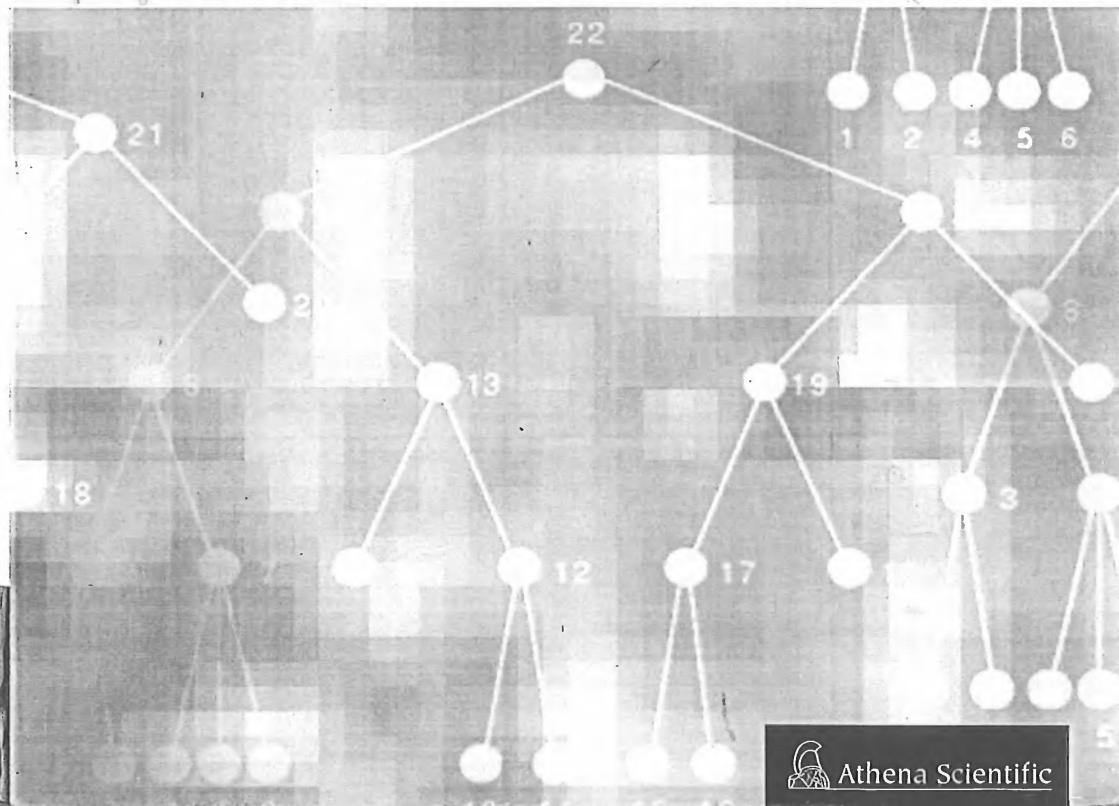


THIRD EDITION

VOLUME 2

Dynamic Programming and Optimal Control

DIMITRI P. BERTSEKAS



Athena Scientific

Dynamic Programming and Optimal Control

Volume II

THIRD EDITION

Dimitri P. Bertsekas

Massachusetts Institute of Technology

WWW site for book information and orders

<http://www.athenasc.com>



Athena Scientific, Belmont, Massachusetts

**Athena Scientific
Post Office Box 805
Nashua, NH 03061-0805
U.S.A.**

Email: info@athenasc.com
WWW: <http://www.athenasc.com>

Cover Design: Ann Gallagher, www.gallagerdesign.com

© 2007, 2001, 1995 Dimitri P. Bertsekas

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

Publisher's Cataloging-in-Publication Data

Bertsekas, Dimitri P.
Dynamic Programming and Optimal Control
Includes Bibliography and Index
1. Mathematical Optimization. 2. Dynamic Programming. I. Title.
QA402.5 .B465 2007 519.703 01-75941

ISBN 1-886529-30-2 (Vol. II)
ISBN 1-886529-26-4 (Vol. I)
ISBN 1-886529-08-6 (Two-volume set – latest editions)

Contents

1. Infinite Horizon – Discounted Problems

1.1.	Minimization of Total Cost – Introduction	p. 3
1.1.1.	The Finite-Horizon DP Algorithm	p. 4
1.1.2.	Shorthand Notation and Monotonicity	p. 6
1.1.3.	A Preview of Infinite Horizon Results	p. 8
1.1.4.	Randomized and History-Dependent Policies	p. 10
1.2.	Discounted Problems with Bounded Cost per Stage	p. 12
1.3.	Finite-State Systems – Computational Methods	p. 19
1.3.1.	Value Iteration and Error Bounds	p. 22
1.3.2.	Variants of Value Iteration	p. 30
1.3.3.	Policy Iteration	p. 38
1.3.4.	Linear Programming	p. 51
1.3.5.	Limited Lookahead and Rollout Policies	p. 53
1.4.	The Role of Contraction Mappings	p. 56
1.4.1.	Sup-Norm Contractions	p. 57
1.4.2.	m -Stage Sup-Norm Contractions	p. 62
1.4.3.	Discounted Problems - Unbounded Cost per Stage	p. 63
1.5.	Scheduling and Multiarmed Bandit Problems	p. 66
1.6.	Notes, Sources, and Exercises	p. 79

2. Stochastic Shortest Path Problems

2.1.	Problem Formulation	p. 94
2.2.	Bellman's Equation	p. 97
2.3.	Value Iteration	p. 105
2.4.	Policy Iteration	p. 108
2.5.	Countable State Problems	p. 112
2.6.	Notes, Sources, and Exercises	p. 114

3. Undiscounted Problems

3.1.	Unbounded Costs per Stage	p. 124
3.2.	Linear Systems and Quadratic Cost	p. 140

3.3. Inventory Control	p. 142
3.4. Optimal Stopping	p. 145
3.5. Optimal Gambling Strategies	p. 150
3.6. Nonstationary and Periodic Problems	p. 157
3.7. Notes, Sources, and Exercises	p. 162
 4. Average Cost per Stage Problems	
4.1. Finite-Spaces Average Cost Models	p. 174
4.1.1. Relation with the Discounted Cost Problem	p. 178
4.1.2. Blackwell Optimal Policies	p. 184
4.1.3. Optimality Equations	p. 194
4.2. Conditions for Equal Average Cost for all Initial States	p. 198
4.3. Value Iteration	p. 204
4.3.1. Single-Chain Value Iteration	p. 207
4.3.2. Multi-Chain Value Iteration	p. 222
4.4. Policy Iteration	p. 229
4.4.1. Single-Chain Policy Iteration	p. 229
4.4.2. Multi-Chain Policy Iteration	p. 235
4.5 Linear Programming	p. 239
4.6. Infinite-Spaces Problems	p. 245
4.6.1. A Sufficient Condition for Optimality	p. 253
4.6.2. Finite State Space and Infinite Control Space	p. 255
4.6.3. Countable States – Vanishing Discount Approach	p. 264
4.6.4. Countable States – Contraction Approach	p. 267
4.6.5. Linear Systems with Quadratic Cost	p. 272
4.7. Notes, Sources, and Exercises	p. 274
 5. Continuous-Time Problems	
5.1. Uniformization	p. 288
5.2. Queueing Applications	p. 295
5.3. Semi-Markov Problems	p. 306
5.4. Notes, Sources, and Exercises	p. 317
 6. Approximate Dynamic Programming	
6.1. Cost Approximation	p. 325
6.2. Approximate Policy Iteration – Direct Policy Evaluation	p. 329
6.2.1. Gradient Methods for Direct Policy Evaluation	p. 332
6.2.2. TD(λ)	p. 336
6.2.3. Optimistic Policy Iteration	p. 337
6.2.4. Approximate Policy Iteration Based on Q -Factors	p. 338
6.3. Indirect Methods for Policy Evaluation	p. 340
6.3.1. Policy Evaluation by Projected Value Iteration	p. 341
6.3.2. Least Squares Policy Evaluation (LSPE)	p. 346

6.3.3. PVI(λ) and LSPE(λ)	p. 348
6.3.4. The LSTD(λ) Algorithm	p. 355
6.3.5. The TD(λ) Algorithm	p. 357
6.3.6. Summary and Examples	p. 359
6.4. Q -Learning	p. 363
6.4.1. Q -Factor Approximations	p. 364
6.4.2. Q -Learning for Optimal Stopping Problems	p. 366
6.5. Stochastic Shortest Path Problems	p. 369
6.6. Average Cost Problems	p. 375
6.6.1. Approximate Policy Evaluation	p. 375
6.6.2. Approximate Policy Iteration	p. 386
6.6.3. Q -Learning for Average Cost Problems	p. 389
6.7. Approximation in Policy Space	p. 392
6.8. Notes, Sources, and Exercises	p. 399
Appendix A: Measure-Theoretic Issues in Dynamic Programming	
A.1. A Two-Stage Example	p. 407
A.2. Resolution of the Measurability Issues	p. 412
References	p. 423
Index	p. 443

CONTENTS OF VOLUME I

1. The Dynamic Programming Algorithm

- 1.1. Introduction
- 1.2. The Basic Problem
- 1.3. The Dynamic Programming Algorithm
- 1.4. State Augmentation and Other Reformulations
- 1.5. Some Mathematical Issues
- 1.6. Dynamic Programming and Minimax Control
- 1.7. Notes, Sources, and Exercises

2. Deterministic Systems and the Shortest Path Problem

- 2.1. Finite-State Systems and Shortest Paths
- 2.2. Some Shortest Path Applications
 - 2.2.1. Critical Path Analysis
 - 2.2.2. Hidden Markov Models and the Viterbi Algorithm
- 2.3. Shortest Path Algorithms
 - 2.3.1. Label Correcting Methods
 - 2.3.2. Label Correcting Variations - A^* Algorithm
 - 2.3.3. Branch-and-Bound
 - 2.3.4. Constrained and Multiobjective Problems
- 2.4. Notes, Sources, and Exercises

3. Deterministic Continuous-Time Optimal Control

- 3.1. Continuous-Time Optimal Control
- 3.2. The Hamilton – Jacobi – Bellman Equation
- 3.3. The Pontryagin Minimum Principle
 - 3.3.1. An Informal Derivation Using the HJB Equation
 - 3.3.2. A Derivation Based on Variational Ideas
 - 3.3.3. The Minimum Principle for Discrete-Time Problems
- 3.4. Extensions of the Minimum Principle
 - 3.4.1. Fixed Terminal State
 - 3.4.2. Free Initial State
 - 3.4.3. Free Terminal Time
 - 3.4.4. Time-Varying System and Cost
 - 3.4.5. Singular Problems
- 3.5. Notes, Sources, and Exercises

4. Problems with Perfect State Information

- 4.1. Linear Systems and Quadratic Cost
- 4.2. Inventory Control
- 4.3. Dynamic Portfolio Analysis
- 4.4. Optimal Stopping Problems
- 4.5. Scheduling and the Interchange Argument

- 4.6. Set-Membership Description of Uncertainty
 - 4.6.1. Set-Membership Estimation
 - 4.6.2. Control with Unknown-but-Bounded Disturbances
 - 4.7. Notes, Sources, and Exercises
- 5. Problems with Imperfect State Information**
- 5.1. Reduction to the Perfect Information Case
 - 5.2. Linear Systems and Quadratic Cost
 - 5.3. Minimum Variance Control of Linear Systems
 - 5.4. Sufficient Statistics and Finite-State Markov Chains
 - 5.4.1. The Conditional State Distribution
 - 5.4.2. Finite-State Systems
 - 5.5. Notes, Sources, and Exercises
- 6. Suboptimal and Adaptive Control**
- 6.1. Certainty Equivalent and Adaptive Control
 - 6.1.1. Caution, Probing, and Dual Control
 - 6.1.2. Two-Phase Control and Identifiability
 - 6.1.3. Certainty Equivalent Control and Identifiability
 - 6.1.4. Self-Tuning Regulators
 - 6.2. Open-Loop Feedback Control
 - 6.3. Limited Lookahead Policies and Applications
 - 6.3.1. Performance Bounds for Limited Lookahead Policies
 - 6.3.2. Computational Issues in Limited Lookahead
 - 6.3.3. Problem Approximation - Enforced Decomposition
 - 6.3.4. Aggregation
 - 6.3.5. Parametric Cost-to-Go Approximation
 - 6.4. Rollout Algorithms
 - 6.4.1. Discrete Deterministic Problems
 - 6.4.2. Q -Factors Evaluated by Simulation
 - 6.4.3. Q -Factor Approximation
 - 6.5. Model Predictive Control and Related Methods
 - 6.5.1. Rolling Horizon Approximations
 - 6.5.2. Stability Issues in Model Predictive Control
 - 6.5.3. Restricted Structure Policies
 - 6.6. Additional Topics in Approximate DP
 - 6.6.1. Discretization
 - 6.6.2. Other Approximation Approaches
 - 6.7. Notes, Sources, and Exercises
- 7. Introduction to Infinite Horizon Problems**
- 7.1. An Overview
 - 7.2. Stochastic Shortest Path Problems
 - 7.3. Discounted Problems
 - 7.4. Average Cost Problems

- 7.5. Semi-Markov Problems
- 7.6. Notes, Sources, and Exercises

Appendix A: Mathematical Review

- A.1. Sets
- A.2. Euclidean Space
- A.3. Matrices
- A.4. Analysis
- A.5. Convex Sets and Functions

Appendix B: On Optimization Theory

- B.1. Optimal Solutions
- B.2. Optimality Conditions
- B.3. Minimization of Quadratic Forms

Appendix C: On Probability Theory

- C.1. Probability Spaces
- C.2. Random Variables
- C.3. Conditional Probability

Appendix D: On Finite-State Markov Chains

- D.1. Stationary Markov Chains
- D.2. Classification of States
- D.3. Limiting Probabilities
- D.4. First Passage Times

Appendix E: Least-Squares Estimation and Kalman Filtering

- E.1. Least-Squares Estimation
- E.2. Linear Least-Squares Estimation
- E.3. State Estimation – Kalman Filter
- E.4. Stability Aspects
- E.5. Gauss-Markov Estimators
- E.6. Deterministic Least-Squares Estimation

Appendix F: Modeling of Stochastic Linear Systems

- F.1. Linear Systems with Stochastic Inputs
- F.2. Processes with Rational Spectrum
- F.3. The ARMAX Model

Appendix G: Formulating Problems of Decision Under Uncertainty

- G.1. The Problem of Decision Under Uncertainty
- G.2. Expected Utility Theory and Risk
- G.3. Stochastic Optimal Control Problems

ABOUT THE AUTHOR

Dimitri Bertsekas studied Mechanical and Electrical Engineering at the National Technical University of Athens, Greece, and obtained his Ph.D. in system science from the Massachusetts Institute of Technology. He has held faculty positions with the Engineering-Economic Systems Dept., Stanford University, and the Electrical Engineering Dept. of the University of Illinois, Urbana. Since 1979 he has been teaching at the Electrical Engineering and Computer Science Department of the Massachusetts Institute of Technology (M.I.T.), where he is currently McAfee Professor of Engineering.

His research spans several fields, including optimization, control, large-scale computation, and data communication networks, and is closely tied to his teaching and book authoring activities. He has written numerous research papers, and thirteen books, several of which are used as textbooks in MIT classes. He consults regularly with private industry and has held editorial positions in several journals.

Professor Bertsekas was awarded the INFORMS 1997 Prize for Research Excellence in the Interface Between Operations Research and Computer Science for his book "Neuro-Dynamic Programming" (co-authored with John Tsitsiklis), the 2000 Greek National Award for Operations Research, and the 2001 ACC John R. Ragazzini Education Award. In 2001, he was elected to the United States National Academy of Engineering.

**ATHENA SCIENTIFIC
OPTIMIZATION AND COMPUTATION SERIES**

1. Convex Analysis and Optimization, by Dimitri P. Bertsekas, with Angelia Nedić and Asuman E. Ozdaglar, 2003, ISBN 1-886529-45-0, 560 pages
2. Introduction to Probability, by Dimitri P. Bertsekas and John N. Tsitsiklis, 2002, ISBN 1-886529-40-X, 430 pages
3. Dynamic Programming and Optimal Control, Two-Volume Set, by Dimitri P. Bertsekas, 2007, ISBN 1-886529-08-6, 1020 pages
4. Nonlinear Programming, 2nd Edition, by Dimitri P. Bertsekas, 1999, ISBN 1-886529-00-0, 791 pages
5. Network Optimization: Continuous and Discrete Models, by Dimitri P. Bertsekas, 1998, ISBN 1-886529-02-7, 608 pages
6. Network Flows and Monotropic Optimization, by R. Tyrrell Rockafellar, 1998, ISBN 1-886529-06-X, 634 pages
7. Introduction to Linear Optimization, by Dimitris Bertsimas and John N. Tsitsiklis, 1997, ISBN 1-886529-19-1, 608 pages
8. Parallel and Distributed Computation: Numerical Methods, by Dimitri P. Bertsekas and John N. Tsitsiklis, 1997, ISBN 1-886529-01-9, 718 pages
9. Neuro-Dynamic Programming, by Dimitri P. Bertsekas and John N. Tsitsiklis, 1996, ISBN 1-886529-10-8, 512 pages
10. Constrained Optimization and Lagrange Multiplier Methods, by Dimitri P. Bertsekas, 1996, ISBN 1-886529-04-3, 410 pages
11. Stochastic Optimal Control: The Discrete-Time Case, by Dimitri P. Bertsekas and Steven E. Shreve, 1996, ISBN 1-886529-03-5, 330 pages

Preface

This two-volume book is based on a first-year graduate course on dynamic programming and optimal control that I have taught for over twenty years at Stanford University, the University of Illinois, and the Massachusetts Institute of Technology. The course has been typically attended by students from engineering, operations research, economics, and applied mathematics. Accordingly, a principal objective of the book has been to provide a unified treatment of the subject, suitable for a broad audience. In particular, problems with a continuous character, such as stochastic control problems, popular in modern control theory, are simultaneously treated with problems with a discrete character, such as Markovian decision problems, popular in operations research. Furthermore, many applications and examples, drawn from a broad variety of fields, are discussed.

The book may be viewed as a greatly expanded and pedagogically improved version of my 1987 book "Dynamic Programming: Deterministic and Stochastic Models," published by Prentice-Hall. I have included much new material on deterministic and stochastic shortest path problems, as well as a new chapter on continuous-time optimal control problems and the Pontryagin Maximum Principle, developed from a dynamic programming viewpoint. I have also added a fairly extensive exposition of simulation-based approximation techniques for dynamic programming. These techniques, which are often referred to as "neuro-dynamic programming" or "reinforcement learning," represent a breakthrough in the practical application of dynamic programming to complex problems that involve the dual curse of large dimension and lack of an accurate mathematical model. Other material was also augmented, substantially modified, and updated.

