

Markov Decision Processes

Markov Decision Processes

Discrete Stochastic
Dynamic Programming

MARTIN L. PUTERMAN

University of British Columbia



A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 1994, 2005 by John Wiley & Sons, Inc. All rights reserved.

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

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008.

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	xv
1. Introduction	1
1.1. The Sequential Decision Model, 1	
1.2. Inventory Management, 3	
1.3. Bus Engine Replacement, 4	
1.4. Highway Pavement Maintenance, 5	
1.5. Communication Models, 8	
1.6. Mate Desertion in Cooper's Hawks, 10	
1.7. So Who's Counting, 13	
Historical Background, 15	
2. Model Formulation	17
2.1. Problem Definition and Notation, 17	
2.1.1. Decision Epochs and Periods, 17	
2.1.2. State and Action Sets, 18	
2.1.3. Rewards and Transition Probabilities, 19	
2.1.4. Decision Rules, 21	
2.1.5. Policies, 22	
2.1.6. Induced Stochastic Processes, Conditional Probabilities, and Expectations, 22	
2.2. A One-Period Markov Decision Problem, 25	
2.3. Technical Considerations, 27	
2.3.1. The Role of Model Assumptions, 28	
2.3.2. The Borel Model, 28	
Bibliographic Remarks, 30	
Problems, 31	
3. Examples	33
3.1. A Two-State Markov Decision Process, 33	
3.2. Single-Product Stochastic Inventory Control, 37	

3.2.1.	Model Formulation,	37
3.2.2.	A Numerical Example,	41
3.3.	Deterministic Dynamic Programs,	42
3.3.1.	Problem Formulation,	42
3.3.2.	Shortest Route and Critical Path Models,	43
3.3.3.	Sequential Allocation Models,	45
3.3.4.	Constrained Maximum Likelihood Estimation,	46
3.4.	Optimal Stopping,	47
3.4.1.	Problem Formulation,	47
3.4.2.	Selling an Asset,	48
3.4.3.	The Secretary Problem,	49
3.4.4.	Exercising an Option,	50
3.5.	Controlled Discrete-Time Dynamic Systems,	51
3.5.1.	Model Formulation,	51
3.5.2.	The Inventory Control Model Revisited,	53
3.5.3.	Economic Growth Models,	55
3.5.4.	Linear Quadratic Control,	56
3.6.	Bandit Models,	57
3.6.1.	Markov Decision Problem Formulation,	57
3.6.2.	Applications,	58
3.6.3.	Modifications,	61
3.7.	Discrete-Time Queueing Systems,	62
3.7.1.	Admission Control,	62
3.7.2.	Service Rate Control,	64
	Bibliographic Remarks,	66
	Problems,	68

4. Finite-Horizon Markov Decision Processes

74

4.1.	Optimality Criteria,	74
4.1.1.	Some Preliminaries,	74
4.1.2.	The Expected Total Reward Criteria,	78
4.1.3.	Optimal Policies,	79
4.2.	Finite-Horizon Policy Evaluation,	80
4.3.	Optimality Equations and the Principle of Optimality,	83
4.4.	Optimality of Deterministic Markov Policies,	88
4.5.	Backward Induction,	92
4.6.	Examples,	94
4.6.1.	The Stochastic Inventory Model,	94
4.6.2.	Routing Problems,	96
4.6.3.	The Sequential Allocation Model,	98
4.6.4.	The Secretary Problem,	100
4.7.	Optimality of Monotone Policies,	102
4.7.1.	Structured Policies,	103
4.7.2.	Superadditive Functions,	103

