Markov Decision Processes

Discrete Stochastic Dynamic Programming

MARTIN L. PUTERMAN
University of British Columbia

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Markov Decision Processes

Markov Decision Processes

Discrete Stochastic Dynamic Programming

MARTIN L. PUTERMAN
University of British Columbia

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representation or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication is available.

ISBN 0-471-72782-2

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To my father-in-law
Dr. Fritz Katzenstein
1908–1993
who never lost his love for learning

Contents

Preface

The past decade has seen a notable resurgence in both applied and theoretical research on Markov decision processes. Branching out from operations research roots of the 1950's, Markov decision process models have gained recognition in such diverse fields as ecology, economics, and communications engineering. These new applications have been accompanied by many theoretical advances. In response to the increased activity and the potential for further advances, I felt that there was a need for an up-to-date, unified and rigorous treatment of theoretical, computational, and applied research on Markov decision process models. This book is my attempt to meet this need.

I have written this book with two primary objectives in mind: to provide a comprehensive reference for researchers, and to serve as a text in an advanced undergraduate or graduate level course in operations research, economics, or control engineering. Further, I hope it will serve as an accessible introduction to the subject for investigators in other disciplines. I expect that the material in this book will be of interest to management scientists, computer scientists, economists, applied mathematicians, control and communications engineers, statisticians, and mathematical ecologists. As a prerequisite, a reader should have some background in real analysis, linear algebra, probability, and linear programming; however, I have tried to keep the book self-contained by including relevant appendices. I hope that this book will inspire readers to delve deeper into this subject and to use these methods in research and application.

Markov decision processes, also referred to as stochastic dynamic programs or stochastic control problems, are models for sequential decision making when outcomes are uncertain. The Markov decision process model consists of decision epochs, states, actions, rewards, and transition probabilities. Choosing an action in a state generates a reward and determines the state at the next decision epoch through a transition probability function. Policies or strategies are prescriptions of which action to choose under any eventuality at every future decision epoch. Decision makers seek policies which are *optimal* in some sense. An analysis of this model includes

1. providing conditions under which there exist easily implementable optimal policies;
2. determining how to recognize these policies;
3. developing and enhancing algorithms for computing them; and
4. establishing convergence of these algorithms.

Surprisingly these analyses depend on the criterion used to compare policies. Because of this, I have organized the book chapters on the basis of optimality criterion.

The primary focus of the book is infinite-horizon discrete-time models with discrete state spaces; however several sections (denoted by *) discuss models with arbitrary state spaces or other advanced topics. In addition, Chap. 4 discusses finite-horizon models and Chap. 11 considers a special class of continuous-time discrete-state models referred to as semi-Markov decision processes.

This book covers several topics which have received little or no attention in other books on this subject. They include modified policy iteration, multichain models with average reward criterion, and sensitive optimality. Further I have tried to provide an in-depth discussion of algorithms and computational issues. The Bibliographic Remarks section of each chapter comments on relevant historical references in the extensive bibliography. I also have attempted to discuss recent research advances in areas such as countable-state space models with average reward criterion, constrained models, and models with risk sensitive optimality criteria. I include a table of symbols to help follow the extensive notation. As far as possible I have used a common framework for presenting results for each optimality criterion which

- explores the relationship between solutions to the optimality equation and the optimal value function;
- establishes the existence of solutions to the optimality equation;
- shows that it characterizes optimal (stationary) policies;
- investigates solving the optimality equation using value iteration, policy iteration, modified policy iteration, and linear programming;
- establishes convergence of these algorithms;
- discusses their implementation; and
- provides an approach for determining the structure of optimality policies.

With rigor in mind, I present results in a "theorem-proof" format. I then elaborate on them through verbal discussion and examples. The model in Sec. 3.1 is analyzed repeatedly throughout the book, and demonstrates many important concepts. I have tried to use simple models to provide counterexamples and illustrate computation; more significant applications are described in Chap. 1, the Bibliographic Remarks sections, and left as exercises in the Problem sections. I have carried out most of the calculations in this book on a PC using the spreadsheet Quattro Pro (Borland International, Scott's Valley, CA), the matrix language GAUSS (Aptech Systems, Inc., Kent, WA), and Bernard Lamond's package MDPS (Lamond and Drouin, 1992). Most of the numerical exercises can be solved without elaborate coding.

For use as a text, I have included numerous problems which contain applications, numerical examples, computational studies, counterexamples, theoretical exercises, and extensions. For a one-semester course, I suggest covering Chap. 1; Secs. 2.1 and 2.2; Chap. 3; Chap. 4; Chap. 5; Secs. 6.1, 6.2.1–6.2.4, 6.3.1–6.3.2, 6.4.1–6.4.2, 6.5.1–6.5.2, 6.6.1–6.6.7, and 6.7; Secs. 8.1, 8.2.1, 8.3, 8.4.1–8.4.3, 8.5.1–8.5.3, 8.6, and 8.8; and Chap. 11. The remaining material can provide the basis for topics courses, projects and independent study.

This book has its roots in conversations with Nico van Dijk in the early 1980's. During his visit to the University of British Columbia, he used my notes for a course on dynamic programming, and suggested that I expand them into a book. Shortly thereafter, Matt Sobel and Dan Heyman invited me to prepare a chapter on Markov decision processes for *The Handbook on Operations Research: Volume II, Stochastic Models*, which they were editing. This was the catalyst. My first version (180 pages single spaced) was closer to a book than a handbook article. It served as an outline for this book, but has undergone considerable revision and enhancement. I have learned a great deal about this subject since then, and have been encouraged by the breadth and depth of renewed research in this area. I have tried to incorporate much of this recent research.

