Universitext

Nicole Bäuerle
Ulrich Rieder

Markov Decision Processes with Applications to Finance

Springer

Markov Decision Processes
with Applications to Finance

Universitext

Universitext is a series of textbooks that presents material from a wide variety of mathematical disciplines at master's level and beyond. The books, often well class-tested by their author, may have an informal, personal even experimental approach to their subject matter. Some of the most successful and established books in the series have evolved through several editions, always following the evolution of teaching curricula, to very polished texts.

Thus as research topics trickle down into graduate-level teaching, first textbooks written for new, cutting-edge courses may make their way into *Universitext*.

For further volumes:
www.springer.com/series/223

Nicole Bäuerle • Ulrich Rieder

Markov Decision Processes with Applications to Finance

Springer

Nicole Bäuerle
Institute for Stochastics
Karlsruhe Institute of Technology
76128 Karlsruhe
Germany
nicole.baeuerle@kit.edu

Ulrich Rieder
Institute of Optimization
and Operations Research
University of Ulm
89069 Ulm
Germany
ulrich.rieder@uni-ulm.de

ISBN 978-3-642-18323-2 e-ISBN 978-3-642-18324-9
DOI 10.1007/978-3-642-18324-9
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011929506

Mathematics Subject Classification (2010): 90C40, 93E20, 60J05, 91G10, 93E35, 60G40

Cover design: deblik

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Für Rolf, Katja und Hannah.
Für Annika, Alexander und Katharina.

Preface

Models in mathematical finance, for example stock price processes, are often defined in continuous-time. Hence optimization problems like consumption-investment problems lead to stochastic control problems in continuous-time. However, only a few of these problems can be solved explicitly. When numerical methods have to be applied, it is sometimes wise to start with a process in discrete-time, as done for example in the *approximating Markov chain approach*. The resulting optimization problem is then a *Markov Decision Problem* and there is a rich toolbox available for solving these kind of problems theoretically and numerically.

The purpose of this book is to present selected parts of the theory of *Markov Decision Processes* and show how they can be applied in particular to problems in finance and insurance. We start by explaining the theory for problems with finite time horizon. Since we have financial applications in mind and since we do not want to restrict to binomial models we have to work with Borel state and action spaces. This framework is also needed for studying *Partially Observable Markov Decision Processes* and *Piecewise Determinsitic Markov Decision Processes*. In contrast to the case of a discrete (finite or countable) state space the theory is more demanding since non-trivial measurability problems have to be solved. However, we have decided to circumvent these kind of problems by introducing a so-called *structure assumption* for the model. The advantage is that in applications this structure assumption is often easily verified and avoids some of the technicalities. This makes the book accessible to readers who are not familiar with general probability and measure theory. Moreover, we present numerous different applications and show how this structure assumption can be verified. Applications range from consumption-investment problems, mean-variance problems, dividend problems in risk theory to indifference pricing and pricing of American options, just to name a few. The book is unique in the presentation and collection of these financial applications. Some of them appear for the first time in a book.

We also consider the theory of infinite horizon *Markov Decision Processes* where we treat so-called *contracting* and *negative* Markov Decision Problems in a unified framework. *Positive* Markov Decision Problems are also presented as well as *stopping problems*. A particular focus is on problems with *partial observation*. These kind of problems cover situations where the decision maker is not able to observe all driving factors of the model. Special cases are Hidden Markov Models and Bayesian Decision Problems. They include statistical aspects, in particular *filtering theory* and can be solved by so-called *filtered Markov Decision Processes*. Moreover *Piecewise Deterministic Markov Decision Processes* are discussed and we give recent applications to finance.