4.7.3.	Optimality of Monotone Policies,	105
4.7.4.	A Price Determination Model,	108
4.7.5.	An Equipment Replacement Model,	109
4.7.6.	Monotone Backward Induction,	111
	Bibliographic Remarks,	112
	Problems,	113
5.	Infinite-Horizon Models: Foundations	119
5.1.	The Value of a Policy,	120
5.2.	The Expected Total Reward Criterion,	123
5.3.	The Expected Total Discounted Reward Criterion,	125
5.4.	Optimality Criteria,	128
5.4.1.	Criteria Based on Policy Value Functions,	128
5.4.2.	Overtaking Optimality Criteria,	130
5.4.3.	Discount Optimality Criteria,	133
5.5.	Markov Policies,	134
5.6.	Vector Notation for Markov Decision Processes,	137
	Bibliographic Remarks,	138
	Problems,	139
6.	Discounted Markov Decision Problems	142
6.1.	Policy Evaluation,	143
6.2.	Optimality Equations,	146
6.2.1.	Motivation and Definitions,	146
6.2.2.	Properties of Solutions of the Optimality Equations,	148
6.2.3.	Solutions of the Optimality Equations,	149
6.2.4.	Existence of Optimal Policies,	152
6.2.5.	General State and Action Spaces,	157
6.3.	Value Iteration and Its Variants,	158
6.3.1.	Rates of Convergence,	159
6.3.2.	Value Iteration,	160
6.3.3.	Increasing the Efficiency of Value Iteration with Splitting Methods,	164
6.4.	Policy Iteration,	174
6.4.1.	The Algorithm	
6.4.2.	Finite State and Action Models,	176
6.4.3.	Nonfinite Models,	177
6.4.4.	Convergence Rates,	181
6.5.	Modified Policy Iteration,	185
6.5.1.	The Modified Policy Iteration Algorithm,	186
6.5.2.	Convergence of Modified Policy Iteration,	188
6.5.3.	Convergence Rates,	192
6.5.4.	Variants of Modified Policy Iteration,	194
6.6.	Spans, Bounds, Stopping Criteria, and Relative Value Iteration,	195

6.6.1.	The Span Seminorm, 195
6.6.2.	Bounds on the Value of a Discounted Markov Decision Processes, 199
6.6.3.	Stopping Criteria, 201
6.6.4.	Value Iteration and Relative Value Iteration, 203
6.7.	Action Elimination Procedures, 206
6.7.1.	Identification of Nonoptimal Actions, 206
6.7.2.	Action Elimination Procedures, 208
6.7.3.	Modified Policy Iteration with Action Elimination and an Improved Stopping Criterion, 213
6.7.4.	Numerical Performance of Modified Policy Iteration with Action Elimination, 216
6.8.	Convergence of Policies, Turnpikes and Planning Horizons, 218
6.9.	Linear Programming, 223
6.9.1.	Model Formation, 223
6.9.2.	Basic Solutions and Stationary Policies, 224
6.9.3.	Optimal Solutions and Optimal Policies, 227
6.9.4.	An Example, 229
6.10.	Countable-State Models, 231
6.10.1.	Unbounded Rewards, 231
6.10.2.	Finite-State Approximations to Countable-State Discounted Models, 239
6.10.3.	Bounds for Approximations, 245
6.10.4.	An Equipment Replacement Model, 248
6.11.	The Optimality of Structured Policies, 255
6.11.1.	A General Framework, 255
6.11.2.	Optimal Monotone Policies, 258
6.11.3.	Continuous and Measurable Optimal Policies, 260
	Bibliographic Remarks, 263
	Problems, 266

7. The Expected Total-Reward Criterion

277

7.1.	Model Classification and General Results, 277
7.1.1.	Existence of the Expected Total Reward, 278
7.1.2.	The Optimality Equation, 280
7.1.3.	Identification of Optimal Policies, 282
7.1.4.	Existence of Optimal Policies, 283
7.2.	Positive Bounded Models, 284
7.2.1.	The Optimality Equation, 285
7.2.2.	Identification of Optimal Policies, 288
7.2.3.	Existence of Optimal Policies, 289
7.2.4.	Value Iteration, 294
7.2.5.	Policy Iteration, 295
7.2.6.	Modified Policy Iteration, 298
7.2.7.	Linear Programming, 299
7.2.8.	Optimal Stopping, 303

