

(3) $$\max_{\{k_{t+1}\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t U[f(k_t) - k_{t+1}]$$

(4) $$\text{s.t.} \quad 0 \le k_{t+1} \le f(k_t), \quad t = 0, \ldots;$$

$$k_0 > 0 \quad \text{given.}$$

Although ultimately we are interested in the case where the planning horizon is infinite, it is instructive to begin with the (much easier!) problem of a finite horizon. If the horizon in (3) were a finite value T instead of infinity, then (3)–(4) would be an entirely standard concave programming problem. With a finite horizon, the set of sequences $\{k_{t+1}\}_{t=0}^{T}$ satisfying (4) is a closed, bounded, and convex subset of $\mathbf{R}^{T+1}$, and the objective function (3) is continuous and strictly concave. Hence there is exactly one solution, and it is completely characterized by the Kuhn-Tucker conditions.

To obtain these conditions note that since $f(0) = 0$ and $U'(0) = \infty$, it is clear that the inequality constraints in (4) do not bind except for k_{T+1}, and it is also clear that $k_{T+1} = 0$. Hence the solution satisfies the first-order and boundary conditions

(5) $$\beta f'(k_t)U'[f(k_t) - k_{t+1}] = U'[f(k_{t-1}) - k_t], \quad t = 1, 2, \ldots, T;$$

(6) $$k_{T+1} = 0, \quad k_0 > 0 \quad \text{given.}$$

Equation (5) is a second-order difference equation in k_t; hence it has a two-parameter family of solutions. The unique optimum for the maximization problem of interest is the one solution in this family that in addition satisfies the two boundary conditions in (6). The following exercise illustrates how (5)–(6) can be used to solve for the optimum in a particular example.

Exercise 2.2 Let $f(k) = k^\alpha$, $0 < \alpha < 1$, and let $U(c) = \ln(c)$. (No, this does not fit all of the assumptions we placed on f and U above, but go ahead anyway.)

a. Write (5) for this case and use the change of variable $z_t = k_t/k_{t-1}^\alpha$ to convert the result into a first-order difference equation in z_t. Plot z_{t+1} against z_t and plot the 45° line on the same diagram.

b. The boundary condition (6) implies that $z_{T+1} = 0$. Using this condition, show that the unique solution is

$$z_t = \alpha\beta \frac{1 - (\alpha\beta)^{T-t+1}}{1 - (\alpha\beta)^{T-t+2}}, \quad t = 1, 2, \ldots, T + 1.$$

c. Check that the path for capital

$$(7) \qquad k_{t+1} = \alpha\beta \frac{1 - (\alpha\beta)^{T-t}}{1 - (\alpha\beta)^{T-t+1}} k_t^\alpha, \quad t = 0, 1, \ldots, T,$$

given k_0, satisfies (5)–(6).

Now consider the infinite-horizon version of the planning problem in Exercise 2.2. Note that if T is large, then the coefficient of k_t^α in (7) is essentially constant at $\alpha\beta$ for a very long time. For the solution to the infinite-horizon problem, can we not simply take the limit of the solutions in (7) as T approaches infinity? After all, we are discussing households that discount the future at a geometric rate! Taking the limit in (7), we find that

$$(8) \qquad k_{t+1} = \alpha\beta k_t^\alpha, \quad t = 0, 1, \ldots.$$

In fact, this conjecture is correct: the limit of the solutions for the finite-horizon problems is the unique solution to the infinite-horizon problem. This is true both for the parametric example in Exercise 2.2 and for the more generally posed problem. But proving it involves establishing the legitimacy of interchanging the operators "max" and "$\lim_{T\to\infty}$"; and doing this is more challenging than one might guess.

Instead we will pursue a different approach. Equation (8) suggests another conjecture: that for the infinite-horizon problem in (3)–(4), for any U and f, the solution takes the form

$$(9) \qquad k_{t+1} = g(k_t), \quad t = 0, 1, \ldots,$$

where $g\colon \mathbf{R}_+ \to \mathbf{R}$ is a fixed savings function. Our intuition suggests that this must be so: since the planning problem takes the same form every period, with only the beginning-of-period capital stock changing from one period to the next, what else but k_t could influence the choice of k_{t+1} and c_t? Unfortunately, Exercise 2.2 does not offer any help in pursuing this conjecture. The change of variable exploited there is obviously specific to the particular functional forms assumed, and a glance at (5) confirms that no similar method is generally applicable.

The strategy we *will* use to pursue this idea involves ignoring (5) and (6) altogether and starting afresh. Although we stated this problem as one of choosing infinite sequences $\{(c_t, k_{t+1})\}_{t=0}^\infty$ for consumption and capital, the problem that in fact faces the planner in period $t = 0$ is that of choosing today's consumption, c_0, and tomorrow's beginning-of-period capital, k_1, and nothing else. The rest can wait until tomorrow. If we knew the planner's preferences over these two goods, we could simply maximize the appropriate function of (c_0, k_1) over the opportunity set defined by (1b), given k_0. But what are the planner's preferences over current consumption and next period's capital?

Suppose that (3)–(4) had already been solved for all possible values of k_0. Then we could define a function $v\colon \mathbf{R}_+ \to \mathbf{R}$ by taking $v(k_0)$ to be the value of the maximized objective function (3), for each $k_0 \geq 0$. A function of this sort is called a *value function.* With v so defined, $v(k_1)$ would give the value of the utility from period 1 on that could be obtained with a beginning-of-period capital stock k_1, and $\beta v(k_1)$ would be the value of this utility discounted back to period 0. Then in terms of this value function v, the planner's problem in period 0 would be

$$(10) \qquad \max_{c_0, k_1} [U(c_0) + \beta v(k_1)]$$

$$\text{s.t.} \quad c_0 + k_1 \leq f(k_0),$$

$$c_0, k_1 \geq 0, \quad k_0 > 0 \quad \text{given.}$$

If the function v were known, we could use (10) to define a function $g\colon \mathbf{R}_+ \to \mathbf{R}_+$ as follows: for each $k_0 \geq 0$, let $k_1 = g(k_0)$ and $c_0 = f(k_0) - g(k_0)$ be the values that attain the maximum in (10). With g so defined, (9) would completely describe the dynamics of capital accumulation from any given initial stock k_0.

We do not at this point "know" v, but we have defined it as the maximized objective function for the problem in (3)–(4). Thus, if solving (10) provides the solution for that problem, then $v(k_0)$ must be the maximized objective function for (10) as well. That is, v must satisfy

$$v(k_0) = \max_{0 \le k_1 \le f(k_0)} \{U[f(k_0) - k_1] + \beta v(k_1)\},$$

where, as before, we have used the fact that goods will not be wasted.

Notice that when the problem is looked at in this recursive way, the time subscripts have become a nuisance: we do not care what the date is. We can rewrite the problem facing a planner with current capital stock k as

$$(11) \qquad v(k) = \max_{0 \le y \le f(k)} \{U[f(k) - y] + \beta v(y)\}.$$

This one equation in the unknown function v is called a *functional equation*, and we will see later that it is a very tractable mathematical object. The study of dynamic optimization problems through the analysis of such functional equations is called *dynamic programming*.

If we knew that the function v was differentiable and that the maximizing value of y—call it $g(k)$—was interior, then the first-order and envelope conditions for (11) would be

$$U'[f(k) - g(k)] = \beta v'[g(k)], \quad \text{and}$$

$$v'(k) = f'(k)U'[f(k) - g(k)],$$

respectively. The first of these conditions equates the marginal utility of consuming current output to the marginal utility of allocating it to capital and enjoying augmented consumption next period. The second condition states that the marginal value of current capital, in terms of total discounted utility, is given by the marginal utility of using the capital in current production and allocating its return to current consumption.

Exercise 2.3 We conjectured that the path for capital given by (8) was optimal for the infinite-horizon planning problem, for the functional forms of Exercise 2.2.

a. Use this conjecture to calculate v by evaluating (2) along the consumption path associated with the path for capital given by (8).

b. Verify that this function v satisfies (11).

c. Is this a proof that (8) gives the optimal policy for this case? What would be needed to make it into one?

Suppose we have established the existence of an optimal savings policy g, either by analyzing conditions (5)–(6) or by analyzing the functional equation (11). What can we do with this information? For the particular parametric example in Exercises 2.2 and 2.3, we can solve for g with pencil-and-paper methods. We can then use the resulting difference equation (8) to compute the optimal sequence of capital stocks $\{k_t\}$. This example is a carefully chosen exception: for most other parametric examples, it is not possible to obtain an explicit analytical solution for the savings function g. In such cases a numerical approach can be used to compute explicit solutions. When all parameters are specified numerically, it is possible to use an algorithm based on (11) to obtain an approximation to g. Then $\{k_t\}$ can be computed using (9), given any initial value k_0.

In addition, there are often qualitative features of the savings function g, and hence of the capital paths generated by (9), that hold under a very wide range of assumptions on f and U. Specifically, we can use either (5)–(6) or the first-order and envelope conditions for (11), together with assumptions on U and f, to characterize the optimal savings function g. We can then, in turn, use the properties of g so established to characterize solutions $\{k_t\}$ to (9). The following exercise illustrates the second of these steps.

Exercise 2.4 a. Let f be as specified in Exercise 2.1, and suppose that the optimal savings function g is characterized by a constant savings rate, $g(k) = sf(k)$, all k, where $s > 0$. Plot g, and on the same diagram plot the 45° line. The points at which $g(k) = k$ are called the *stationary solutions, steady states, rest points,* or *fixed points* of g. Prove that there is exactly one positive stationary point k^*.

b. Use the diagram to show that if $k_0 > 0$, then the sequence $\{k_t\}$ given by (9) converges to k^* as $t \to \infty$. That is, let $\{k_t\}_{t=0}^{\infty}$ be a sequence satisfying (9), given some $k_0 \ge 0$. Prove that $\lim_{t\to\infty} k_t = k^*$, for any $k_0 > 0$. Show that this convergence is monotonic. Can it occur in a finite number of periods?

This exercise contains most of the information that can be established about the qualitative behavior of a sequence generated by a deterministic dynamic model. The stationary points have been located and characterized, their stability properties established, and the motion of the system has been described qualitatively for all possible initial positions. We take this example as a kind of image of what one might hope to establish for more complicated models, or as a source of reasonable conjectures. (Information about the rate of convergence to the steady state k^*, for k_t near k^*, can be obtained by taking a linear approximation to g in a neighborhood of k^*. Alternatively, numerical simulations can be used to study the rate of convergence over any range of interest.)

From the discussion above, we conclude that a fruitful way of analyzing a stationary, infinite-horizon optimization problem like the one in (3)–(4) is by examining the associated functional equation (11) for this example—and the difference equation (9) involving the associated policy function. Several steps are involved in carrying out this analysis.

First we need to be sure that the solution(s) to a problem posed in terms of infinite sequences are also the solution(s) to the related functional equation. That is, we need to show that by using the functional equation we have not changed the problem. Then we must develop tools for studying equations like (11). We must establish the existence and uniqueness of a value function v satisfying the functional equation and, where possible, to develop qualitative properties of v. We also need to establish properties of the associated policy function g. Finally we must show how qualitative properties of g are translated into properties of the sequences generated by g.

Since a wide variety of problems from very different substantive areas of economics all have this same mathematical structure, we want to develop these results in a way that is widely applicable. Doing this is the task of Part II.

2.2 A Stochastic Model of Optimal Growth

The deterministic model of optimal growth discussed above has a variety of stochastic counterparts, corresponding to different assumptions about the nature of the uncertainty. In this section we consider a model in which the uncertainty affects the technology only, and does so in a specific way.

Assume that output is given by $y_t = z_t f(k_t)$ where $\{z_t\}$ is a sequence of

independently and identically distributed (i.i.d.) random variables, and f is defined as it was in the last section. The shocks may be thought of as arising from crop failures, technological breakthroughs, and so on. The feasibility constraints for the economy are then

$$(1) \qquad k_{t+1} + c_t \leq z_t f(k_t), \quad c_t, k_{t+1} \geq 0, \quad \text{all } t, \text{ all } \{z_t\}.$$

Assume that the households in this economy rank stochastic consumption sequences according to the expected utility they deliver, where their underlying (common) utility function takes the same additively separable form as before:

$$(2) \qquad \mathrm{E}[u(c_0, c_1, \ldots)] = \mathrm{E}\left[\sum_{t=0}^{\infty} \beta^t U(c_t)\right].$$

Here $\mathrm{E}(\cdot)$ denotes expected value with respect to the probability distribution of the random variables $\{c_t\}_{t=0}^{\infty}$.

Now consider the problem facing a benevolent social planner in this stochastic environment. As before, his objective is to maximize the objective function in (2) subject to the constraints in (1). Before proceeding, we need to be clear about the timing of information, actions, and decisions, about the objects of choice for the planner, and about the distribution of the random variables $\{c_t\}_{t=0}^{\infty}$.

Assume that the timing of information and actions in each period is as follows. At the beginning of period t the current value z_t of the exogenous shock is realized. Thus, the pair (k_t, z_t), and hence the value of total output $z_t f(k_t)$, are known when consumption c_t takes place and end-of-period capital k_{t+1} is accumulated. The pair (k_t, z_t) is called the *state* of the economy at date t.

As we did in the deterministic case, we can think of the planner in period 0 as choosing, in addition to the pair (c_0, k_1), an infinite sequence $\{(c_t, k_{t+1})\}_{t=1}^{\infty}$ describing all future consumption and capital pairs. In the stochastic case, however, this is not a sequence of numbers but a sequence of *contingency plans*, one for each period. Specifically, consumption c_t and end-of-period capital k_{t+1} in each period $t = 1, 2, \ldots$ are contingent on the realizations of the shocks $z_1, z_2, \ldots, z_t$. This sequence of realizations is information that is available when the decision is being carried out but is unknown in period 0 when the decision is being made.