It is our aim to present the material in a mathematically rigorous framework. This is not always easy. For example, the last-mentioned problems with partial observation need a lot of definitions and notation. However each chapter on theory is followed by a chapter with applications and we give examples throughout the text which we hope shed some light on the theory. Also at the end of each chapter on theory we provide a list of exercises where the reader can test her knowledge.
Having said all this, not much general probability and optimization theory is necessary to enjoy this book. In particular we do not need the theory of stochastic calculus which is a necessary tool for continuous-time problems. The reader should however be familiar with concepts like *conditional expectation* and *transition kernels*. The only exception is Section 2.4 which is a little bit more demanding. Special knowledge in finance and insurance is not necessary. Some fundamentals are covered in the appendix. As outlined above we provide an example-driven approach. The book is intended for graduate students, researchers and practitioners in mathematics, finance, economics and operations research. Some of the chapters have been tried out in courses for masters students and in seminars.
Last but not least we would like to thank our friends and colleagues Alfred Müller, Jörn Sass, Manfred Schäl and Luitgard Veraart who have carefully read parts of an earlier version and provided helpful comments and suggestions. We are also grateful to our students Stefan Ehrenfried, Dominik Joos and André Mundt who gave significant input and corrected errors, as well as to the students at Ulm University and KIT who struggled with the text in their seminars. Special thanks go to Rolf Bäuerle and Sebastian Urban for providing some of the figures.

Bretten and Ulm, *Nicole Bäuerle*
September 2010 *Ulrich Rieder*

Contents

List of Symbols

Markov Decision Model (non-stationary)

N	finite time horizon
E, $\mathfrak{E}$	state space with σ-algebra
A, $\mathfrak{A}$	action space with σ-algebra
D_n	admissible state-action pairs at time n
$Q_n(\cdot\|x,a)$	stochastic transition kernel from D_n to E
$r_n(x,a)$	one-stage reward at time n
$g_N(x)$	terminal reward
$\mathcal{Z}, \mathfrak{Z}$	disturbance space with σ-algebra
$T_n(x,a,z)$	transition function of the state process
$Q_n^Z(\cdot\|x,a)$	stochastic transition kernel from D_n to $\mathcal{Z}$
(X_n)	state process
(Z_n)	process of disturbances
F_n	set of decision rules at time n
π	$= (f_n)$ policy
h_n	$= (x_0, a_0, x_1, \ldots, x_n)$ history up to time n
$\mathit{\Pi}_N$	set of history-dependent N-stage policies
$\mathbb{P}_{nx}^{\pi}$	probability measure under policy π given $X_n = x$
$\mathbb{E}_{nx}^{\pi}$	expectation operator
$V_{n\pi}(x)$	expected total reward from n to N under policy π
$V_n(x)$	maximal expected total reward from n to N
$\delta_n^N(x)$	upper bound for $V_n(x)$
$(L_n v)(x,a)$	$= r_n(x,a) + \int v(x') Q_n(dx'\|x,a)$ reward operator
$(\mathcal{T}_{nf} v)(x)$	$= (L_n v)(x, f(x))$ reward operator of f
$(\mathcal{T}_n v)(x)$	$= \sup_{a \in D_n(x)} (L_n v)(x,a)$ maximal reward operator

Markov Decision Model (stationary)

E, $\mathfrak{E}$	state space with σ-algebra

A, $\mathfrak{A}$	action space with σ-algebra
D	admissible state-action pairs
$Q(\cdot\|x,a)$	stochastic transition kernel from D to E
$r(x,a)$	one-stage reward
$g(x)$	terminal reward
β	discount factor
$\mathcal{Z}, \mathfrak{Z}$	disturbance space with σ-algebra
$T(x,a,z)$	transition function of the state process
$Q^Z(\cdot\|x,a)$	stochastic transition kernel from D to $\mathcal{Z}$
F	set of decision rules
$J_{n\pi}(x)$	expected discounted reward over n stages under policy π
$J_n(x)$	maximal expected discounted reward over n stages
$\delta_N(x)$	upper bound for $J_N(x)$
$b(x)$	(upper) bounding function
$(Lv)(x,a)$	$= r(x,a) + \int v(x')Q(dx'\|x,a)$ reward operator
$(\mathcal{T}_f v)(x)$	$= (Lv)(x, f(x))$ reward operator of f
$(\mathcal{T}v)(x)$	$= \sup_{a\in D(x)}(Lv)(x,a)$ maximal reward operator