- 7.3. Negative Models, 309
 - 7.3.1. The Optimality Equation, 309
 - 7.3.2. Identification and Existence of Optimal Policies, 311
 - 7.3.3. Value Iteration, 313
 - 7.3.4. Policy Iteration, 316
 - 7.3.5. Modified Policy Iteration, 318
 - 7.3.6. Linear Programming, 319
 - 7.3.7. Optimal Parking, 319
- 7.4. Comparison of Positive and Negative Models, 324
- Bibliographic Remarks, 325
- Problems, 326

8. Average Reward and Related Criteria

331

- 8.1. Optimality Criteria, 332
 - 8.1.1. The Average Reward of a Fixed Policy, 332
 - 8.1.2. Average Optimality Criteria, 333
- 8.2. Markov Reward Processes and Evaluation Equations, 336
 - 8.2.1. The Gain and Bias, 337
 - 8.2.2. The Laurent Series Expansion, 341
 - 8.2.3. Evaluation Equations, 343
- 8.3. Classification of Markov Decision Processes, 348
 - 8.3.1. Classification Schemes, 348
 - 8.3.2. Classifying a Markov Decision Process, 350
 - 8.3.3. Model Classification and the Average Reward Criterion, 351
- 8.4. The Average Reward Optimality Equation—Unichain Models, 353
 - 8.4.1. The Optimality Equation, 354
 - 8.4.2. Existence of Solutions to the Optimality Equation, 358
 - 8.4.3. Identification and Existence of Optimal Policies, 360
 - 8.4.4. Models with Compact Action Sets, 361
- 8.5. Value Iteration in Unichain Models, 364
 - 8.5.1. The Value Iteration Algorithm, 364
 - 8.5.2. Convergence of Value Iteration, 366
 - 8.5.3. Bounds on the Gain, 370
 - 8.5.4. An Aperiodicity Transformation, 371
 - 8.5.5. Relative Value Iteration, 373
 - 8.5.6. Action Elimination, 373
- 8.6. Policy Iteration in Unichain Models, 377
 - 8.6.1. The Algorithm, 378
 - 8.6.2. Convergence of Policy Iteration for Recurrent Models, 379
 - 8.6.3. Convergence of Policy Iteration for Unichain Models, 380
- 8.7. Modified Policy Iteration in Unichain Models, 385
 - 8.7.1. The Modified Policy Iteration Algorithm, 386

- 8.7.2. Convergence of the Algorithm, 387
- 8.7.3. Numerical Comparison of Algorithms, 388
- 8.8. Linear Programming in Unichain Models, 391
 - 8.8.1. Linear Programming for Recurrent Models, 392
 - 8.8.2. Linear Programming for Unichain Models, 395
- 8.9. State Action Frequencies, Constrained Models and Models with Variance Criteria, 398
 - 8.9.1. Limiting Average State Action Frequencies, 398
 - 8.9.2. Models with Constraints, 404
 - 8.9.3. Variance Criteria, 408
- 8.10. Countable-State Models, 412
 - 8.10.1. Counterexamples, 413
 - 8.10.2. Existence Results, 414
 - 8.10.3. A Communications Model, 421
 - 8.10.4. A Replacement Model, 424
- 8.11. The Optimality of Structured Policies, 425
 - 8.11.1. General Theory, 425
 - 8.11.2. Optimal Monotone Policies, 426
- Bibliographic Remarks, 429
- Problems, 433

9. The Average Reward Criterion—Multichain and Communicating Models 441

- 9.1. Average Reward Optimality Equations: Multichain Models, 442
 - 9.1.1. Multichain Optimality Equations, 443
 - 9.1.2. Property of Solutions of the Optimality Equations, 445
 - 9.1.3. Existence of Solutions to the Optimality Equations, 448
 - 9.1.4. Identification and Existence of Optimal Policies, 450
- 9.2. Policy Iteration for Multichain Models, 451
 - 9.2.1. The Algorithm, 452
 - 9.2.2. An Example, 454
 - 9.2.3. Convergence of the Policy Iteration in Multichain Models, 455
 - 9.2.4. Behavior of the Iterates of Policy Iteration, 458
- 9.3. Linear Programming in Multichain Models, 462
 - 9.3.1. Dual Feasible Solutions and Randomized Decision Rules, 463
 - 9.3.2. Basic Feasible Solutions and Deterministic Decision Rules, 467
 - 9.3.3. Optimal Solutions and Policies, 468
- 9.4. Value Iteration, 472
 - 9.4.1. Convergence of $v^n - ng^*$, 472
 - 9.4.2. Convergence of Value Iteration, 476
- 9.5. Communicating Models, 478
 - 9.5.1. Policy Iteration, 478
 - 9.5.2. Linear Programming, 482
 - 9.5.3. Value Iteration, 483