With the new material, however, the book grew so much in size that it became necessary to divide it into two volumes: one on finite horizon, and the other on infinite horizon problems. This division was not only natural in terms of size, but also in terms of style and orientation. The first volume is more oriented towards modeling, and the second is more oriented towards mathematical analysis and computation. I have included in the first volume a final chapter that provides an introductory treatment of infinite horizon problems. The purpose is to make the first volume self-

contained for instructors who wish to cover a modest amount of infinite horizon material in a course that is primarily oriented towards modeling, conceptualization, and finite horizon problems,

Many topics in the book are relatively independent of the others. For example Chapter 2 of Vol. I on shortest path problems can be skipped without loss of continuity, and the same is true for Chapter 3 of Vol. I, which deals with continuous-time optimal control. As a result, the book can be used to teach several different types of courses.

- (a) A two-semester course that covers both volumes.
- (b) A one-semester course primarily focused on finite horizon problems that covers most of the first volume.
- (c) A one-semester course focused on stochastic optimal control that covers Chapters 1, 4, 5, and 6 of Vol. I, and Chapters 1, 2, and 4 of Vol. II.
- (d) A one-semester course that covers Chapter 1, about 50% of Chapters 2 through 6 of Vol. I, and about 70% of Chapters 1, 2, and 4 of Vol. II. This is the course I usually teach at MIT.
- (e) A one-quarter engineering course that covers the first three chapters and parts of Chapters 4 through 6 of Vol. I.
- (f) A one-quarter mathematically oriented course focused on infinite horizon problems that covers Vol. II.

The mathematical prerequisite for the text is knowledge of advanced calculus, introductory probability theory, and matrix-vector algebra. A summary of this material is provided in the appendixes. Naturally, prior exposure to dynamic system theory, control, optimization, or operations research will be helpful to the reader, but based on my experience, the material given here is reasonably self-contained.

The book contains a large number of exercises, and the serious reader will benefit greatly by going through them. Solutions to all exercises are compiled in a manual that is available to instructors from the author. Many thanks are due to the several people who spent long hours contributing to this manual, particularly Steven Shreve, Eric Loiederman, Lakis Polymenakos, and Cynara Wu.

Dynamic programming is a conceptually simple technique that can be adequately explained using elementary analysis. Yet a mathematically rigorous treatment of general dynamic programming requires the complicated machinery of measure-theoretic probability. My choice has been to bypass the complicated mathematics by developing the subject in generality, while claiming rigor only when the underlying probability spaces are countable. A mathematically rigorous treatment of the subject is carried out in my monograph "Stochastic Optimal Control: The Discrete Time

Case," Academic Press, 1978,† coauthored by Steven Shreve. This monograph complements the present text and provides a solid foundation for the subjects developed somewhat informally here.

Finally, I am thankful to a number of individuals and institutions for their contributions to the book. My understanding of the subject was sharpened while I worked with Steven Shreve on our 1978 monograph. My interaction and collaboration with John Tsitsiklis on stochastic shortest paths and approximate dynamic programming have been most valuable. Michael Caramanis, Emmanuel Fernandez-Gaucherand, Pierre Humbert, Lennart Ljung, and John Tsitsiklis taught from versions of the book, and contributed several substantive comments and homework problems. A number of colleagues offered valuable insights and information, particularly David Castanon, Eugene Feinberg, and Krishna Pattipati. NSF provided research support. Prentice-Hall graciously allowed the use of material from my 1987 book. Teaching and interacting with the students at MIT have kept up my interest and excitement for the subject.

Dimitri P. Bertsekas
bertsekas@lids.mit.edu
<http://web.mit.edu/dimitrib/www/home.html>

—

† Note added in the 2nd edition: This monograph was republished by Athena Scientific in 1996.

Preface to the Second Edition

This second edition of Vol. II should be viewed as a relatively minor revision of the original. The coverage was expanded in a few areas as follows:

- (a) In Chapter 1, material was added on variants of the policy iteration method.
- (b) In Chapter 2, the material on neuro-dynamic programming methods was updated and expanded to reflect some recent developments.
- (c) In Chapter 4, material was added on some new value iteration methods.
- (d) In Chapter 5, the material on semi-Markov problems was revised, with a significant portion simplified and shifted to Volume I.

There are also a few miscellaneous additions and improvements scattered throughout the text. Finally, a new internet-based feature was added to the book, which extends its scope and coverage. Many of the theoretical exercises have been solved in detail and their solutions have been posted in the book's [www](#) page

<http://www.athenasc.com/dpbook.html>

These exercises have been marked with the symbol 

I would like to express my thanks to the many colleagues who contributed suggestions for improvement of the second edition.

Dimitri P. Bertsekas
bertsekas@lids.mit.edu
<http://web.mit.edu/dimitrib/www/home.html>
June, 2001

Preface to the Third Edition

This is a major revision of the 2nd edition, and contains a substantial amount of new material, as well as a reorganization of old material. The length of the text has increased by more than 50%, and more than half of the old material has been restructured and/or revised. Most of the added material is in four areas.

- (a) The coverage of the average cost problem of Chapter 4 has greatly increased in scope and depth. In particular, there is now a full analysis of multi-chain problems, as well as a more extensive analysis of infinite-spaces problems (Section 4.6).
- (b) The material on approximate dynamic programming has been collected in Chapter 6. It has been greatly expanded to include new research, thereby supplementing the 1996 book “*Neuro-Dynamic Programming*.”
- (c) Contraction mappings and their role in various analyses have been highlighted in new material on infinite state space problems (Sections 1.4, 2.5, and 4.6), and in their use in the approximate dynamic programming material of Chapter 6.
- (d) The mathematical measure-theoretic issues that must be addressed for a rigorous theory of stochastic dynamic programming have been illustrated and summarized in an appendix for the benefit of the mathematically oriented reader.

Also some exercises were added and a few sections were revised while preserving their essential content.

I would like to express my thanks to many colleagues who contributed valuable comments. I am particularly thankful to Ciamac Moallemi, Steven Shreve, John Tsitsiklis, and Ben Van Roy, who reviewed some of the new material and each contributed several substantial suggestions. I wish to thank especially Janey Yu for her extraordinary help. Janey read with great care and keen eye large parts of the book, contributed important analysis and many incisive, substantive comments, and also collaborated with me in research that was included in Chapter 6.

Dimitri P. Bertsekas

<http://web.mit.edu/dimitrib/www/home.html>
Fall 2006

Infinite Horizon – Discounted Problems

Contents

1.1.	Minimization of Total Cost – Introduction	p. 3
1.1.1.	The Finite-Horizon DP Algorithm	p. 4
1.1.2.	Shorthand Notation and Monotonicity	p. 6
1.1.3.	A Preview of Infinite Horizon Results	p. 8
1.1.4.	Randomized and History-Dependent Policies	p. 10
1.2.	Discounted Problems with Bounded Cost per Stage	p. 12
1.3.	Finite-State Systems – Computational Methods	p. 19
1.3.1.	Value Iteration and Error Bounds	p. 22
1.3.2.	Variants of Value Iteration	p. 30
1.3.3.	Policy Iteration	p. 38
1.3.4.	Linear Programming	p. 51
1.3.5.	Limited Lookahead and Rollout Policies	p. 53
1.4.	The Role of Contraction Mappings	p. 56
1.4.1.	Sup-Norm Contractions	p. 57
1.4.2.	m -Stage Sup-Norm Contractions	p. 62
1.4.3.	Discounted Problems - Unbounded Cost per Stage .	p. 63
1.5.	Scheduling and Multiarmed Bandit Problems	p. 66
1.6.	Notes, Sources, and Exercises	p. 79

This volume focuses on stochastic optimal control problems with an infinite number of decision stages (an infinite horizon). An introduction to these problems was presented in Chapter 7 of Vol. I. Here, we provide a more comprehensive analysis. In particular, we do not assume a finite number of states and we also discuss the associated analytical and computational issues in much greater depth.

We recall from Chapter 7 of Vol. I the following four classes of infinite horizon problems of major interest:

- (a) Discounted problems with bounded cost per stage.
- (b) Stochastic shortest path problems.
- (c) Discounted and undiscounted problems with unbounded cost per stage.
- (d) Average cost per stage problems.

Each of the first four chapters of the present volume considers one of the above problem classes, while the fifth chapter extends some of the analysis to continuous-time problems with a countable number of states. The final chapter discusses *approximate dynamic programming*, a methodology that aims to compute approximations to the optimal cost-to-go function by using Monte-Carlo simulation and parametric architectures (such as neural networks, or feature-based architectures of the type discussed in Chapter 6 of Vol. I). We note that extensive treatments of this subject have been given in the books by Bertsekas and Tsitsiklis [BeT96], and Sutton and Barto [SuB98]. However, the present volume focuses to a great extent on new research that became available after the appearance of these books.

Throughout this volume we concentrate on the perfect information case, where each decision is made with exact knowledge of the current system state. Imperfect state information problems can be treated, as in Chapter 5 of Vol. I, by reformulation into perfect information problems involving a sufficient statistic. History-dependent policies, where the control may depend on the entire system history up to the current stage, have been excluded from our development. The reason is that they cannot result in cost reduction, as we show in Section 1.1.4.

For the sake of mathematical rigor, we explicitly assume that the disturbance space is countable, so that the calculus of discrete probability applies throughout our development. In particular, every expected value arising in our analysis is defined as an infinite sum of a countable number of terms. However, on occasion we pause to discuss how some of our results can be used to solve problems with an uncountable disturbance space. For the benefit of the mathematically advanced reader, we have also provided in Appendix A an orientation on the central mathematical issues for a rigorous theory of dynamic programming and stochastic control in general spaces. For a detailed development, we refer to the research monograph by Bertsekas and Shreve [BeS78], which can be freely downloaded from the internet.

1.1 MINIMIZATION OF TOTAL COST – INTRODUCTION

We now formulate the total cost minimization problem, which is the subject of this chapter and the next two. This is an infinite horizon, stationary version of the basic problem of Chapter 1 of Vol. I.

Total Cost Infinite Horizon Problem

Consider the stationary discrete-time dynamic system

$$x_{k+1} = f(x_k, u_k, w_k), \quad k = 0, 1, \dots, \quad (1.1)$$

where for all k , the state x_k is an element of a space S , the control u_k is an element of a space C , and the random disturbance w_k is an element of a space D . We assume that D is a countable set. The control u_k is constrained to take values in a given nonempty subset $U(x_k)$ of C , which depends on the current state x_k [$u_k \in U(x_k)$, for all $x_k \in S$]. The random disturbances w_k , $k = 0, 1, \dots$, are characterized by probability distributions $P(\cdot | x_k, u_k)$ that are independent of k , where $P(w_k | x_k, u_k)$ is the probability of occurrence of w_k , when the current state and control are x_k and u_k , respectively. Thus the probability of w_k may depend explicitly on x_k and u_k , but not on values of prior disturbances w_{k-1}, \dots, w_0 .

Given an initial state x_0 , we want to find a policy $\pi = \{\mu_0, \mu_1, \dots\}$, where $\mu_k : S \mapsto C$, $\mu_k(x_k) \in U(x_k)$, for all $x_k \in S$, $k = 0, 1, \dots$, that minimizes the cost function

$$J_\pi(x_0) = \lim_{N \rightarrow \infty} E_{w_k \atop k=0,1,\dots} \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}, \quad (1.2)$$

subject to the system equation constraint (1.1).† The cost per stage $g : S \times C \times D \mapsto \mathbb{R}$ is given, and α is a positive scalar referred to as the *discount factor*.

† In what follows we will generally impose appropriate assumptions on the cost per stage g and the discount factor α that guarantee that the limit defining the total cost $J_\pi(x_0)$ exists. If this limit is not known to exist, we implicitly assume that $J_\pi(x_0)$ is defined as

$$J_\pi(x_0) = \limsup_{N \rightarrow \infty} E_{w_k \atop k=0,1,\dots} \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}.$$

Note that the expected value of the N -stages cost of π is defined as a (possibly infinite) sum, since the disturbances w_k , $k = 0, 1, \dots$, take values in a countable set. Thus there is no need to impose measurability assumptions on g ; see also the discussion in Appendix A. Indeed, the reader may verify that all the subsequent mathematical expressions that involve an expected value can be written as summations over a finite or a countable set, so they make sense without resort to measure-theoretic integration concepts.

We denote by Π the set of all *admissible* policies π , i.e., the set of all sequences of functions $\pi = \{\mu_0, \mu_1, \dots\}$ with $\mu_k : S \mapsto C$, $\mu_k(x) \in U(x)$ for all $x \in S$, $k = 0, 1, \dots$. The optimal cost function J^* is defined by

$$J^*(x) = \min_{\pi \in \Pi} J_\pi(x), \quad x \in S.$$

An optimal policy, for a given initial state x , is one that attains the optimal cost $J^*(x)$. This policy may depend on x , but we will generally find that for most problems, an optimal policy, when it exists, may be chosen to be independent of the initial state. Very often, such a policy may be taken to be *stationary*, i.e., have the form $\pi = \{\mu, \mu, \dots\}$, in which case it is referred to as the stationary policy μ . We say that μ is optimal if $J_\mu(x) = J^*(x)$ for all states x .

The cost $J_\pi(x_0)$ given by Eq. (1.2) represents the limit of expected finite horizon costs. These costs are well defined as discussed in Section 1.5 of Vol. I. Another possibility would be to minimize over π the expected infinite horizon cost

$$\underset{\substack{w_k \\ k=0,1,\dots}}{E} \left\{ \sum_{k=0}^{\infty} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}.$$

Such a cost would require a far more complex mathematical formulation (a probability measure on the space of all disturbance sequences; see [BeS78]). However, we mention that, under the assumptions that we will be using, the preceding expression is equal to the cost given by Eq. (1.2). This may be proved by using the monotone convergence theorem (see Section 3.1) and other stochastic convergence theorems, which allow interchange of limit and expectation under appropriate conditions.

We finally note that while we have restricted the disturbances to take values in a countable set, our model is considerably more general than a model where the system is a controlled Markov chain with a countable number of states. For example our model includes as a special case deterministic problems with arbitrary state and control spaces.

1.1.1 The Finite-Horizon DP Algorithm

Consider any admissible policy $\pi = \{\mu_0, \mu_1, \dots\}$, any positive integer N , and any function $J : S \mapsto \mathbb{R}$. Suppose that we accumulate the costs of the first N stages, and we add to them some terminal cost of the form $\alpha^N J(x_N)$, where J is some function, for a total expected cost

$$\underset{\substack{w_k \\ k=0,1,\dots}}{E} \left\{ \alpha^N J(x_N) + \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}.$$

The minimum of this cost over π can be calculated by starting with $\alpha^N J(x)$ and by carrying out N iterations of the corresponding DP algorithm, as given in Section 1.3 of Vol. I. This algorithm is given for, $k = 1, \dots, N$, by

$$J_{N-k}(x) = \min_{u \in U(x)} E\{\alpha^{N-k} g(x, u, w) + J_{N-k+1}(f(x, u, w))\}, \quad (1.3)$$

with the initial condition

$$J_N(x) = \alpha^N J(x),$$

where $J_{N-k}(x)$ denotes the optimal cost of the last k stages starting from state x . For each initial state x , the optimal N -stage cost is $J_0(x)$, obtained from the last step of the algorithm.

To rewrite the DP algorithm in more convenient form, consider for all k and x , the functions V_k given by

$$V_k(x) = \frac{J_{N-k}(x)}{\alpha^{N-k}}.$$

Then $V_N(x)$ is the optimal N -stage cost $J_0(x)$, while the DP recursion (1.3) can be equivalently be written in terms of the functions V_k as

$$V_{k+1}(x) = \min_{u \in U(x)} E\{g(x, u, w) + \alpha V_k(f(x, u, w))\}, \quad k = 0, 1, \dots, N-1,$$

with the initial condition

$$V_0(x) = J(x).$$

The above algorithm can be used to calculate *all* the optimal finite horizon cost functions with a *single* DP recursion. In particular, suppose that we have computed the optimal $(N-1)$ -stage cost function V_{N-1} . Then, to calculate the optimal N -stage cost function V_N , we do not need to execute the N -stage DP algorithm. Instead, we can calculate V_N using the one-stage iteration

$$V_N(x) = \min_{u \in U(x)} E\{g(x, u, w) + \alpha V_{N-1}(f(x, u, w))\}.$$

More generally, starting with some terminal cost function, we can consider applying repeatedly the DP iteration as above. With each application, we will be obtaining the optimal cost function of some finite horizon problem, whose horizon will be longer by one stage over the horizon of the preceding problem. Note that this convenience is possible only because we are dealing with a stationary system and a common cost function g for all stages.

1.1.2 Shorthand Notation and Monotonicity

The preceding method of calculating finite horizon optimal costs motivates the introduction of two mappings that play an important theoretical role, and provide a convenient shorthand notation in expressions that would be too complicated to write otherwise.

For any function $J : S \mapsto \mathbb{R}$, we consider the function obtained by applying the DP mapping to J , and we denote it by†

$$(TJ)(x) = \min_{u \in U(x)} E \{g(x, u, w) + \alpha J(f(x, u, w))\}, \quad x \in S, \quad (1.4)$$

where $E\{\cdot\}$ denotes expected value over w with respect to the distribution $P(w | x, u)$. Since $(TJ)(\cdot)$ is itself a function defined on the state space S , we view T as a mapping that transforms the function J on S into the function TJ on S . Note that TJ is the optimal cost function for the one-stage problem that has stage cost g and terminal cost αJ .

Similarly, for any function $J : S \mapsto \mathbb{R}$ and any control function $\mu : S \mapsto C$, we denote

$$(T_\mu J)(x) = E \{g(x, \mu(x), w) + \alpha J(f(x, \mu(x), w))\}, \quad x \in S. \quad (1.5)$$

Again, $T_\mu J$ may be viewed as the cost function associated with μ for the one-stage problem that has stage cost g and terminal cost αJ .

We will denote by T^k the composition of the mapping T with itself k times; i.e., for all k we write

$$(T^k J)(x) = (T(T^{k-1} J))(x), \quad x \in S.$$

Thus $T^k J$ is the function obtained by applying the mapping T to the function $T^{k-1} J$. For convenience, we also write

$$(T^0 J)(x) = J(x), \quad x \in S.$$

Similarly, $T_\mu^k J$ is defined by

$$(T_\mu^k J)(x) = (T_\mu(T_\mu^{k-1} J))(x), \quad x \in S,$$

and

$$(T_\mu^0 J)(x) = J(x), \quad x \in S.$$

It can be verified by induction that $(T^k J)(x)$ is the optimal cost for the k -stage, α -discounted problem with initial state x , cost per stage g , and terminal cost function $\alpha^k J$. Similarly, $(T_\mu^k J)(x)$ is the cost of a policy

† Whenever we use the mapping T , we will impose sufficient assumptions to guarantee that the expected value involved in Eq. (1.4) is well defined.