Financial Markets

S_n^0	bond price at time n
i_{n+1}	interest rate in $[n, n+1)$
S_n	$= (S_n^1, \ldots, S_n^d)$ stock prices at time n
$\tilde{R}_{n+1}^k$	$= \frac{S_{n+1}^k}{S_n^k}$ relative price change of asset k
R_n^k	$= \frac{\tilde{R}_n^k}{1+i_n} - 1$ relative risk process of asset k
$(\mathcal{F}_n)$	market filtration
ϕ	$= (\phi_n)$ self-financing portfolio strategy
$VaR_\gamma(X)$	Value-at-Risk at level γ
$AVaR_\gamma(X)$	Average-Value-at-Risk at level γ
$dom\, U$	domain of utility function U
CARA	constant absolute risk aversion
HARA	hyperbolic absolute risk aversion
MV	mean variance
MR	mean risk

Partially Observable Markov Decision Model

E_X	observable part of the state space
E_Y	unobservable part of the state space
A	action space
D	admissible state-action pairs

$Q(\cdot\|x,y,a)$	stochastic transition kernel from $E_Y \times D$ to $E_X \times E_Y$
Q_0	initial distribution of Y_0
$r(x,y,a)$	one-stage reward
$g(x)$	terminal reward
β	discount factor
$\mathcal{Z}$	disturbance space
$Q^{Z,Y}(\cdot\|x,y,a)$	stochastic transition kernel from $E_Y \times D$ to $\mathcal{Z} \times E_Y$
T_X	transition function of the observable state process
(X_n)	observable state process
(Y_n)	unobservable state process
(Z_n)	process of disturbances
Φ	Bayes operator
μ_n	conditional distribution of Y_n
h_n	$= (x_0, a_0, x_1, \ldots, x_n)$ history up to time n
$\tilde{h}_n$	$= (x_0, a_0, z_1, x_1, \ldots, z_n, x_n)$
Π_N	set of history-dependent N-stage policies
(t_n)	sequential sufficient statistic
$\hat{\Phi}$	information update operator

Markov Decision Model with infinite horizon

F^∞	set of infinite-stage policies
f^∞	$= (f, f, \ldots)$ stationary policy
$J_{\infty\pi}(x)$	expected discounted reward under policy π
$J_\infty(x)$	maximal expected discounted reward
$J(x)$	$= \lim_{n\to\infty} J_n(x)$ limit value function
$\delta(x)$	upper bound for $J_\infty(x)$
$\varepsilon(x)$	upper bound for the negative part of the rewards
$b(x)$	(upper) bounding function
α_b	contraction module
$(\mathcal{T}_\circ v)(x)$	$= \sup_{a\in D(x)} \beta \int v(x')Q(dx'\|x,a)$ shift operator
LsA_n	upper limit of the set sequence (A_n)

Stopping Problems

τ	stopping time
τ_π	stopping time induced by policy π
R_τ	reward under stopping time τ
$V_N^*(x)$	maximal reward of N-period stopping problem
$V_\infty^*(x)$	maximal reward of unbounded stopping problem
$G_\pi(x)$	$= \liminf_{n\to\infty} J_{n\pi}(x)$
$G(x)$	$= \sup_\pi G_\pi$
S_n^*	optimal stopping set at time n

Special Symbols

δ_x	one-point measure in x
$x^{\pm}$	$= \max\{0, \pm x\}$
$\mathbb{M}(E)$	$= \{v : E \to [-\infty, \infty),\ \text{measurable}\}$
$\mathbb{B}_b$	$= \{v \in \mathbb{M}(E) \mid \|v\|_b < \infty\}$
$\mathbb{B}_b^+$	$= \{v \in \mathbb{M}(E) \mid \|v^+\|_b < \infty\}$
$\mathbb{B}$	$= \{v \in \mathbb{M}(E) \mid v(x) \leq \delta(x) \text{ for all } x \in E\}$
$\|v\|_b$	weighted supremum norm
$\mathcal{B}(E)$	Borel σ-algebra in E
MTP_2	multivariate total positivity of order 2
$\leq_{st}$	stochastic order
$\leq_{lr}$	likelihood ratio order
$\leq_{cx}$	convex order
$\leq_{icx}$	increasing convex order
$\mathcal{N}(\mu, \sigma^2)$	normal distribution
$Exp(\lambda)$	exponential distribution
$Be(\alpha, \beta)$	beta distribution
$B(n, p)$	binomial distribution
$Poi(\lambda)$	Poisson distribution
$f(x) \propto g(x)$	f equals g up to a constant
$x \wedge y$	$:= (\min\{x_1, y_1\}, \ldots, \min\{x_d, y_d\})$
$x \vee y$	$:= (\max\{x_1, y_1\}, \ldots, \max\{x_d, y_d\})$
∇f	gradient of f
$\square$	end of proof
$\lozenge$	end of remark
$\blacklozenge$	end of example