Bibliographic Remarks, 484

Problems, 487

10. Sensitive Discount Optimality

492

- 10.1. Existence of Optimal Policies, 493
 - 10.1.1. Definitions, 494
 - 10.1.2. Blackwell Optimality, 495
 - 10.1.3. Stationary n -Discount Optimal Policies, 497
 - 10.2. Optimality Equations, 501
 - 10.2.1. Derivation of Sensitive Discount Optimality Equations, 501
 - 10.2.2. Lexicographic Ordering, 503
 - 10.2.3. Properties of Solutions of the Sensitive Optimality Equations, 505
 - 10.3. Policy Iteration, 511
 - 10.3.1. The Algorithm, 511
 - 10.3.2. An Example, 513
 - 10.3.3. Convergence of the Algorithm, 515
 - 10.3.4. Finding Blackwell Optimal Policies, 517
 - 10.4. The Expected Total-Reward Criterion Revisited, 519
 - 10.4.1. Relationship of the Bias and Expected Total Reward, 519
 - 10.4.2. Optimality Equations and Policy Iteration, 520
 - 10.4.3. Finding Average Overtaking Optimal Policies, 525
- Bibliographic Remarks, 526
- Problems, 528

11. Continuous-Time Models

530

- 11.1. Model Formulation, 531
 - 11.1.1. Probabilistic Structure, 531
 - 11.1.2. Rewards or Costs, 533
 - 11.1.3. Decision Rules and Policies, 534
 - 11.1.4. Induced Stochastic Processes, 534
- 11.2. Applications, 536
 - 11.2.1. A Two-State Semi-Markov Decision Process, 536
 - 11.2.2. Admission Control for a G/M/1 Queueing System, 537
 - 11.2.3. Service Rate Control in an M/G/1 Queueing System, 539
- 11.3. Discounted Models, 540
 - 11.3.1. Model Formulation, 540
 - 11.3.2. Policy Evaluation, 540
 - 11.3.3. The Optimality Equation and Its Properties, 545
 - 11.3.4. Algorithms, 546
 - 11.3.5. Unbounded Rewards, 547

11.4. Average Reward Models, 548	
11.4.1. Model Formulation, 548	
11.4.2. Policy Evaluation, 550	
11.4.3. Optimality Equations, 554	
11.4.4. Algorithms, 558	
11.4.5. Countable-State Models, 559	
11.5. Continuous-Time Markov Decision Processes, 560	
11.5.1. Continuous-Time Markov Chains, 561	
11.5.2. The Discounted Model, 563	
11.5.3. The Average Reward Model, 567	
11.5.4. Queueing Admission Control, 568	
Bibliographic Remarks, 573	
Problems, 574	
Afterword	579
Notation	581
Appendix A. Markov Chains	587
A.1. Basic Definitions, 587	
A.2. Classification of States, 588	
A.3. Classifying the States of a Finite Markov Chain, 589	
A.4. The Limiting Matrix, 591	
A.5. Matrix Decomposition, the Drazin Inverse and the Deviation Matrix, 594	
A.6. The Laurent Series Expansion of the Resolvent, 599	
A.7. A. A. Markov, 600	
Appendix B. Semicontinuous Functions	602
Appendix C. Normed Linear Spaces	605
C.1. Linear Spaces, 605	
C.2. Eigenvalues and Eigenvectors, 607	
C.3. Series Expansions of Inverses, 607	
C.4. Continuity of Inverses and Products, 609	
Appendix D. Linear Programming	610
Bibliography	613
Index	643

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 Spijksma and Adam Schwartz 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