$\{\mu, \mu, \dots\}$ for the same problem. To illustrate the case where $k = 2$, note that

$$\begin{aligned} (T^2 J)(x) &= \min_{u \in U(x)} E\{g(x, u, w) + \alpha(TJ)(f(x, u, w))\} \\ &= \min_{u_0 \in U(x)} \min_{w_0} E \left\{ g(x, u_0, w_0) + \alpha \min_{u_1 \in U(f(x, u_0, w_0))} \min_{w_1} E \left\{ g(f(x, u_0, w_0), u_1, w_1) \right. \right. \\ &\quad \left. \left. + \alpha J(f(f(x, u_0, w_0), u_1, w_1)) \right\} \right\} \\ &= \min_{u_0 \in U(x)} \min_{w_0} E \left\{ g(x, u_0, w_0) + \min_{u_1 \in U(f(x, u_0, w_0))} \min_{w_1} E \left\{ \alpha g(f(x, u_0, w_0), u_1, w_1) \right. \right. \\ &\quad \left. \left. + \alpha^2 J(f(f(x, u_0, w_0), u_1, w_1)) \right\} \right\}. \end{aligned}$$

The last expression can be recognized as the DP algorithm for the 2-stage, α -discounted problem with initial state x , cost per stage g , and terminal cost function $\alpha^2 J$.

Finally, consider a k -stage policy $\pi = \{\mu_0, \mu_1, \dots, \mu_{k-1}\}$. Then, the expression $(T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{k-1}} J)(x)$ is defined recursively for $i = 0, \dots, k-2$ by

$$(T_{\mu_i} T_{\mu_{i+1}} \cdots T_{\mu_{k-1}} J)(x) = (T_{\mu_i}(T_{\mu_{i+1}} \cdots T_{\mu_{k-1}} J))(x)$$

and represents the cost of the policy π for the k -stage, α -discounted problem with initial state x , cost per stage g , and terminal cost function $\alpha^k J$.

The following monotonicity property plays a fundamental role in the developments of this volume.

Lemma 1.1.1: (Monotonicity Lemma) For any functions $J : S \mapsto \mathbb{R}$ and $J' : S \mapsto \mathbb{R}$, such that

$$J(x) \leq J'(x), \quad \text{for all } x \in S,$$

and for any stationary policy $\mu : S \mapsto C$, we have

$$(T^k J)(x) \leq (T^k J')(x), \quad \text{for all } x \in S, k = 1, 2, \dots,$$

$$(T_\mu^k J)(x) \leq (T_\mu^k J')(x), \quad \text{for all } x \in S, k = 1, 2, \dots$$

Proof: If we view $(T^k J)(x)$ and $(T_\mu^k J)(x)$ as k -stage problem costs with the terminal cost function $\alpha^k J$, the result becomes clear: as the terminal cost function increases uniformly so do the k -stage costs. (Alternatively,

one may prove the lemma by using a straightforward induction argument.)
Q.E.D.

For any two functions $J : S \mapsto \mathbb{R}$ and $J' : S \mapsto \mathbb{R}$, we write

$$J \leq J' \quad \text{if } J(x) \leq J'(x) \text{ for all } x \in S.$$

With this notation, Lemma 1.1.1 is stated as

$$J \leq J' \quad \Rightarrow \quad T^k J \leq T^k J', \quad k = 1, 2, \dots,$$

$$J \leq J' \quad \Rightarrow \quad T_\mu^k J \leq T_\mu^k J', \quad k = 1, 2, \dots$$

Let us also denote by $e : S \mapsto \mathbb{R}$ the unit function that takes the value 1 identically on S :

$$e(x) = 1, \quad \text{for all } x \in S.$$

We have from the definitions (1.4) and (1.5) of T and T_μ , for any function $J : S \mapsto \mathbb{R}$ and scalar r

$$(T(J + re))(x) = (TJ)(x) + \alpha r, \quad x \in S,$$

$$(T_\mu(J + re))(x) = (T_\mu J)(x) + \alpha r, \quad x \in S.$$

More generally, the following lemma can be verified by induction using the preceding two relations.

Lemma 1.1.2: For every k , function $J : S \mapsto \mathbb{R}$, stationary policy μ , and scalar r , we have

$$(T^k(J + re))(x) = (T^k J)(x) + \alpha^k r, \quad \text{for all } x \in S,$$

$$(T_\mu^k(J + re))(x) = (T_\mu^k J)(x) + \alpha^k r, \quad \text{for all } x \in S.$$

1.1.3 A Preview of Infinite Horizon Results

Let us speculate on the type of results that we will be aiming for.

- (a) *Convergence of the DP Algorithm.* Let J_0 denote the zero function [$J_0(x) = 0$ for all x]. Since the infinite horizon cost of a policy is, by definition, the limit of its k -stage costs as $k \rightarrow \infty$, it is reasonable to speculate that the optimal infinite horizon cost is equal to the limit of the optimal k -stage costs; i.e.,

$$J^*(x) = \lim_{k \rightarrow \infty} (T^k J_0)(x), \quad x \in S.$$

This means that if we start with the zero function J_0 and iterate with the DP algorithm indefinitely, we will get in the limit the optimal cost function J^* . Also, for $\alpha < 1$ and a bounded function J , a terminal cost $\alpha^k J$ diminishes with k , so it is reasonable to speculate that, if $\alpha < 1$,

$$J^*(x) = \lim_{k \rightarrow \infty} (T^k J)(x), \quad \text{for all } x \in S \text{ and bounded functions } J.$$

- (b) *Bellman's Equation.* Since by definition we have for all $x \in S$

$$(T^{k+1} J_0)(x) = \min_{u \in U(x)} E_w \{g(x, u, w) + \alpha(T^k J_0)(f(x, u, w))\},$$

it is reasonable to speculate that if $\lim_{k \rightarrow \infty} T^k J_0 = J^*$ as in (a) above, then we must have by taking limit as $k \rightarrow \infty$,

$$J^*(x) = \min_{u \in U(x)} E_w \{g(x, u, w) + \alpha J^*(f(x, u, w))\}, \quad x \in S,$$

or, equivalently,

$$J^* = TJ^*.$$

This is known as *Bellman's equation* and asserts that the optimal cost function J^* is a fixed point of the mapping T . We will see that Bellman's equation holds for all the total cost minimization problems that we will consider, although depending on our assumptions, its proof can be quite complex.

- (c) *Characterization of Optimal Stationary Policies.* If we view Bellman's equation as the DP algorithm taken to its limit as $k \rightarrow \infty$, it is reasonable to speculate that if $\mu(x)$ attains the minimum in the right-hand side of Bellman's equation for all x , then the stationary policy μ is optimal.

Most of the analysis of total cost infinite horizon problems revolves around the above three issues and also around the issue of efficient computation of J^* and an optimal stationary policy. For the discounted cost problems with bounded cost per stage considered in this chapter, and for stochastic shortest path problems under our assumptions of Chapter 2, the preceding conjectures are correct. For problems with unbounded costs per stage and for stochastic shortest path problems where our assumptions of Chapter 2 are violated, there may be counterintuitive mathematical phenomena that invalidate some of the preceding conjectures. This illustrates that infinite horizon problems should be approached carefully and with mathematical precision.

1.1.4 Randomized and History-Dependent Policies

Our formulation of the total cost infinite horizon problem involves certain restrictions on the admissible policies that facilitate the analysis. In particular, we assume that at each time k , the control is applied with knowledge of the current state x_k . Such policies are called *Markov* because they do not involve dependence on states beyond the current. However, what if the control were allowed to depend on the entire past history

$$h_k = \{x_0, u_0, \dots, x_{k-1}, u_{k-1}, x_k\},$$

which ordinarily would be available at time k ? Is it possible that better performance can be achieved in this way?

Another related question is whether we can achieve better performance with *randomized* policies where instead of choosing a single control to apply at time k , we select a probability distribution over the control constraint set, and choose a control randomly according to this distribution.

To address this question, let us consider *randomized history-dependent policies* $\pi = \{\mu_0, \mu_1, \dots\}$, where μ_k is a function that maps a history h_k into a probability distribution $\mu_k(u_k | h_k)$ over $U(x_k)$. For mathematical simplicity, in this section we will assume that in addition to the disturbance space, the control space is also countable. As a result, for a fixed initial state, the set of possible histories h_k is countable, so the distributions $\mu_k(u_k | h_k)$ are defined on countable sets and can be manipulated without the need for tools from measure theoretic probability theory.

Let us also consider a special case, *randomized Markov policies* $\pi = \{\mu_0, \mu_1, \dots\}$, where μ_k is a function that maps the state x_k into a probability distribution $\mu_k(u_k | x_k)$ over the control constraint set $U(x_k)$.

A given distribution over a countable subset of initial states and a randomized history-dependent policy define a probability distribution on the countable set of state-control pair (x_k, u_k) of each stage k that will occur with positive probability. An important result is that any such probability distribution can also be generated by a randomized Markov policy, as shown by the following proposition.

Proposition 1.1.1: (Adequacy of Markov Policies) Assume that the control space is countable, and consider an initial state distribution that takes values over a countable set. The probability distribution of each pair (x_k, u_k) and the expected cost of each stage corresponding to a randomized history-dependent policy can also be obtained with a randomized Markov policy.

Proof: Let $\pi = \{\mu_0, \mu_1, \dots\}$ be a randomized history-dependent policy, and let $\xi_k(x_k)$ and $\zeta_k(x_k, u_k)$ be the corresponding distributions of x_k

and (x_k, u_k) , respectively. Consider a randomized Markov policy $\bar{\pi} = \{\bar{\mu}_0, \bar{\mu}_1, \dots\}$, where $\bar{\mu}_k$ is defined for all x_k with $\xi_k(x_k) > 0$ by

$$\bar{\mu}_k(u_k | x_k) = \frac{\zeta_k(x_k, u_k)}{\xi_k(x_k)}.$$

Let $\bar{\xi}_k(x_k)$ and $\bar{\zeta}_k(x_k, u_k)$ be the corresponding distributions of x_k and (x_k, u_k) , respectively. We will show by induction that for all k , x_k , and u_k , we have

$$\xi_k(x_k) = \bar{\xi}_k(x_k), \quad \zeta_k(x_k, u_k) = \bar{\zeta}_k(x_k, u_k). \quad (1.6)$$

It is sufficient to show this for all k , x_k , and u_k such that $\zeta_k(x_k, u_k) > 0$.

Indeed, for $k = 0$, $\xi_0(x_0)$ and $\bar{\xi}_0(x_0)$ are both equal to the distribution of the initial state, while

$$\bar{\zeta}_0(x_0, u_0) = \bar{\xi}_0(x_0) \bar{\mu}_0(u_0 | x_0) = \bar{\xi}_0(x_0) \frac{\zeta_0(x_0, u_0)}{\xi_0(x_0)} = \zeta_0(x_0, u_0).$$

Suppose that Eq. (1.6) holds for some k . Then, we have

$$\begin{aligned} \bar{\xi}_{k+1}(x_{k+1}) &= \sum_{x_k, u_k} \bar{\zeta}_k(x_k, u_k) p_{x_{k+1}|x_k}(u_k) \\ &= \sum_{x_k, u_k} \bar{\xi}_k(x_k) \bar{\mu}_k(u_k | x_k) p_{x_{k+1}|x_k}(u_k) \\ &= \sum_{x_k, u_k} \bar{\xi}_k(x_k) \frac{\zeta_k(x_k, u_k)}{\xi_k(x_k)} p_{x_{k+1}|x_k}(u_k) \\ &= \sum_{x_k, u_k} \zeta_k(x_k, u_k) p_{x_{k+1}|x_k}(u_k) \\ &= \xi_{k+1}(x_{k+1}), \end{aligned}$$

where $p_{x_{k+1}|x_k}(u_k)$ are the transition probabilities of the system, and the summation is over all pairs (x_k, u_k) such that $\zeta_k(x_k, u_k) > 0$. Furthermore,

$$\begin{aligned} \bar{\zeta}_{k+1}(x_{k+1}, u_{k+1}) &= \bar{\xi}_{k+1}(x_{k+1}) \bar{\mu}_k(u_{k+1} | x_{k+1}) \\ &= \bar{\xi}_{k+1}(x_{k+1}) \frac{\zeta_{k+1}(x_{k+1}, u_{k+1})}{\xi_{k+1}(x_{k+1})} \\ &= \zeta_{k+1}(x_{k+1}, u_{k+1}), \end{aligned}$$

thereby completing the induction. Thus π and $\bar{\pi}$ generate the same state-control pair distributions. From this it also follows that their corresponding expected costs of every stage are equal. Q.E.D.

The preceding proposition shows that the expected cost of any history-dependent randomized policy over a finite horizon can be replicated with

a Markov randomized policy. This implies that for a finite horizon problem, one can safely restrict attention to Markov policies, and need not consider history-dependent policies. Furthermore, the same is true for an infinite horizon problem, provided the N -stage costs of a history-dependent randomized policy converge to its infinite horizon cost as $N \rightarrow \infty$. In particular this is true for the three major classes of total cost problems that we will discuss: discounted problems with bounded cost per stage (the present chapter), problems with nonnegative cost per stage (Chapter 3), and problems with nonpositive cost per stage (Chapter 3).

Is it possible to dispense with randomized policies and restrict oneself to deterministic Markov policies? This is true very often. By this we mean, that for many classes of interesting total cost problems, it can be shown that the optimal cost using randomized policies is the same as the optimal cost using deterministic policies, and that if there exists an optimal (possibly randomized) policy, there exists an optimal deterministic policy. Included are finite horizon problems, the discounted cost problems with bounded cost per stage of the present chapter, and the finite state and control spaces models of Chapters 2 and 4; in fact for all these problems, it will be shown that one may restrict attention to *stationary* deterministic Markov policies. The exceptions arise primarily in the unbounded cost per stage models of Chapter 3, and also in some models not considered in this book, such as constrained DP problems. The book by Bertsekas and Shreve [BeS78] delineates some situations where randomized policies may be of genuine interest. Our approach in this book is to formulate problems in terms of deterministic Markov policies (which may depend, however, on the initial state), to discuss the existence of optimal policies within this class (or a subset thereof, such as stationary policies), and to comment selectively on what may be possible with randomized policies.

1.2 DISCOUNTED PROBLEMS WITH BOUNDED COST PER STAGE

We now discuss the simplest type of infinite horizon problem. We assume the following:

Assumption D (Discounted Cost – Bounded Cost per Stage):
The cost per stage g satisfies

$$|g(x, u, w)| \leq M, \quad \text{for all } (x, u, w) \in S \times C \times D,$$

where M is some scalar. Furthermore, $0 < \alpha < 1$.

Boundedness of the cost per stage is not as restrictive as might appear. It holds for problems where the spaces S , C , and D are finite sets. Even if they are not finite, in a computational solution of the problem they will ordinarily be approximated by finite sets. Also, when S , C , and D are subsets of Euclidean spaces, it is often possible to reformulate the problem so that they are bounded, and as a result the cost is bounded.

The following proposition shows that the DP algorithm converges to the optimal cost function J^* for an arbitrary bounded starting function J . This will follow as a consequence of Assumption D, which implies that the “tail” of the cost after stage N ,

$$\lim_{K \rightarrow \infty} E \left\{ \sum_{k=N}^K \alpha^k g(x_k, \mu_k(x_k), w_k) \right\},$$

diminishes to zero as $N \rightarrow \infty$. Furthermore, when a terminal cost $\alpha^N J(x_N)$ is added to the N -stage cost, its effect diminishes to zero as $N \rightarrow \infty$ if J is bounded.

Proposition 1.2.1: (Convergence of the DP Algorithm) For any bounded function $J : S \mapsto \mathbb{R}$, the optimal cost function satisfies

$$J^*(x) = \lim_{N \rightarrow \infty} (T^N J)(x), \quad \text{for all } x \in S.$$

Proof: For every positive integer K , initial state $x_0 \in S$, and policy $\pi = \{\mu_0, \mu_1, \dots\}$, we break down the cost $J_\pi(x_0)$ into the portions incurred over the first K stages and over the remaining stages

$$\begin{aligned} J_\pi(x_0) &= \lim_{N \rightarrow \infty} E \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\} \\ &= E \left\{ \sum_{k=0}^{K-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\} \\ &\quad + \lim_{N \rightarrow \infty} E \left\{ \sum_{k=K}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}. \end{aligned}$$

Since by Assumption D we have $|g(x_k, \mu_k(x_k), w_k)| \leq M$, we also obtain

$$\left| \lim_{N \rightarrow \infty} E \left\{ \sum_{k=K}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\} \right| \leq M \sum_{k=K}^{\infty} \alpha^k = \frac{\alpha^K M}{1 - \alpha}.$$

Using these relations, it follows that

$$\begin{aligned} J_\pi(x_0) - \frac{\alpha^K M}{1-\alpha} - \alpha^K \max_{x \in S} |J(x)| \\ \leq E \left\{ \alpha^K J(x_K) + \sum_{k=0}^{K-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\} \\ \leq J_\pi(x_0) + \frac{\alpha^K M}{1-\alpha} + \alpha^K \max_{x \in S} |J(x)|. \end{aligned}$$

By taking the minimum over π , we obtain for all x_0 and K ,

$$\begin{aligned} J^*(x_0) - \frac{\alpha^K M}{1-\alpha} - \alpha^K \max_{x \in S} |J(x)| \\ \leq (T^K J)(x_0) \\ \leq J^*(x_0) + \frac{\alpha^K M}{1-\alpha} + \alpha^K \max_{x \in S} |J(x)|, \end{aligned} \tag{1.7}$$

and by taking the limit as $K \rightarrow \infty$, the result follows. **Q.E.D.**

Note that based on the preceding proposition, the DP algorithm may be used to compute at least an approximation to J^* . This computational method together with some additional methods will be examined in the next section.

Given any stationary policy μ , we can consider a modified discounted problem, which is the same as the original except that the control constraint set contains only one element for each state x , the control $\mu(x)$; i.e., the control constraint set is $\tilde{U}(x) = \{\mu(x)\}$ instead of $U(x)$. Proposition 1.2.1 applies to this modified problem and yields the following corollary:

Corollary 1.2.1.1: For every stationary policy μ , the associated cost function satisfies

$$J_\mu(x) = \lim_{N \rightarrow \infty} (T_\mu^N J)(x), \quad \text{for all } x \in S.$$

The next proposition shows that J^* is the unique solution of Bellman's equation.