Chapter 1
Introduction and First Examples

Suppose a system is given which can be controlled by sequential decisions. The state transitions are random and we assume that the *system state process* is *Markovian* which means that previous states have no influence on future states. Given the current state of the system (which could be for example the wealth of an investor) the controller or decision maker has to choose an *admissible action* (for example a possible investment). Once an action is chosen there is a random system transition according to a stochastic law (for example a change in the asset value) which leads to a new state. The task is to control the process in an optimal way. In order to formulate a reasonable optimization criterion we assume that each time an action is taken, the controller obtains a certain *reward*. The aim is then to control the system in such a way that the expected total discounted rewards are maximized. All these quantities together which have been described in an informal way, define a so-called *Markov Decision Process*. The Markov Decision Process is the sequence of random variables (X_n) which describes the stochastic evolution of the system states. Of course the distribution of (X_n) depends on the chosen actions. Figure 1.1 shows the schematic evolution of a Markov Decision Process.

We summarize the main model data in the following list:

- E denotes the *state space* of the system. A state $x \in E$ is the information which is available for the controller at time n. Given this information an action has to be selected.
- A denotes the *action space*. Given a specific state $x \in E$ at time n, a certain subclass $D_n(x) \subset A$ of actions may only be admissible.
- $Q_n(B|x,a)$ is a *stochastic transition kernel* which gives the probability that the next state at time $n+1$ is in the set B if the current state is x and action a is taken at time n.
- $r_n(x,a)$ gives the (discounted) *one-stage reward* of the system at time n if the current state is x and action a is taken.

N. Bäuerle and U. Rieder, *Markov Decision Processes with Applications to Finance*, Universitext, DOI 10.1007/978-3-642-18324-9_1,

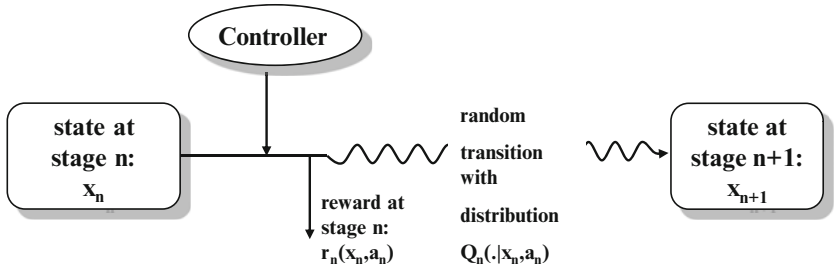

Fig. 1.1 General evolution of a Markov Decision Model.

- $g_N(x)$ gives the (discounted) *terminal reward* of the system at the end of the planning horizon.

An important assumption about these processes is that the evolution is Markovian. Since the system obeys a stochastic transition law, the sequence of visited states is not known at the beginning. Thus, a control π is a sequence of decision rules (f_n) with $f_n : E \to A$ where $f_n(x) \in D_n(x)$ determines for each possible state $x \in E$ the next action $f_n(x)$ at time n. Such a sequence $\pi = (f_n)$ is called *policy* or *strategy*. Formally the *Markov Decision Problem* is given by

$$V_0(x) := \sup_{\pi} \mathbb{E}_x^{\pi} \left[\sum_{k=0}^{N-1} r_k\big(X_k, f_k(X_k)\big) + g_N(X_N) \right], \quad x \in E,$$