Many individuals have provided valuable input and/or reviews of portions of this book. Of course, all errors remain my responsibility. I want to thank Hong Chen, Eugene Feinberg, and Bernard Lamond for their input, comments and corrections. I especially want to thank Laurence Baxter, Moshe Haviv, Floske Spieksma and Adam Shwartz for their invaluable comments on several chapters of this book. I am indebted to Floske for detecting several false theorems and unequal equalities. Adam used the first 6 chapters while in proof stage as a course text. My presentation benefited greatly from his insightful critique of this material. Linn Sennott deserves special thanks for her numerous reviews of Sects. 6.10 and 8.10, and I want to thank Pat Kennedy for reviewing my presentation of her research on Cooper's hawk mate desertion, and providing the beautiful slide which appears as Fig. 1.6.1. Bob Foley, Kamal Golabi, Tom McCormick, Evan Porteus, Maurice Queyranne, Matt Sobel, and Pete Veinott have also provided useful input. Several generations of UBC graduate students have read earlier versions of the text. Tim Lauck, Murray Carlson, Peter Roorda, and Kaan Katiriciougulu have all made significant contributions. Tim Lauck wrote preliminary drafts of Sects. 1.4, 1.6, and 8.7.3, provided several problems, and pointed out many inaccuracies and typos. I could not have completed this book without the support of my research assistant, Noel Paul, who prepared all figures and tables, most of the Bibliography, tracked down and copied many of the papers cited in the book, and obtained necessary permissions. I especially wish to thank the Natural Sciences and Engineering Research Council for supporting this project through Operating Grant A5527, The University of British Columbia Faculty of Commerce for ongoing support during the book's development and the Department of Statistics at The University of Newcastle (Australia) where I completed the final version of this book. My sincere thanks also go to Kimi Sugeno of John Wiley and Sons for her editorial assistance and to Kate Roach of John Wiley and Sons who cheerfully provided advice and encouragement.

Finally, I wish to express my appreciation to my wife, Dodie Katzenstein, and my children, Jenny and David, for putting up with my divided attention during this book's six year gestation period.

MARTIN L. PUTERMAN

Markov Decision Processes

CHAPTER 1

Introduction

Each day people make many decisions; decisions which have both immediate and long-term consequences. Decisions must not be made in isolation; today's decision impacts on tomorrow's and tomorrow's on the next day's. By not accounting for the relationship between present and future decisions, and present and future outcomes, we may not achieve good overall performance. For example, in a long race, deciding to sprint at the beginning may deplete energy reserves quickly and result in a poor finish.

This book presents and studies a model for sequential decision making under uncertainty, which takes into account both the outcomes of current decisions and future decision making opportunities. While this model may appear quite simple, it encompasses a wide range of applications and has generated a rich mathematical theory.

1.1 THE SEQUENTIAL DECISION MODEL

We describe the sequential decision making model which we symbolically represent in Figure 1.1.1. At a specified point in time, a decision maker, agent, or controller observes the state of a system. Based on this state, the decision maker chooses an action. The action choice produces two results: the decision maker receives an immediate reward (or incurs an immediate cost), and the system evolves to a new state at a subsequent point in time according to a probability distribution determined by the action choice. At this subsequent point in time, the decision maker faces a similar problem, but now the system may be in a different state and there may be a different set of actions to choose from.

The key ingredients of this sequential decision model are the following.

1. A set of decision epochs.
2. A set of system states.
3. A set of available actions
4. A set of state and action dependent immediate rewards or costs.
5. A set of state and action dependent transition probabilities.

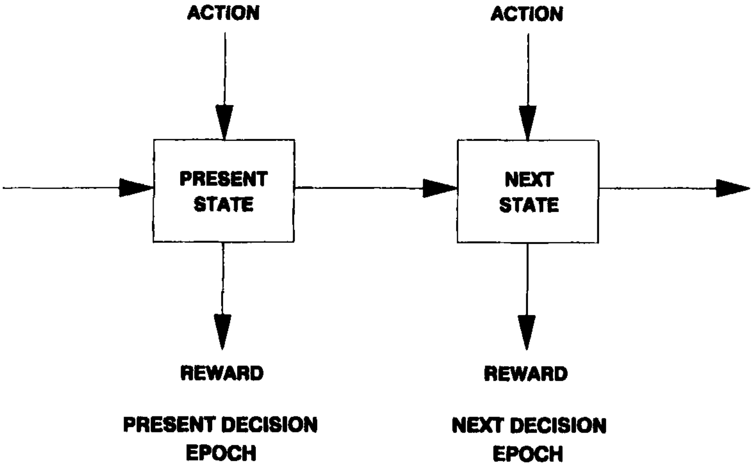

Figure 1.1.1 Symbolic representation of a sequential decision problem.

With the exception of some models which we refer to in the Afterword, we assume that all of these elements are known to the decision maker at the time of each decision.

Using this terminology, we describe the probabilistic sequential decision model as follows. At each decision epoch (or time), the system state provides the decision maker with all necessary information for choosing an action from the set of available actions in that state. As a result of choosing an action in a state, two things happen: the decision maker receives a reward, and the system evolves to a possibly different state at the next decision epoch. Both the rewards and transition probabilities depend on the state and the choice of action. As this process evolves through time, the decision maker receives a sequence of rewards.

At each decision epoch, the decision maker chooses an action in the state occupied by the system at that time. A *policy* provides the decision maker with a prescription for choosing this action in any possible future state. A *decision rule* specifies the action to be chosen at a particular time. It may depend on the present state alone or together with all previous states and actions. A policy is a sequence of decision rules. Implementing a policy generates a sequence of rewards. The sequential decision problem is to choose, prior to the first decision epoch, a policy to maximize a function of this reward sequence. We choose this function to reflect the decision maker's intertemporal tradeoffs. Possible choices for these functions include the expected total discounted reward or the long-run average reward.

This book focuses on a particular sequential decision model which we refer to as a *Markov decision process* model. In it, the set of available actions, the rewards, and the transition probabilities depend only on the current state and action and not on states occupied and actions chosen in the past. The model is sufficiently broad to allow modeling most realistic sequential decision-making problems.

We àddress the following questions in this book.

1. When does an optimal policy exist?
2. When does it have a particular form?
3. How do we determine or compute an optimal policy efficiently?

We will see that the choice of the optimality criterion and the form of the basic model elements has significant impact on the answers to these questions.

Often you see these models referred to as *dynamic programming models* or *dynamic programs*. We reserve the expression "dynamic programming" to describe an approach for solving sequential decision models based on inductive computation.

In the remainder of this chapter, we illustrate these concepts with significant and colorful applications from several disciplines. The Bibliographic Remarks section provides a brief historical review.

1.2 INVENTORY MANAGEMENT

Sequential decision models have been widely applied to inventory control problems and represent one of the earliest areas of application. The scope of these applications ranges from determining reorder points for a single product to controlling a complex multiproduct multicenter supply network. Some of the earliest and most noteworthy results in stochastic operations research concern the form of the optimal policy under various assumptions about the economic parameters. We describe an application of a model of this type.