Proposition 1.2.2: (Bellman's Equation) The optimal cost function J^* satisfies

$$J^*(x) = \min_{u \in U(x)} E_w \{g(x, u, w) + \alpha J^*(f(x, u, w))\}, \quad \text{for all } x \in S, \quad (1.8)$$

or, equivalently,

$$J^* = TJ^*.$$

Furthermore, J^* is the unique solution of this equation within the class of bounded functions.

Proof: From Eq. (1.7), we have for all $x \in S$ and N ,

$$J^*(x) - \frac{\alpha^N M}{1-\alpha} \leq (T^N J_0)(x) \leq J^*(x) + \frac{\alpha^N M}{1-\alpha},$$

where J_0 is the zero function [$J_0(x) = 0$ for all $x \in S$]. Applying the mapping T to this relation and using the Monotonicity Lemma 1.1.1 as well as Lemma 1.1.2, we obtain for all $x \in S$ and N

$$(TJ^*)(x) - \frac{\alpha^{N+1} M}{1-\alpha} \leq (T^{N+1} J_0)(x) \leq (TJ^*)(x) + \frac{\alpha^{N+1} M}{1-\alpha}.$$

By taking the limit as $N \rightarrow \infty$ in the preceding relation and using the fact

$$\lim_{N \rightarrow \infty} (T^{N+1} J_0)(x) = J^*(x)$$

(cf. Prop. 1.2.1), we obtain $J^* = TJ^*$.

To show uniqueness, observe that if J is bounded and satisfies $J = TJ$, then $J = \lim_{N \rightarrow \infty} T^N J$, so by Prop. 1.2.1, we have $J = J^*$. Q.E.D.

Based on the same reasoning we used to obtain Cor. 1.2.1.1 from Prop. 1.2.1, we have:

Corollary 1.2.2.1: For every stationary policy μ , the associated cost function satisfies

$$J_\mu(x) = E_w \{g(x, \mu(x), w) + \alpha J_\mu(f(x, \mu(x), w))\}, \quad \text{for all } x \in S,$$

or, equivalently,

$$J_\mu = T_\mu J_\mu.$$

Furthermore, J_μ is the unique solution of this equation within the class of bounded functions.

The next proposition characterizes stationary optimal policies.

Proposition 1.2.3: (Necessary and Sufficient Condition for Optimality) A stationary policy μ is optimal if and only if $\mu(x)$ attains the minimum in Bellman's equation (1.8) for each $x \in S$; i.e.,

$$TJ^* = T_\mu J^*.$$

Proof: If $TJ^* = T_\mu J^*$, then using Bellman's equation ($J^* = TJ^*$), we have $J^* = T_\mu J^*$, so by the uniqueness part of Cor. 1.2.2.1, we obtain $J^* = J_\mu$; i.e., μ is optimal. Conversely, if the stationary policy μ is optimal, we have $J^* = J_\mu$, which by Cor. 1.2.2.1, yields $J^* = T_\mu J^*$. Combining this with Bellman's equation ($J^* = TJ^*$), we obtain $TJ^* = T_\mu J^*$. Q.E.D.

Note that Prop. 1.2.3 implies the existence of an optimal stationary policy when the minimum in the right-hand side of Bellman's equation is attained for all $x \in S$. In particular, when $U(x)$ is finite for each $x \in S$, an optimal stationary policy is guaranteed to exist.

We finally show the following convergence rate estimate for any function J that is bounded:

$$\max_{x \in S} |(T^k J)(x) - J^*(x)| \leq \alpha^k \max_{x \in S} |J(x) - J^*(x)|, \quad k = 0, 1, \dots$$

This relation is obtained by using the fact $T^k J^* = J^*$ (which follows from Bellman's equation) and the following result:

Proposition 1.2.4: For any two bounded functions $J : S \mapsto \mathbb{R}$, $J' : S \mapsto \mathbb{R}$, and for all $k = 0, 1, \dots$, there holds

$$\max_{x \in S} |(T^k J)(x) - (T^k J')(x)| \leq \alpha^k \max_{x \in S} |J(x) - J'(x)|. \quad (1.9)$$

Proof: Denote

$$c = \max_{x \in S} |J(x) - J'(x)|.$$

Then we have

$$J(x) - c \leq J'(x) \leq \overline{J(x)} + c, \quad x \in S.$$

Applying T^k in this relation and using the Monotonicity Lemma 1.1.1 as well as Lemma 1.1.2, we obtain

$$(T^k J)(x) - \alpha^k c \leq (T^k J')(x) \leq (T^k J)(x) + \alpha^k c, \quad x \in S.$$

It follows that

$$|(T^k J)(x) - (T^k J')(x)| \leq \alpha^k c, \quad x \in S,$$

which proves the result. **Q.E.D.**

As earlier, we have:

Corollary 1.2.4.1: For any two bounded functions $J : S \mapsto \mathbb{R}$, $J' : S \mapsto \mathbb{R}$, and any stationary policy μ , we have

$$\max_{x \in S} |(T_\mu^k J)(x) - (T_\mu^k J')(x)| \leq \alpha^k \max_{x \in S} |J(x) - J'(x)|, \quad k = 0, 1, \dots$$

Example 1.2.1 (Machine Replacement)

Consider an infinite horizon discounted version of a problem we formulated in Section 1.1 of Vol. I. Here, we want to operate efficiently a machine that can be in any one of n states, denoted $1, 2, \dots, n$. State 1 corresponds to a machine in perfect condition. The transition probabilities p_{ij} are given. There is a cost $g(i)$ for operating for one time period the machine when it is in state i . The options at the start of each period are to (a) let the machine operate one more period in the state it currently is, or (b) replace the machine with a new machine (state 1) at a cost R . Once replaced, the machine is guaranteed to stay in state 1 for one period; in subsequent periods, it may deteriorate to states $j \geq 1$ according to the transition probabilities p_{1j} . We assume an infinite horizon and a discount factor $\alpha \in (0, 1)$, so the theory of this section applies.

Bellman's equation (cf. Prop. 1.2.2) takes the form

$$J^*(i) = \min \left[R + g(1) + \alpha J^*(1), g(i) + \alpha \sum_{j=1}^n p_{ij} J^*(j) \right], \quad i = 1, \dots, n.$$

By Prop. 1.2.3, a stationary policy is optimal if it replaces at states i where

$$R + g(1) + \alpha J^*(1) < g(i) + \alpha \sum_{j=1}^n p_{ij} J^*(j),$$

and it does not replace at states i where

$$R + g(1) + \alpha J^*(1) > g(i) + \alpha \sum_{j=1}^n p_{ij} J^*(j).$$

We can use the convergence of the DP algorithm (cf. Prop. 1.2.1) to characterize the optimal cost function using properties of the finite horizon cost functions. In particular, the DP algorithm starting from the zero function takes the form

$$\begin{aligned} J_0(i) &= 0, \\ (TJ_0)(i) &= \min[R + g(1), g(i)], \\ (T^k J_0)(i) &= \min \left[R + g(1) + \alpha(T^{k-1} J_0)(1), g(i) + \alpha \sum_{j=1}^n p_{ij}(T^{k-1} J_0)(j) \right]. \end{aligned}$$

Assume that $g(i)$ is nondecreasing in i , and that the transition probabilities satisfy

$$\sum_{j=1}^n p_{ij} J(j) \leq \sum_{j=1}^n p_{(i+1)j} J(j), \quad i = 1, \dots, n-1, \quad (1.10)$$

for all functions $J(i)$, which are monotonically nondecreasing in i . This assumption is satisfied if

$$p_{ij} = 0, \quad \text{if } j < i,$$

i.e., the machine cannot go to a better state with usage, and

$$p_{ij} \leq p_{(i+1)j}, \quad \text{if } i < j,$$

i.e., the chance of going to a given bad state j from a better state $i < j$ increases as i gets worse. Since $g(i)$ is nondecreasing in i , we have that $(TJ_0)(i)$ is nondecreasing in i , and in view of the assumption (1.10), the same is true for $(T^2 J_0)(i)$. Similarly, it is seen that, for all k , $(T^k J_0)(i)$ is nondecreasing in i and so is its limit

$$J^*(i) = \lim_{k \rightarrow \infty} (T^k J_0)(i).$$

This is intuitively clear: the optimal cost should not decrease as the machine starts at a worse initial state. It follows that the function

$$g(i) + \alpha \sum_{j=1}^n p_{ij} J^*(j)$$

is nondecreasing in i . Consider the set of states

$$S_R = \left\{ i \mid R + g(1) + \alpha J^*(1) \leq g(i) + \alpha \sum_{j=1}^n p_{ij} J^*(j) \right\},$$

and let

$$i^* = \begin{cases} \text{smallest state in } S_R & \text{if } S_R \text{ is nonempty,} \\ n+1 & \text{otherwise.} \end{cases}$$

Then, an optimal policy takes the form

$$\text{replace if and only if } i \geq i^*,$$

as shown in Fig. 1.2.1.

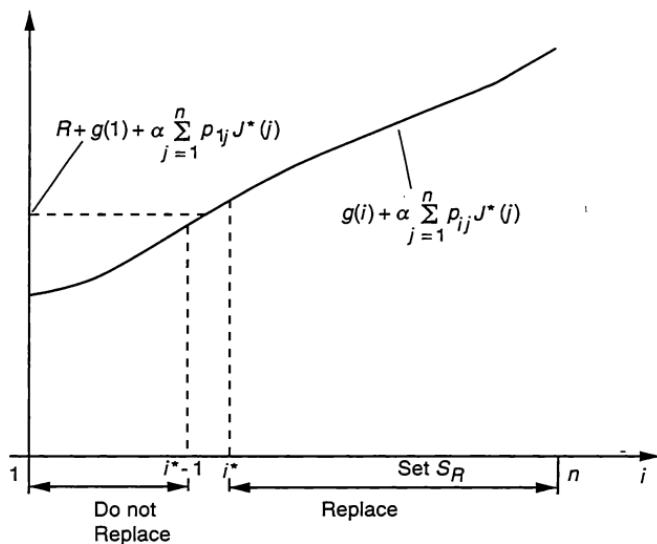


Figure 1.2.1 Determining the optimal policy in the machine replacement example.

1.3 FINITE-STATE SYSTEMS – COMPUTATIONAL METHODS

In this section we discuss several alternative approaches for numerical solution of the discounted problem with bounded cost per stage. The first approach, *value iteration*, is essentially the DP algorithm and yields in the limit the optimal cost function and an optimal policy, as discussed in the preceding section. We will describe some variations aimed at accelerating convergence. Two other approaches, *policy iteration* and *linear programming*, terminate in a finite number of iterations (assuming the number of states and controls are finite). However, when the number of states is large, these approaches are impractical because of large overhead per iteration. Several variants of these methods have been proposed to overcome this difficulty, including some that are based on approximations.

In Chapter 6, we will consider some additional methods, which are well-suited for systems that are hard to model but relatively easy to simulate. In particular, we will assume there that the transition probabilities of the problem are unknown, but the system's dynamics and cost structure can be observed through simulation. We will discuss various approximate forms of value iteration and policy iteration, which use an approximation architecture.

Throughout this section we assume a discounted problem (Assumption D holds). We further assume that the state, control, and disturbance

spaces underlying the problem are finite sets, so that we are dealing in effect with the control of a finite-state Markov chain.

We first translate some of our earlier analysis in a notation that is more convenient for Markov chains. Let the state space S consist of n states denoted by $1, 2, \dots, n$:

$$S = \{1, 2, \dots, n\}.$$

The transition probabilities are denoted by

$$p_{ij}(u) = P(x_{k+1} = j \mid x_k = i, u_k = u), \quad i, j \in S, u \in U(i).$$

They may be given a priori or they may be calculated from the system equation

$$x_{k+1} = f(x_k, u_k, w_k)$$

and the known probability distribution $P(\cdot \mid x, u)$ of the input disturbance w_k . Indeed, we have

$$p_{ij}(u) = P(W_{ij}(u) \mid i, u),$$

where $W_{ij}(u)$ is the (finite) set

$$W_{ij}(u) = \{w \in D \mid f(i, u, w) = j\}.$$

To simplify notation, we assume that the cost per stage does not depend on w . This amounts to using expected cost per stage in all calculations, which makes no essential difference in the definitions of the mappings T and T_μ of Eqs. (1.4) and (1.5), and in the subsequent analysis. Thus, if $\tilde{g}(i, u, j)$ is the cost of using u at state i and moving to state j , we use as cost per stage the expected cost $g(i, u)$ given by

$$g(i, u) = \sum_{j=1}^n p_{ij}(u) \tilde{g}(i, u, j).$$

The mappings T and T_μ of Eqs. (1.4) and (1.5) can be written as

$$(TJ)(i) = \min_{u \in U(i)} \left[g(i, u) + \alpha \sum_{j=1}^n p_{ij}(u) J(j) \right]; \quad i = 1, 2, \dots, n,$$

$$(T_\mu J)(i) = g(i, \mu(i)) + \alpha \sum_{j=1}^n p_{ij}(\mu(i)) J(j), \quad i = 1, 2, \dots, n.$$

Any function J on S , as well as the functions TJ and $T_\mu J$ may be represented by the n -dimensional vectors

$$J = \begin{bmatrix} J(1) \\ \vdots \\ J(n) \end{bmatrix}, \quad TJ = \begin{bmatrix} (TJ)(1) \\ \vdots \\ (TJ)(n) \end{bmatrix}, \quad T_\mu J = \begin{bmatrix} (T_\mu J)(1) \\ \vdots \\ (T_\mu J)(n) \end{bmatrix}.$$

For a stationary policy μ , we denote by P_μ the transition probability matrix

$$P_\mu = \begin{bmatrix} p_{11}(\mu(1)) & \cdots & p_{1n}(\mu(1)) \\ \vdots & \ddots & \vdots \\ p_{n1}(\mu(n)) & \cdots & p_{nn}(\mu(n)) \end{bmatrix},$$

and by g_μ the cost vector

$$g_\mu = \begin{bmatrix} g(1, \mu(1)) \\ \vdots \\ g(n, \mu(n)) \end{bmatrix}.$$

We can then write in vector notation

$$T_\mu J = g_\mu + \alpha P_\mu J.$$

The cost function J_μ corresponding to a stationary policy μ is, by Cor. 1.2.2.1, the unique solution of the equation

$$J_\mu = T_\mu J_\mu = g_\mu + \alpha P_\mu J_\mu.$$

This equation should be viewed as a system of n linear equations with n unknowns, the components $J_\mu(i)$ of the n -dimensional vector J_μ . The equation can also be written as

$$(I - \alpha P_\mu) J_\mu = g_\mu,$$

or, equivalently,

$$J_\mu = (I - \alpha P_\mu)^{-1} g_\mu,$$

where I denotes the $n \times n$ identity matrix. The invertibility of the matrix $I - \alpha P_\mu$ is assured since we have proved that the system of equations representing $J_\mu = T_\mu J_\mu$ has a unique solution for any vector g_μ (cf. Cor. 1.2.2.1). For another way to see that $I - \alpha P_\mu$ is an invertible matrix, note that the eigenvalues of any transition probability matrix lie within the unit circle of the complex plane. Thus no eigenvalue of αP_μ can be equal to 1, which is the necessary and sufficient condition for $I - \alpha P_\mu$ to be invertible.

1.3.1 Value Iteration and Error Bounds

Here we start with any n -dimensional vector J and successively compute TJ, T^2J, \dots . By Prop. 1.2.1, we have for all i

$$\lim_{k \rightarrow \infty} (T^k J)(i) = J^*(i).$$

Furthermore, by Prop. 1.2.4 [using $J' = J^*$ in Eq. (1.9)], the error sequence $|(T^k J)(i) - J^*(i)|$ is bounded by a constant multiple of α^k , for all $i \in S$. This method is called *value iteration* (sometimes it is also called *successive approximation*). The method can be substantially improved thanks to certain monotonic error bounds, which are easily obtained as a byproduct of the computation.

The following argument is helpful in understanding the nature of these bounds. Let us first break down the cost of a stationary policy μ into the first stage cost and the remainder:

$$J_\mu(i) = g(i, \mu(i)) + \sum_{k=1}^{\infty} \alpha^k E\{g(x_k, \mu(x_k)) \mid x_0 = i\}.$$

It follows that

$$g_\mu + \left(\frac{\alpha \underline{\beta}}{1 - \alpha} \right) e \leq J_\mu \leq g_\mu + \left(\frac{\alpha \bar{\beta}}{1 - \alpha} \right) e, \quad (1.11)$$

where e is the unit vector, $e = (1, 1, \dots, 1)'$, and $\underline{\beta}$ and $\bar{\beta}$ are the minimum and maximum cost per stage:

$$\underline{\beta} = \min_i g(i, \mu(i)), \quad \bar{\beta} = \max_i g(i, \mu(i)).$$

Using the definition of $\underline{\beta}$ and $\bar{\beta}$, we can strengthen the bounds (1.11) as follows:

$$\left(\frac{\beta}{1 - \alpha} \right) e \leq g_\mu + \left(\frac{\alpha \underline{\beta}}{1 - \alpha} \right) e \leq J_\mu \leq g_\mu + \left(\frac{\alpha \bar{\beta}}{1 - \alpha} \right) e \leq \left(\frac{\bar{\beta}}{1 - \alpha} \right) e. \quad (1.12)$$

These bounds will now be applied in the context of the value iteration method.

Suppose that we have a vector J and we compute

$$T_\mu J = \widehat{g_\mu} + \alpha P_\mu J.$$

By subtracting this equation from the relation

$$J_\mu = g_\mu + \alpha P_\mu J_\mu,$$

we obtain

$$J_\mu - J = T_\mu J - J + \alpha P_\mu (J_\mu - J).$$

This equation can be viewed as a *variational form* of the equation $J_\mu = T_\mu J_\mu$, and implies that $J_\mu - J$ is the cost vector associated with the stationary policy μ and a cost per stage vector equal to $T_\mu J - J$. Therefore, the bounds (1.12) apply with J_μ replaced by $J_\mu - J$ and g_μ replaced by $T_\mu J - J$. It follows that

$$\begin{aligned} \left(\frac{\gamma}{1-\alpha} \right) e &\leq T_\mu J - J + \left(\frac{\alpha\gamma}{1-\alpha} \right) e \\ &\leq J_\mu - J \\ &\leq T_\mu J - J + \left(\frac{\alpha\bar{\gamma}}{1-\alpha} \right) e \\ &\leq \left(\frac{\bar{\gamma}}{1-\alpha} \right) e, \end{aligned}$$

where

$$\underline{\gamma} = \min_i [(T_\mu J)(i) - J(i)], \quad \bar{\gamma} = \max_i [(T_\mu J)(i) - J(i)].$$

Equivalently, for every vector J , we have

$$J + \frac{\underline{c}}{\alpha} e \leq T_\mu J + \underline{c} e \leq J_\mu \leq T_\mu J + \bar{c} e \leq J + \frac{\bar{c}}{\alpha} e,$$

where

$$\underline{c} = \frac{\alpha\underline{\gamma}}{1-\alpha}, \quad \bar{c} = \frac{\alpha\bar{\gamma}}{1-\alpha}.$$

The following proposition is obtained by a more sophisticated application of the preceding argument.