where the supremum is taken over all admissible policies. Obviously the optimization problem is non-trivial since a decision today does not only determine the current reward but also has complicated influences on future states. The policies which have been defined above are Markovian by definition since the decision depends only on the current state. Indeed it suffices to search for the optimal control among all Markovian policies, though sometimes it is convenient to consider history-dependent policies.
The theory of Markov Decision Processes deals with stochastic optimization problems in *discrete time*. The time steps do not have to be equal but this is often assumed. Sometimes problems which are formulated in continuous-time can be reduced to a discrete-time model by considering an embedded state process. The theory of stochastic control problems in continuous time is quite different and not a subject of this book. However, when continuous-time problems have to be solved numerically, one way is to consider an approximation

of the state process in discrete time. This procedure is called the *approximating Markov chain approach*. The resulting problem can then be solved with the techniques presented here. Still the theory on Markov Decision Processes encompasses a lot of different models and formulations, and we will not deal with all of them. Since we have mainly applications in finance in mind we consider Markov Decision Processes with arbitrary state and action spaces (more precisely Borel spaces). Note that in these applications the spaces are often not discrete. Markov Decision Processes are also called *Markov Control Processes* or *Stochastic Dynamic Programs*. We distinguish problems with

- finite horizon $N < \infty$ – infinite horizon $N = \infty$,
- complete state observation – partial state observation,
- problems with constraints – without constraints,
- total (discounted) cost criterion – average cost criterion.

We will consider Markov Decision Problems with *finite horizon* in Chapter 2 and models with *infinite horizon* in Chapter 7. Sometimes models with infinite horizon appear in a natural way (for example when the original problem has a random horizon or when the original problem has a fixed time horizon but random time steps are used to solve it) and often these models can be seen as approximations of Markov Decision Problems with finite but large horizon. We will encounter different approaches to problems with infinite horizon in Chapter 7 and in Section 10.2 where unbounded stopping problems are treated. In any case, some convergence assumptions are needed to ensure that the infinite horizon Markov Decision Problem is well-defined. The solution of such optimization problems is then often easier because the value function can be characterized as the unique fixed point or as the smallest superharmonic function of an operator, and moreover the optimal policy is stationary. We will treat so-called *negative* Markov Decision Problems where the reward functions are negative (or zero) and *contracting* Markov Decision Problems where the maximal reward operator is contracting, in a unified framework. Besides these models we also consider so-called *positive* Markov Decision Problems where the reward functions are positive (or zero).
Another way to distinguish Markov Decision Problems is according to what can be observed by the controller. This is in contrast to deterministic control problems where the information for the controller is determined by the deterministic transition law. In stochastic control models also statistical aspects come into play. For example when a part of the state cannot be observed (e.g. some fundamental economic indicators which influence the asset price change), however due to observations of the state some information about the unobservable part is obtained. Such a model is called a *Partially Observable Markov Decision Problem*. In this case the statistical *filtering* theory has to be combined with the optimization problem. It will turn out that the Partially Observable Markov Decision Problem can be reformulated as a Markov Decision Problem with complete observation by enlarging the state space. Indeed an estimate of relevant information has to be added to the state and

updated during each new observation. Given this information a decision is taken. This is called the *separation principle of estimation and control*.
Markov Decision Problems already take some constraints about admissible actions into account. However, sometimes optimization problems arise where there are additional constraints. We do not treat these kind of problems in a systematic way but we consider for example *mean-variance problems* or *mean-risk problems* where the portfolio strategy of the investor has to satisfy some risk constraints. In Chapter 8 we consider *Piecewise Deterministic Markov Decision Processes*. These continuous-time optimization problems can be solved by discrete-time Markov Decision Processes with an action space consisting of functions. More precisely we have to introduce relaxed control functions. In Chapter 10 we deal with discrete-time *stopping problems*. Given that the underlying process is Markovian we show that stopping problems can be solved by Markov Decision Processes.