Technically, then, the planner chooses among sequences of functions, where the tth function in the sequence has as its arguments the history $(z_1, \ldots, z_t)$ of shocks realized between the time the plan is drawn up and the time the decision is carried out. The feasible set for the planner is the set of pairs (c_0, k_1) and sequences of functions $\{[c_t(\cdot), k_{t+1}(\cdot)]\}_{t=1}^{\infty}$ that satisfy (1) for all periods and all realizations of the shocks.

For any element of this set of feasible contingency plans, the exogenously given probability distribution of the shocks determines the distribution of future consumptions, so the expectation in (2) is well defined. The next exercise indicates the issues involved when one views the problem directly as one of choosing a sequence of contingency plans.

Exercise 2.5 Consider the finite-horizon version of the planning problem, with the objective function in (2), the constraints in (1), and the horizon T. Assume that the shocks $\{z_t\}_{t=0}^{T}$ take on only the finite list of values $a_1, \ldots, a_n$; and assume that the probabilities of these outcomes are $\pi_1, \ldots, \pi_n$ respectively in each period. State the first-order conditions for this problem. (This is mainly bookkeeping, but working out the details is instructive. Begin by making a list of *all* decision variables. In what Euclidean space does the planner's feasible set lie?)

This is one way of setting out the problem of optimal growth under uncertainty. There is another way, the analogue of the recursive formulation for the deterministic case. Here we let $v(k, z)$ be the value of the maximized objective (2) when the initial state is (k, z). Then a choice (c, y) of current consumption c and end-of-period capital y yields current utility $U(c)$ and implies that the system next period will be in the state (y, z'), where z' will be chosen by "nature" according to the fixed distribution governing the exogenous shocks. The maximum expected utility that can be obtained from this position is $v(y, z')$; so its discounted value as viewed in the current period, with z' unknown, is $\beta E[v(y, z')]$. These considerations motivate the functional equation

$$(3) \qquad v(k, z) = \max_{0 \le y \le f(k)} \{U[zf(k) - y] + \beta E[v(y, z')]\}.$$

The study of (3) yields the optimal choice of capital $y^* = g(k, z)$ as a function of the state (k, z) at the time the decision is taken. From this recursive point of view, then, the stochastic optimal growth problem is formally very similar to the deterministic one.

The methods used to characterize the optimal policy in the stochastic case are completely analogous to those used for the deterministic case. If we assume differentiability and an interior optimum, the first-order condition for (3) is

$$U'[zf(k) - g(k, z)] = \beta E\{v_1[g(k, z), z']\}.$$

This condition implicitly defines a policy function g that has as its arguments the two state variables k and z. Then the optimal capital path is given by the stochastic difference equation

$$(4) \qquad k_{t+1} = g(k_t, z_t),$$

where $\{z_t\}$ is an i.i.d. sequence of random shocks. The following exercise looks at (3)–(4) for the special case of log utility and Cobb-Douglas technology studied in the last section.

Exercise 2.6 Let $U(c) = \ln(c)$ and $f(k) = k^{\alpha}$, $0 < \alpha < 1$, as we did in Exercises 2.2–2.4. Conjecture that an optimal policy is, as before,

$$(5) \qquad k_{t+1} = \alpha\beta z_t k_t^{\alpha}, \quad \text{all } t, \text{ all } \{z_t\}.$$

Calculate the value of the objective function (2) under this policy, given $k_0 = k$ and $z_0 = z$, and call this value $v(k, z)$. Verify that the function v so defined satisfies (3).

Working out the dynamics of the state variable k_t that are implied by the policy function g is quite different in the stochastic case. Equation (4) and its specialization (5) are called (first-order) *stochastic difference equations,* and the random variables $\{k_t\}$ generated by such equations are called a (first-order) *Markov process.* It is useful to recall the results obtained for the deterministic difference equation in Exercise 2.4 and to think about possible analogues for the stochastic case. Clearly, the sequence $\{k_t\}$ described by (5) is not going to converge to any single value in the presence of the recurring shocks z_t. Can anything be said about its behavior?

Taking logs in (5), we obtain

$$\ln(k_{t+1}) = \ln(\alpha\beta) + \alpha \ln(k_t) + \ln(z_t).$$

Since the shocks $\{z_t\}$ are i.i.d. random variables, so are the logs $\{\ln(z_t)\}$. Now suppose that the latter are normally distributed, with common mean μ and variance σ^2.

Exercise 2.7 Given k_0, $\{\ln(k_t)\}_{t=1}^{\infty}$ is a sequence of normally distributed random variables with means $\{\mu_t\}_{t=1}^{\infty}$ and variances $\{\sigma_t^2\}_{t=1}^{\infty}$. Find these means and variances and calculate their limiting values as $t \to \infty$.

In this example, then, the sequence of probability distributions for the random variables $\{k_t\}$ converges as t increases without bound. Moreover, the combination of linearity and normality permits explicit pencil-and-paper calculation of the distributions of all the k_t's. This type of calculation is not possible in general, but convergence of the sequence of distribution functions for the k_t's to a limiting distribution can be verified under much broader assumptions. The basic idea is as follows.

Let the sequence $\{k_t\}$ be described by (5) but drop the assumption that the z_t's are log-normally distributed. Instead let G be the (common) cumulative distribution function for the z_t's. Then given the initial capital stock $k_0 > 0$, next period's stock k_1 is a random variable whose cumulative distribution function—call it ψ_1—is determined by G. In particular, for any $a > 0$,

$$\begin{aligned}\psi_1(a) &= \Pr\{k_1 \leq a\} = \Pr\{\alpha\beta z_0 k_0^{\alpha} \leq a\}\\ &= \Pr\{z_0 \leq a/\alpha\beta k_0^{\alpha}\} = G(a/\alpha\beta k_0^{\alpha}).\end{aligned}$$

Thus k_0 and G determine the distribution function ψ_1 of k_1.

Since the same logic holds for any successive pair of periods, we can define the function

$$H(a, b) = \Pr\{k_{t+1} \leq a | k_t = b\} = G(a/\alpha\beta b^{\alpha}), \quad \text{all } a, b > 0. \tag{6}$$

H is called a *transition function*. With H so defined, the sequence of distribution functions $\{\psi_t\}_{t=1}^{\infty}$ for the k_t's is given inductively by

$$\psi_{t+1}(a) = \Pr\{k_{t+1} \leq a\} = \int H(a, b)d\psi_t(b), \quad t = 0, 1, \ldots, \tag{7}$$

where the distribution ψ_0 is simply a mass point at the given initial value k_0.

More generally, given a stochastic difference equation of the form in (4) and given a distribution function G for the exogenous shocks, we can define a transition function H as we did in (6). Then for any initial value $k_0 > 0$, the sequence $\{\psi_t\}$ of distribution functions for the k_t's is given by (7). Exercise 2.7 suggests that if g and G are in some suitable families, then H is such that this sequence converges (in some sense) to a limiting distribution function ψ satisfying

$$\psi(k') = \int H(k', k)\, d\psi(k). \tag{8}$$

A distribution function ψ satisfying (8) is called an *invariant distribution* for the transition function H. The idea is that if the distribution ψ gives a probabilistic description of the capital stock k_t in any period t, then ψ also describes the distribution of the capital stock in periods $t + 1, t + 2, \ldots$. An invariant distribution is thus a stochastic analogue to a stationary point of a deterministic system.

Now suppose that g and G are given and that the associated transition function H has a unique stationary distribution ψ. Suppose further that for any $k_0 > 0$, the sequence $\{\psi_t\}$ defined by (7) converges to ψ. Let ϕ be a continuous function and consider the sample average $(1/T)\Sigma_{t=1}^{T}\phi(k_t)$ for this function, along some sample path. One might expect that this sample average is, for long time horizons, approximately equal to the mathematical expectation of ϕ taken with respect to the limiting distribution ψ. That is, one might expect that

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \phi(k_t) = \int \phi(k)d\psi(k), \tag{9}$$

at least along most sample paths. A statement of this sort is called a *law of large numbers*. Later we will specify precisely what is meant by "most sample paths" and will develop conditions under which (9) holds.

When (9) does hold, we can calculate the sample average on the left in (9) from observed time series, calculate the integral on the right in (9) from the theory, and use a comparison of the two as a test of the theory. The first calculation is easy. Much of this book is concerned with methods for carrying out the second.

As the discussion above suggests, the techniques of dynamic programming are, if anything, even more useful for analyzing stochastic models than they are for looking at deterministic problems. Exercise 2.5 illustrates the complexity of looking at stochastic, dynamic problems in terms of sequences, even when the horizon is finite. On the other hand, we will see later that functional equations like (3) are no more difficult to handle than their deterministic counterparts. The main ingredients are a convenient language for talking about distributions for stochastic shocks and a few basic results about expectation operators like the one in (3).

The solution to a functional equation like (3) involves an optimal policy function g like the one in (4), and hence we are interested in studying the properties of time series produced by systems like (4). This analysis is significantly harder than the analysis of solutions to deterministic difference equations, but it is not unmanageable. Clearly a stability theory for stochastic systems requires several things. First we must define precisely what convergence means for a sequence of distribution functions. Then we need to develop sufficient conditions on transition functions, like the function H above, to ensure that H has a unique invariant distribution and that the sequence of distribution functions given by (7) converges, in the desired sense, to that invariant distribution. Finally, to connect the theory to observed behavior, we must develop conditions under which a law of large numbers holds.

The reader should not be surprised that carrying out this agenda requires laying some preliminary groundwork. Some definitions, notation, and basic results from modern probability theory are needed, as well as some basic information about Markov processes. This preliminary material, as well as the analysis of stochastic recursive models, is the content of Part III.

2.3 Competitive Equilibrium Growth

In the last two sections we were concerned exclusively with the allocation problem faced by a hypothetical social planner. In this section we show that the solutions to planning problems of this type can, under appropriate conditions, be interpreted as predictions about the behavior of market economies. The argument establishing this is based, of course, on the classical connection between competitive equilibria and Pareto optima. These connections hold under fairly broad assumptions, and in later chapters we will establish them in a very general setting. At that time we will also show that in situations where the connection between competitive equilibria and Pareto optima breaks down, as it does in the presence of taxes or other distortions, the study of competitive equilibria can be carried out by a direct analysis of the appropriate first-order conditions.

Recall that in the models discussed above there were many identical households, and we took the (common) preferences of these households to be the preferences attributed to the social planner. In addition, there were many identical firms, all with the same constant-returns-to-scale technology, so the technology available to the economy was the same as that available to each firm. Thus, the planning problems considered in Sections 2.1 and 2.2 were Pareto problems for economies with many agents. That is, they can be viewed as problems of maximizing a weighted average of households' utilities, specialized to a case where all households had identical tastes and were given equal weight, and hence received identical allocations. Thus the solutions to planning problems of the type we considered were Pareto-optimal allocations. In this section we show that these allocations are exactly the ones that correspond to competitive equilibria. For simplicity we restrict attention here to the case of certainty and of a finite time horizon.

Suppose that we have solved the finite-horizon optimal growth problem of Section 2.1 and that $\{(c_t^*, k_{t+1}^*)\}_{t=0}^T$ is the solution. Our goal is to find prices that support these quantities as a competitive equilibrium. However, we must first specify the ownership rights of households and firms, as well as the structure of markets. It is crucial to be specific on these matters.

Assume that households own all factors of production and all shares in firms and that these endowments are equally distributed across households. Each period households sell factor services to firms and buy the goods produced by firms, consuming some and accumulating the rest as capital. Assume that firms own nothing; they simply hire capital and labor on a rental basis to produce output each period, sell the output produced back to households, and return any profits that result to shareholders. Finally, assume that all transactions take place in a single once-and-for-all market that meets in period 0. All trading takes place at that time, so all prices and quantities are determined simultaneously. No further trades are negotiated later. After this market has closed, in periods $t = 0, 1, \ldots, T$, agents simply deliver the quantities of factors and goods they have contracted to sell and receive those they have contracted to buy.

Assume that the convention for prices in this one big market is as follows. Let p_t be the price of a unit of output delivered in period t, for $t = 0, 1, \ldots, T$, expressed in abstract units-of-account. Let w_t be the price of a unit of labor delivered in period t, expressed in units of goods in period t, so that w_t is the real wage. Similarly let r_t be the real rental price of capital in period t.

Given the prices $\{(p_t, r_t, w_t)\}_{t=0}^{T}$, the problem faced by the representative firm is to choose input demands and output supplies $\{(k_t, n_t, y_t)\}_{t=0}^{T}$ that maximize net discounted profits. Thus its decision problem is

$$\max \pi = \sum_{t=0}^{T} p_t[y_t - r_t k_t - w_t n_t] \tag{1}$$

$$\text{s.t.} \quad y_t \leq F(k_t, n_t), \quad t = 0, 1, \ldots, T. \tag{2}$$

Given the same price sequence, the typical household must choose demand for consumption and investment, and supplies of current capital and labor, $\{(c_t, i_t, x_{t+1}, k_t, n_t)\}_{t=0}^{T}$, given initial capital holdings x_0. In making these choices the household faces several constraints. First, the total value of goods purchased cannot exceed the total value of wages plus rental income plus profits the household receives. Second, the household's holdings of real capital in each period $t + 1$ are equal to its holdings in period t, net of depreciation, plus any new investment. Third, the quantity of each factor supplied by the household in each period must be nonnegative but cannot exceed the quantity available to it in that period. Finally, consumption and capital holdings must be nonnegative. Thus its decision problem is

$$\max \sum_{t=0}^{T} \beta^t U(c_t) \tag{3}$$

$$\text{s.t.} \quad \sum_{t=0}^{T} p_t[c_t + i_t] \leq \sum_{t=0}^{T} p_t[r_t k_t + w_t n_t] + \pi; \tag{4}$$

$$x_{t+1} = (1 - \delta)x_t + i_t, \quad t = 0, 1, \ldots, T, \text{ given } x_0; \tag{5}$$

$$0 \leq n_t \leq 1, \; 0 \leq k_t \leq x_t, \quad t = 0, 1, \ldots, T; \tag{6}$$

$$c_t \geq 0, \; x_{t+1} \geq 0, \quad t = 0, 1, \ldots, T. \tag{7}$$

Note that capital stocks owned, x_{t+1}, and capital supplied to firms, k_{t+1} are required to be nonnegative. However, gross investment, i_t, may be negative. This assumption is the one that was made, implicitly, in Section 2.1.