Through local dealerships, Canadian Tire, Inc. operates a chain of automotive supply stores throughout Canada. The 21 stores in the Pacific region are operated by a single management group. Backup inventory for these 21 stores is maintained at a central warehouse in Burnaby, British Columbia. It stocks roughly 29,000 products. Periodically, inventory is delivered from the central warehouse to each of its stores to maintain target stock levels.

The timing of inventory replenishment varies with store size. At stores designated as "small," the inventory position of each product is reviewed once a week. For each product the inventory position (stock on hand) at the time of review determines the quantity, if any, to order. Orders arrive in about three days. Associated with an order for a particular product is a fixed charge associated with the time spent locating the item in the warehouse and shelving the item at the store. In addition to the fixed charge for filling the order, there is a daily carrying charge for keeping an item in inventory at a store. Management policy also dictates that at least 97.5% of demand be satisfied from stock on hand.

We now describe a sequential decision model for determining optimal reorder points and reorder levels for a single product at a single store. Decision epochs are the weekly review periods, and the system state is the product inventory at the store at the time of review. In a given state, actions correspond to the amount of stock to order from the warehouse for delivery at the store. Transition probabilities depend on the quantity ordered and the random customer demand for the product throughout the week. A decision rule specifies the quantity to be ordered as a function of the stock on hand at the time of review, and a policy consists of a sequence of such restocking functions. Management seeks a reordering policy which minimizes long-run average ordering and inventory carrying costs subject to the above constraint on the probability of being unable to satisfy customer demand.

Desirable properties for optimal policies in this setting are that they be simple to implement and not vary with time. Without the constraint on the probability of satisfying customer demand, the optimal policy may be shown to be of the following type: when the stock level falls below a certain threshold, order up to a target level;

otherwise do not order. With the inclusion of such a constraint, a policy of this form may not be optimal.

The importance of effective inventory control to effective cost management cannot be overemphasized. Sir Graham Day, chairman of Britain's Cadbury-Schweppes PLC notes (*The Globe and Mail*, October 20, 1992, p. C24):

> "I believe that the easiest money any business having any inventory can save lies with the minimization of that inventory."

The roots of sequential decision making lie in this discipline. The book by Arrow, Karlin, and Scarf (1958) provides a good overview of the foundations of mathematical inventory theory; Porteus (1991) provides a recent review.

1.3 BUS ENGINE REPLACEMENT

Markov decision process models have been applied to a wide range of equipment maintenance and replacement problems. In these settings, a decision maker periodically inspects the condition of the equipment, and based on its age or condition decides on the extent of maintenance, if any, to carry out. Choices may vary from routine maintenance to replacement. Costs are associated with maintenance and operating the equipment in its current status. The objective is to balance these two cost components to minimize a measure of long-term operating costs.

Howard (1960) provided a prototype for such models with his "automobile replacement problem." In it, an individual periodically decides whether or not to trade in an automobile and, if so, with what age automobile to replace it. Subsequently, many variants of this model have been studied and analyzed. In this section and the next, we describe two applications of such models.

Rust (1987) formulates and analyzes the following problem. Harold Zurcher, superintendent of maintenance at the Madison (Wisconsin) Metropolitan Bus Company, has the responsibility of keeping a fleet of buses in good working condition. One aspect of the job is deciding when to replace the bus engines.

Zurcher's replacement problem may be formulated as a Markov decision process model as follows. Replacement decisions are made monthly and the system state represents the accumulated engine mileage since the last replacement. Costs include an age-dependent monthly operating cost and a replacement cost. The monthly operating costs include a routine operating and maintenance cost component and an unexpected failure cost component. The failure cost accounts for the probability of breakdown for a bus of a given age and costs associated with towing, repair, and lost goodwill. If Zurcher decides to replace an engine, then the company incures a (large) replacement cost and, subsequently, the routine maintenance and operating cost associated with the replacement engine. Transition probabilities describe changes in accumulated mileage and the chance of an unplanned failure for a bus engine of a particular age. For each engine, Zurcher seeks an age-dependent replacement policy to minimize expected total discounted or long-run average costs.

The algorithms in Chaps. 4, 6, or 9 can be used to compute such an optimal policy for Harold Zurcher. However, the theory shows that, under reasonable assumptions, an optimal policy has a particularly simple and appealing form; at the first monthly inspection at which the mileage exceeds a certain level, referred to as a *control limit*, the engine must be replaced; otherwise it is not. Rust examines whether Zurcher adopts such a policy using data from the Madison Metropolitan Bus Company.

Operating, maintenance, and replacement costs vary with engine type. The table below summarizes Zurcher's data on replacement costs and average mileage at replacement for two main engine types.

Engine Type	Replacement Cost	Average Mileage at Replacement
1979 GMC T8H203	$9499	199,733
1975 GMC 5308A	$7513	257,336

This data shows that, although the replacement cost for a 1979 engine exceeded that of a 1975 engine by $2000, Zurcher decided to replace the 1979 engines 57,600 miles and 14 months earlier than the 1975 engines. This suggests that routine maintenance and operating costs differ for these two engine types and that they increase faster with mileage in the 1979 engines. Rust's analysis of the data suggests that these costs may be modeled by linear or "square-root" functions of age.

Further data suggests that Zurcher's decisions departed from a simple control limit policy. Between 1974 and 1985, 27 T8H203 engines and 33 5308A engines were replaced. The mileage at replacement varied from 124,800 to 273,400 for the T8H203 engine and between 121,200 and 387,300 for the 5308A engine. Thus we might infer that Zurcher is making his decisions suboptimally. Rust adopts a different viewpoint. He hypothesizes that Zurcher's decisions coincide with an optimal policy of a Markov decision process model; however, Zurcher takes into account many measurements and intangibles that are not known by the problem solver. In his extensive paper, Rust (1987) provides an approach for accounting for these factors, estimating model parameters, and testing this hypothesis. He concludes that, after taking these unobservables into account, Zurcher's behavior is consistent with minimizing long-run average operating cost.