The theory of Markov Decision Processes which is outlined in this book addresses questions like: Does an optimal policy exist? Has it a particular form? Can an optimal policy be computed efficiently? Is it possible to derive properties of the optimal value function analytically? Besides developing the theory of Markov Decision Problems a main aim of this book is to show Markov Decision Problems in action. The applications are mostly taken from finance and insurance but are not limited to these areas. This book focuses on Markov Decision Processes with the total reward criterion. Problems with average-reward and risk-sensitive criteria are not treated in this book.

1.1 Applications

We will mainly focus on applications in finance, however the areas where Markov Decision Processes are used to solve problems are quite diverse. They appear in production planning, inventory control, operations management, engineering, biology and statistics, just to name a few. Let us consider some applications.

Example 1.1.1 (Consumption Problem). Suppose there is an investor with given initial capital. At the beginning of each of N periods she can decide how much of the capital she consumes and how much she invests into a risky asset. The amount she consumes is evaluated by a utility function U as well as the terminal wealth. The remaining capital is invested into a risky asset where we assume that the investor is *small* and thus not able to influence the asset price and the asset is *liquid*. How should she consume/invest in order to maximize the sum of her expected discounted utility?
The state x of the system is here the available capital. The action $a = f(x)$ is the amount of money which is consumed, where it is reasonable to assume

that $0 \leq a \leq x$. The reward is given by $U(a)$ and the terminal reward by $U(x)$. Hence the aim is to maximize

$$\mathbb{E}_x^\pi \left[\sum_{k=0}^{N-1} U\big(f_k(X_k)\big) + U(X_N) \right]$$

where the maximization is over all policies $\pi = (f_0, \ldots, f_{N-1})$. This problem is solved in Section 4.3. ⧫

Example 1.1.2 (Cash Balance or Inventory Problem). Imagine a company which tries to find the optimal level of cash over a finite number of N periods. We assume that there is a random stochastic change in the cash reserve each period (due to withdrawal or earnings). Since the firm does not earn interest from the cash position, there are holding cost for the cash reserve if it is positive, but also interest (cost) in case it is negative. The cash reserve can be increased or decreased by the management at each decision epoch which implies transfer costs. What is the optimal cash balance policy?
The state x of the system is here the current cash reserve. The action $a = f(x)$ is either the new cash reserve or the amount of money which is transferred from the cash reserve to assets. The reward is a negative cost determined by the transfer cost and the holding or understocking cost. This example is treated in Sections 2.6.2 as a finite horizon problem and in Section 7.6.2 as an infinite horizon problem. ⧫

Example 1.1.3 (Mean-Variance Problem). Consider a small investor who acts on a given financial market. Her aim is to choose among all portfolios which yield at least a certain expected return (benchmark) after N periods, the one with smallest portfolio variance. What is the optimal investment strategy?
This is an optimization problem with an additional constraint. As in the first example the state x of the system is the available capital. The action $a = f(x)$ is the investment decision. When we assume that there are d different assets available, then $a = (a_1, \ldots, a_d) \in \mathbb{R}^d$ and a_k gives the amount of money which is invested in asset k. The aim is to solve

$$(MV) \quad \begin{cases} \mathrm{Var}_{x_0}^\pi[X_N] \to \min \\ \\ \mathbb{E}_{x_0}^\pi[X_N] \geq \mu \end{cases}$$

where the minimization is over all policies $\pi = (f_0, \ldots, f_{N-1})$. In order to get rid of the constraint and to define the one-stage reward there is some work needed. This problem is investigated intensively in Section 4.6 and with partial observation in Section 6.2. ⧫

Example 1.1.4 (Dividend Problem in Risk Theory). Imagine we consider the risk reserve of an insurance company which earns some premia on the one hand but has to pay out possible claims on the other hand. At the beginning

of each period the insurer can decide upon paying a dividend. A dividend can only be paid when the risk reserve at that time point is positive. Once the risk reserve got negative we say that the company is ruined and has to stop its business. Which dividend pay-out policy maximizes the expected discounted dividends until ruin?
The state x of the system is here the current risk reserve. The action $a = f(x)$ is the dividend which is paid out where $a \leq x$. The one-stage reward is the dividend which is paid. This problem has to be dealt with as one with infinite horizon since the time horizon is not fixed in advance. This example is treated in Section 9.2. It can be shown that the optimal policy is stationary and has a certain structure which is called *band-policy*. ⧫