A competitive equilibrium is a set of prices $\{(p_t, r_t, w_t)\}_{t=0}^{T}$, an allocation $\{(k_t^d, n_t^d, y_t)\}_{t=0}^{T}$ for the typical firm, and an allocation $\{(c_t, i_t, x_{t+1}, k_t^s, n_t^s)\}_{t=0}^{T}$ for the typical household, such that

a. $\{(k_t^d, n_t^d, y_t)\}$ solves (1)–(2) at the stated prices;
b. $\{(c_t, i_t, x_{t+1}, k_t^s, n_t^s)\}$ solves (3)–(7) at the stated prices;
c. all markets clear: $k_t^d = k_t^s$, $n_t^d = n_t^s$, $c_t + i_t = y_t$, all t.

To find a competitive equilibrium, we begin by conjecturing that it has certain features. Later we will verify that these conjectures are correct. First, since the representative household's preferences are strictly monotone, we conjecture that goods prices are strictly positive for each period: $p_t > 0$, all t. Also, since both factors have strictly positive marginal products, we conjecture that both factor prices are strictly positive for all periods: $w_t > 0$ and $r_t > 0$, all t. Finally, since in equilibrium markets clear, we let $k_t = k_t^s = k_t^d$ and $n_t = n_t^s = n_t^d$, all t, denote the quantities of capital and labor traded.

Now consider the typical firm. If the price of goods is strictly positive in each period, then the firm supplies to the market all of the output that it produces each period. That is, (2) holds with equality, for all t. Also, note that since the firm simply rents capital and hires labor for each period, its problem is equivalent to a series of one-period maximization problems. Hence its input demands solve

$$\max_{k_t, n_t} p_t[F(k_t, n_t) - r_t k_t - w_t n_t], \quad t = 0, 1, \ldots, T. \tag{8}$$

It then follows that (real) factor prices must be equal to marginal products:

$$r_t = F_k(k_t, n_t), \quad t = 0, 1, \ldots, T; \tag{9}$$

$$w_t = F_n(k_t, n_t), \quad t = 0, 1, \ldots, T. \tag{10}$$

Since F is homogeneous of degree one, when we substitute from (9) and (10) into (8), we find that $\pi = 0$. Note, too, that $k_{T+1} = 0$.

Next consider the typical household. Since supplying available factors causes no disutility to the household, in every period it supplies all that is available. That is, $n_t = 1$ and $k_t = x_t$, all t. Using these facts and substituting from (5) to eliminate i_t, we can write the household's problem as

$$\text{(11)} \qquad \max \sum_{t=0}^{T} \beta^t U(c_t)$$

$$\text{(12)} \qquad \text{s.t.} \sum_{t=0}^{T} p_t[c_t + k_{t+1} - (r_t + 1 - \delta)k_t - w_t] \leq 0,$$

$$\text{(13)} \qquad c_t \geq 0,\ k_{t+1} \geq 0, \quad t = 0, 1, \ldots, T;$$

given $k_0 = x_0$.

Since $\lim_{c \to 0} U'(c) = \infty$, the nonnegativity constraints on the c_t's in (13) are never binding. Hence the first-order conditions for the household are

$$\text{(14)} \qquad \beta^t U'(c_t) - \lambda p_t = 0,$$

$$\text{(15)} \qquad \lambda[(r_{t+1} + 1 - \delta)p_{t+1} - p_t] \leq 0,$$

with equality if $k_{t+1} > 0$, $t = 0, 1, \ldots, T$;

where λ is the multiplier associated with the budget constraint (12).

Therefore a competitive equilibrium is characterized by quantities and prices $\{(c_t^e, k_{t+1}^e, p_t^e, r_t^e, w_t^e)\}_{t=0}^T$, with all goods and factor prices strictly positive, such that $\{(k_t^e, n_t = 1)\}_{t=0}^T$ solves (8) at the given prices, $\{(c_t^e, k_{t+1}^e)\}_{t=0}^T$ solves (11)–(13) at the given prices, $k_0 = x_0$, $k_{T+1} = 0$, and in addition

$$\text{(16)} \qquad F(k_t^e, 1) = c_t^e + k_{t+1}^e - (1 - \delta)k_t^e, \quad \text{all } t.$$

Now that we have defined and partially characterized a competitive equilibrium for the economy of Section 2.1, we can be more specific about the connections between equilibrium and optimal allocations that we referred to earlier. First note that if $\{(c_t^e, k_{t+1}^e, p_t^e, w_t^e, r_t^e)\}_{t=0}^T$ is an equilibrium, then $\{(c_t^e, k_{t+1}^e)\}_{t=0}^T$ is a solution to the planning problem discussed in Section 2.1. To prove this we need only show that $\{(c_t^e, k_{t+1}^e)\}$ is Pareto optimal. Suppose to the contrary that $\{(c_t', k_{t+1}')\}$ is a feasible allocation and that $\{c_t'\}$ yields higher total utility in the objective function (11). Then this allocation must violate (12), or the household would have chosen it. But if (12) is violated, then (16) implies that

$$\pi' = \sum_{t=0}^{T} p_t^e[F(k_t', 1) - r_t^e k_t' - w_t^e] > 0 = \pi^e,$$

contradicting the hypothesis that $\{(k_t^e, n_t = 1)\}_{t=0}^T$ was a profit-maximizing choice of inputs. This result is a version of the first fundamental theorem of welfare economics.

Conversely, suppose that $\{(c_t^*, k_{t+1}^*)\}_{t=0}^T$ is a solution to the planner's problem in Section 2.1. Then $\{k_{t+1}^*\}_{t=0}^T$ is the unique sequence satisfying the first-order and boundary conditions

$$\text{(17)} \qquad \beta f'(k_t^*)U'[f(k_t^*) - k_{t+1}^*] = U'[f(k_{t-1}^*) - k_t^*], \quad t = 1, 2, \ldots, T;$$

$$\text{(18)} \qquad k_{T+1}^* = 0,\ k_0^* = x_0;$$

and $\{c_t^*\}$ is given by

$$\text{(19)} \qquad c_t^* = f(k_t^*) - k_{t+1}^*, \quad t = 0, 1, \ldots, T;$$

where the function $f(k) = F(k, 1) + (1 - \delta)k$ is as defined in Section 2.1. To construct a competitive equilibrium with these quantities, we must find supporting prices $\{(p_t^*, r_t^*, w_t^*)\}_{t=0}^T$.

To do this, note that (9) and (15) together suggest that goods prices must satisfy

$$\text{(20)} \qquad p_t^* = p_{t-1}^*/f'(k_t^*), \quad t = 1, 2, \ldots, T;$$

where $p_0 > 0$ is arbitrary, and (9) and (10) imply that real wage and rental rates must satisfy

$$\text{(21)} \qquad r_t^* = f'(k_t^*) - (1 - \delta), \quad t = 1, 2, \ldots, T;$$

$$\text{(22)} \qquad w_t^* = f(k_t^*) - k_t^* f'(k_t^*), \quad t = 1, 2, \ldots, T.$$

It is not difficult to verify that these prices together with the quantities in (17)–(19) constitute a competitive equilibrium, and we leave the proof as an exercise. This result is a version of the second fundamental theorem of welfare economics.

Exercise 2.8 Show that at the prices given in (20)–(22), the allocation $\{(c_t^*, k_{t+1}^*)\}_{t=0}^T$ defined in (17)–(19) is utility maximizing for the household [solves (11)–(13)]; that the allocation $\{(k_t^*, n_t^* = 1)\}_{t=0}^T$ is profit maximizing for the firm [solves (8)]; and that $\{(c_t^*, k_{t+1}^*)\}_{t=0}^T$ satisfies (16).

We also leave it as an exercise to show that the same quantities and prices constitute a competitive equilibrium if firms instead of households are the owners of capital.

Exercise 2.9 Suppose that households are prohibited from owning capital directly. Instead, firms own all of the initial capital stock k_0 and also make all future investments in capital. Households own all shares in firms, and returns to the latter now include returns to capital. Modify the statements of the firm's and the household's problems to fit these arrangements and show that the quantities in (17)–(19) together with the prices in (20)–(22) still constitute a competitive equilibrium.

We have interpreted these equilibrium prices and quantities as being determined in a single market-clearing operation. But there is another way to think of an economy as arriving at the quantities and prices calculated above. Suppose that the agents meet in a market at the beginning of every period, not just in period 0. In the market held in period t, agents trade current-period labor, rental services of existing capital, and final output. In addition, one security is traded: a claim to one unit of final output in the subsequent period. In each period, factor and bond prices are expressed in terms of current-period goods.

Notice that with a sequence of markets the household must form expectations about future prices in order to arrive at its decisions in the market in period t. In particular, its expectations about future consumption goods prices and future rental rates on capital affect its current consumption-savings decision. Thus some assumption is needed about how these expectations are formed. Suppose, for example, that the household has perfect foresight about all future prices. (This assumption is the specialization for a deterministic context of the more general notion of rational expectations.) Although we do not carry out the proof here, it is not hard to show that, under the assumption of perfect foresight, this set of markets is equivalent to the one above in the sense that the competitive equilibrium allocation is the same for the two settings, and the prices are closely related.

Exercise 2.10 Suppose that the market structure is as described above. Modify the statements of the firm's and the household's problems to fit these arrangements. Show that under perfect foresight the quantities in (17)–(19), the factor prices in (21)–(22), and the bond prices

$$q_t = \beta U'(c_{t+1}^*)/U'(c_t^*) = 1/f'(k_{t+1}^*), \quad t = 0, 1, \ldots, T-1;$$

constitute a competitive equilibrium.

(In fact, for the representative household economy here, the sequential market structure can be even further simplified by eliminating securities markets. Since the net supply of such securities is zero, in equilibrium each household has a net demand of zero for each of the securities. Hence, if these markets are simply shut down, the remaining prices and the real allocation are unaltered. This conclusion does *not* hold, however, in an economy with heterogeneous households.)

We have, then, two examples of market economies: one with complete markets in the Arrow-Debreu sense, the other with markets limited to spot transactions in factors of production, goods, and one-period securities. Both economies reproduce the optimal path of capital accumulation discussed in Section 2.1, provided agents in the sequence economy have perfect foresight about future prices.

There is yet a third way in which the solution to the optimal growth model of Section 2.1 can be interpreted as a competitive equilibrium, one that is closely related to the dynamic programming approach of Section 2.1 and to the sequence of markets interpretation of equilibrium above. The general idea is to characterize equilibrium prices as functions of the single state variable k, the economy-wide capital stock, and to view individual households as dynamic programmers.

To develop this idea, first note that since firms in the sequence economy solve the sequence of one-period problems in (8), factor prices can be expressed as functions of the state:

$$(23) \qquad R(k) = F_k(k, 1) \quad \text{and} \quad \omega(k) = F_n(k, 1), \quad \text{all } k > 0.$$

At these prices, which in each period are taken by the firms as fixed numbers, the quantities $(k, 1)$ are profit maximizing and lead to zero profits.

Next we must develop a dynamic program representing the decision problem faced by a typical household. To do so we need a notation that distinguishes between the economy-wide capital stock, k, over which the household has no control, and its own capital stock, K, over which it has complete control. Looking ahead, we know that in equilibrium it must be the case that $K = k$, but it is important to keep in mind that the representative household does not behave as it does out of a sense of social responsibility: it must be induced by prices to do so.

Let the individual household's state variable be the pair (K, k), let $V(K, k)$ denote its optimum value function, and suppose that it expects the economy-wide capital stock next period to be $k' = h(k)$. Then the household's decision problem is represented by the functional equation

$$(24) \qquad V(K, k) = \max_{C, Y} \{U(C) + \beta V[Y, h(k)]\}$$

$$\text{s.t.} \quad C + [Y - (1 - \delta)K] \leq KR(k) + \omega(k).$$

Let $H(K, k)$ be the optimal policy function for this problem.

Under the assumption of perfect foresight, the fact that all households are identical implies that in equilibrium it must be the case that $h(k) = H(k, k)$. Thus we define a recursive competitive equilibrium to be a value function $V: \mathbf{R}^2_+ \rightarrow \mathbf{R}$, a policy function $H: \mathbf{R}^2_+ \rightarrow \mathbf{R}_+$ for the representative household, an economy-wide law of motion $h: \mathbf{R}_+ \rightarrow \mathbf{R}_+$ for capital, and factor price functions $R: \mathbf{R}_+ \rightarrow \mathbf{R}_+$ and $\omega: \mathbf{R}_+ \rightarrow \mathbf{R}_+$, such that V satisfies (24), given h; H is the optimal policy function for (24); $H(k, k) = h(k)$, all k; and R and ω satisfy (23).

From this recursive point of view, the statement that competitive equilibrium allocations are Pareto optimal means that if (V, H, h, R, ω) is an equilibrium, then the function $v(k) = V(k, k)$ is the value function for the planner's problem, and $g = h$ is the planner's optimal policy function. Conversely, the fact that Pareto-optimal allocations can be supported as competitive equilibria means the following. If v is the value function for the planner's problem and g is the planner's optimal policy function, then the value and policy function for the individual household, the pair (V, H) satisfying (24), have the property that $V(k, k) = v(k)$ and $H(k, k) = g(k)$, all k.