1.4 HIGHWAY PAVEMENT MAINTENANCE

The Arizona Department of Transportation (ADOT) manages a 7,400 mile road network. Up to the mid 1970s its primary activity was construction of new roadways. As the Arizona roadway system neared completion, and because of changing federal guidelines, ADOT's emphasis shifted in the late 1970's to maintaining existing roads. Between 1975 and 1979, highway preservation expenditures doubled from $25 million to $52 million, and evidence suggested that such an increase would continue. By this time it was evident to ADOT management that a systematic centralized procedure for allocation of these funds was needed. In 1978, in conjunction with Woodward-Clyde Consultants of San Francisco, ADOT developed a pavement management system based on a Markov decision process model to improve allocation of its limited resources while ensuring that the quality of its roadways was preserved. In 1980, the first year of implementation, this system saved $14 million, nearly a third of Arizona's maintenance budget, with no decline in road quality. Cost savings over the next four years were predicted to be $101 million. Subsequently, this model was modified for use in Kansas, Finland, and Saudi Arabia. Related models have been developed for bridge and pipeline management. In this section, we describe the Arizona pavement management model. We base our presentation on Golabi, Kulkarni, and Way (1982), and additional information provided by Golabi in a personal communication.

The pavement management system relies on a dynamic long-term model to identify maintenance policies which minimize long-run average costs subject to constraints on road quality. To apply the model, the Arizona highway network was divided into 7,400 one-mile sections and nine subnetworks on the basis of road type, traffic density, and regional environment. For each category, a dynamic model was developed that specified the conditions of road segments, maintenance actions that could be used under each condition, and the expected yearly deterioration or improvement in pavement conditions resulting from each such action. In addition, costs associated with each maintenance action were determined. Developing categories for system states, actions, costs, and the state-to-state dynamics under different actions was a nontrivial task requiring data, models of road conditions, statistical analysis, and subject matter expertise.

We describe the management model for asphalt concrete highways; that for Portland cement concrete roadways had different states and actions. Decisions were made annually. The system state characterized the pavement condition of a one-mile segment by its roughness (three levels), its percentage of cracking (three levels), the change in cracking from the previous year (three levels), and an index which measured the time since the last maintenance operation and the nature of the operation (five levels). Consequently, a road segment could be described by one of 135 (3 × 3 × 3 × 5) possible states, but, since some combinations were not possible, 120 states were used.

Actions corresponded to available pavement rehabilitation activities. These ranged from relatively inexpensive routine maintenance to costly actions such as thick resurfacing or recycling of the entire roadway. A list of possible actions and associated construction costs appear in Table 1.4.1 below. For each state, however, only about six of the actions were considered feasible.

Costs consisted of the action-dependent construction costs (Table 1.4.1) and annual routine maintenance costs (Table 1.4.2). Annual routine maintenance costs varied with the road condition and rehabilitation action. When only routine maintenance was carried out, these costs varied with the roughness and degree of cracking of the road segment; when a seal coat was applied, these costs varied only with roughness; and if any other rehabilitation action was taken, maintenance costs were independent of previous road condition. These costs were determined through a regression model based on existing data.

Transition probabilities specify the likelihood of yearly changes in road condition under the various maintenance actions. These were estimated using existing data, under the assumption that each dimension of the state description varied independently. Since in each state only a limited number of subsequent states could occur, most of the transition probabilities (97%) were zero.

The performance criteria was cost minimization subject to constraints on the proportion of roads in acceptable and unacceptable states. For example, ADOT policy requires that at least 80% of high traffic roadways must have a roughness level not exceeding 165 inches/mile, while at most 5% of these roads could have roughness exceeding 256 inches/mile. Similar constraints applied to levels of cracking.

Table 1.4.1 Rehabilitation Actions and Construction Costs

Action Index	Action Description[a]	Construction Cost $/yd^2
1	Routine Maintenance	0
2	Seal Coat	0.55
3	ACFC	0.75
4	ACFC + AR	2.05
5	ACFC + HS	1.75
6	1.5 inch AC	1.575
7	1.5 inch AC + AR	2.875
8	1.5 inch AC + HS	2.575
9	2.5 inch AC	2.625
10	2.5 inch AC + AR	3.925
11	2.5 inch AC + HS	3.625
12	3.5 inch AC	3.675
13	3.5 inch AC + AR	4.975
14	3.5 inch AC + HS	4.675
15	4.5 inch AC	4.725
16	5.5 inch AC	5.775
17	Recycling (equivalent to 6 inch AC)	6.3

[a] Abbreviations used in table: ACFC-Asphalt concrete fine coat, AR-Asphalt Rubber, HS-Heater Scarifier, AC-Asphalt concrete

Table 1.4.2 Annual Routine Maintenance Costs

State After Rehabilitation Action		Rehabilitation	Cost
Roughness (in/mile)	Percentage of Cracking	Action[a]	$/yd^2
120 (±45)	5 (±5)	RM	0.066
120 (±45)	20 (±10)	RM	0.158
120 (±45)	45 (±15)	RM	0.310
120 (±45)	Any	SC	0.036
210 (±45)	5 (±5)	RM	0.087
210 (±45)	20 (±10)	RM	0.179
210 (±45)	45 (±15)	RM	0.332
210 (±45)	Any	SC	0.057
300 (±45)	5 (±5)	RM	0.102
300 (±45)	20 (±10)	RM	0.193
300 (±45)	45 (±15)	RM	0.346
300 (±45)	Any	SC	0.071
Any	Any	OT	0.036

[a] Action Abbreviations; RM-routine maintenance, SC-seal coat, OT-any other

This model is an example of a *constrained* average reward Markov decision process model and can be solved using the linear programming methodology in Chaps. 8 and 9. This model was designed not only to yield a single solution but also to interactively examine the consequences of regulatory policies and budget changes. Examples of solutions are too lengthy to be presented here, but one aspect of the solution is worth noting. Because of the addition of constraints, the optimal policy may be *randomized.* This means that in some states, it may be optimal to use a chance mechanism to determine the course of action. For example, if the road segment is cracked, 40% of the time it should be resurfaced with one inch of asphalt concrete (AC) and 60% of the time with two inches of AC. This caused no difficulty because the model was applied to individual one-mile road segments so that this randomized policy could be implemented by repairing 40% of them with one inch of AC and 60% with two inches of AC. Also, in a few instances, the model recommended applying a different maintenance action to a road segment than to its two adjacent segments. In such cases the solution was modified to simplify implementation yet maintain the same level of overall cost and satisfy road quality constraints.