Proposition 1.3.1: For every vector J , state i , and k , we have

$$\begin{aligned} (T^k J)(i) + \underline{c}_k &\leq (T^{k+1} J)(i) + \underline{c}_{k+1} \\ &\leq J^*(i) \\ &\leq (T^{k+1} J)(i) + \bar{c}_{k+1} \\ &\leq (T^k J)(i) + \bar{c}_k, \end{aligned} \tag{1.13}$$

where

$$\underline{c}_k = \frac{\alpha}{1-\alpha} \min_{i=1,\dots,n} [(T^k J)(i) - (T^{k-1} J)(i)], \tag{1.14}$$

$$\bar{c}_k = \frac{\alpha}{1-\alpha} \max_{i=1,\dots,n} [(T^k J)(i) - (T^{k-1} J)(i)]. \tag{1.15}$$

Proof: Denote

$$\underline{\gamma} = \min_{i=1,\dots,n} [(TJ)(i) - J(i)].$$

We have

$$J + \underline{\gamma}e \leq TJ. \quad (1.16)$$

Applying T to both sides and using the monotonicity of T , we have

$$TJ + \alpha\underline{\gamma}e \leq T^2J,$$

and, combining this relation with Eq. (1.16), we obtain

$$J + (1 + \alpha)\underline{\gamma}e \leq TJ + \alpha\underline{\gamma}e \leq T^2J. \quad (1.17)$$

This process can be repeated, first applying T to obtain

$$TJ + (\alpha + \alpha^2)\underline{\gamma}e \leq T^2J + \alpha^2\underline{\gamma}e \leq T^3J,$$

and then using Eq. (1.16) to write

$$J + (1 + \alpha + \alpha^2)\underline{\gamma}e \leq TJ + (\alpha + \alpha^2)\underline{\gamma}e \leq T^2J + \alpha^2\underline{\gamma}e \leq T^3J.$$

After k steps, this results in the inequalities

$$\begin{aligned} J + \left(\sum_{i=0}^k \alpha^i \right) \underline{\gamma}e &\leq TJ + \left(\sum_{i=1}^k \alpha^i \right) \underline{\gamma}e \\ &\leq T^2J + \left(\sum_{i=2}^k \alpha^i \right) \underline{\gamma}e \\ &\quad \dots \\ &\leq T^{k+1}J. \end{aligned}$$

Taking the limit as $k \rightarrow \infty$ and using the equality $\underline{c}_1 = \alpha\underline{\gamma}/(1 - \alpha)$, we obtain

$$J + \left(\frac{\underline{c}_1}{\alpha} \right) e \leq TJ + \underline{c}_1 e \leq T^2J + \alpha\underline{c}_1 e \leq J^*, \quad (1.18)$$

where \underline{c}_1 is defined by Eq. (1.14). Replacing J by $T^k J$ in this inequality, we have

$$T^{k+1}J + \underline{c}_{k+1}e \leq J^*,$$

which is the second inequality in Eq. (1.13).

From Eq. (1.17), we obtain

$$\alpha\underline{\gamma} \leq \min_{i=1,\dots,n} [(T^2J)(i) - (TJ)(i)],$$

and consequently

$$\alpha c_1 \leq c_2.$$

Using this relation in Eq. (1.18) yields

$$TJ + \underline{c}_1 e \leq T^2 J + \underline{c}_2 e,$$

and replacing J by $T^{k-1}J$, we have the first inequality in Eq. (1.13). An analogous argument shows the last two inequalities in Eq. (1.13). **Q.E.D.**

We note that the preceding proof does not rely on the finiteness of the state space, and indeed Prop. 1.3.1 can be proved for an infinite state space (see also Exercise 1.9). The following example demonstrates the nature of the error bounds.

Example 1.3.1 (Illustration of the Error Bounds)

Consider a problem where there are two states and two controls

$$S = \{1, 2\}, \quad C = \{u^1, u^2\}.$$

The transition probabilities corresponding to the controls u^1 and u^2 are as shown in Fig. 1.3.1; that is, the transition probability matrices are

$$P(u^1) = \begin{bmatrix} p_{11}(u^1) & p_{12}(u^1) \\ p_{21}(u^1) & p_{22}(u^1) \end{bmatrix} = \begin{bmatrix} 3/4 & 1/4 \\ 3/4 & 1/4 \end{bmatrix},$$

$$P(u^2) = \begin{bmatrix} p_{11}(u^2) & p_{12}(u^2) \\ p_{21}(u^2) & p_{22}(u^2) \end{bmatrix} = \begin{bmatrix} 1/4 & 3/4 \\ 1/4 & 3/4 \end{bmatrix}.$$

The transition costs are

$$g(1, u^1) = 2, \quad g(1, u^2) = 0.5, \quad g(2, u^1) = 1, \quad g(2, u^2) = 3,$$

and the discount factor is $\alpha = 0.9$. The mapping T is given for $i = 1, 2$ by

$$(TJ)(i) = \min \left\{ g(i, u^1) + \alpha \sum_{j=1}^2 p_{ij}(u^1) J(j), g(i, u^2) + \alpha \sum_{j=1}^2 p_{ij}(u^2) J(j) \right\}.$$

The scalars \underline{c}_k and \bar{c}_k of Eqs. (1.14) and (1.15) are given by

$$\underline{c}_k = \frac{\alpha}{1 - \alpha} \min \{ (T^k J)(1) - (T^{k-1} J)(1), (T^k J)(2) - (T^{k-1} J)(2) \},$$

$$\bar{c}_k = \frac{\alpha}{1 - \alpha} \max \{ (T^k J)(1) - (T^{k-1} J)(1), (T^k J)(2) - (T^{k-1} J)(2) \}.$$

The results of the value iteration method starting with the zero function J_0 [$J_0(1) = J_0(2) = 0$] are shown in Fig. 1.3.2 and illustrate the power of the error bounds.

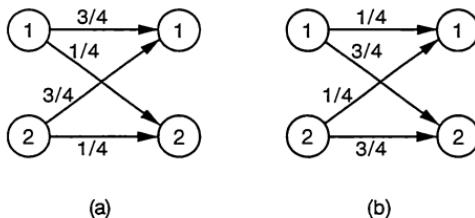


Figure 1.3.1 State transition diagram for Example 1.3.1: (a) $u = u^1$; (b) $u = u^2$.

k	$(T^k J_0)(1)$	$(T^k J_0)(2)$	$(T^k J_0)(1) + c_k$	$(T^k J_0)(1) + \bar{c}_k$	$(T^k J_0)(2) + c_k$	$(T^k J_0)(2) + \bar{c}_k$
0	0	0				
1	0.500	1.000	5.000	9.500	5.500	10.000
2	1.287	1.562	6.350	8.375	6.625	8.650
3	1.844	2.220	6.856	7.767	7.232	8.144
4	2.414	2.745	7.129	7.540	7.460	7.870
5	2.896	3.247	7.232	7.417	7.583	7.768
6	3.343	3.686	7.287	7.371	7.629	7.712
7	3.740	4.086	7.308	7.345	7.654	7.692
8	4.099	4.444	7.319	7.336	7.663	7.680
9	4.422	4.767	7.324	7.331	7.669	7.676
10	4.713	5.057	7.326	7.329	7.671	7.674
11	4.974	5.319	7.327	7.328	7.672	7.673
12	5.209	5.554	7.327	7.328	7.672	7.673
13	5.421	5.766	7.327	7.328	7.672	7.673
14	5.612	5.957	7.328	7.328	7.672	7.672
15	5.783	6.128	7.328	7.328	7.672	7.672

Figure 1.3.2 Performance of the value iteration method with and without the error bounds of Prop. 1.3.1 for the problem of Example 1.3.1.

Termination Issues – Optimality of the Obtained Policy

Let us now discuss how to use the error bounds to obtain an optimal or

near-optimal policy in a finite number of value iterations. We first note that given any J , if we compute TJ and a policy μ attaining the minimum in the calculation of TJ , i.e., $T_\mu J = TJ$, then we can obtain the following bound on the suboptimality of μ :

$$\max_i [J_\mu(i) - J^*(i)] \leq \frac{\alpha}{1-\alpha} \left(\max_i [(TJ)(i) - J(i)] - \min_i [(TJ)(i) - J(i)] \right). \quad (1.19)$$

To see this, apply Eq. (1.13) with $k = 1$ to obtain for all i

$$\underline{c}_1 \leq J^*(i) - (TJ)(i) \leq \bar{c}_1,$$

and also apply Eq. (1.13) with $k = 1$ and with T_μ replacing T to obtain

$$\underline{c}_1 \leq J_\mu(i) - (T_\mu J)(i) = J_\mu(i) - (TJ)(i) \leq \bar{c}_1.$$

Subtracting the above two equations, we obtain the estimate (1.19).

In practice, one terminates the value iteration method when the difference $(\bar{c}_k - \underline{c}_k)$ of the error bounds becomes sufficiently small. One can then take as final estimate of J^* the “median”

$$\tilde{J}_k = T^k J + \left(\frac{\bar{c}_k + \underline{c}_k}{2} \right) e \quad (1.20)$$

or the “average”

$$\hat{J}_k = T^k J + \frac{\alpha}{n(1-\alpha)} \sum_{i=1}^n ((T^k J)(i) - (T^{k-1} J)(i)) e. \quad (1.21)$$

Both of these vectors lie in the region delineated by the error bounds. Then, the estimate (1.19) provides a bound on the suboptimality of the policy μ attaining the minimum in the calculation of $T^k J$.

The bound (1.19) can also be used to show that after a sufficiently large number of value iterations, the stationary policy μ^k that attains the minimum in the k th value iteration [i.e. $(T_{\mu^k} T^{k-1}) J = T^k J$] is optimal. Indeed, since the number of stationary policies is finite, there exists an $\bar{\epsilon} > 0$ such that if a stationary policy μ is known to satisfy

$$\max_i [J_\mu(i) - J^*(i)] < \bar{\epsilon},$$

then μ must be optimal, i.e., $J_\mu = J^*$. Now let \bar{k} be such that for all $k \geq \bar{k}$ we have

$$\frac{\alpha}{1-\alpha} \left(\max_i [(T^k J)(i) - (T^{k-1} J)(i)] - \min_i [(T^k J)(i) - (T^{k-1} J)(i)] \right) < \bar{\epsilon}.$$

Then from Eq. (1.19) we see that for all $k \geq \bar{k}$, the stationary policy that attains the minimum in the k th value iteration is optimal.

Rate of Convergence

To analyze the rate of convergence of value iteration with error bounds, assume that there is a stationary policy μ^* that attains the minimum over μ in the relation

$$\min_{\mu} T_{\mu} T^{k-1} J = T^k J$$

for all k sufficiently large, so that eventually the method reduces to the linear iteration

$$J := g_{\mu^*} + \alpha P_{\mu^*} J.$$

In view of our preceding discussion, this is true for example if μ^* is a unique optimal stationary policy. Generally the rate of convergence of linear iterations is governed by the maximum eigenvalue modulus of the matrix of the iteration [which is α in our case, since any transition probability matrix has a unit eigenvalue with corresponding eigenvector $e = (1, 1, \dots, 1)'$, while all other eigenvalues lie within the unit circle of the complex plane].

It turns out, however, that when error bounds are used, the rate at which the iterates \hat{J}_k and \tilde{J}_k of Eqs. (1.20) and (1.21) approach the optimal cost vector J^* is governed by the modulus of the *subdominant* eigenvalue of the transition probability matrix P_{μ^*} , i.e., the eigenvalue with second largest modulus. The proof of this is outlined in Exercise 1.8. For a sketch of the ideas involved, let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of P_{μ^*} ordered according to decreasing modulus; that is

$$|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|,$$

with λ_1 equal to 1 and λ_2 being the subdominant eigenvalue. Assume that there is a set of linearly independent eigenvectors e_1, e_2, \dots, e_n corresponding to $\lambda_1, \lambda_2, \dots, \lambda_n$ with $e_1 = e = (1, 1, \dots, 1)'$. Then the initial error $J - J_{\mu^*}$ can be expressed as a linear combination of the eigenvectors

$$J - J_{\mu^*} = \xi_1 e + \sum_{j=2}^n \xi_j e_j$$

for some scalars $\xi_1, \xi_2, \dots, \xi_n$. Since $T_{\mu^*} J = g_{\mu^*} + \alpha P_{\mu^*} J$ and $J_{\mu^*} = g_{\mu^*} + \alpha P_{\mu^*} J_{\mu^*}$, successive errors are related by

$$T_{\mu^*} J - J_{\mu^*} = \alpha P_{\mu^*} (J - J_{\mu^*}), \quad \text{for all } J.$$

Thus the error after k iterations can be written as

$$T_{\mu^*}^k J - J_{\mu^*} = \alpha^k \xi_1 e + \alpha^k \sum_{j=2}^n \lambda_j^k \xi_j e_j.$$

Using the error bounds of Prop. 1.3.1 amounts to a translation of $T_{\mu^*}^k J$ along the vector e . Thus, at best, the error bounds are tight enough to eliminate the component $\alpha^k \xi_1 e$ of the error, but cannot affect the remaining term $\alpha^k \sum_{j=2}^n \lambda_j^k \xi_j e_j$, which diminishes like $\alpha^k |\lambda_2|^k$ with λ_2 being the subdominant eigenvalue.

Problems where Convergence is Slow

In Example 1.3.1, the convergence of value iteration with the error bounds is very fast. For this example, it can be verified that $\mu^*(1) = u^2$, $\mu^*(2) = u^1$, and that

$$P_{\mu^*} = \begin{bmatrix} 1/4 & 3/4 \\ 3/4 & 1/4 \end{bmatrix}.$$

The eigenvalues of P_{μ^*} can be calculated to be $\lambda_1 = 1$ and $\lambda_2 = -\frac{1}{2}$, which explains the fast convergence, since the modulus 1/2 of the subdominant eigenvalue λ_2 is considerably smaller than one. On the other hand, there are situations where convergence of the method even with the use of error bounds is very slow. For example, suppose that P_{μ^*} is block diagonal with two or more blocks, or more generally, that P_{μ^*} corresponds to a system with multiple recurrent classes of states (see Appendix D of Vol. I). Then it can be shown that the subdominant eigenvalue λ_2 is equal to 1, and convergence is typically slow when α is close to 1.

As an example, consider the following three simple deterministic problems, each having a single policy and multiple recurrent classes of states:

Problem 1: $n = 3, P_\mu$ = three-dimensional identity, $g(i, \mu(i)) = i$.

Problem 2: $n = 5, P_\mu$ = five-dimensional identity, $g(i, \mu(i)) = i$.

Problem 3: $n = 6, g(i, \mu(i)) = i$ and

$$P_\mu = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

Figure 1.3.3 shows the number of iterations needed by the value iteration method with and without the error bounds of Prop. 1.3.1, for the three preceding problems. In all cases, the starting function in all cases was taken to be zero, and J_μ was found within an error per coordinate of $10^{-6} \max_i |J_\mu(i)|$. The performance is rather unsatisfactory but, nonetheless, is typical of situations where the subdominant eigenvalue modulus of the optimal transition probability matrix is close to 1.

Elimination of Nonoptimal Actions in Value Iteration

We know from Prop. 1.2.3 that if $\tilde{u} \in U(i)$ is such that

$$g(i, \tilde{u}) + \alpha \sum_{j=1}^n p_{ij}(\tilde{u}) J^*(j) > J^*(i),$$

	Probl. 1 $\alpha = .9$	Probl. 1 $\alpha = .99$	Probl. 2 $\alpha = .9$	Probl. 2 $\alpha = .99$	Probl. 3 $\alpha = .9$	Probl. 3 $\alpha = .99$
W/out bounds	131	1374	131	1374	132	1392
With bounds	127	1333	129	1352	131	1374

Figure 1.3.3 Number of iterations for the value iteration method with and without error bounds. The problems are deterministic. Because the subdominant eigenvalue of the transition probability matrix is equal to 1, the error bounds are ineffective.

then \tilde{u} cannot be optimal at state i , i.e., for every optimal stationary policy μ , we have $\mu(i) \neq \tilde{u}$. Therefore, if we are sure that the above inequality holds, we can safely eliminate \tilde{u} from the admissible set $U(i)$. While we cannot check this inequality, since we do not know the optimal cost function J^* , we can guarantee that it holds if

$$g(i, \tilde{u}) + \alpha \sum_{j=1}^n p_{ij}(\tilde{u}) \underline{J}(j) > \overline{J}(i), \quad (1.22)$$

where \overline{J} and \underline{J} are upper and lower bounds satisfying

$$\underline{J}(i) \leq J^*(i) \leq \overline{J}(i), \quad i = 1, \dots, n.$$

The preceding observation is the basis for a useful application of the error bounds given earlier in Prop. 1.3.1. As these bounds are computed in the course of the value iteration method, the inequality (1.22) can be simultaneously checked and nonoptimal actions can be eliminated from the admissible set with attendant savings in subsequent computations. Since the upper and lower bound functions \overline{J} and \underline{J} converge to J^* , it can be seen [taking into account the finiteness of the constraint set $U(i)$] that eventually all nonoptimal $\tilde{u} \in U(i)$ will be eliminated, thereby reducing the set $U(i)$ after a finite number of iterations to the set of controls that are optimal at i . In this manner the computational requirements of value iteration may be substantially reduced, at the expense of the extra overhead needed to maintain the set of controls not as yet eliminated at each $i \in S$.

1.3.2 Variants of Value Iteration

We will now consider a few variants of value iteration. Some of these variants are aimed at accelerating convergence, while others involve approximations. Value iteration methods will be combined with policy iteration

methods in the next section. Additional value iteration methods, aimed at complex problems with a very large number of states, will be discussed in Section 6.3.

Gauss-Seidel Version of Value Iteration

In the value iteration method described earlier, the estimate of the cost function is iterated for all states simultaneously. An alternative is to iterate one state at a time, while incorporating into the computation the interim results. This corresponds to using what is known as the *Gauss-Seidel method* for solving the nonlinear system of equations $J = TJ$ (see, e.g., Bertsekas and Tsitsiklis [BeT89] or Ortega and Rheinboldt [OrR70]).