The sense in which these statements are economically reasonable conjectures can be spelled out by a comparison of the first-order and envelope conditions for the two dynamic programs. For the planner's problem these are

$$(25) \qquad U'[f(k) - g(k)] = \beta v'[g(k)];$$

$$(26) \qquad v'(k) = U'[f(k) - g(k)]f'(k).$$

For the household's problem, the analogous conditions are

$$U'[f(k) + f'(k)(K - k) - H(K, k)] = \beta V_1[H(K, k), h(k)] \quad \text{and}$$

$$V_1(K, k) = U'[f(k) + f'(k)(K - k) - H(K, k)]f'(k).$$

In equilibrium, $K = k$; since $H(k, k) = h(k)$, these conditions can then be written as

$$(27) \qquad U'[f(k) - h(k)] = \beta V_1[h(k), h(k)] \quad \text{and}$$

$$(28) \qquad V_1(k, k) = U'[f(k) - h(k)]f'(k).$$

Thus, if $v'(k) = V_1(k, k)$, then the equilibrium conditions (27)–(28) match the conditions (25)–(26) for the planner's problem.

We will repeatedly exploit these classic connections between competitive equilibria and Pareto optima as a device for proving the existence of equilibria in market economies and for characterizing them. That is, we will solve planning problems, not for the normative purpose of prescribing outcomes, but for the positive purpose of predicting market outcomes from a given set of preferences and technology.

This device is useful in situations where the two fundamental theorems of welfare economics apply, but these theorems fail in the presence of increasing returns to scale, externalities, distorting taxes, and so on. Suppose that we want to consider the competitive equilibrium in an economy in which there is, say, a flat-rate tax on labor income. We know that in the presence of such a tax the competitive equilibrium allocation is not, in general, Pareto optimal. That is, in general we cannot describe the competitive equilibrium of such an economy by describing an associated social planning problem.

In such cases, establishing the existence and qualitative properties of a competitive equilibrium requires looking directly at the equilibrium conditions. In the recursive context above this approach means that we must establish the existence of functions (V, H, h) such that V and H are the value and policy functions for the household's dynamic programming problem, given the economy-wide law of motion h for the state variable; and $H(k, k) = h(k)$, all k. These considerations lead us to look directly at the analogues of (27)–(28). Establishing the existence of a competitive equilibrium involves establishing directly the existence of functions $\phi(k) = V_1(k, k)$ and $h(k)$ satisfying those equations. Given h, the functions V and H can then be found by solving (24).

In this section, we have illustrated that the methods for studying recursive optimization problems developed in Parts II and III can be used in two ways in the analysis of competitive equilibria. In situations where the connections between competitive equilibria and Pareto optima hold, equilibria can be studied by analyzing the associated Pareto problem. In situations where those connections fail, the methods of dynamic programming can still be used to study the problems facing individual agents in the economy. However, new arguments are needed to establish the existence of equilibria. These two approaches to the study of competitive equilibria are the subject of Part IV.

2.4 Conclusions and Plans

We began this chapter with a deterministic model of optimal growth and then explored a number of variations on it. In the course of the discussion, we have raised a wide variety of substantive and technical issues, passing over questions in both categories lightly with promises of better treatments to come. It is time to spell out these promises in more detail. We will do this by describing briefly the plan for the rest of the book.

Deterministic systems like those discussed in Section 2.1 are the subject of Part II (Chapters 3–6). Chapter 3 contains some preliminary mathematical material needed to study functional equations like those discussed earlier. This background allows us to develop the needed results for deterministic models and also lays the foundation for the study of stochastic problems.

Chapter 4 then deals with dynamic programming in a deterministic context. The optimal growth model of Section 2.1 is a typical example

and illustrates the necessary ingredients of such a treatment. First we must show that for stationary, infinite-horizon optimization problems like the social planner's problem, the problem stated as one of choosing an optimal sequence of decisions is equivalent (in some sense) to the problem stated in the form of a functional equation. With this established, we can then study functional equations for bounded, constant-returns-to-scale, and unbounded problems.

In Chapter 5 we turn to substantive economic models that are amenable to analysis using these tools. These applications, which are drawn from a variety of substantive areas of economics, are intended to give some idea of the broad applicability of these methods.

Chapter 6 treats methods for characterizing the behavior of deterministic, recursive systems over time: the theory of stability for autonomous difference equations. We first review results on global stability and then treat local stability. We conclude with several economic applications of these methods and with some examples that illustrate the types of behavior possible in unstable systems.

Stochastic systems, like those we saw in Section 2.2, are treated in Part III (Chapters 7–14). In generalizing the analysis of Chapters 4–6 to include stochastic shocks, a variety of approaches are possible. We have chosen to take a modern attack, one that allows us to deal with very general classes of stochastic shocks when looking at dynamic programming problems, and that yields additional benefits later when we study the stochastic counterpart of stability theory. To take this approach, we must first develop some of the basic tools of the theory of measure and integration.

This background is presented in Chapters 7 and 8. Chapter 7 is a self-contained treatment of the definitions and results from measure theory that are needed in later chapters; and Chapter 8 contains an introduction to Markov processes, the natural generalization of the stochastic difference equations discussed above.

With these mathematical preliminaries in place, Chapter 9 deals with stochastic dynamic programming, paralleling Chapter 4 as closely as possible. The rewards from Chapters 7 and 8 are apparent here (we hope!). With the appropriate notation and results in hand, the arguments used in Chapter 9 to study stochastic models are fairly simple extensions of those in Chapter 4.

Chapter 10 then provides a variety of economic applications, drawn from a number of different substantive areas. Some of these are stochastic analogues to models discussed in Chapter 5; others are entirely new.

Chapters 11 and 12 survey results on convergence, in various senses, for Markov processes: extensions of the ideas sketched in Section 2.2 to a much wider variety of problems. This material is the body of theory suited to characterizing the dynamics for state variables generated by optimal policy functions for stochastic dynamic programs. Substantive economic applications of these methods are discussed in Chapter 13. Some of these applications are continuations of those discussed in Chapter 10, others are new. Chapter 14 provides a law of large numbers for Markov processes.

The use of recursive systems within a general equilibrium framework, as illustrated in Section 2.3 above, is the subject of Part IV (Chapters 15–18). Chapter 15 returns at a more abstract level to the connections between Pareto-optimal and competitive equilibrium allocations. In particular, we there review the two fundamental theorems of welfare economics in a way that applies to the kinds of infinite dimensional commodity spaces that arise in dynamic applications. We also treat the issue of constructing prices for problems involving infinite time horizons and/or uncertainty. Chapter 16 then contains a number of applications, designed to illustrate how a variety of planning problems can be interpreted as market equilibria.

When a market equilibrium is also the solution to a benevolent social planner's problem, this fact vastly simplifies the analysis. However, there are many market situations of great interest—situations in which markets are subject to distortions due to taxes, external effects, or various kinds of market imperfections—that cannot be analyzed in this way. In many such cases it is still possible to construct recursive equilibria directly, using the line of argument discussed briefly in Section 2.3. Chapter 17 presents several mathematical results, fixed-point theorems, that have proved useful in such cases, and illustrates their application. In Chapter 18 we conclude with further illustrations of these methods.

2.5 Bibliographic Notes

Modern growth theory began with Frank Ramsey's (1928) classic paper and then lay dormant for almost 30 years. (Although a substantial body of literature on growth developed during the 1930s and 1940s, this work is quite different from the neoclassical theory of growth both in motivation and in terms of the specific models used: its goal was to show that high, persistent rates of unemployment are a necessary feature of long-run growth, and the models used generally featured fixed-proportions technologies.) The field was reawakened by the work of Solow (1956) and Swan (1956) and has been active ever since. The work by Solow and Swan, and much that followed immediately, relied on the assumption that households save a fixed proportion of their income. These models were meant to be descriptive rather than prescriptive, and no attempt was made to model households' preferences and expectations.

Households' preferences finally reentered the discussion when economists looked at the issue of growth from a normative point of view. The deterministic theory of optimal growth, of which the one-sector model discussed in Section 2.1 is the simplest case, was developed independently and simultaneously by Cass (1965) and Koopmans (1965). A stochastic model that incorporated shocks to production, like the one discussed in Section 2.2, was first studied by Brock and Mirman (1972) and by Mirman and Zilcha (1975).

The first modern treatment of the connections between Pareto optima and competitive equilibria was provided by Arrow (1951) for the case where the commodity space is a finite-dimensional Euclidean space. This treatment applies, for example, to the finite-horizon optimal growth problem discussed in Section 2.3. Debreu (1954) showed that the same line of argument holds in certain infinite-dimensional spaces, and his is the treatment that we will need later to deal with infinite-horizon models.

The interpretation of a competitive equilibrium in terms of a sequence of markets can also be made for stochastic models. To make this interpretation, it must be assumed that agents have *rational expectations* in the sense of Muth (1961). See Radner (1972) for a pioneering general equilibrium application of this idea.

PART II

Deterministic Models

3 Mathematical Preliminaries

In Chapter 2 the optimal growth problem

$$\max_{\{(c_t, k_{t+1})\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t U(c_t)$$

$$\text{s.t.} \quad c_t + k_{t+1} \leq f(k_t),$$

$$c_t, k_{t+1} \geq 0, \quad t = 0, 1, \ldots,$$

$$\text{given } k_0,$$

was seen to lead to the functional equation

$$(1) \qquad v(k) = \max_{c,y} [U(c) + \beta v(y)]$$

$$\text{s.t.} \quad c + y \leq f(k),$$

$$c, y \geq 0.$$

The purpose of this chapter and the next is to show precisely the relationship between these two problems and others like them and to develop the mathematical methods that have proved useful in studying the latter. In Section 2.1 we argued in an informal way that the solutions to the two problems should be closely connected, and this argument will be made rigorous later. In the rest of this introduction we consider alternative methods for finding solutions to (1), outline the one to be pursued, and describe the mathematical issues it raises. In the remaining sections of the chapter we deal with these issues in turn. We draw upon this material extensively in Chapter 4, where functional equations like (1) are analyzed.

In (1) the functions U and f are given—they take specific forms known to us—and the value function v is unknown. Our task is to prove the existence and uniqueness of a function v satisfying (1) and to deduce its properties, given those of U and f. The classical (nineteenth-century) approach to this problem was the *method of successive approximations,* and it works in the following very commonsensical way. Begin by taking an initial guess that a specific function, call it v_0, satisfies (1). Then define a new function, v_1, by

$$v_1(k) = \max_{0 \le y \le f(k)} \{U[f(k) - y] + \beta v_0(y)\}. \tag{2}$$

If it should happen that $v_1(k) = v_0(k)$, for all $k \ge 0$, then clearly v_0 is a solution to (1). Lucky guessing (cf. Exercise 2.3) is one way to establish the existence of a function satisfying (1), but it is notoriously unreliable. The method of successive approximations proceeds in a more systematic way.

Suppose, as is usually the case, that $v_1 \ne v_0$. Then use v_1 as a new guess and define the sequence of functions $\{v_n\}$ recursively by

$$v_{n+1}(k) = \max_{0 \le y \le f(k)} \{U[f(k) - y] + \beta v_n(y)\}, \quad n = 0, 1, 2, \ldots. \tag{3}$$

The hope behind this iterative process is that as n increases, the successive approximations v_n get closer to a function v that actually satisfies (1). That is, the hope is that the limit of the sequence $\{v_n\}$ is a solution v. Moreover, if it can be shown that $\lim_{n\to\infty} v_n$ is the same for any initial guess v_0, then it will follow that this limit is the only function satisfying (1). (Why?)

Is there any reason to hope for success in this analytical strategy? Recall that our reason for being interested in (1) is to use it to locate the optimal capital accumulation policy for a one-sector economy. Suppose we begin by choosing any feasible capital accumulation policy, that is, any function g_0 satisfying $0 \le g_0(k) \le f(k)$, all $k \ge 0$. [An example is the policy of saving a constant fraction of income: $g_0(k) = \theta f(k)$, where $0 < \theta < 1$.] The lifetime utility yielded by this policy, as a function of the initial capital stock k_0, is

$$w_0(k_0) = \sum_{t=0}^{\infty} \beta^t U[f(k_t) - g_0(k_t)],$$

where

$$k_{t+1} = g_0(k_t),\ t = 0, 1, 2, \ldots.$$

The following exercise develops a result about (g_0, w_0) that is used later.

Exercise 3.1 Show that

$$w_0(k) = U[f(k) - g_0(k)] + \beta w_0[g_0(k)], \quad \text{all } k \ge 0.$$

If the utility from the policy g_0 is used as the initial guess for a value function—that is, if $v_0 = w_0$—then (2) is the problem facing a planner who can choose capital accumulation optimally for one period but must follow the policy g_0 in all subsequent periods. Thus $v_1(k)$ is the level of lifetime utility attained, and the maximizing value of y—call it $g_1(k)$—is the optimal level for end-of-period capital. Both v_1 and g_1 are functions of beginning-of-period capital k.

Notice that since $g_0(k)$ is a feasible choice in the first period, the planner will do no worse than he would by following the policy g_0 from the beginning, and in general he will be able to do better. That is, for any feasible policy g_0 and associated initial value function v_0,

$$\begin{aligned} v_1(k) &= \max_{0 \le y \le f(k)} \{U[f(k) - y] + \beta v_0(y)\} \\ &\ge \{U[f(k) - g_0(k)] + \beta v_0[g_0(k)]\} \\ &= v_0(k), \end{aligned} \tag{4}$$

where the last line follows from Exercise 3.1.

Now suppose the planner has the option of choosing capital accumulation optimally for two periods but must follow the policy g_0 thereafter. If y is his choice for end-of-period capital in the first period, then from the second period on the best he can do is to choose $g_1(y)$ for end-of-period capital and enjoy total utility $v_1(y)$. His problem in the first period is thus $\max[U(c) + \beta v_1(y)]$, subject to the constraints in (1). The maximized value of this objective function was defined, in (3), as $v_2(k)$. Hence it follows from (4) that

$$\begin{aligned} v_2(k) &= \max_{0\le y\le f(k)} \{U[f(k) - y] + \beta v_1(y)\} \\ &\ge \max_{0\le y\le f(k)} \{U[f(k) - y] + \beta v_0(y)\} \\ &= v_1(k). \end{aligned}$$

Continuing in this way, one establishes by induction that $v_{n+1}(k) \ge v_n(k)$, all k, $n = 0, 1, 2, \ldots$. The successive approximations defined in (3) are improvements, reflecting the fact that planning flexibility over longer and longer finite horizons offers new options without taking any other options away. Consequently it seems reasonable to suppose that the sequence of functions $\{v_n\}$ defined in (3) might converge to a solution v to (1). That is, the method of successive approximations seems to be a reasonable way to locate and characterize solutions.