In addition to producing significant cost reductions, the model showed that

> "...corrective actions in the past were too conservative; it was common to resurface a road with five inches of asphalt concrete.... The policies recommended by the pavement management system...are less conservative; for example, a recommendation of three inches of overlay is rather rare and is reserved for the worst conditions. (Golabi, Kulkarni, and Way, 1982, p. 16)."

Observations such as this are consistent with the findings of many operations research studies. For example, preliminary results in the inventory control study described in Sect. 1.2 suggest that current in store inventory levels are 50% too high.

1.5 COMMUNICATIONS MODELS

A wide range of computer, manufacturing, and communications systems can be modeled by networks of interrelated queues (waiting lines) and servers. Efficient operation of these systems leads to a wide range of dynamic optimization problems. Control actions for these systems include rejecting arrivals, choosing routings, and varying service rates. These decisions are made frequently and must take into account the likelihood of future events to avoid congestion.

These models are widely applied and have had significant impact as noted by the following article in *The New York Times*, May 12, 1992, p. C2.

"More Dial Mom Than Expected"

> Even greater numbers of people called their mothers on Mother's Day than AT&T had expected.
>
> ...A call-routing computer technique enabled the American Telephone and Telegraph Company to complete more calls than last year, when it logged 93.4 million calls.

> On Sunday, there were about 1.5 million uncompleted calls, where customers got a recorded announcement advising them to call later, compared with 3.9 million last year.
>
> A new computer technique called real-time network routing helped AT&T shepherd a larger number of calls through the labyrinth of telephone computers known as switches. By creative zigzagging around the country, AT&T could direct calls so that they were more likely to avoid congestion, especially in suburbs, which do not have the high-capacity telephone lines that big cities do.

We now describe a sequential decision process model for a particular communication system. Many packet communications systems are configured so that multiple terminals generating low rate, bursty traffic and must share a single channel to communicate with each other or with a central hub (Fig. 1.5.1).

This system architecture is typical of satellite broadcast networks where multiple earth stations communicate over the same radio frequency, and computer local area networks (LAN's) where many computers send job requests to a central file server over a single coaxial cable. Since a single channel may only carry one stream of traffic, the problem arises as to how to coordinate the traffic from the terminals to make the most efficient use of the channel.

The *Slotted ALOHA Protocol* is a popular and especially simple technique for providing such coordination. We describe the slotted ALOHA channel model and the mechanism by which it controls channel access. Stations communicate over a slotted ALOHA channel through equal-length packets of data. Time on the channel is divided into slots of the same length as the packets, and all terminals are synchronized so that packet transmissions always begin at the leading edge of a time slot and occupy exactly one slot. New packets are randomly generated at any idle terminal during a slot, and are transmitted in the following slot. If no other stations transmit a packet in that slot, the transmission is considered successful and the terminal returns to idle mode. If more than one terminal generates a packet, a collision occurs, the data become garbled, and the station goes into retransmission mode and must retransmit the packet in a future slot. If a collision occurs and all involved terminals always retransmit in the next slot, collisions will continue endlessly. To avoid this situation, the slotted ALOHA protocol specifies that stations in retransmission mode transmit in the next slot with a specified retransmission probability, thus achieving a random backoff between retransmission attempts. When a terminal successfully retransmits the packet, it returns to idle mode and waits for a new packet to be

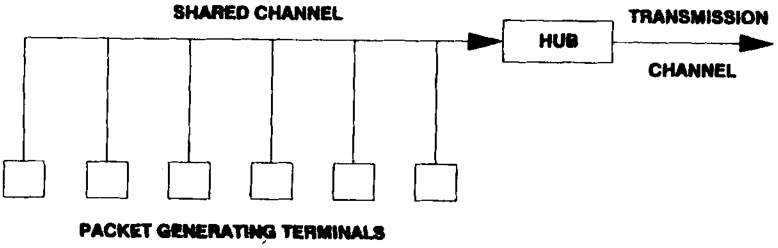

Figure 1.5.1 Multiple access channel configuration.

generated. We see then that, although the slotted ALOHA protocol does not avoid collisions on the channel, the use of a random retransmission backoff provides a scheme for effective contention resolution among terminals.

Since the message generating probability is fixed, the only means available to control channel access within this model is by regulating the retransmission probability. If it is held constant and the number of terminals in retransmission mode becomes large, the probability of a collision in the next slot will also become large. As the collisions become more frequent, newly arriving packets tend to become backlogged, increasing the number of terminals in retransmission mode. Thus, with a fixed retransmission probability, the system is prone to become highly congested, reducing the chance of a successful transmission to close to zero. This instability may be alleviated (for certain values of the packet generation probability) by taking into account the current number of terminals in retransmission mode when choosing a retransmission probability.

We now describe a Markov decision process model for this control problem. Decision epochs correspond to time slots, and the system state is the number of terminals in retransmission mode. Actions correspond to choosing a retransmission probability. The system generates a reward of one unit for each packet successfully transmitted, and transition probabilities combine the probabilities that new packets are generated in a time slot and a successful packet transmission occurs when the retransmission probability has been set at a particular level. The objective is to choose a retransmission probability-setting policy which maximizes the long-run average expected channel throughput (rate of successful packets per slot).

Feinberg, Kogan, and Smirnov (1985) show that the optimal retransmission probability is a monotonically decreasing function of the system state whenever the mean packet arrival rate (number of terminals times packet generation probability) is less than one. If this rate exceeds one, the system will become congested and other optimality criteria may be used. This control policy agrees with intuition in that the system will react to increasing congestion by decreasing retransmission probabilities and thus maintaining a reasonable probability of successful packet transmission.

In practical applications, the number of stations in retransmission mode and the packet generation probability are rarely known. They must be estimated on the basis of the history of channel observations (idle, successful, and collision slots). In this case, incorporation of both state and parameter estimation into the Markov decision process model is necessary to find the optimal retransmission policy. We provide references for models of this type in the Afterword.

1.6 MATE DESERTION IN COOPER'S HAWKS

Markov decision process models are becoming increasingly popular in behavioral ecology. They have been used in a wide range of contexts to gain insight into factors influencing animal behavior. Examples include models of social and hunting behavior of lions (Clark, 1987; Mangel and Clark, 1988), site selection and number of eggs laid by apple maggots and medflys (Mangel, 1987), daily vertical migration of sockeye salmon and zooplankton (Levy and Clark, 1988; Mangel and Clark 1988), changes in mobility of spiders in different habitats (Gallespie and Caraco, 1987), and singing versus foraging tradeoffs in birds (Houston and McNamara, 1986).