For n -dimensional vectors J , define the mapping F by

$$(FJ)(1) = \min_{u \in U(1)} \left[g(1, u) + \alpha \sum_{j=1}^n p_{1j}(u) J(j) \right] \quad (1.23)$$

and, for $i = 2, \dots, n$,

$$(FJ)(i) = \min_{u \in U(i)} \left[g(i, u) + \alpha \sum_{j=1}^{i-1} p_{ij}(u) (FJ)(j) + \alpha \sum_{j=i}^n p_{ij}(u) J(j) \right]. \quad (1.24)$$

In words, $(FJ)(i)$ is computed by the same equation as $(TJ)(i)$ except that the previously calculated values $(FJ)(1), \dots, (FJ)(i-1)$ are used in place of $J(1), \dots, J(i-1)$. Note that the computation of FJ is as easy as the computation of TJ (unless a parallel computer is used, in which case the computation of TJ may potentially be obtained much faster than FJ ; see Tsitsiklis [Tsi89], and Bertsekas and Tsitsiklis [BeT91a] for a comparative analysis).

Consider now the value iteration method whereby we compute J, FJ, F^2J, \dots . The following propositions show that the method is valid and provide an indication of better performance over the earlier value iteration method.

Proposition 1.3.2: Let J and J' be n -dimensional vectors. Then for any $k = 0, 1, \dots$,

$$\max_{i \in S} |(F^k J)(i) - (F^k J')(i)| \leq \alpha^k \max_{i \in S} |J(i) - J'(i)|. \quad (1.25)$$

Furthermore,

$$(FJ^*)(i) = J^*(i), \quad i \in S, \quad (1.26)$$

$$\lim_{k \rightarrow \infty} (F^k J)(i) = J^*(i), \quad i \in S. \quad (1.27)$$

Proof: It is sufficient to prove Eq. (1.25) for $k = 1$. We have by the definition of F and Prop. 1.2.4,

$$|(FJ)(1) - (FJ')(1)| \leq \alpha \max_{i \in S} |J(i) - J'(i)|.$$

Also, using this inequality,

$$\begin{aligned} |(FJ)(2) - (FJ')(2)| &\leq \alpha \max \{ |(FJ)(1) - (FJ')(1)|, |J(2) - J'(2)|, \dots, \\ &\quad |J(n) - J'(n)| \} \\ &\leq \alpha \max_{i \in S} |J(i) - J'(i)|. \end{aligned}$$

Proceeding similarly, we have, for every i and $j \leq i$,

$$|(FJ)(j) - (FJ')(j)| \leq \alpha \max_{i \in S} |J(i) - J'(i)|,$$

so Eq. (1.25) is proved for $k = 1$. The equation $FJ^* = J^*$ follows from the definition (1.23) and (1.24) of F , and Bellman's equation $J^* = TJ^*$. The convergence property (1.27) follows from Eqs. (1.25) and (1.26). **Q.E.D.**

Proposition 1.3.3: If an n -dimensional vector J satisfies

$$J(i) \leq (TJ)(i) \leq J^*(i), \quad i = 1, \dots, n,$$

then

$$(T^k J)(i) \leq (F^k J)(i) \leq J^*(i), \quad i = 1, \dots, n, \quad k = 1, 2, \dots \quad (1.28)$$

Proof: The proof follows by using the definition (1.23) and (1.24) of F , and the monotonicity property of T (Lemma 1.1.1). **Q.E.D.**

The preceding proposition indicates that the Gauss-Seidel version converges faster than the ordinary value iteration method, and provides the main motivation for using the mapping F in place of T in the value iteration method. The faster convergence property can be substantiated by further analysis (see e.g., Bertsekas and Tsitsiklis [BeT89]) and has been confirmed in practice through extensive experimentation. This comparison is not entirely fair, however, because the ordinary method will normally be used in conjunction with the error bounds of Prop. 1.3.1. One may also employ error bounds in the Gauss-Seidel version (see Exercise 1.9).

However, there is no clear superiority of one method over the other when bounds are introduced. One possibility is to use the Gauss-Seidel version in most iterations, but occasionally introduce an iteration of the ordinary method in order to obtain the corresponding error bounds. This mixed type of method usually performs better than the ordinary method alone. A final point in favor of the ordinary method is that it is better suited for parallel computation than the Gauss-Seidel version.

We note that there is a more flexible form of the Gauss-Seidel method, which selects states in arbitrary order to update their costs. This method maintains an approximation J to the optimal vector J^* , and at each iteration, it selects a state i and replaces $J(i)$ by $(TJ)(i)$. The remaining values $J(j)$, $j \neq i$, are left unchanged. The choice of the state i at each iteration is arbitrary, except for the restriction that all states are selected infinitely often. This method is an example of an *asynchronous fixed point iteration* and can be shown to converge to J^* starting from any initial J . Analyses of this type of method are given in Bertsekas [Ber82a], and in Chapter 6 of Bertsekas and Tsitsiklis [BeT89]; see also Section 1.3.2 and Exercise 1.20.

Generic Rank-One Corrections

We may view value iteration coupled with the error bounds of Prop. 1.3.1 as a method that makes a correction to the results of value iteration along the unit vector e . It is possible to generalize the idea of correction along a fixed vector so that it works for any type of convergent linear iteration.

Let us consider the case of a single stationary policy μ and an iteration of the form $J := FJ$, where

$$FJ = h_\mu + Q_\mu J.$$

Here, Q_μ is a matrix with eigenvalues strictly within the unit circle, and h_μ is a vector such that

$$J_\mu = FJ_\mu.$$

An example is the Gauss-Seidel iteration of Section 1.3.1, and some other examples are given in Exercises 1.4, 1.5, and 1.7. Also, the value iteration method for stochastic shortest path problems and a single stationary policy, to be discussed in Section 2.2, is of the above form.

Consider in place of $J := FJ$, an iteration of the form

$$J := F\tilde{J},$$

where \tilde{J} is related to J by

$$\tilde{J} = J + \tilde{\gamma}d,$$

with d a fixed vector and $\tilde{\gamma}$ a scalar to be selected in some optimal manner. In particular, consider choosing $\tilde{\gamma}$ by minimizing over γ

$$\|J + \gamma d - F(J + \gamma d)\|^2,$$

which, by denoting

$$z = Q_\mu d,$$

can be written as

$$\|J - FJ + \gamma(d - z)\|^2.$$

By setting to zero the derivative of this expression with respect to γ , it is straightforward to verify that the optimal solution is

$$\tilde{\gamma} = \frac{(d - z)'(FJ - J)}{\|d - z\|^2}.$$

Thus the iteration $J := F\tilde{J}$ can be written as

$$J := MJ,$$

where

$$MJ = FJ + \tilde{\gamma}z.$$

We note that this iteration requires only slightly more computation than the iteration $J := FJ$, since the vector z is computed once and the computation of $\tilde{\gamma}$ is simple.

A key question of course is under what circumstances the iteration $J := MJ$ converges faster than the iteration $J := FJ$, and whether indeed it converges at all to J_μ . It is straightforward to verify that in the case where $Q_\mu = \alpha P_\mu$ and $d = e$, the iteration $J := MJ$ can be written as

$$J := T_\mu J + \frac{\alpha}{n(1-\alpha)} \sum_{i=1}^n ((T_\mu J)(i) - J(i))e,$$

[compare with Eq. (1.21)]. Thus in this case the iteration $J := M(J)$ shifts the result $T_\mu J$ of value iteration to a vector that lies somewhere in the middle of the error bound range given by Prop. 1.3.1. By the result of this proposition it follows that the iteration converges to J_μ .

Generally, however, the iteration $J := MJ$ need not converge in the case where the direction vector d is chosen arbitrarily. If on the other hand d is chosen to be an eigenvector of Q_μ , convergence can be proved. This is shown in Exercise 1.8, where it is also proved that if d is an eigenvector corresponding to the dominant eigenvalue of Q_μ (the one with largest modulus), the convergence rate of the iteration $J := MJ$ is governed by the subdominant eigenvalue of Q_μ (the one with second largest modulus). One possibility for finding approximately such an eigenvector is to apply F a sufficiently large number of times to a-vector J . In particular, suppose that the initial error $J - J_\mu$ can be decomposed as

$$J - J_\mu = \sum_{j=1}^n \xi_j e_j$$

for some scalars ξ_1, \dots, ξ_n , where e_1, \dots, e_n are eigenvectors of Q_μ , and $\lambda_1, \dots, \lambda_n$ are corresponding eigenvalues. Suppose also that λ_1 is the unique dominant eigenvalue, i.e., $|\lambda_j| < |\lambda_1|$ for $j = 2, \dots, n$. Then the difference $F^{k+1}J - F^k J$ is nearly equal to $\xi_1(\lambda_1^{k+1} - \lambda_1^k)e_1$ for large k and can be used to estimate the dominant eigenvector e_1 . In order to decide whether k has been chosen large enough, one can test to see if the angle between the successive differences $F^{k+1}J - F^k J$ and $F^k J - F^{k-1}J$ is very small; if this is so, the components of $F^{k+1}J - F^k J$ along the eigenvectors e_2, \dots, e_n must also be very small. (For a more sophisticated version of this argument, see Bertsekas [Ber95a], where the generic rank-one correction method is developed in more general form.)

We can thus consider a two-phase approach: in the first phase, we apply several times the regular iteration $J := FJ$ both to improve our estimate of J and also to obtain an estimate d of an eigenvector corresponding to a dominant eigenvalue; in the second phase we use the modified iteration $J := MJ$ that involves extrapolation along d . It can be shown that the two-phase method converges to J_μ provided the error in the estimation of d is small enough, i.e., the cosine of the angle between d and $Q_\mu d$ as measured by the ratio

$$\frac{(F^k J - F^{k-1}J)'(F^{k-1}J - F^{k-2}J)}{\|F^k J - F^{k-1}J\| \cdot \|F^{k-1}(J) - F^{k-2}J\|}$$

is sufficiently close to one.

Note that the computation of the first phase is not wasted since it uses the iteration $J := FJ$ that we are trying to accelerate. Furthermore, since the second phase involves the calculation of FJ at the current iterate J , any error bounds or termination criteria based on FJ can be used to terminate the algorithm. As a result, the same finite termination mechanism can be used for both iterations $J := FJ$ and $J := MJ$.

One difficulty of the correction method outlined above is that the appropriate vector d depends on Q_μ and therefore also on μ . In the case of optimization over several policies, the mapping F is defined by

$$(FJ)(i) = \min_{u \in U(i)} \left[h_i(u) + \sum_{j=1}^n q_{ij}(u)J(j) \right], \quad i = 1, \dots, n. \quad (1.29)$$

One can then use the rank-one correction approach in two different ways:

- (1) Iteratively compute the cost vectors of the policies generated by a policy iteration scheme of the type discussed in the next subsection.
- (2) Guess at an optimal policy within the first phase, switch to the second phase, and then return to the first phase if the policy changes “substantially” during the second phase. In particular, in the first phase,

the iteration $J := FJ$ is used, where F is the nonlinear mapping of Eq. (1.29). Upon switching to the second phase, the vector z is taken to be equal to $Q_{\mu^*}d$, where μ^* is the policy that attains the minimum in Eq. (1.29) at the time of the switch. The second phase consists of the iteration

$$J := MJ = FJ + \tilde{\gamma}z,$$

where F is the nonlinear mapping of Eq. (1.29), and $\tilde{\gamma}$ is again given by

$$\tilde{\gamma} = \frac{(d - z)'(FJ - J)}{\|d - z\|^2}.$$

To guard against subsequent changes in policy, which induce corresponding changes in the matrix Q_{μ^*} , one should ensure that the method is working properly, for example, by recomputing d if the policy changes and/or the error $\|FJ - J\|$ is not reduced at a satisfactory rate. This method is generally effective because the value iteration method typically finds an optimal policy well before it finds the optimal cost vector.

Multiple-Rank Corrections

From the preceding analysis, we see that the rank-one correction method is ineffective in problems where there is little or no separation between the dominant and the subdominant eigenvalue moduli. This may happen both because the convergence rate of the method for obtaining d is slow, and also because the convergence rate of the modified iteration $J := MJ$ is not much faster than the one of the regular iteration $J := FJ$.

To address this difficulty, one may try corrections over subspaces of dimension larger than one. For example, we may consider an iteration of the form

$$J := F(J + Wy),$$

where F is a linear mapping as earlier, W is some matrix, and y is some vector. This amounts to applying F to a correction of J along the subspace spanned by the columns of W . Once the matrix W is selected, we may select y so that

$$\|J + Wy - F(J + Wy)\|^2$$

is minimized. By setting to zero the gradient with respect to y of the above expression, we can verify that the optimal vector is given by

$$\tilde{y} = (Z'Z)^{-1}\bar{Z}'(\bar{F}\bar{J} - J),$$

where $Z = (I - \alpha P_\mu)W$. The corresponding iteration then becomes

$$J := F(J + W\tilde{y}).$$

Some alternative possibilities for choosing the matrix W have been proposed in Bertsekas and Castanon [BeC89], and Bertsekas [Ber95a].

Much of our discussion regarding the rank-one correction method also applies to this generalized version. In particular, we can use a two-phase implementation, which allows a return from phase two to phase one whenever the progress of phase two is unsatisfactory. Furthermore, a version of the method that works in the case of multiple policies is possible.

Infinite State Space – Approximate Value Iteration

The value iteration method is valid under the assumptions of Prop. 1.2.1, so it is guaranteed to converge to J^* for problems with infinite state and control spaces. However, for such problems, the method may be implementable only through approximations. In particular, given a function J , one may only be able to calculate a function \tilde{J} such that

$$\max_{x \in S} |\tilde{J}(x) - (TJ)(x)| \leq \epsilon, \quad (1.30)$$

where ϵ is a given positive scalar. A similar situation may occur even when the state space is finite but the number of states is very large. Then instead of calculating $(TJ)(x)$ for all states x , one may do so only for some states and estimate $(TJ)(x)$ for the remaining states x by some form of interpolation, or by a least-squares error fit of $(TJ)(x)$ with a function from a suitable parametric class (compare with the discussion of Chapter 6). Then the function \tilde{J} thus obtained will satisfy a relation such as (1.30).

We are thus led to consider the approximate value iteration method that generates a sequence $\{J_k\}$ satisfying

$$\max_{x \in S} |J_{k+1}(x) - (TJ_k)(x)| \leq \epsilon, \quad k = 0, 1, \dots \quad (1.31)$$

starting from an arbitrary bounded function J_0 . Generally, such a sequence “converges” to J^* to within an error of $\epsilon/(1 - \alpha)$. To see this, note that Eq. (1.31) yields

$$TJ_0 - \epsilon e \leq J_1 \leq TJ_0 + \epsilon e.$$

By applying T to this relation, we obtain

$$T^2J_0 - \alpha\epsilon e \leq TJ_1 \leq T^2J_0 + \alpha\epsilon e,$$

so by using Eq. (1.31) to write

$$TJ_1 - \epsilon e \leq J_2 \leq TJ_1 + \epsilon e,$$

we have

$$T^2J_0 - \epsilon(1 + \alpha)e \leq J_2 \leq T^2J_0 + \epsilon(1 + \alpha)e.$$

Proceeding similarly, we obtain for all $k \geq 1$,

$$T^{k-1}J_0 - \epsilon(1 + \alpha + \cdots + \alpha^{k-1})e \leq J_k \leq T^{k-1}J_0 + \epsilon(1 + \alpha + \cdots + \alpha^{k-1})e.$$

By taking the limit superior and the limit inferior as $k \rightarrow \infty$, and by using the fact $\lim_{k \rightarrow \infty} T^k J_0 = J^*$, we see that

$$J^* - \frac{\epsilon}{1-\alpha}e \leq \liminf_{k \rightarrow \infty} J_k \leq \limsup_{k \rightarrow \infty} J_k \leq J^* + \frac{\epsilon}{1-\alpha}e.$$

It is also possible to obtain versions of the error bounds of Prop. 1.3.1 for the approximate value iteration method. We have from that proposition

$$\begin{aligned} TJ_k - \frac{\alpha}{1-\alpha} \min_{x \in S} [(TJ_k)(x) - J_k(x)]e &\leq J^* \\ &\leq TJ_k + \frac{\alpha}{1-\alpha} \max_{x \in S} [(TJ_k)(x) - J_k(x)]e. \end{aligned}$$

By using Eq. (1.31) in the above relation, we obtain

$$\begin{aligned} J_{k+1} - \epsilon e - \frac{\alpha}{1-\alpha} \min_{x \in S} [J_{k+1}(x) + \epsilon - J_k(x)]e &\leq J^* \\ &\leq J_{k+1} + \epsilon e + \frac{\alpha}{1-\alpha} \max_{x \in S} [J_{k+1}(x) + \epsilon - J_k(x)]e, \end{aligned}$$

or

$$\begin{aligned} J_{k+1} - \frac{\epsilon + \alpha \min_{x \in S} [J_{k+1}(x) - J_k(x)]}{1-\alpha}e &\leq J^* \\ &\leq J_{k+1} + \frac{\epsilon + \alpha \max_{x \in S} [J_{k+1}(x) - J_k(x)]}{1-\alpha}e. \end{aligned}$$

These bounds hold even when the state space is infinite because the bounds of Prop. 1.3.1 can be shown for an infinite state space as well. However, for these bounds to be useful, one should know ϵ .

1.3.3 Policy Iteration

The policy iteration algorithm generates a sequence of stationary policies, each with improved cost over the preceding one. Given the stationary policy μ , and the corresponding cost function J_μ , an improved policy $\{\bar{\mu}, \bar{\mu}, \dots\}$ is computed by minimization in the DP equation corresponding to J_μ , i.e., $T_{\bar{\mu}} J_\mu = TJ_\mu$, and the process is repeated.

The algorithm is based on the following proposition.

Proposition 1.3.4: Let μ and $\bar{\mu}$ be stationary policies such that $T_{\bar{\mu}}J_{\mu} = TJ_{\mu}$, or equivalently, for $i = 1, \dots, n$,

$$g(i, \bar{\mu}(i)) + \alpha \sum_{j=1}^n p_{ij}(\bar{\mu}(i)) J_{\mu}(j) = \min_{u \in U(i)} \left[g(i, u) + \alpha \sum_{j=1}^n p_{ij}(u) J_{\mu}(j) \right].$$

Then we have

$$J_{\bar{\mu}}(i) \leq J_{\mu}(i), \quad i = 1, \dots, n. \quad (1.32)$$

Furthermore, if μ is not optimal, strict inequality holds in the above equation for at least one state i .