This method can be described in a somewhat different and much more convenient language. As we showed in the discussion above, for any function $w: \mathbf{R}_+ \to \mathbf{R}$, we can define a new function—call it $Tw: \mathbf{R}_+ \to \mathbf{R}$—by

$$(5) \qquad (Tw)(k) = \max_{0\le y\le f(k)} \{U[f(k) - y] + \beta w(y)\}.$$

When we use this notation, the method of successive approximations amounts to choosing a function v_0 and studying the sequence $\{v_n\}$ defined by $v_{n+1} = Tv_n$, $n = 0, 1, 2, \ldots$. The goal then is to show that this sequence converges and that the limit function v satisfies (1). Alternatively, we can simply view the operator T as a mapping from some set C of functions into itself: $T: C \to C$. In this notation solving (1) is equivalent to locating a *fixed point* of the mapping T, that is, a function $v \in C$ satisfying $v = Tv$, and the method of successive approximations is viewed as a way to construct this fixed point.

To study operators T like the one defined in (5), we need to draw on several basic mathematical results. To show that T maps an appropriate space C of functions into itself, we must decide what spaces of functions are suitable for carrying out our analysis. In general we want to limit attention to continuous functions. This choice raises the issue of whether, given a continuous function w, the function Tw defined by (5) is also continuous. Finally, we need a fixed-point theorem that applies to operators like T on the space C we have selected. The rest of the chapter deals with these issues.

In Section 3.1 we review the basic facts about metric spaces and normed vector spaces and define the space C that will be used repeatedly later. In Section 3.2 we prove the Contraction Mapping Theorem, a fixed-point theorem of vast usefulness. In Section 3.3 we review the main facts we will need about functions, like Tw above, that are defined by maximization problems.

3.1 Metric Spaces and Normed Vector Spaces

The preceding section motivates the study of certain functional equations as a means of finding solutions to problems posed in terms of infinite sequences. To pursue the study of these problems, as we will in Chapter 4, we need to talk about infinite sequences $\{x_t\}_{t=0}^{\infty}$ of states, about candidates for the value function v, and about the convergence of sequences of various sorts. To do this, we will find it convenient to think of both infinite sequences and certain classes of functions as elements of infinite-dimensional normed vector spaces. Accordingly, we begin here with the definitions of vector spaces, metric spaces, and normed vector spaces. We then discuss the notions of convergence and Cauchy convergence, and define the notion of completeness for a metric space. Theorem 3.1 then establishes that the space of bounded, continuous, real-valued functions on a set $X \subseteq \mathbf{R}^l$ is complete.

We begin with the definition of a vector space.

DEFINITION *A (real)* ***vector space*** *X is a set of elements (vectors) together with two operations, addition and scalar multiplication. For any two vectors $x, y \in X$, addition gives a vector $x + y \in X$; and for any vector $x \in X$ and any real number $\alpha \in \mathbf{R}$, scalar multiplication gives a vector $\alpha x \in X$. These operations obey the usual algebraic laws; that is, for all $x, y, z \in X$, and $\alpha, \beta \in \mathbf{R}$:*

a. $x + y = y + x$;

b. $(x + y) + z = x + (y + z)$;

c. $\alpha(x + y) = \alpha x + \alpha y$;
d. $(\alpha + \beta)x = \alpha x + \beta x$; *and*
e. $(\alpha\beta)x = \alpha(\beta x)$.
Moreover, there is a zero vector $\theta \in X$ *that has the following properties:*
f. $x + \theta = x$; *and*
g. $0x = \theta$.
Finally,
h. $1x = x$.

The adjective "real" simply indicates that scalar multiplication is defined taking the real numbers, not elements of the complex plane or some other set, as scalars. All of the vector spaces used in this book are real, and the adjective will not be repeated. Important features of a vector space are that it has a "zero" element and that it is closed under addition and scalar multiplication. Vector spaces are also called *linear spaces.*

Exercise 3.2 Show that the following are vector spaces:
a. any finite-dimensional Euclidean space $\mathbf{R}^l$;
b. the set $X = \{x \in \mathbf{R}^2: x = \alpha z, \text{ some } \alpha \in \mathbf{R}\}$, where $z \in \mathbf{R}^2$;
c. the set X consisting of all infinite sequences $(x_0, x_1, x_2, \ldots)$, where $x_i \in \mathbf{R}$, all i;
d. the set of all continuous functions on the interval $[a, b]$.
Show that the following are not vector spaces:
e. the unit circle in $\mathbf{R}^2$;
f. the set of all integers, $I = \{\ldots, -1, 0, +1, \ldots\}$;
g. the set of all nonnegative functions on $[a, b]$.

To discuss convergence in a vector space or in any other space, we need to have the notion of distance. The notion of distance in Euclidean space is generalized in the abstract notion of a *metric,* a function defined on any two elements in a set the value of which has an interpretation as the distance between them.

DEFINITION *A **metric space** is a set S, together with a metric (distance function)* $\rho: S \times S \to \mathbf{R}$, *such that for all* $x, y, z \in S$:
a. $\rho(x, y) \geq 0$, *with equality if and only if* $x = y$;
b. $\rho(x, y) = \rho(y, x)$; *and*
c. $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$.

The definition of a metric thus abstracts the four basic properties of Euclidean distance: the distance between distinct points is strictly positive; the distance from a point to itself is zero; distance is symmetric; and the triangle inequality holds.

Exercise 3.3 Show that the following are metric spaces.
a. Let S be the set of integers, with $\rho(x, y) = |x - y|$.
b. Let S be the set of integers, with $\rho(x, y) = 0$ if $x = y$, 1 if $x \neq y$.
c. Let S be the set of all continuous, strictly increasing functions on $[a, b]$, with $\rho(x, y) = \max_{a \leq t \leq b} |x(t) - y(t)|$.
d. Let S be the set of all continuous, strictly increasing functions on $[a, b]$, with $\rho(x, y) = \int_a^b |x(t) - y(t)| dt$.
e. Let S be the set of all rational numbers, with $\rho(x, y) = |x - y|$.
f. Let $S = \mathbf{R}$, with $\rho(x, y) = f(|x - y|)$, where $f: \mathbf{R}_+ \to \mathbf{R}_+$ is continuous, strictly increasing, and strictly concave, with $f(0) = 0$.

For vector spaces, metrics are usually defined in such a way that the distance between any two points is equal to the distance of their difference from the zero point. That is, since for any points x and y in a vector space S, the point $x - y$ is also in S, the metric on a vector space is usually defined in such a way that $\rho(x, y) = \rho(x - y, \theta)$. To define such a metric, we need the concept of a norm.

DEFINITION *A **normed vector space** is a vector space S, together with a norm* $\|\cdot\|: S \to \mathbf{R}$, *such that for all* $x, y \in S$ *and* $\alpha \in \mathbf{R}$:
a. $\|x\| \geq 0$, *with equality if and only if* $x = \theta$;
b. $\|\alpha x\| = |\alpha| \cdot \|x\|$; *and*
c. $\|x + y\| \leq \|x\| + \|y\|$ *(the triangle inequality).*

Exercise 3.4 Show that the following are normed vector spaces.
a. Let $S = \mathbf{R}^l$, with $\|x\| = [\sum_{i=1}^l x_i^2]^{1/2}$ (Euclidean space).
b. Let $S = \mathbf{R}^l$, with $\|x\| = \max_i |x_i|$.
c. Let $S = \mathbf{R}^l$, with $\|x\| = \sum_{i=1}^l |x_i|$.
d. Let S be the set of all bounded infinite sequences $(x_1, x_2, \ldots)$, $x_k \in \mathbf{R}$, all k, with $\|x\| = \sup_k |x_k|$. (This space is called l_∞.)
e. Let S be the set of all continuous functions on $[a, b]$, with $\|x\| = \sup_{a \leq t \leq b} |x(t)|$. (This space is called $C[a, b]$.)
f. Let S be the set of all continuous functions on $[a, b]$, with $\|x\| = \int_a^b |x(t)| dt$.

It is standard to view any normed vector space $(S, \|\cdot\|)$ as a metric space, where the metric is taken to be $\rho(x, y) = \|x - y\|$, all $x, y \in S$.

The notion of convergence of a sequence of real numbers carries over without change to any metric space.

DEFINITION *A sequence $\{x_n\}_{n=0}^{\infty}$ in S **converges** to $x \in S$, if for each $\varepsilon > 0$, there exists N_ε such that*

(1) $\rho(x_n, x) < \varepsilon$, all $n \geq N_\varepsilon$.

Thus a sequence $\{x_n\}$ in a metric space (S, ρ) converges to $x \in S$ if and only if the sequence of distances $\{\rho(x_n, x)\}$, a sequence in $\mathbf{R}_+$, converges to zero. In this case we write $x_n \to x$.

Verifying convergence directly involves having a "candidate" for the limit point x so that the inequality (1) can be checked. When a candidate is not immediately available, the following alternative criterion is often useful.

DEFINITION *A sequence $\{x_n\}_{n=0}^{\infty}$ in S is a **Cauchy sequence** (satisfies the **Cauchy criterion**) if for each $\varepsilon > 0$, there exists N_ε such that*

(2) $\rho(x_n, x_m) < \varepsilon$, all $n, m \geq N_\varepsilon$.

Thus a sequence is Cauchy if the points get closer and closer to each other. The following exercise illustrates some basic facts about convergence and the Cauchy criterion.

Exercise 3.5 a. Show that if $x_n \to x$ and $x_n \to y$, then $x = y$. That is, if $\{x_n\}$ has a limit, then that limit is unique.

b. Show that if a sequence $\{x_n\}$ is convergent, then it satisfies the Cauchy criterion.

c. Show that if a sequence $\{x_n\}$ satisfies the Cauchy criterion, then it is bounded.

d. Show that $x_n \to x$ if and only if every subsequence of $\{x_n\}$ converges to x.

The advantage of the Cauchy criterion is that, in contrast to (1), (2) can be checked with knowledge of $\{x_n\}$ only. For the Cauchy criterion to be useful, however, we must work with spaces where it implies the existence of a limit point.

DEFINITION *A metric space (S, ρ) is **complete** if every Cauchy sequence in S converges to an element in S.*

In complete metric spaces, then, verifying that a sequence satisfies the Cauchy criterion is a way of verifying the existence of a limit point in S.

Verifying the completeness of particular spaces can take some work. We take as given the following

FACT *The set of real numbers $\mathbf{R}$ with the metric $\rho(x, y) = |x - y|$ is a complete metric space.*

Exercise 3.6 a. Show that the metric spaces in Exercises 3.3a,b and 3.4a–e are complete and that those in Exercises 3.3c–e and 3.4f are not. Show that the space in 3.3c is complete if "strictly increasing" is replaced with "nondecreasing."

b. Show that if (S, ρ) is a complete metric space and S' is a closed subset of S, then (S', ρ) is a complete metric space.

A complete normed vector space is called a ***Banach space.***

The next example is no more difficult than some of those in Exercise 3.6, but since it is important in what follows and illustrates clearly each of the steps involved in verifying completeness, we present the proof here.

THEOREM 3.1 *Let $X \subseteq \mathbf{R}^l$, and let $C(X)$ be the set of bounded continuous functions $f: X \to \mathbf{R}$ with the sup norm, $\|f\| = \sup_{x \in X} |f(x)|$. Then $C(X)$ is a complete normed vector space. (Note that if X is compact then every continuous function is bounded. Otherwise the restriction to bounded functions must be added.)*

Proof. That $C(X)$ is a normed vector space follows from Exercise 3.4e. Hence it suffices to show that if $\{f_n\}$ is a Cauchy sequence, there exists $f \in C(X)$ such that

$$\text{for any } \varepsilon > 0 \text{ there exists } N_\varepsilon \text{ such that } \|f_n - f\| \leq \varepsilon, \quad \text{all } n \geq N_\varepsilon.$$

Three steps are involved: to find a "candidate" function f; to show that $\{f_n\}$ converges to f in the sup norm; and to show that $f \in C(X)$ (that f is bounded and continuous). Each step involves its own entirely distinct logic.

Fix $x \in X$; then the sequence of real numbers $\{f_n(x)\}$ satisfies

$$|f_n(x) - f_m(x)| \leq \sup_{y \in X} |f_n(y) - f_m(y)| = \|f_n - f_m\|.$$

Therefore it satisfies the Cauchy criterion; and by the completeness of the real numbers, it converges to a limit point—call it $f(x)$. The limiting values define a function $f: X \to \mathbf{R}$ that we take to be our candidate.

Next we must show that $\|f_n - f\| \to 0$ as $n \to \infty$. Let $\varepsilon > 0$ be given and choose N_ε so that $n, m \geq N_\varepsilon$ implies $\|f_n - f_m\| \leq \varepsilon/2$. Since $\{f_n\}$ satisfies the Cauchy criterion, this can be done. Now for any fixed $x \in X$ and all $m \geq n \geq N_\varepsilon$,

$$\begin{aligned} |f_n(x) - f(x)| &\leq |f_n(x) - f_m(x)| + |f_m(x) - f(x)| \\ &\leq \|f_n - f_m\| + |f_m(x) - f(x)| \\ &\leq \varepsilon/2 + |f_m(x) - f(x)|. \end{aligned}$$

Since $\{f_m(x)\}$ converges to $f(x)$, we can choose m separately for each fixed $x \in X$ so that $|f_m(x) - f(x)| \leq \varepsilon/2$. Since the choice of x was arbitrary, it follows that $\|f_n - f\| \leq \varepsilon$, all $n \geq N_\varepsilon$. Since $\varepsilon > 0$ was arbitrary, the desired result then follows.