The theory of natural selection suggests that organisms predisposed to behavioral characteristics that allow them to adapt most efficiently to their environment have the greatest chance of reproduction and survival. Since any organism alive today has a substantial evolutionary history, we might infer that this organism has adopted optimal or near-optimal survival strategies which can be observed in day-to-day activity.

Models have been based on regarding the behavior of an organism as its reaction or response to its environment, conditional on its state of well being. Throughout its life, it makes behavioral choices which affect its chances of survival and successful reproduction. Investigators have used probabilistic sequential decision process models to determine state- and time-dependent strategies which maximizes a function of its survival and reproductive success probabilities and then compared model results to observed behavior. If there is "reasonable agreement," then the derived optimal policy may provide insight into the behavioral strategy of the organism.

We describe Kelly and Kennedy's (1993) use of this methodology in their study of mate desertion in Cooper's hawks (Acceipiter cooperii). Over a five-year period, they studied nesting behavior of several birds near Los Alamos National Laboratory in north-central New Mexico (Fig. 1.6.1.) They observed that more than 50% of the females deserted their nests before the young reached independence, and noted that the male of this species continued to feed the young regardless of whether or not a female was present. At issue was determining factors that influenced the female's decision to desert and the female's tradeoffs between her survival and that of her offspring.

In the study, the physical conditions of both the nestlings (young birds) and the female were monitored, assisted by the use of radiotelemetry. Females were trapped and tagged early in the breeding season, providing an opportunity to assess the initial health of the females. Birds with greater body mass had larger energy reserves and were considered healthier. Rather than disturb the nestlings, their health was determined by assuming that nestlings were initially healthy. A developmental model was

Figure 1.6.1 A female Cooper's hawk and her brood. (Photograph courtesy of Patricia L. Kennedy.)

used to account for parental hunting behavior, captures, and nestling growth rates.

Kelly and Kennedy developed a model for a single nesting season based on the assumption that behavioral choices of the female hawk maximized a weighted average of the probability of nestling survival and the probability of the female's survival to the next breeding season. The sequential decision model was used to determine an optimal behavioral strategy.

The nesting season was divided into four periods representing basic stages of development of the young.

1. Early nestling period.
2. Late nestling period.
3. Early fledgling dependence period.
4. Late fledgling dependence period.

The end of the late fledgling period marked the point at which the brood reaches independence.

The system state is a two-dimensional health index representing female and brood energy reserves. The states were constrained to lie between lower levels which represented the minimum physical condition for survival, and upper levels corresponding to limiting physical attributes of the birds.

Three basic behavioral strategies were observed for the female.

1. Stay at the nest to protect the young.
2. Hunt to supplement the food supplied by the male.
3. Desert the nest.

Decisions were assumed to have been made at the start of each of the above periods.

From one developmental stage to the next, the change in energy reserves of both the female and the young depends on the female's behavioral strategy and the amount of food captured, a random quantity. At the time of independence, the female's and brood's states of health depend on their initial energy reserves, the female's behavior, and the availability of food. The respective health indices at the end of the decision making period determine the probability of survival of the female and of the brood to the subsequent nesting period.

Using data estimated from the five-year study, results in the literature, and some intelligent guesswork, Kelly and Kennedy determined transition probabilities for the above model. They then solved the model to determine the optimal policy using inductive methods we describe in Chap. 4. Figure 1.6.2, which we adopt from their paper, shows the optimal policy under a specified degree of tradeoff between female and brood survival.

The four graphs show the optimal behavioral action as a function of the health of the female and the brood at each of the four decision periods. The vertical axis represents the female's health index and the horizontal axis represents the brood's health index. Low values indicate states of low-energy reserve.

Observe that, in all periods, if both the female's and the brood's health index exceed 4, the optimal strategy for the female is to stay at the nest and protect the young. At the other extreme, if the health index of both the female and the brood is at its lowest value, the optimal strategy for the female is to desert the nest.

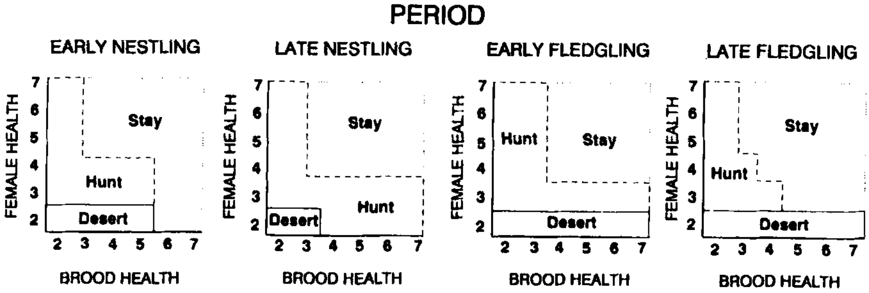

Figure 1.6.2 Symbolic representation of optimal desertion strategy under a specific tradeoff parameter choice. Quantities on axes denote health indices. (Adapted from Kelly and Kennedy, 1993.)

There are other patterns to this optimal strategy. For a fixed value of the brood health index, as the female's health index increases the optimal strategy changes from desert to hunt to stay. Similarly, if the female's health index is fixed, as the brood's energy reserves increase the strategy changes from desert to hunt to stay. Thus there is a form of monotonicity in the optimal strategy. One might conjecture that the behavioral strategy of the female will have this form under any parameter values. Observing such patterns can sometimes yield insight into theoretical results beyond a specific numerical scenario. In subsequent chapters, we will provide methods for identifying models in which optimal strategies have a particular form.

Kelly and Kennedy (p. 360–361) conclude

> "The agreement of model predictions and observed strategies supported, but did not prove, the modelling hypotheses that:
>
> 1. a female's strategy during brood rearing maximizes the weighted average of the expected probability of survival of her current offspring and her future reproductive potential, and
> 2. the female's strategy choices were influenced by multiple factors including her state, the state of her brood, the risks to nestlings associated with each strategy, and the male's and female's foraging capabilities.
>
> ...dynamic state variable models are powerful tools for studying the complexities of animal behavior from an evolutionary standpoint because they lead to quantitative testable predictions about behavioral strategies."