Proof: Since $J_{\mu} = T_{\mu}J_{\mu}$ (Cor. 1.2.2.1) and, by hypothesis, $T_{\bar{\mu}}J_{\mu} = TJ_{\mu}$, we have for every i ,

$$\begin{aligned} J_{\mu}(i) &= g(i, \mu(i)) + \alpha \sum_{j=1}^n p_{ij}(\mu(i)) J_{\mu}(j) \\ &\geq g(i, \bar{\mu}(i)) + \alpha \sum_{j=1}^n p_{ij}(\bar{\mu}(i)) J_{\mu}(j) \\ &= (T_{\bar{\mu}}J_{\mu})(i). \end{aligned}$$

Applying repeatedly $T_{\bar{\mu}}$ on both sides of this inequality and using the monotonicity of $T_{\bar{\mu}}$ (Lemma 1.1.1) and Cor. 1.2.1.1, we obtain

$$J_{\mu} \geq T_{\bar{\mu}}J_{\mu} \geq \dots \geq T_{\bar{\mu}}^k J_{\mu} \geq \dots \geq \lim_{N \rightarrow \infty} T_{\bar{\mu}}^N J_{\mu} = J_{\bar{\mu}},$$

proving Eq. (1.32).

If $J_{\mu} = J_{\bar{\mu}}$, then from the preceding relation it follows that $J_{\mu} = T_{\bar{\mu}}J_{\mu}$ and since by hypothesis we have $T_{\bar{\mu}}J_{\mu} = TJ_{\mu}$, we obtain $J_{\mu} = TJ_{\mu}$, implying that $J_{\mu} = J^*$ by Prop. 1.2.2. Thus μ must be optimal. It follows that if μ is not optimal, then $J_{\bar{\mu}}(i) < J_{\mu}(i)$ for some state i . Q.E.D.

Policy Iteration Algorithm

Step 1: (Initialization) Guess an initial stationary policy μ^0 .

Step 2: (Policy Evaluation) Given the stationary policy μ^k , compute the corresponding cost function J_{μ^k} from the linear system of equations

$$(I - \alpha P_{\mu^k}) J_{\mu^k} = g_{\mu^k}.$$

Step 3: (Policy Improvement) Obtain a new stationary policy μ^{k+1} satisfying

$$T_{\mu^{k+1}} J_{\mu^k} = TJ_{\mu^k}.$$

If $J_{\mu^k} = TJ_{\mu^k}$ stop; else return to Step 2 and repeat the process.

Since the collection of all stationary policies is finite (by the finiteness of S and C) and an improved policy is generated at every iteration, it follows that the algorithm will find an optimal stationary policy in a finite number of iterations. This property is the main advantage of policy iteration over value iteration, which in general converges in an infinite number of iterations. On the other hand, finding the exact value of J_{μ^k} in Step 2 of the algorithm requires solving the system of linear equations $(I - \alpha P_{\mu^k})J_{\mu^k} = g_{\mu^k}$. The dimension of this system is equal to the number of states, and thus when this number is very large, the method is not attractive.

Figure 1.3.4 provides a geometric interpretation of policy iteration and compares it with value iteration.

We note that in some cases, one can exploit the special structure of the problem at hand to accelerate policy iteration. For example, sometimes we can show that if μ belongs to some restricted subset M of admissible control functions, then J_μ has a form guaranteeing that $\bar{\mu}$ will also belong to the subset M . In this case, policy iteration will be confined within the subset M , if the initial policy belongs to M . Furthermore, the policy evaluation step may be facilitated. For an example, see Exercise 1.14.

We now demonstrate policy iteration by means of the example considered earlier in this section.

Example 1.3.1 (continued)

Let us go through the calculations of the policy iteration method:

Initialization: We select the initial stationary policy

$$\mu^0(1) = u^1, \quad \mu^0(2) = u^2.$$

Policy Evaluation: We obtain J_{μ^0} through the equation $J_{\mu^0} = TJ_{\mu^0}$ or, equivalently, the linear system of equations

$$J_{\mu^0}(1) = g(1, u^1) + \alpha p_{11}(u^1)J_{\mu^0}(1) + \alpha p_{12}(u^1)J_{\mu^0}(2),$$

$$J_{\mu^0}(2) = g(2, u^2) + \alpha p_{21}(u^2)J_{\mu^0}(1) + \alpha p_{22}(u^2)J_{\mu^0}(2).$$

Substituting the data of the problem, we have

$$J_{\mu^0}(1) = 2 + 0.9 \cdot \frac{3}{4} \cdot J_{\mu^0}(1) + 0.9 \cdot \frac{1}{4} \cdot J_{\mu^0}(2),$$

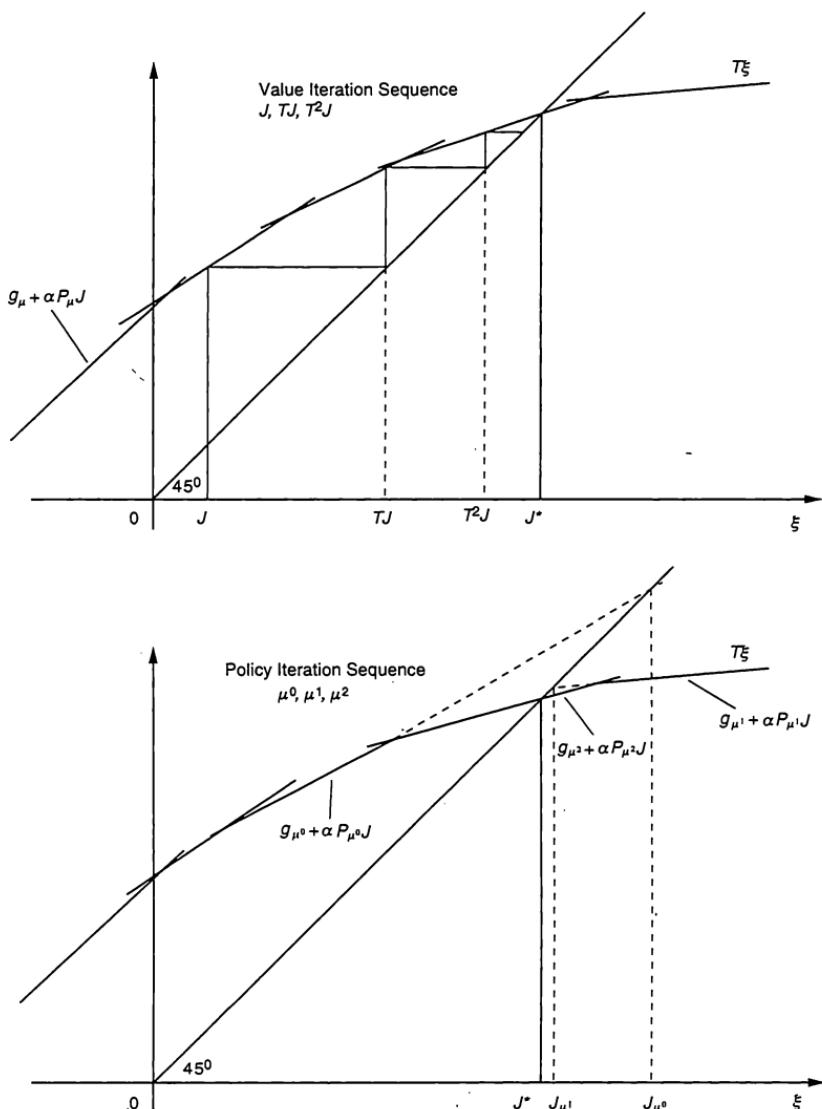


Figure 1.3.4 Geometric interpretation of policy iteration and value iteration. Each stationary policy μ defines the linear function $g_\mu + \alpha P_\mu J$ of the vector J , and TJ is the piecewise linear function $\min_\mu [g_\mu + \alpha P_\mu J]$. The optimal cost J^* satisfies $J^* = TJ^*$, so it is obtained from the intersection of the graph of TJ and the 45 degree line shown. The value iteration sequence is indicated in the top figure by the staircase construction, which asymptotically leads to J^* . The policy iteration sequence terminates when the correct linear segment of the graph of TJ (i.e., the optimal stationary policy) is identified, as shown in the bottom figure.

$$J_{\mu^0}(2) = 3 + 0.9 \cdot \frac{1}{4} \cdot J_{\mu^0}(1) + 0.9 \cdot \frac{3}{4} \cdot J_{\mu^0}(2).$$

Solving this system of linear equations for $J_{\mu^0}(1)$ and $J_{\mu^0}(2)$, we obtain

$$J_{\mu^0}(1) \simeq 24.12, \quad J_{\mu^0}(2) \simeq 25.96.$$

Policy Improvement: We now find $\mu^1(1)$ and $\mu^1(2)$ satisfying $T_{\mu^1} J_{\mu^0} = TJ_{\mu^0}$. We have

$$\begin{aligned} (TJ_{\mu^0})(1) &= \min \left\{ 2 + 0.9 \left(\frac{3}{4} \cdot 24.12 + \frac{1}{4} \cdot 25.96 \right), \right. \\ &\quad \left. 0.5 + 0.9 \left(\frac{1}{4} \cdot 24.12 + \frac{3}{4} \cdot 25.96 \right) \right\}, \\ &= \min \{24.12, 23.45\} = 23.45, \\ (TJ_{\mu^0})(2) &= \min \left\{ 1 + 0.9 \left(\frac{3}{4} \cdot 24.12 + \frac{1}{4} \cdot 25.96 \right), \right. \\ &\quad \left. 3 + 0.9 \left(\frac{1}{4} \cdot 24.12 + \frac{3}{4} \cdot 25.96 \right) \right\} \\ &= \min \{23.12, 25.95\} = 23.12. \end{aligned}$$

The minimizing controls are

$$\mu^1(1) = u^2, \quad \mu^1(2) = u^1.$$

Policy Evaluation: We obtain J_{μ^1} through the equation $J_{\mu^1} = T_{\mu^1} J_{\mu^0}$:

$$J_{\mu^1}(1) = g(1, u^2) + \alpha p_{11}(u^2) J_{\mu^1}(1) + \alpha p_{12}(u^2) J_{\mu^1}(2),$$

$$J_{\mu^1}(2) = g(2, u^1) + \alpha p_{21}(u^1) J_{\mu^1}(1) + \alpha p_{22}(u^1) J_{\mu^1}(2).$$

Substitution of the data of the problem and solution of the system of equations yields

$$J_{\mu^1}(1) \simeq 7.33, \quad J_{\mu^1}(2) \simeq 7.67.$$

Policy Improvement: We perform the minimization required to find TJ_{μ^1} :

$$\begin{aligned} (TJ_{\mu^1})(1) &= \min \left\{ 2 + \underline{0.9} \left(\frac{3}{4} \cdot 7.33 + \frac{1}{4} \cdot 7.67 \right), \right. \\ &\quad \left. 0.5 + 0.9 \left(\frac{1}{4} \cdot 7.33 + \frac{3}{4} \cdot 7.67 \right) \right\} \\ &= \min \{8.67, 7.33\} = 7.33, \end{aligned}$$

$$\begin{aligned}
 (TJ_{\mu^1})(2) &= \min \left\{ 1 + 0.9 \left(\frac{3}{4} \cdot 7.33 + \frac{1}{4} \cdot 7.67 \right), \right. \\
 &\quad \left. 3 + 0.9 \left(\frac{1}{4} \cdot 7.33 + \frac{3}{4} \cdot 7.67 \right) \right\} \\
 &= \min\{7.67, 9.83\} = 7.67.
 \end{aligned}$$

Hence we have $J_{\mu^1} = TJ_{\mu^1}$, which implies that μ^1 is optimal and $J_{\mu^1} = J^*$:

$$\mu^*(1) = u^2, \quad \mu^*(2) = u^1, \quad J^*(1) \simeq 7.33, \quad J^*(2) \simeq 7.67.$$

Modified Policy Iteration

When the number of states is large, solving the linear system

$$(I - \alpha P_{\mu^k})J_{\mu^k} = g_{\mu^k}$$

in the policy evaluation step by direct methods such as Gaussian elimination can be prohibitively time-consuming. One way to get around this difficulty is to solve this system iteratively by using value iteration. In fact, we may consider solving the system only approximately by executing a limited number of value iterations. This is called the *modified policy iteration algorithm*.

To formalize this method, let J_0 be an arbitrary n -dimensional vector. Let m_0, m_1, \dots be positive integers, and let the vectors J_1, J_2, \dots and the stationary policies μ_0, μ_1, \dots be defined by

$$T_{\mu^k} J_k = TJ_k, \quad J_{k+1} = T_{\mu^k}^{m_k} J_k, \quad k = 0, 1, \dots$$

Thus, a stationary policy μ^k is defined from J_k according to the policy improvement equation $T_{\mu^k} J_k = TJ_k$, and the cost J_{μ^k} is approximately evaluated by $m_k - 1$ additional value iterations, yielding the vector J_{k+1} , which is used in turn to define μ^{k+1} .

Note that if $m_k = 1$ for all k in the modified policy iteration algorithm, we obtain the value iteration method, while if $m_k = \infty$ we obtain the policy iteration method, where the policy evaluation step is performed iteratively by means of value iteration. Analysis and computational experience suggest that it is usually best to take m_k larger than 1 according to some heuristic scheme. A key idea here is that a value iteration involving a single policy (evaluating $T_\mu J$ for some μ and J) is much less expensive than an iteration involving all policies (evaluating TJ for some J), when the number of controls available at each state is large. Note that error bounds such as the ones of Prop. 1.3.1 can be used to improve the approximation process.

Furthermore, Gauss-Seidel iterations can be used in place of the usual value iterations.

The convergence properties of the modified policy iteration method will be discussed in the context of a more general algorithm, which we will introduce next. The convergence result for this algorithm includes the assumption $T_{\mu^0} J_0 \leq J_0$. However, for the modified policy iteration algorithm this assumption is not necessary (see Exercise 1.15).

Asynchronous Policy Iteration

Let us consider a more general policy iteration algorithm, whereby value iterations and policy updates are executed selectively, for only some of the states. This type of algorithm generates a sequence J_k of cost-to-go estimates and a corresponding sequence μ^k of stationary policies. Given (J_k, μ^k) , we select a subset S_k of the states, and we generate the new pair (J_{k+1}, μ^{k+1}) in one of two possible ways: either we update J_k according to

$$J_{k+1}(i) = \begin{cases} (T_{\mu^k} J_k)(i), & \text{if } i \in S_k, \\ J_k(i), & \text{otherwise,} \end{cases} \quad (1.33)$$

while we leave the policy unchanged by setting $\mu^k = \mu^{k+1}$, or else we update μ^k according to

$$\mu^{k+1}(i) = \begin{cases} \arg \min_{u \in U(i)} \left[g(i, u) + \alpha \sum_{j=1}^n p_{ij}(u) J_k(j) \right], & \text{if } i \in S_k, \\ \mu^k(i), & \text{otherwise,} \end{cases} \quad (1.34)$$

while we leave the cost-to-go estimate unchanged by setting $J_k = J_{k+1}$.

Note that by combining value and policy updates in various ways, we may synthesize more structured methods. In particular, if a policy update is followed by a value update for the same set of states S_k , the latter update is equivalent to the value iteration update

$$J_{k+1}(i) = \begin{cases} (T J_k)(i), & \text{if } i \in S_k, \\ J_k(i), & \text{otherwise.} \end{cases} \quad (1.35)$$

Furthermore, if $S_k = \{1, \dots, n\}$ and an “infinite” number of value updates (1.33) are done before doing a single policy update (1.34), we essentially obtain the policy iteration method. If $S_k = \{1, \dots, n\}$ and m_k value updates (1.33) are done followed by a policy update (1.34), we obtain the modified policy iteration method (or the value iteration method if $m_k = 1$). On the other hand, if S_k always consists of a single state and each policy update (1.34) is followed by a value update (1.33) [or equivalently by an update (1.35)], the algorithm reduces to the asynchronous value iteration method.

The asynchronous policy iteration algorithm also contains as a special case a version of the policy iteration method that uses a partition of the

state space into subsets, and performs “partial” policy iterations that are restricted to one of the subsets at a time. In particular, at the k th iteration, this method selects one of the subsets, call it S_k , obtains a new policy μ^{k+1} via the policy update (1.34), and “evaluates” this policy only for the states in S_k . The evaluation can be performed by using an infinite number of value updates for all states in S_k , with the cost-to-go values of the states $j \notin S_k$ held fixed at $J_k(j)$. Alternatively, the iteration can be performed by solving a restricted system of linear equations whose variables are the cost-to-go values of the states $i \in S_k$ only.

Let us now consider the convergence of the asynchronous policy iteration method. For this, we will need the assumption that the initial conditions J_0 and μ^0 satisfy $T_{\mu^0} J_0 \leq J_0$ (one possibility to obtain such initial conditions is to select μ^0 in some way and then take $J_0 = J_{\mu^0}$). Alternatively, we may select any J and use $J_0 = J + re$, where r is a sufficiently large scalar. In particular, since

$$T_{\mu^0}(J + re) = T_{\mu^0}J + \alpha re = (J + re) + (T_{\mu^0}J - J - (1 - \alpha)re),$$

it is sufficient to take r large enough so that the last term on the right above is a vector with nonpositive components. We have the following proposition.

Proposition 1.3.5: Assume that the value update (1.33) and the policy update (1.34) are executed infinitely often for all states, and that the initial conditions J_0 and μ^0 are such that $T_{\mu^0} J_0 \leq J_0$. Let (J_k, μ^k) be the sequence generated by the asynchronous policy iteration algorithm. Then J_k converges to J^* .

Proof: The idea of the proof is to first use the assumption $T_{\mu^0} J_0 \leq J_0$ to show that $J^* \leq J_{k+1} \leq J_k$ for all k , so that the sequence J_k converges to a limit \bar{J} . We will then use the assumption that the value and policy updates are executed infinitely often for all states to show that $\bar{J} = T\bar{J}$, so that $\bar{J} = J^*$.

We first show that for all k , we have

$$T_{\mu^k} J_k \leq J_k \quad \Rightarrow \quad T_{\mu^{k+1}} J_{k+1} \leq J_{k+1} \leq J_k. \quad (1.36)$$

Indeed assume that at iteration k we have $T_{\mu^k} J_k \leq J_k$, and consider two cases:

- (1) *The value update (1.33) is executed next.* Then we have

$$J_{k+1}(i) = (T_{\mu^k} J_k)(i) \leq J_k(i), \quad \text{if } i \in S_k, \quad (1.37)$$

and

$$J_{k+1}(i) = J_k(i), \quad \text{if } i \notin S_k, \quad (1.38)$$

so that $J_{k+1} \leq J_k$. Using this inequality, the monotonicity of T_{μ^k} , and the fact $\mu^{k+1} = \mu^k$, we obtain

$$T_{\mu^{k+1}} J_{k+1} = T_{\mu^k} J_{k+1} \leq T_{\mu^k} J_k. \quad (1.39)$$

On the other hand, from Eq. (1.37) we have

$$(T_{\mu^k} J_k)(i) = J_{k+1}(i), \quad \text{if } i \in S_k,$$

while from Eq. (1.38) and the hypothesis $T_{\mu^k} J_k \leq J_k$, we have

$$(T_{\mu^k} J_k)(i) \leq J_k(i) = J_{k+1}(i), \quad \text{if } i \notin S_k.$$

The above two relations imply that $T_{\mu^k} J_k \leq J_{k+1}$, which when combined with Eq. (1.39), shows that $T_{\mu^{k+1}} J_{k+1} \leq J_{k+1}$ and completes the proof of Eq. (1.36).