Finally, we must show that f is bounded and continuous. Boundedness is obvious. To prove that f is continuous, we must show that for every $\varepsilon > 0$ and every $x \in X$, there exists $\delta > 0$ such that

$$|f(x) - f(y)| < \varepsilon \text{ if } \|x - y\|_E < \delta,$$

where $\|\cdot\|_E$ is the Euclidean norm on $\mathbf{R}^l$. Let ε and x be given. Choose k so that $\|f - f_k\| < \varepsilon/3$; since $f_n \to f$ (in the sup norm), such a choice is possible. Then choose δ so that

$$\|x - y\|_E < \delta \text{ implies } |f_k(x) - f_k(y)| < \varepsilon/3.$$

Since f_k is continuous, such a choice is possible. Then

$$\begin{aligned} |f(x) - f(y)| &\leq |f(x) - f_k(x)| + |f_k(x) - f_k(y)| + |f_k(y) - f(y)| \\ &\leq 2\|f - f_k\| + |f_k(x) - f_k(y)| \\ &< \varepsilon. \quad \blacksquare \end{aligned}$$

Although we have organized these component arguments into a theorem about a function space, each should be familiar to students of calculus. Convergence in the sup norm is simply uniform convergence. The proof above is then just an amalgam of the standard proofs that a sequence of functions that satisfies the Cauchy criterion uniformly converges uniformly and that uniform convergence "preserves continuity."

Exercise 3.7 a. Let $C^1[a, b]$ be the set of all continuously differentiable functions on $[a, b] = X \subset \mathbf{R}$, with the norm $\|f\| = \sup_{x \in X}\{|f(x)| + |f'(x)|\}$. Show that $C^1[a, b]$ is a Banach space. [*Hint.* Notice that

$$\sup_{x \in X}|f(x)| + \sup_{x \in X}|f'(x)| \geq \|f\| \geq \max\{\sup_{x \in X}|f(x)|, \sup_{x \in X}|f'(x)|\}.]$$

b. Show that this set of functions with the norm $\|f\| = \sup_{x \in X}|f(x)|$ is not complete. That is, give an example of a sequence of functions that is Cauchy in the given norm that does not converge to a function in the set. Is this sequence Cauchy in the norm of part (a)?

c. Let $C^k[a, b]$ be the set of all k times continuously differentiable functions on $[a, b] = X \subset \mathbf{R}$, with the norm $\|f\| = \sum_{i=0}^{k} \alpha_i \max_{x \in X}|f^i(x)|$, where $f^i = d^i f(x)/dx^i$. Show that this space is complete if and only if $\alpha_i > 0$, $i = 0, 1, \ldots, k$.

3.2 The Contraction Mapping Theorem

In this section we prove two main results. The first is the Contraction Mapping Theorem, an extremely simple and powerful fixed point theorem. The second is a set of sufficient conditions, due to Blackwell, for establishing that certain operators are contraction mappings. The latter are useful in a wide variety of economic applications and will be drawn upon extensively in the next chapter.

We begin with the following definition.

DEFINITION *Let (S, ρ) be a metric space and $T: S \to S$ be a function mapping S into itself. T is a* **contraction mapping** *(with* **modulus** *β) if for some $\beta \in (0, 1)$, $\rho(Tx, Ty) \leq \beta\rho(x, y)$, for all $x, y \in S$.*

Perhaps the most familiar examples of contraction mappings are those on a closed interval $S = [a, b]$, with $\rho(x, y) = |x - y|$. Then $T: S \to S$ is a contraction if for some $\beta \in (0, 1)$.

$$\frac{|Tx - Ty|}{|x - y|} \leq \beta < 1, \quad \text{all } x, y \in S \text{ with } x \neq y.$$

That is, T is a contraction mapping if it is a function with slope uniformly less than one in absolute value.

Exercise 3.8 Show that if T is a contraction on S, then T is uniformly continuous on S.

The ***fixed points*** of T, the elements of S satisfying $Tx = x$, are the intersections of Tx with the 45° line, as shown in Figure 3.1. Hence it is clear that any contraction on this space has a unique fixed point. This conclusion is much more general.

THEOREM 3.2 (*Contraction Mapping Theorem*) *If (S, ρ) is a complete metric space and $T: S \to S$ is a contraction mapping with modulus β, then*

a. T has exactly one fixed point v in S, and

b. for any $v_0 \in S$, $\rho(T^n v_0, v) \leq \beta^n \rho(v_0, v)$, $n = 0, 1, 2, \ldots$.

Proof. To prove (a), we must find a candidate for v, show that it satisfies $Tv = v$, and show that no other element $\hat{v} \in S$ does.

Define the iterates of T, the mappings $\{T^n\}$, by $T^0x = x$, and $T^n x = T(T^{n-1}x)$, $n = 1, 2, \ldots$. Choose $v_0 \in S$, and define $\{v_n\}_{n=0}^{\infty}$ by $v_{n+1} = Tv_n$, so that $v_n = T^n v_0$. By the contraction property of T,

$$\rho(v_2, v_1) = \rho(Tv_1, Tv_0) \leq \beta\rho(v_1, v_0).$$

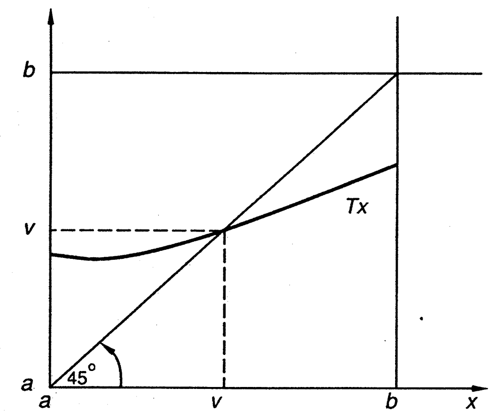

Figure 3.1

Continuing by induction, we get

$$(1) \qquad \rho(v_{n+1}, v_n) \leq \beta^n \rho(v_1, v_0), \quad n = 1, 2, \ldots.$$

Hence, for any $m > n$,

$$\begin{aligned}\rho(v_m, v_n) &\leq \rho(v_m, v_{m-1}) + \cdots + \rho(v_{n+2}, v_{n+1}) + \rho(v_{n+1}, v_n)\\ &\leq [\beta^{m-1} + \cdots + \beta^{n+1} + \beta^n]\,\rho(v_1, v_0)\\ &= \beta^n[\beta^{m-n-1} + \cdots + \beta + 1]\rho(v_1, v_0)\\ (2) \qquad &\leq \frac{\beta^n}{1-\beta}\rho(v_1, v_0),\end{aligned}$$

where the first line uses the triangle inequality and the second follows from (1). It is clear from (2) that $\{v_n\}$ is a Cauchy sequence. Since S is complete, it follows that $v_n \to v \in S$.

To show that $Tv = v$, note that for all n and all $v_0 \in S$,

$$\begin{aligned}\rho(Tv, v) &\leq \rho(Tv, T^n v_0) + \rho(T^n v_0, v)\\ &\leq \beta\rho(v, T^{n-1}v_0) + \rho(T^n v_0, v).\end{aligned}$$

We have demonstrated that both terms in the last expression converge to zero as $n \to \infty$; hence $\rho(Tv, v) = 0$, or $Tv = v$.

Finally, we must show that there is no other function $\hat{v} \in S$ satisfying $T\hat{v} = \hat{v}$. Suppose to the contrary that $\hat{v} \neq v$ is another solution. Then

$$0 < a = \rho(\hat{v}, v) = \rho(T\hat{v}, Tv) \leq \beta\rho(\hat{v}, v) = \beta a,$$

which cannot hold, since $\beta < 1$. This proves part (a).

To prove part (b), observe that for any $n \geq 1$

$$\rho(T^n v_0, v) = \rho[T(T^{n-1}v_0), Tv] \leq \beta\rho(T^{n-1}v_0, v),$$

so that (b) follows by induction. ■

Recall from Exercise 3.6b that if (S, ρ) is a complete metric space and S' is a closed subset of S, then (S', ρ) is also a complete metric space. Now suppose that $T: S \to S$ is a contraction mapping, and suppose further that T maps S' into itself, $T(S') \subseteq S'$ (where $T(S')$ denotes the image of S' under T). Then T is also a contraction mapping on S'. Hence the unique fixed point of T on S lies in S'. This observation is often useful for establishing qualitative properties of a fixed point. Specifically, in some situations we will want to apply the Contraction Mapping Theorem twice: once on a large space to establish uniqueness, and again on a smaller space to characterize the fixed point more precisely.

The following corollary formalizes this argument.

COROLLARY 1 *Let (S, ρ) be a complete metric space, and let $T: S \to S$ be a contraction mapping with fixed point $v \in S$. If S' is a closed subset of S and $T(S') \subseteq S'$, then $v \in S'$. If in addition $T(S') \subseteq S'' \subseteq S'$, then $v \in S''$.*

Proof. Choose $v_0 \in S'$, and note that $\{T^n v_0\}$ is a sequence in S' converging to v. Since S' is closed, it follows that $v \in S'$. If in addition $T(S') \subseteq S''$, then it follows that $v = Tv \in S''$. ■

Part (b) of the Contraction Mapping Theorem bounds the distance $\rho(T^n v_0, v)$ between the nth approximation and the fixed point in terms of the distance $\rho(v_0, v)$ between the initial approximation and the fixed point. However, if v is not known (as is the case if one is computing v), then neither is the magnitude of the bound. Exercise 3.9 gives a computationally useful inequality.

Exercise 3.9 Let (S, ρ), T, and v be as given above, let β be the modulus of T, and let $v_0 \in S$. Show that

$$\rho(T^n v_0, v) \leq \frac{1}{1-\beta}\rho(T^n v_0, T^{n+1}v_0).$$

The following result is a useful generalization of the Contraction Mapping Theorem.

COROLLARY 2 *(N-Stage Contraction Theorem) Let (S, ρ) be a complete metric space, let $T: S \to S$, and suppose that for some integer N, $T^N: S \to S$ is a contraction mapping with modulus β. Then*

a. T has exactly one fixed point in S, and

b. for any $v_0 \in S$, $\rho(T^{kN}v_0, v) \leq \beta^k\rho(v_0, v)$, $k = 0, 1, 2, \ldots$.

Proof. We will show that the unique fixed point v of T^N is also the unique fixed point of T. We have

$$\rho(Tv, v) = \rho[T(T^N v), T^N v] = \rho[T^N(Tv), T^N v] \leq \beta\rho(Tv, v).$$

Since $\beta \in (0, 1)$, this implies that $\rho(Tv, v) = 0$, so v is a fixed point of T. To establish uniqueness, note that any fixed point of T is also a fixed point of T^N. Part (b) is established using the same argument as in the proof of Theorem 3.2. ■

The next exercise shows how the Contraction Mapping Theorem is used to prove existence and uniqueness of a solution to a differential equation.

Exercise 3.10 Consider the differential equation and boundary condition $dx(s)/ds = f[x(s)]$, all $s \geq 0$, with $x(0) = c \in \mathbf{R}$. Assume that $f: \mathbf{R} \to \mathbf{R}$ is continuous, and for some $B > 0$ satisfies the Lipschitz condition $|f(a) - f(b)| \leq B|a - b|$, all $a, b \in \mathbf{R}$. For any $t > 0$, consider $C[0, t]$, the space of bounded continuous functions on $[0, t]$, with the sup norm. Recall from Theorem 3.1 that this space is complete.

a. Show that the operator T defined by

$$(Tv)(s) = c + \int_0^s f[v(z)]dz, \quad 0 \leq s \leq t,$$

maps $C[0, t]$ into itself. That is, show that if v is bounded and continuous on $[0, t]$, then so is Tv.

b. Show that for some $\tau > 0$, T is a contraction on $C[0, \tau]$.

c. Show that the unique fixed point of T on $C[0, \tau]$ is a differentiable function, and hence that it is the unique solution on $[0, \tau]$ to the given differential equation.

Another useful route to verifying that certain operators are contractions is due to Blackwell.

THEOREM 3.3 *(Blackwell's sufficient conditions for a contraction) Let* $X \subseteq \mathbf{R}^l$, *and let* $B(X)$ *be a space of bounded functions* $f: X \to \mathbf{R}$, *with the sup norm. Let* $T: B(X) \to B(X)$ *be an operator satisfying*

a. *(monotonicity)* $f, g \in B(X)$ *and* $f(x) \leq g(x)$, *for all* $x \in X$, *implies* $(Tf)(x) \leq (Tg)(x)$, *for all* $x \in X$;

b. *(discounting) there exists some* $\beta \in (0, 1)$ *such that*

$$[T(f + a)](x) \leq (Tf)(x) + \beta a, \quad \text{all } f \in B(X),\ a \geq 0,\ x \in X.$$

[Here $(f + a)(x)$ *is the function defined by* $(f + a)(x) = f(x) + a$.*] Then* T *is a contraction with modulus* β.

Proof. If $f(x) \leq g(x)$ for all $x \in X$, we write $f \leq g$. For any $f, g \in B(X)$, $f \leq g + \|f - g\|$. Then properties (a) and (b) imply that

$$Tf \leq T(g + \|f - g\|) \leq Tg + \beta\|f - g\|.$$

Reversing the roles of f and g gives by the same logic

$$Tg \leq Tf + \beta\|f - g\|.$$

Combining these two inequalities, we find that $\|Tf - Tg\| \leq \beta\|f - g\|$, as was to be shown. ■

In many economic applications the two hypotheses of Blackwell's theorem can be verified at a glance. For example, in the one-sector optimal growth problem, an operator T was defined by

$$(Tv)(k) = \max_{0 \leq y \leq f(k)} \{U[f(k) - y] + \beta v(y)\}.$$

If $v(y) \leq w(y)$ for all values of y, then the objective function for which Tw is the maximized value is uniformly higher than the function for which Tv is the maximized value; so the monotonicity hypothesis (a) is obvious. The discounting hypothesis (b) is equally easy, since

$$\begin{aligned} T(v + a)(k) &= \max_{0 \leq y \leq f(k)} \{U[f(k) - y] + \beta[v(y) + a]\} \\ &= \max_{0 \leq y \leq f(k)} \{U[f(k) - y] + \beta v(y)\} + \beta a \\ &= (Tv)(k) + \beta a. \end{aligned}$$

Blackwell's result will play a key role in our analysis of dynamic programs.