1.7 SO WHO'S COUNTING

Games of chance and strategy provide natural settings for applying sequential decision models. Dubins and Savage (1965) in their monograph *How to Gamble if You Must* developed a model for gambling, not unlike the sequential decision model herein, and developed a rich mathematical theory for analyzing it. Their basic observation was that, even in an unfair game, some betting strategies might be better than others. Markov decision process models apply to such games of chance and also to a wide range of board and computer games. In this section we show how such a

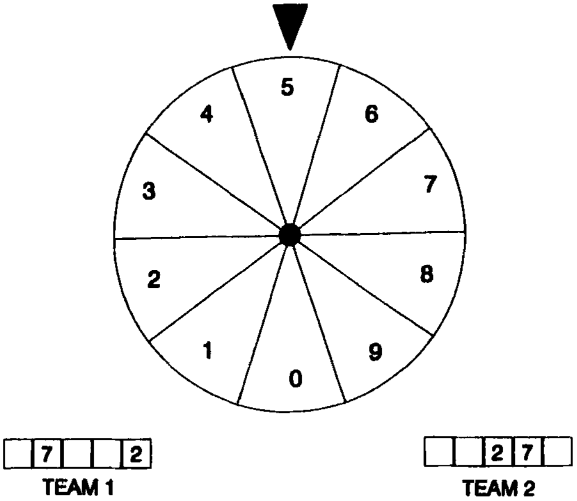

Figure 1.7.1 Spinner for "So Who's Counting."

model can be used to determine an optimal strategy in a challenging yet easy to describe television game show.

The mock game show *But Who's Counting* appeared on *Square One*, a mathematically oriented educational television program on the U.S. public broadcasting network. The game is played as follows. There are two teams of players. At each of five consecutive rounds of the game, a spinner (Fig. 1.7.1) produces a number between 0 and 9, each with equal probability. After each spin, the teams select an available digit of a five-digit number to place the number produced by the spinner. The team which generates the largest number wins the game.

Figure 1.7.1 illustrates the status of the game immediately following the third spin. At the two previous spins, the numbers 2 and 7 had appeared. At this point in time, the order in which they previously appeared is immaterial. Team 1 placed the 7 in the 1,000s place and the 2 in 1s place, while team 2 placed the 2 in the 100s place and the 7 in the 10s place. Each team must now decide where to place the 5. What would you do in this case if you were on team 1? On team 2?

Now ignore the competitive aspect of this game and suppose you were in the game alone with the objective of obtaining the highest five-digit number. Reflect on what strategy you would use. If you observed a 9, surely you would want to get the best out of it, and, thus, place it in the highest digit available. However, what would you do if you had a 5 or a 6?

We formulate the single-player game as a sequential decision problem. Decision epochs correspond to the instant immediately after the spinner identifies a number. We take as the system state the locations of the unoccupied digits and the number which has appeared on the spinner. Actions correspond to placing the number into one of the available digits and the reward equals the number times the place value for that digit. The objective is to choose a digit-placing strategy which maximizes the expected value of the five-digit number.

We can use the methods of Chap. 4 directly to derive an optimal policy. It has the property that the decision into which unoccupied digit to place the observed number

Table 1.7.1 Optimal Policy for "But Who's Counting."

Observed Number	Optimal Digit Locations				
	Spin 1	Spin 2	Spin 3	Spin 4	Spin 5
0	5	4	3	2	1
1	5	4	3	2	1
2	5	4	3	2	1
3	4	3	3	2	1
4	3	3	2	2	1
5	3	2	2	1	1
6	2	2	1	1	1
7	1	1	1	1	1
8	1	1	1	1	1
9	1	1	1	1	1

should be based on the number of unoccupied positions remaining, and not on their place values or the values of the previously placed digits. This observation enables us to summarize succinctly the optimal policy as in Table 1.7.1.

In this table, the entries represent the location of the unoccupied digit (counting from the left) into which to place the observed number. For example, consider the decision faced by a player on team 1 in Fig. 1.7.1 after observing a 5 on spin 3. Table 1.7.1 prescribes that the 5 should be placed in the 100's position. To proceed optimally, a player on team 2 should place the 5 in the 1000's position.

Furthermore, using methods in Chap. 4, we can show that using this policy yields an expected score of 78,734.12, compared to that of a random-digit choice policy which would result in an expected score of 49,999.5. Of course, in a particular game, this strategy may not always yield the greatest score, but in the long run, it will do best on average.

This problem is a special case of a *sequential assignment problem*. Ross (1983, p. 124) provides a clever approach for solving these problems in general. He establishes existence of and gives a method for computing a set of *critical levels* which in this context determine the optimal placement of the number. If the number is above the highest critical level, then it should be placed in the leftmost digit available. If it is between the highest and second highest, it should be placed in the second-leftmost unoccupied digit, and so on. His approach shows that the optimal policy would still be the same if we had any other increasing values for the contribution of digits to the total reward instead of 1, 10, 100, 1000, or 10,000.

It is not hard to think of variations of this problem. We may view it from a game theoretic point of view in which the objective is to derive a strategy which maximizes the probability of winning the game, or we may consider a single-person game in which the numbers have unequal probabilities.

HISTORICAL BACKGROUND

The books by Bellman (1957) and Howard (1960) popularized the study of sequential decision processes; however, this subject had earlier roots. Certainly some of the basic concepts date back to the calculus of variations problems of the 17th century. Cayley's paper (Cayley, 1875), which did not resurface until the 1960s, proposed an interesting problem which contains many of the key ingredients of a stochastic sequential decision problem. We describe and analyze this problem in detail in Chaps. 3 and 4.

The modern study of stochastic sequential decision problems began with Wald's work on sequential statistical problems during the Second World War. Wald embarked on this research in the early 1940's, but did not publish in until later because of wartime security requirements. His book (1947) presents the essence of this theory.

Pierre Massé, director of 17 French electric companies and minister in charge of French electrical planning, introduced many of the basic concepts in his extensive analysis of water resource management models (1946). Statistician Lucien Le Cam (1990), reflecting on his early days at Electricité de France, noted

> "Massé had developed a lot of mathematics about programming for the future. What had become known in this country (the United States) as "dynamic programming," invented by Richard Bellman, was very much alive in Massé's work, long before Bellman had a go at it."