Example 1.1.5 (Bandit Problem). Suppose we have two slot machines with unknown success probability θ_1 and θ_2. At each stage we have to choose one of the arms. We receive one Euro if the arm wins, else no cash flow appears. How should we play in order to maximize our expected total reward over N trials?
This problem is a *Partially Observable Markov Decision Problem* since the success probabilities are not known. Hence the state of the system must here be interpreted as the available information of the decision maker. This information can be represented as the number of successes and failures at both arms up to this time point. Here $x = (m_1, n_1, m_2, n_2) \in \mathbb{N}_0^4$ denotes the number of successes m_i and failures n_i at arm i. An estimate for the win probability at arm i is then $\frac{m_i}{m_i+n_i}$. The action is obviously to choose one of the arms. The one-stage reward is the expected one-stage reward under the given information. This problem is treated in Section 5.5. Under some assumptions it can be shown that a so-called *index-policy* is optimal.
Bandit problems are generic problems which have a number of serious applications, for example medical trials of a new drug. ⧫

Example 1.1.6 (Pricing of American Options). In order to find the fair price of an American option and its optimal exercise time, one has to solve an optimal stopping problem. In contrast to a European option the buyer of an American option can choose to exercise any time up to and including the expiration time. In Section 11.1 we show how such an optimal stopping problem can be solved in the framework of Markov Decision Processes. ⧫

1.2 Organization of the Book

The book consists of eleven chapters which can be roughly grouped into four parts. The first part from Chapter 2 to 4 deals with the theory of Markov Decision Problems with finite time horizon, introduces the financial markets which are used later and provides some applications. The second part, which consists of Chapters 5 and 6, presents the theory of Partially Observable

Markov Decision Processes and provides some applications. Part III, which consists of Chapters 7, 8 and 9, investigates Markov Decision Problems with infinite time horizon, Piecewise Deterministic Markov Decision Processes, as well as applications. The last part – Chapters 10 and 11 – deals with stopping problems. Chapters with theory and applications alternate. The theory of Markov Decision Problems is presented in a self-contained way in Chapters 2, 5, 7, 8 and 10. Section 2.4 deals with conditions under which Markov Decision Problems satisfy the structure assumption. This part is slightly more advanced than the other material and might be skipped at first reading. Chapters 5 and 6 are not necessary for the understanding of the remaining chapters of the book (despite two examples in Chapters 10 and 11).

1.3 Notes and References

Historical Notes:
The first important books on Markov Decision Processes are Bellman (1957) (for a reprint see Bellman (2003)) and Howard (1960). The term 'Markov Decision Process' was coined by Bellman (1954). Shapley (1953) (for a reprint see Shapley (2003)) was the first study of Markov Decision Processes in the context of stochastic games. For more information on the origins of this research area see Puterman (1994) and Feinberg and Shwartz (2002). Later a more mathematical rigorous treatment of this theory appeared in Dubins and Savage (1965), Blackwell (1965), Shiryaev (1967) and Hinderer (1970). The fascinating book of Dubins and Savage (1965) deals with gambling models, however the underlying ideas are essentially the same. Blackwell (1965) introduces the model description which is used up to now. He was the first to give a rigorous treatment of discounted problems with general state spaces. Hinderer (1970) deals with general non-stationary models where reward functions and transition kernels may depend on the whole history of the underlying process. Another step towards generalizing the models are the books of Bertsekas and Shreve (1978) and Dynkin and Yushkevich (1979). There also the basic measurability questions are investigated.

Related Textbooks:
Nowadays a lot of excellent textbooks and handbooks on Markov Decision Processes exist and we are not able to give a complete list here. Thus we restrict to those books which we have frequently consulted or which are a reasonable addition and contain supplementary material.
By now, classical textbooks on Markov Decision Processes (besides the ones we have already mentioned in the 'Historical Notes') are Derman (1970), Ross (1970, 1983), Hordijk (1974), Whittle (1982, 1983), Schäl (1990), White (1993), Puterman (1994), Hernández-Lerma and Lasserre (1996), Filar and