3.3 The Theorem of the Maximum

We will want to apply the Contraction Mapping Theorem to analyze dynamic programming problems that are much more general than the examples that have been discussed to this point. If x is the beginning-of-period state variable, an element of $X \subseteq \mathbf{R}^l$, and $y \in X$ is the end-of-period state to be chosen, we would like to let the current period return $F(x, y)$ and the set of feasible y values, given x, be specified as generally as possible. On the other hand, we want the operator T defined by

$$(Tv)(x) = \sup_y [F(x, y) + \beta v(y)]$$

$$\text{s.t.} \quad y \text{ feasible given } x,$$

to take the space $C(X)$ of bounded continuous functions of the state into itself. We would also like to be able to characterize the set of maximizing values of y, given x.

To describe the feasible set, we use the idea of a ***correspondence*** from a set X into a set Y: a relation that assigns a set $\Gamma(x) \subseteq Y$ to each $x \in X$. In the case of interest here, $Y = X$. Hence we seek restrictions on the correspondence $\Gamma: X \to X$ describing the feasibility constraints and on the return function F, which together ensure that if $v \in C(X)$ and $(Tv)(x) =$

$\sup_{y\in\Gamma(x)}[F(x, y) + \beta v(y)]$ then $Tv \in C(X)$. Moreover, we wish to determine the implied properties of the correspondence $G(x)$ containing the maximizing values of y for each x. The main result in this section is the Theorem of the Maximum, which accomplishes both tasks.

Let $X \subseteq \mathbf{R}^l$; let $Y \subseteq \mathbf{R}^m$; let $f: X \times Y \to \mathbf{R}$ be a (single-valued) function; and let $\Gamma: X \to Y$ be a (nonempty, possibly multivalued) correspondence. Our interest is in problems of the form $\sup_{y\in\Gamma(x)} f(x, y)$. If for each x, $f(x, \cdot)$ is continuous in y and the set $\Gamma(x)$ is nonempty and compact, then for each x the maximum is attained. In this case the function

$$(1) \qquad h(x) = \max_{y\in\Gamma(x)} f(x, y)$$

is well defined, as is the nonempty set

$$(2) \qquad G(x) = \{y \in \Gamma(x): f(x, y) = h(x)\}$$

of y values that attain the maximum. In this section further restrictions on f and Γ will be added, to ensure that the function h and the set G vary in a continuous way with x.

There are several notions of continuity for correspondences, and each can be characterized in a variety of ways. For our purposes it is convenient to use definitions stated in terms of sequences.

DEFINITION *A correspondence $\Gamma: X \to Y$ is* ***lower hemi-continuous*** *(l.h.c.) at x if $\Gamma(x)$ is nonempty and if, for every $y \in \Gamma(x)$ and every sequence $x_n \to x$, there exists $N \geq 1$ and a sequence $\{y_n\}_{n=N}^{\infty}$ such that $y_n \to y$ and $y_n \in \Gamma(x_n)$, all $n \geq N$. [If $\Gamma(x')$ is nonempty for all $x' \in X$, then it is always possible to take $N = 1$.]*

DEFINITION *A compact-valued correspondence $\Gamma: X \to Y$ is* ***upper hemi-continuous*** *(u.h.c.) at x if $\Gamma(x)$ is nonempty and if, for every sequence $x_n \to x$ and every sequence $\{y_n\}$ such that $y_n \in \Gamma(x_n)$, all n, there exists a convergent subsequence of $\{y_n\}$ whose limit point y is in $\Gamma(x)$.*

Figure 3.2 displays a correspondence that is l.h.c. but not u.h.c. at x_1; is u.h.c. but not l.h.c. at x_2; and is both u.h.c. and l.h.c. at all other points. Note that our definition of u.h.c. applies only to correspondences that are compact-valued. Since all of the correspondences we will be dealing

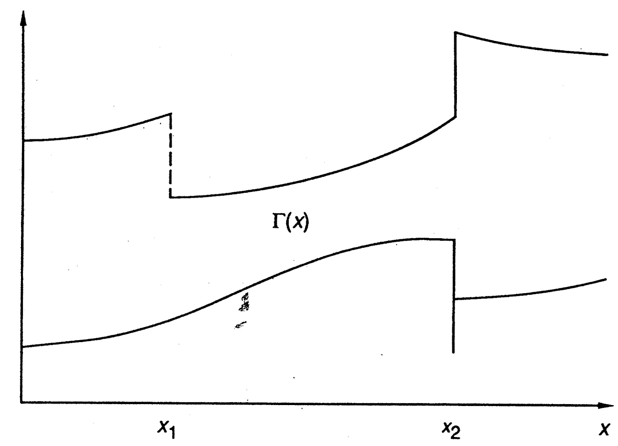

Figure 3.2

with satisfy this requirement, the restriction will not be binding. (A definition of u.h.c. for all correspondences is available, but it is stated in terms of images of open sets. For our purposes this definition is much less convenient, and its wider scope is never useful.)

DEFINITION *A correspondence $\Gamma: X \to Y$ is* ***continuous*** *at $x \in X$ if it is both u.h.c. and l.h.c. at x.*

A correspondence $\Gamma: X \to Y$ is called l.h.c., u.h.c., or continuous if it has that property at every point $x \in X$. The following exercises highlight some important facts about upper and lower hemi-continuity. Note that if $\Gamma: X \to Y$, then for any set $\hat{X} \subset X$, we define

$$\Gamma(\hat{X}) = \{y \in Y: y \in \Gamma(x), \text{ for some } x \in \hat{X}\}.$$

Exercise 3.11 a. Show that if Γ is single-valued and u.h.c., then it is continuous.

b. Let $\Gamma: \mathbf{R}^k \to \mathbf{R}^{l+m}$, and define $\phi: \mathbf{R}^k \to \mathbf{R}^l$ by

$$\phi(x) = \{y_1 \in \mathbf{R}^l: (y_1, y_2) \in \Gamma(x) \text{ for some } y_2 \in \mathbf{R}^m\}.$$

Show that if Γ is compact-valued and u.h.c., then so is ϕ.

c. Let $\phi: X \to Y$ and $\psi: X \to Y$ be compact-valued and u.h.c., and define $\Gamma = \phi \cup \psi$ by

$$\Gamma(x) = \{y \in Y: y \in \phi(x) \cup \psi(x)\}, \quad \text{all } x \in X.$$

Show that Γ is compact-valued and u.h.c.

d. Let $\phi: X \to Y$ and $\psi: X \to Y$ be compact-valued and u.h.c., and suppose that

$$\Gamma(x) = \{y \in Y: y \in \phi(x) \cap \psi(x)\} \neq \emptyset, \quad \text{all } x \in X.$$

Show that Γ is compact-valued and u.h.c.

e. Show that if $\phi: X \to Y$ and $\psi: Y \to Z$ are compact-valued and u.h.c., then the correspondence $\psi \circ \phi = \Gamma: X \to Z$ defined by

$$\Gamma(x) = \{z \in Z: z \in \psi(y), \text{ for some } y \in \phi(x)\}$$

is also compact-valued and u.h.c.

f. Let $\Gamma_i: X \to Y_i$, $i = 1, \ldots, k$, be compact-valued and u.h.c. Show that $\Gamma: X \to Y = Y_1 \times \ldots \times Y_k$ defined by

$$\Gamma(x) = \{y \in Y: y = (y_1, \ldots, y_k), \text{ where } y_i \in \Gamma_i(x), i = 1, \ldots, k\},$$

is also compact-valued and u.h.c.

g. Show that if $\Gamma: X \to Y$ is compact-valued and u.h.c., then for any compact set $K \subseteq X$, the set $\Gamma(K) \subseteq Y$ is also compact. [*Hint.* To show that $\Gamma(K)$ is bounded, suppose the contrary. Let $\{y_n\}$ be a divergent sequence in $\Gamma(K)$, and choose $\{x_n\}$ such that $y_n \in \Gamma(x_n)$, all n.]

Exercise 3.12 a. Show that if Γ is single-valued and l.h.c., then it is continuous.

b. Let $\Gamma: \mathbf{R}^k \to \mathbf{R}^{l+m}$, and define $\phi: \mathbf{R}^k \to \mathbf{R}^l$ by

$$\phi(x) = \{y_1 \in \mathbf{R}^l: (y_1, y_2) \in \Gamma(x), \text{ for some } y_2 \in \mathbf{R}^m\}.$$

Show that if Γ is l.h.c., then so is ϕ.

c. Let $\phi: X \to Y$ and $\psi: X \to Y$ be l.h.c., and define $\Gamma = \phi \cup \psi$ by

$$\Gamma(x) = \{y \in Y: y \in \phi(x) \cup \psi(x)\}, \quad \text{all } x \in X.$$

Show that Γ is l.h.c.

d. Let $\phi: X \to Y$ and $\psi: X \to Y$ be l.h.c., and suppose that

$$\Gamma(x) = \{y \in Y: y \in \phi(x) \cap \psi(x)\} \neq \emptyset, \quad \text{all } x \in X.$$

Show by example that Γ need not be l.h.c. Show that if ϕ and ψ are both convex-valued, and if $\text{int } \phi(x) \cap \text{int } \psi(x) \neq \emptyset$, then Γ is l.h.c. at x.

e. Show that if $\phi: X \to Y$ and $\psi: Y \to Z$ are l.h.c., then the correspondence $\psi \circ \phi = \Gamma: X \to Z$ defined by

$$\Gamma(x) = \{z \in Z: z \in \psi(y), \text{ for some } y \in \phi(x)\}$$

is also l.h.c.

f. Let $\Gamma_i: X \to Y_i$, $i = 1, \ldots, k$, be l.h.c. Show that $\Gamma: X \to Y = Y_1 \times \ldots \times Y_k$ defined by

$$\Gamma(x) = \{y \in Y: y = (y_1, \ldots, y_k), \text{ where } y_i \in \Gamma_i(x), i = 1, \ldots, k\}$$

is l.h.c.

The next two exercises show some of the relationships between constraints stated in terms of inequalities involving continuous functions and those stated in terms of continuous correspondences. These relationships are extremely important for many problems in economics, where constraints are often stated in terms of production functions, budget constraints, and so on.

Exercise 3.13 a. Let $\Gamma: \mathbf{R}_+ \to \mathbf{R}_+$ be defined by $\Gamma(x) = [0, x]$. Show that Γ is continuous.

b. Let $f: \mathbf{R}^l_+ \to \mathbf{R}_+$ be a continuous function, and define the correspondence $\Gamma: \mathbf{R}^l_+ \to \mathbf{R}_+$ by $\Gamma(x) = [0, f(x)]$. Show that Γ is continuous.

c. Let $f_i: \mathbf{R}^l_+ \times \mathbf{R}^m \to \mathbf{R}_+$, $i = 1, \ldots, l$, be continuous functions. Define $\Gamma: \mathbf{R}^l_+ \times \mathbf{R}^m \to \mathbf{R}^l_+$ by

$$\Gamma(x, z) = \left\{y \in \mathbf{R}^l_+: 0 \leq y_i \leq f_i(x^i, z), i = 1, \ldots, l; \text{ and } \sum_{i=1}^{l} x^i \leq x\right\}.$$

Show that Γ is continuous.

Exercise 3.14 a. Let $H(x, y)$: $\mathbf{R}_+^l \times \mathbf{R}_+^m \rightarrow \mathbf{R}$ be continuous, strictly increasing in its first l arguments, strictly decreasing in its last m arguments, with $H(0, 0) = 0$. Define Γ: $\mathbf{R}^l \rightarrow \mathbf{R}^m$ by $\Gamma(x) = \{y \in \mathbf{R}^m: H(x, y) \geq 0\}$. Show that if $\Gamma(x)$ is compact-valued, then Γ is continuous at x.

b. Let $H(x, y)$: $\mathbf{R}^l \times \mathbf{R}^m \rightarrow \mathbf{R}$ be continuous and concave, and define Γ as in part (a). Show that if $\Gamma(x)$ is compact-valued and there exists some $\hat{y} \in \Gamma(x)$ such that $H(x, \hat{y}) > 0$, then Γ is continuous at x.

c. Define $H: \mathbf{R} \times \mathbf{R} \rightarrow \mathbf{R}$ by $H(x, y) = 1 - \max\{|x|, |y|\}$, and define $\Gamma(x)$ as in part (a). Where does Γ fail to be l.h.c.?

When trying to establish properties of a correspondence $\Gamma: X \rightarrow Y$, it is sometimes useful to deal with its ***graph***, the set

$$A = \{(x, y) \in X \times Y: y \in \Gamma(x)\}.$$

The next two results provide conditions on A that are sufficient to ensure the upper and lower hemi-continuity respectively of Γ.

THEOREM 3.4 *Let $\Gamma: X \rightarrow Y$ be a nonempty-valued correspondence, and let A be the graph of Γ. Suppose that A is closed, and that for any bounded set $\hat{X} \subseteq X$, the set $\Gamma(\hat{X})$ is bounded. Then Γ is compact-valued and u.h.c.*

Proof. For each $x \in X$, $\Gamma(x)$ is closed (since A is closed) and is bounded (by hypothesis). Hence Γ is compact-valued.