A description of Massé's reservoir management model appears in Gessford and Karlin (1958).

Arrow (1958, p. 13), in his colorful description of the economic roots of the dynamic stochastic inventory model, comments

> "... it was Wald's work (rather than Massé's, which was unknown in this country at the time) which directly led to later work in multi-period inventory."

A precise time line with proper antecedants is difficult to construct. Heyman and Sobel (1984, p. 192) note

> "The modern foundations were laid between 1949 and 1953 by people who spent at least part of that period as staff members at the RAND Corporation in Santa Monica, California. Dates of actual publication are not reliable guides to the order in which ideas were discovered during this period."

Investigators associated with this path breaking work include Arrow, Bellman, Blackwell, Dvoretsky, Girschik, Isaacs, Karlin, Kiefer, LaSalle, Robbins, Shapley, and Wolfowitz. Their work on games (Bellman and Blackwell, 1949; Bellman and LaSalle, 1949; Shapley, 1953), stochastic inventory models (Arrow, Harris, and Marschak,1951; Dvoretsky, Kiefer, and Wolfowitz, 1952), pursuit problems (Isaacs, 1955, 1965) and sequential statistical problems (Arrow, Blackwell, and Girshick, 1949; Robbins, 1952; Kiefer, 1953) laid the groundwork for subsequent developments.

Bellman in numerous papers identified common ingredients to these problems and through his work on functional equations, dynamic programming, and the principle of optimality, became the first major player. Bellman (1954) contains a concise presentation of many of his main ideas and a good bibliography of early work. His 1957 book contains numerous references to his own and other early research and is must reading for all investigators in the field. Karlin (1955) recognized and began studying the rich mathematical foundations of this subject.

CHAPTER 2

Model Formulation

This chapter introduces the basic components of a Markov decision process and discusses some mathematical and notational subtleties. Chapters 1 and 3 contain many examples of Markov decision processes. We encourage you to refer to those examples often to gain a clear understanding of the Markov decision process model. Section 2.2 illustrates these concepts and their interrelationship in the context of a one-period model.

A Markov decision process model consists of five elements: decision epochs, states, actions, transition probabilities, and rewards. We describe these in detail below.

2.1 PROBLEM DEFINITION AND NOTATION

A decision maker, agent, or controller (who we refer to as he with no sexist overtones intended) is faced with the problem, or some might say, the opportunity, of influencing the behavior of a probabilistic system as it evolves through time. He does this by making decisions or choosing actions. His goal is to choose a sequence of actions which causes the system to perform optimally with respect to some predetermined performance criterion. Since the system we model is ongoing, the state of the system prior to tomorrow's decision depends on today's decision. Consequently, decisions must not be made myopically, but must anticipate the opportunities and costs (or rewards) associated with future system states.

2.1.1 Decision Epochs and Periods

Decisions are made at points of time referred to as *decision epochs*. Let T denote the set of decision epochs. This subset of the non-negative real line may be classified in two ways: as either a discrete set or a continuum, and as either a finite or an infinite set. When discrete, decisions are made at all decision epochs. When a continuum, decisions may be made at

1. all decision epochs (continuously),
2. random points of time when certain events occur, such as arrivals to a queueing system, or
3. opportune times chosen by the decision maker.

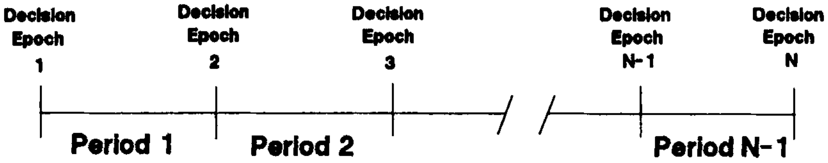

Figure 2.1.1 Decision Epochs and Periods.

When decisions are made continuously, the sequential decision problems are best analyzed using control theory methods based on dynamic system equations.

In discrete time problems, time is divided into *periods* or *stages*. We formulate models so that a decision epoch corresponds to the beginning of a period (see Fig. 2.1.1). The set of decision epochs is either finite, in which case $T \equiv \{1, 2, \ldots, N\}$ for some integer $N < \infty$, or infinite, in which case $T \equiv \{1, 2, \ldots\}$. We write $T = \{1, 2, \ldots, N\}$, $N \leq \infty$ to include both cases. When T is an interval, we denote it by either $T = [0, N]$ or $T = [0, \infty)$. Elements of T (decision epochs) will be denoted by t and usually referred to as "time t." When N is finite, the decision problem will be called a *finite horizon* problem; otherwise it will be called an *infinite horizon* problem. Most of this book will focus on infinite horizon models. We adopt the convention that, in finite horizon problems, decisions are *not* made at decision epoch N: we include it for evaluation of the final system state. Consequently, the last decision is made at decision epoch N-1. Frequently we refer to this as an N-1 period problem.

The primary focus of this book will be models with discrete T. A particular continuous time model (a semi-Markov decision process) will be discussed (Chapter 11).

2.1.2 State and Action Sets

At each decision epoch, the system occupies a *state*. We denote the set of possible system states by S. If, at some decision epoch, the decision maker observes the system in state $s \in S$, he may choose action a from the set of allowable actions in state s, A_s. Let $A = \bigcup_{s \in S} A_s$ (Fig. 2.1.2.) Note we assume that S and A_s do not vary with t. We expand on this point below.

The sets S and A_s may each be either

1. arbitrary finite sets,
2. arbitrary countably infinite sets,
3. compact subsets of finite dimensional Euclidean space, or
4. non-empty Borel subsets of complete, separable metric spaces.

In nondiscrete settings, many subtle mathematical issues arise which, while interesting, detract from the main ideas of Markov decision process theory. We expand on such issues in Section 2.3 and other sections of this book. These more technical sections will be indicated by asterisks. Otherwise, we assume that S and A_s are *discrete* (finite or countably infinite) unless explicitly noted.

Actions may be chosen either randomly or deterministically. Denote by $\mathscr{P}(A_s)$ the collection of probability distributions on (Borel) subsets of A_s and by $\mathscr{P}(A)$ the set of probability distributions on (Borel) subsets of A. (We may regard $q(\cdot) \in \mathscr{P}(A_s)$ as an