Let $\hat{x} \in X$, and let $\{x_n\} \subseteq X$ with $x_n \rightarrow \hat{x}$. Since Γ is nonempty-valued, we can choose $y_n \in \Gamma(x_n)$, all n. Since $x_n \rightarrow \hat{x}$, there is a bounded set $\hat{X} \subset X$ containing $\{x_n\}$ and $\hat{x}$. Then by hypothesis $\Gamma(\hat{X})$ is bounded. Hence $\{y_n\} \subset \Gamma(\hat{X})$ has a convergent subsequence, call it $\{y_{n_k}\}$; let $\hat{y}$ be the limit point of this subsequence. Then $\{(x_{n_k}, y_{n_k})\}$ is a sequence in A converging to $(\hat{x}, \hat{y})$; since A is closed, it follows that $(\hat{x}, \hat{y}) \in A$. Hence $\hat{y} \in \Gamma(\hat{x})$, so Γ is u.h.c. at $\hat{x}$. Since $\hat{x}$ was arbitrary, this establishes the desired result. ∎

To see why the hypothesis of boundedness is required in Theorem 3.4, consider the correspondence $\Gamma: \mathbf{R}_+ \rightarrow \mathbf{R}_+$ defined by

$$\Gamma(0) = 0, \quad \text{and} \quad \Gamma(x) = \{0, 1/x\}, \quad \text{all } x > 0.$$

The graph of Γ is closed, but Γ is not u.h.c. at $x = 0$.

The next exercise is a kind of converse to Theorem 3.4.

Exercise 3.15 Let $\Gamma: X \rightarrow Y$ be a compact-valued u.h.c. correspondence with graph A. Show that if X is compact then A is compact.

The next theorem deals with lower hemi-continuity. For any $x \in \mathbf{R}^l$ and any $\varepsilon > 0$, let $B(x, \varepsilon)$ denote the closed ball of radius ε about x: $B(x, \varepsilon) = \{x' \in X: \|x - x'\| \leq \varepsilon\}$.

THEOREM 3.5 *Let $\Gamma: X \rightarrow Y$ be a nonempty-valued correspondence, and let A be the graph of Γ. Suppose that A is convex; that for any bounded set $\hat{X} \subseteq X$, there is a bounded set $\hat{Y} \subseteq Y$ such that $\Gamma(x) \cap \hat{Y} \neq \emptyset$, all $x \in \hat{X}$; and that for every $x \in X$, there exists some $\varepsilon > 0$ such that the set $B(x, \varepsilon) \cap X$ is closed and convex. Then Γ is l.h.c.*

Proof. Choose $\hat{x} \in X$; $\hat{y} \in \Gamma(\hat{x})$; and $\{x_n\} \subset X$ with $x_n \rightarrow \hat{x}$. Choose $\varepsilon > 0$ such that the set $\hat{X} = B(\hat{x}, \varepsilon) \cap X$ is closed and convex. Note that for some $N \geq 1$, $x_n \in \hat{X}$, all $n \geq N$; without loss of generality we take $N = 1$.

Let D denote the boundary of the set $\hat{X}$. Every point x_n has at least one representation as a convex combination of $\hat{x}$ and a point in D. For each n, choose $\alpha_n \in [0, 1]$ and $d_n \in D$ such that

$$x_n = \alpha_n d_n + (1 - \alpha_n)\hat{x}.$$

Since D is a bounded set and $x_n \rightarrow \hat{x}$, it follows that $\alpha_n \rightarrow 0$. Choose $\hat{Y}$ such that $\Gamma(x) \cap \hat{Y} \neq \emptyset$, all $x \in \hat{X}$. Then for each n, choose $\hat{y}_n \in \Gamma(d_n) \cap \hat{Y}$, and define

$$y_n = \alpha_n \hat{y}_n + (1 - \alpha_n)\hat{y}, \quad \text{all } n.$$

Since $(d_n, \hat{y}_n) \in A$, all n, $(\hat{x}, \hat{y}) \in A$, and A is convex, it follows that $(x_n, y_n) \in A$, all n. Moreover, since $\alpha_n \rightarrow 0$ and all of the $\hat{y}_n$'s lie in the bounded set $\hat{Y}$, it follows that $y_n \rightarrow \hat{y}$. Hence $\{(x_n, y_n)\}$ lies in A and converges to $(\hat{x}, \hat{y})$, as was to be shown. ∎

We are now ready to answer the questions we posed at the beginning of this section: Under what conditions do the function $h(x)$ defined by the maximization problem in (1) and the associated set of maximizing y values $G(x)$ defined in (2) vary continuously with x? An answer is provided in the following theorem, which will repeatedly be applied later.

THEOREM 3.6 (*Theorem of the Maximum*) *Let $X \subseteq \mathbf{R}^l$ and $Y \subseteq \mathbf{R}^m$, let $f: X \times Y \to \mathbf{R}$ be a continuous function, and let $\Gamma: X \to Y$ be a compact-valued and continuous correspondence. Then the function $h: X \to \mathbf{R}$ defined in (1) is continuous, and the correspondence $G: X \to Y$ defined in (2) is nonempty, compact-valued, and u.h.c.*

Proof. Fix $x \in X$. The set $\Gamma(x)$ is nonempty and compact, and $f(x, \cdot)$ is continuous; hence the maximum in (1) is attained, and the set $G(x)$ of maximizers is nonempty. Moreover, since $G(x) \subseteq \Gamma(x)$ and $\Gamma(x)$ is compact, it follows that $G(x)$ is bounded. Suppose $y_n \to y$, and $y_n \in G(x)$, all n. Since $\Gamma(x)$ is closed, $y \in \Gamma(x)$. Also, since $h(x) = f(x, y_n)$, all n, and f is continuous, it follows that $f(x, y) = h(x)$. Hence $y \in G(x)$; so $G(x)$ is closed. Thus $G(x)$ is nonempty and compact, for each x.

Next we will show that $G(x)$ is u.h.c. Fix x, and let $\{x_n\}$ be any sequence converging to x. Choose $y_n \in G(x_n)$, all n. Since Γ is u.h.c., there exists a subsequence $\{y_{n_k}\}$ converging to $y \in \Gamma(x)$. Let $z \in \Gamma(x)$. Since Γ is l.h.c., there exists a sequence $z_{n_k} \to z$, with $z_{n_k} \in \Gamma(x_{n_k})$, all k. Since $f(x_{n_k}, y_{n_k}) \geq f(x_{n_k}, z_{n_k})$, all k, and f is continuous, it follows that $f(x, y) \geq f(x, z)$. Since this holds for any $z \in \Gamma(x)$, it follows that $y \in G(x)$. Hence G is u.h.c.

Finally, we will show that h is continuous. Fix x, and let $\{x_n\}$ be any sequence converging to x. Choose $y_n \in G(x_n)$, all n. Let $\bar{h} = \lim\sup h(x_n)$ and $\underline{h} = \lim\inf h(x_n)$. Then there exists a subsequence $\{x_{n_k}\}$ such that $\bar{h} = \lim f(x_{n_k}, y_{n_k})$. But since G is u.h.c., there exists a subsequence of $\{y_{n_k}\}$, call it $\{y_j'\}$, converging to $y \in G(x)$. Hence $\bar{h} = \lim f(x_j, y_j') = f(x, y) = h(x)$. An analogous argument establishes that $h(x) = \underline{h}$. Hence $\{h(x_n)\}$ converges, and its limit is $h(x)$. ■

The following exercise illustrates through concrete examples what this theorem does and does not say.

Exercise 3.16 a. Let $X = \mathbf{R}$, and let $\Gamma(x) = Y = [-1, +1]$, all $x \in X$. Define $f: X \times Y \to \mathbf{R}$ by $f(x, y) = xy^2$. Graph $G(x)$; show that $G(x)$ is u.h.c. but not l.h.c. at $x = 0$.

b. Let $x = \mathbf{R}$, and let $\Gamma(x) = [0, 4]$, all $x \in X$. Define

$$f(x, y) = \max\{2 - (y - 1)^2, x + 1 - (y - 2)^2\}.$$

Graph $G(x)$ and show that it is u.h.c. Exactly where does it fail to be l.h.c.?

c. Let $X = \mathbf{R}$, $\Gamma(x) = \{y \in \mathbf{R}: -x \leq y \leq x\}$, and $f(x, y) = \cos(y)$. Graph $G(x)$ and show that it is u.h.c. Exactly where does it fail to be l.h.c.?

Suppose that in addition to the hypotheses of the Theorem of the Maximum the correspondence Γ is convex-valued and the function f is strictly concave in y. Then G is single-valued, and by Exercise 3.11a it is a continuous function—call it g. The next two results establish properties of g. Lemma 3.7 shows that if $f(x, y)$ is close to the maximized value $f[x, g(x)]$, then y is close to $g(x)$. Theorem 3.8 draws on this result to show that if $\{f_n\}$ is a sequence of continuous functions, each strictly concave in y, converging uniformly to f, then the sequence of maximizing functions $\{g_n\}$ converges pointwise to g. The latter convergence is uniform if X is compact.

LEMMA 3.7 *Let $X \subseteq \mathbf{R}^l$ and $Y \subseteq \mathbf{R}^m$. Assume that the correspondence $\Gamma: X \to Y$ is nonempty, compact- and convex-valued, and continuous, and let A be the graph of Γ. Assume that the function $f: A \to \mathbf{R}$ is continuous and that $f(x, \cdot)$ is strictly concave, for each $x \in X$. Define the function $g: X \to Y$ by*

$$g(x) = \underset{y \in \Gamma(x)}{\operatorname{argmax}} f(x, y).$$

Then for each $\varepsilon > 0$ and $x \in X$, there exists $\delta_x > 0$ such that

$$y \in \Gamma(x) \text{ and } |f[x, g(x)] - f(x, y)| < \delta_x \text{ implies } \|g(x) - y\| < \varepsilon.$$

If X is compact, then $\delta > 0$ can be chosen independently of x.

Proof. Note that under the stated assumptions g is a well-defined, continuous (single-valued) function. We first prove the claim for the case where X is compact. Note that in this case A is a compact set by Exercise 3.15. For each $\varepsilon > 0$, define

$$A_\varepsilon = \{(x, y) \in A: \|g(x) - y\| \geq \varepsilon\}.$$

If $A_\varepsilon = \emptyset$, all $\varepsilon > 0$, then Γ is single-valued and the result is trivial. Otherwise there exists $\hat{\varepsilon} > 0$ sufficiently small such that for all $0 < \varepsilon < \hat{\varepsilon}$, the set A_ε is nonempty and compact. For any such ε, let

$$\delta = \min_{(x,y) \in A_\varepsilon} |f[x, g(x)] - f(x, y)|.$$

Since the function being minimized is continuous and A_ε is compact, the minimum is attained. Moreover, since $[x, g(x)] \notin A_\varepsilon$, all $x \in X$, it follows that $\delta > 0$. Then

$$y \in \Gamma(x) \text{ and } \|g(x) - y\| \geq \varepsilon \text{ implies } |f[x, g(x)] - f(x, y)| \geq \delta,$$

as was to be shown.

If X is not compact, the argument above can be applied separately for each fixed $x \in X$. ■

THEOREM 3.8 *Let X, Y, Γ, and A be as defined in Lemma 3.7. Let $\{f_n\}$ be a sequence of continuous (real-valued) functions on A; assume that for each n and each $x \in X$, $f_n(x, \cdot)$ is strictly concave in its second argument. Assume that f has the same properties and that $f_n \to f$ uniformly (in the sup norm). Define the functions g_n and g by*

$$g_n(x) = \underset{y \in \Gamma(x)}{\operatorname{argmax}} f_n(x, y), \quad n = 1, 2, \ldots, \text{ and}$$

$$g(x) = \underset{y \in \Gamma(x)}{\operatorname{argmax}} f(x, y).$$

Then $g_n \to g$ pointwise. If X is compact, $g_n \to g$ uniformly.

Proof. First note that since $g_n(x)$ is the unique maximizer of $f_n(x, \cdot)$ on $\Gamma(x)$, and $g(x)$ is the unique maximizer of $f(x, \cdot)$ on $\Gamma(x)$, it follows that

$$\begin{aligned} 0 &\leq f[x, g(x)] - f[x, g_n(x)] \\ &\leq f[x, g(x)] - f_n[x, g(x)] + f_n[x, g_n(x)] - f[x, g_n(x)] \\ &\leq 2\|f - f_n\|, \quad \text{all } x \in X. \end{aligned}$$

Since $f_n \to f$ uniformly, it follows immediately that for any $\delta > 0$, there exists $M_\delta \geq 1$ such that

$$(3) \qquad 0 \leq f[x, g(x)] - f[x, g_n(x)] \leq 2\|f - f_n\| < \delta, \quad \text{all } x \in X, \text{ all } n \geq M_\delta.$$

To show that $g_n \to g$ pointwise, we must establish that for each $\varepsilon > 0$ and $x \in X$, there exists $N_x \geq 1$ such that

$$(4) \qquad \|g(x) - g_n(x)\| < \varepsilon, \quad \text{all } n \geq N_x.$$

By Lemma 3.7, it suffices to show that for any $\delta_x > 0$ and $x \in X$ there exists $N_x \geq 1$ such that

$$(5) \qquad |f[x, g(x)] - f[x, g_n(x)]| < \delta_x, \quad \text{all } n \geq N_x.$$

From (3), it follows that any $N_x \geq M_{\delta_x}$ has the required property.

Suppose X is compact. To establish that $g_n \to g$ uniformly, we must show that for each $\varepsilon > 0$ there exists $N \geq 1$ such that (4) holds for all $x \in X$. By Lemma 3.7, it suffices to show that for any $\delta > 0$, there exists $N \geq 1$, such that (5) holds for all $x \in X$. From (3) it follows that any $N \geq M_\delta$ has the required property. ■

3.4 Bibliographic Notes

For a more detailed discussion of metric spaces, see Kolmogorov and Fomin (1970, chap. 2) or Royden (1968, chap. 7). Good discussions of normed vector spaces can be found in Kolmogorov and Fomin (1970, chap. 4) and Luenberger (1969, chap. 2), both of which also treat the Contraction Mapping Theorem. Blackwell's sufficient condition is Theorem 5 in Blackwell (1965). The Theorem of the Maximum dates from Berge (1963, chap. 6), and can also be found in Hildenbrand (1974, pt. I.B). Both of these also contain excellent treatments of upper and lower hemi-continuity.