- (2) *The policy update (1.34) is executed next.* Then we have $J_{k+1} = J_k$, and by using the relation $T_{\mu^k} J_k \leq J_k$, we obtain for $i \in S_k$

$$\begin{aligned} (T_{\mu^{k+1}} J_{k+1})(i) &= (T_{\mu^{k+1}} J_k)(i) \\ &= (T J_k)(i) \\ &\leq (T_{\mu^k} J_k)(i) \\ &\leq J_k(i) \\ &= J_{k+1}(i), \end{aligned} \quad (1.40)$$

and for $i \notin S_k$

$$\begin{aligned} (T_{\mu^{k+1}} J_{k+1})(i) &= (T_{\mu^{k+1}} J_k)(i) \\ &= (T_{\mu^k} J_k)(i) \\ &\leq J_k(i) \\ &= J_{k+1}(i), \end{aligned} \quad (1.41)$$

so that $T_{\mu^{k+1}} J_{k+1} \leq J_{k+1}$, and Eq. (1.36) is shown.

Equation (1.36) and the hypothesis $T_{\mu^0} J_0 \leq J_0$ imply that

$$J_{k+1} \leq J_k, \quad T J_k \leq T_{\mu^k} J_k \leq J_k, \quad k = 0, 1, \dots \quad (1.42)$$

By using the monotonicity of T , we also have $T^m J_k \leq J_k$ for all m , and by taking the limit as $m \rightarrow \infty$, we obtain

$$J^* \leq J_k, \quad \overline{k} = 0, 1, \dots$$

From this equation and Eq. (1.42), we see that J_k converges to some limit \overline{J} satisfying

$$T \overline{J} \leq \overline{J} \leq J_k, \quad k = 0, 1, \dots \quad (1.43)$$

Furthermore, from Eqs. (1.39)-(1.41), we have

$$T_{\mu^{k+1}} J_{k+1} \leq T_{\mu^k} J_k, \quad k = 0, 1, \dots \quad (1.44)$$

Suppose, to arrive at a contradiction, that there exists a state i such that $(T\bar{J})(i) < \bar{J}(i)$, and let \bar{k} be such that $(TJ_k)(i) < \bar{J}(i)$ for all $k \geq \bar{k}$ (such an integer \bar{k} exists by the continuity of T). Let k be an iteration index that satisfies $k \geq \bar{k}$ and is such that the policy update (1.34) is executed for state i , and let k' be the first iteration index with $k' > k$ such that the value update (1.33) is executed for state i . Then we have

$$\begin{aligned} J_{k'+1}(i) &= (T_{\mu^{k'}} J_{k'})(i) \\ &\leq (T_{\mu^{k+1}} J_{k+1})(i) \\ &\leq (T_{\mu^{k+1}} J_k)(i) \\ &= (TJ_k)(i) \\ &< \bar{J}(i), \end{aligned} \quad (1.45)$$

where the first equality follows from the value update (1.33), the first inequality follows from Eq. (1.44), the second inequality follows from the relation $J_{k+1} \leq J_k$ and the monotonicity of $T_{\mu^{k+1}}$, and the second equality follows from the policy update (1.34). The relation (1.45) contradicts the relation (1.43). Thus we must have $(T\bar{J})(i) = \bar{J}(i)$ for all i , which implies that $\bar{J} = J^*$, since J^* is the unique solution of Bellman's equation. Q.E.D.

Infinite State Space – Approximate Policy Iteration

The policy iteration method can be defined for problems with infinite state and control spaces by means of the relation

$$T_{\mu^{k+1}} J_{\mu^k} = TJ_{\mu^k}, \quad k = 0, 1, \dots$$

The proof of Prop. 1.3.4 can then be used to show that the generated sequence of policies $\{\mu^k\}$ is improving in the sense that $J_{\mu^{k+1}} \leq J_{\mu^k}$ for all k . However, for infinite state space problems, the policy evaluation step and/or the policy improvement step of the method may be implementable only through approximations. A similar situation may occur even when the state space is finite but the number of states is very large.

We are thus led to consider an approximate policy iteration method that generates a sequence of stationary policies $\{\mu^k\}$ and a corresponding sequence of approximate cost functions $\{J_k\}$ satisfying

$$\max_{x \in S} |J_k(x) - J_{\mu^k}(x)| \leq \delta, \quad k = 0, 1, \dots \quad (1.46)$$

and

$$\max_{x \in S} |(T_{\mu^{k+1}} J_k)(x) - (T J_k)(x)| \leq \epsilon, \quad k = 0, 1, \dots \quad (1.47)$$

where δ and ϵ are some positive scalars, and μ^0 is an arbitrary stationary policy. We call this the *approximate policy iteration algorithm*. The following proposition provides error bounds for this algorithm.

Proposition 1.3.6: The sequence $\{\mu^k\}$ generated by the approximate policy iteration algorithm satisfies

$$\limsup_{k \rightarrow \infty} \max_{x \in S} (J_{\mu^k}(x) - J^*(x)) \leq \frac{\epsilon + 2\alpha\delta}{(1 - \alpha)^2}. \quad (1.48)$$

Proof: From Eqs. (1.46) and (1.47), we have for all k

$$T_{\mu^{k+1}} J_{\mu^k} - \alpha\delta e \leq T_{\mu^{k+1}} J_k \leq T J_k + \epsilon e,$$

where $e = (1, 1, \dots, 1)'$ is the unit vector, while from Eq. (1.46), we have for all k

$$T J_k \leq T J_{\mu^k} + \alpha\delta e.$$

By combining these two relations, we obtain for all k

$$T_{\mu^{k+1}} J_{\mu^k} \leq T J_{\mu^k} + (\epsilon + 2\alpha\delta)e \leq T_{\mu^k} J_{\mu^k} + (\epsilon + 2\alpha\delta)e. \quad (1.49)$$

From Eq. (1.49) and the equation $T_{\mu^k} J_{\mu^k} = J_{\mu^k}$, we have

$$T_{\mu^{k+1}} J_{\mu^k} \leq J_{\mu^k} + (\epsilon + 2\alpha\delta)e.$$

By subtracting from this relation the equation $T_{\mu^{k+1}} J_{\mu^{k+1}} = J_{\mu^{k+1}}$, we obtain

$$T_{\mu^{k+1}} J_{\mu^k} - T_{\mu^{k+1}} J_{\mu^{k+1}} \leq J_{\mu^k} - J_{\mu^{k+1}} + (\epsilon + 2\alpha\delta)e,$$

which can be written as

$$J_{\mu^{k+1}} - J_{\mu^k} \leq \alpha F_k + (\epsilon + 2\alpha\delta)e, \quad (1.50)$$

where F_k is the function given by

$$\begin{aligned} F_k(x) &= \alpha^{-1} (T_{\mu^{k+1}} J_{\mu^{k+1}})(x) - \alpha^{-1} (T_{\mu^{k+1}} J_{\mu^k})(x) \\ &= E_w \{ J_{\mu^{k+1}}(f(x, \mu^{k+1}(x), w)) - J_{\mu^k}(f(x, \mu^{k+1}(x), w)) \}. \end{aligned}$$

Let

$$\xi_k = \max_{x \in S} (J_{\mu^{k+1}}(x) - J_{\mu^k}(x)).$$

Then we have $F_k(x) \leq \xi_k$ for all $x \in S$, and Eq. (1.50) yields

$$\xi_k \leq \alpha \xi_k + \epsilon + 2\alpha\delta,$$

or

$$\xi_k \leq \frac{\epsilon + 2\alpha\delta}{1 - \alpha}. \quad (1.51)$$

Let

$$\zeta_k = \max_{x \in S} (J_{\mu^k}(x) - J^*(x)).$$

From Eq. (1.49) and the relation

$$\max_{x \in S} ((TJ_{\mu^k})(x) - J^*(x)) \leq \alpha \zeta_k,$$

which follows from Prop. 1.2.4, we have

$$\begin{aligned} T_{\mu^{k+1}} J_{\mu^k} &\leq TJ_{\mu^k} + (\epsilon + 2\alpha\delta)e \\ &= J^* + TJ_{\mu^k} - TJ^* + (\epsilon + 2\alpha\delta)e \\ &\leq J^* + \alpha \zeta_k + (\epsilon + 2\alpha\delta)e. \end{aligned}$$

We also have

$$\begin{aligned} T_{\mu^{k+1}} J_{\mu^k} &= T_{\mu^{k+1}} J_{\mu^{k+1}} + T_{\mu^{k+1}} J_{\mu^k} - T_{\mu^{k+1}} J_{\mu^{k+1}} \\ &= J_{\mu^{k+1}} - \alpha P_{\mu^{k+1}}(J_{\mu^{k+1}} - J_{\mu^k}), \end{aligned}$$

and by subtracting the last two relations and rearranging terms, we obtain

$$\begin{aligned} J_{\mu^{k+1}} - J^* &\leq \alpha \zeta_k e + \alpha P_{\mu^{k+1}}(J_{\mu^{k+1}} - J_{\mu^k}) + (\epsilon + 2\alpha\delta)e \\ &\leq \alpha \zeta_k e + \alpha F_k + (\epsilon + 2\alpha\delta)e. \end{aligned}$$

From this relation we see that

$$\zeta_{k+1} \leq \alpha \zeta_k + \alpha \xi_k + \epsilon + 2\alpha\delta.$$

By taking the limit superior as $k \rightarrow \infty$ and by using Eq. (1.51), we obtain

$$(1 - \alpha) \limsup_{k \rightarrow \infty} \zeta_k \leq \alpha \frac{\epsilon + 2\alpha\delta}{1 - \alpha} + \epsilon + 2\alpha\delta.$$

This relation simplifies to

$$\limsup_{k \rightarrow \infty} \zeta_k \leq \frac{\epsilon + 2\alpha\delta}{(1 - \alpha)^2},$$

which was to be proved. Q.E.D.

Proposition 1.3.6 suggests that the approximate policy iteration method makes steady progress up to a point and then the iterates J_{μ^k} oscillate within a neighborhood of the optimum J^* . This behavior appears to be typical in practice. Note that for $\delta = 0$ and $\epsilon = 0$, Prop. 1.3.6 shows that the cost sequence $\{J_{\mu^k}\}$ generated by the (exact) policy iteration algorithm converges to J^* , even when the state space is infinite. More generally, the proposition implies that if δ and ϵ are small, the approximate policy iteration method will also yield a policy that is near-optimal. Unfortunately, the error bound is inversely proportional to $(1 - \alpha)^2$, which is not very reassuring when α is close to 1. Yet an example given in Bertsekas and Tsitsiklis [BeT96], p. 283, shows that the error bound cannot be improved in the sense that for any $\alpha < 1$, there is a problem where the error bound is attained within an arbitrarily small tolerance.

Policy Iteration as an Actor-Critic System

There is an interesting interpretation of policy iteration methods that will prove useful when we will consider approximate versions of policy iteration later. In this interpretation, the policy evaluation step is viewed as the work of a *critic*, who evaluates the performance of the current policy, i.e., calculates an estimate of J_{μ^k} . The policy improvement step is viewed as the work of an *actor*, who takes into account the latest evaluation of the critic, i.e., the estimate of J_{μ^k} , and acts out the improved policy μ^{k+1} (see Fig. 1.3.5). Note that the exact, the modified, and the asynchronous policy iteration algorithms can all be viewed as actor-critic systems. In particular, modified policy iteration amounts to an incomplete evaluation of the current policy by the critic, using just a few value iteration steps. Asynchronous policy iteration can be similarly viewed as a process where the critic's policy evaluation and the critic's feedback to the actor are irregular and take place at different times for different states.

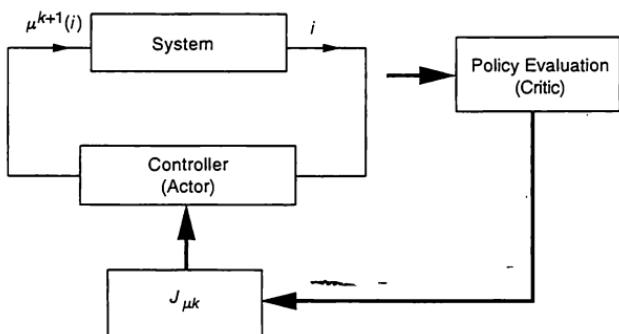


Figure 1.3.5 Interpretation of policy iteration as an actor-critic system.

1.3.4 Linear Programming

Since $\lim_{N \rightarrow \infty} T^N J = J^*$ for all J (cf. Prop. 1.2.1), we have

$$J \leq TJ \quad \Rightarrow \quad J \leq J^* = TJ^*.$$

Thus J^* is the “largest” J that satisfies the constraint $J \leq TJ$. This constraint can be written as a finite system of linear inequalities

$$J(i) \leq g(i, u) + \alpha \sum_{j=1}^n p_{ij}(u) J(j), \quad i = 1, \dots, n, \quad u \in U(i),$$

and delineates a polyhedron in \Re^n . The optimal cost vector J^* is the “northeast” corner of this polyhedron, as illustrated in Fig. 1.3.6. In particular, $J^*(1), \dots, J^*(n)$ solve the following problem (in $\lambda_1, \dots, \lambda_n$):

$$\text{maximize} \quad \sum_{i \in \tilde{S}} \lambda_i$$

$$\text{subject to} \quad \lambda_i \leq g(i, u) + \alpha \sum_{j=1}^n p_{ij}(u) \lambda_j, \quad i = 1, \dots, n, \quad u \in U(i),$$

where \tilde{S} is any nonempty subset of the state space $S = (1, \dots, n)$. This is a linear program with n variables and as many as $n \times q$ constraints, where q is the maximum number of elements of the sets $U(i)$. As n increases, its solution becomes more complex. For very large n and q , the linear programming approach can be practical only with the use of special large-scale linear programming methods.

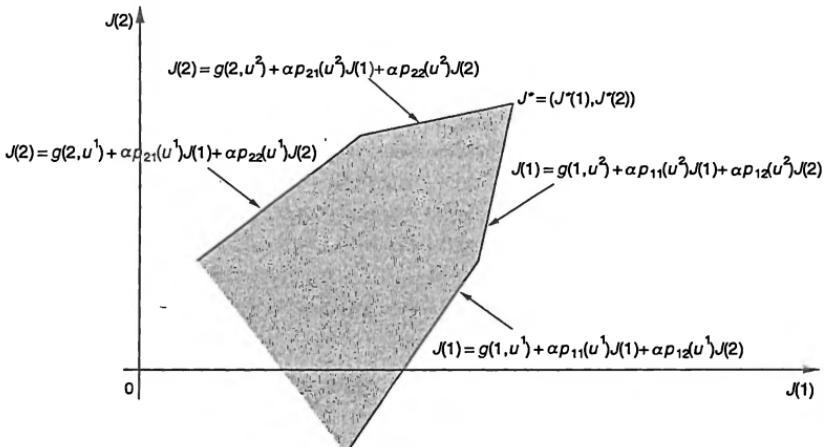


Figure 1.3.6 Linear programming problem associated with the discounted infinite horizon problem. The constraint set is shaded and the objective to maximize is $J(1) + J(2)$.

Example 1.3.1 (continued)

For the example considered earlier in this section, the linear programming problem takes the form

$$\begin{aligned} & \text{maximize } \lambda_1 + \lambda_2 \\ \text{subject to } & \lambda_1 \leq 2 + 0.9 \left(\frac{3}{4} \lambda_1 + \frac{1}{4} \lambda_2 \right), \quad \lambda_1 \leq 0.5 + 0.9 \left(\frac{1}{4} \lambda_1 + \frac{3}{4} \lambda_2 \right), \\ & \lambda_2 \leq 1 + 0.9 \left(\frac{3}{4} \lambda_1 + \frac{1}{4} \lambda_2 \right), \quad \lambda_2 \leq 3 + 0.9 \left(\frac{1}{4} \lambda_1 + \frac{3}{4} \lambda_2 \right). \end{aligned}$$

Cost Approximation Based on Linear Programming

When the number of states is very large or infinite, we may consider finding an approximation to the optimal cost function, which can be used in turn to obtain a (suboptimal) policy by minimization in Bellman's equation. One possibility is to approximate $J^*(x)$ with the *linear* form .

$$\tilde{J}(x, r) = \sum_{k=1}^m r_k w_k(x), \quad (1.52)$$

where $r = (r_1, \dots, r_m)$ is a vector of parameters, and for each state x , $w_k(x)$ are some fixed and known scalars. This amounts to approximating the cost function $J^*(x)$ by a linear combination of m given functions $w_k(x)$, where $k = 1, \dots, m$. These functions play the role of a *basis* for the space of cost function approximations $\tilde{J}(x, r)$ that can be generated with different choices of r (see also the discussion of approximations in Chapter 6).

It is then possible to determine r by using $\tilde{J}(x, r)$ in place of J^* in the preceding linear programming approach. In particular, we compute r as the solution of the program

$$\begin{aligned} & \text{maximize } \sum_{x \in \tilde{S}} \tilde{J}(x, r) \\ \text{subject to } & \tilde{J}(x, r) \leq g(x, u) + \alpha \sum_{y \in S} p_{xy}(u) \tilde{J}(y, r), \quad x \in \tilde{S}, u \in \tilde{U}(x), \end{aligned}$$

where \tilde{S} is either the state space S or a suitably chosen finite subset of S , and $\tilde{U}(x)$ is either $U(x)$ or a suitably chosen finite subset of $U(x)$. Because $\tilde{J}(x, r)$ is linear in the parameter vector r , the above program is linear in the parameters r_1, \dots, r_m . Thus if m is small, the number of variables of the linear program is small. The number of constraints is as large as $s \cdot q$, where s is the number of elements of \tilde{S} and q is the maximum number of elements of the sets $\tilde{U}(x)$. However, linear programs with a small number of

variables and a large number of constraints can often be solved relatively quickly with the use of special large-scale linear programming methods known as cutting plane or column generation methods (see e.g. Dantzig [Dan63], Bertsimas and Tsitsiklis [BeT97], Bertsekas [Ber99]). Thus, the preceding linear programming approach may be practical even for problems with a very large number of states.

1.3.5 Limited Lookahead and Rollout Policies

Similar to finite horizon problems (cf. Vol. I, Section 6.3), an effective suboptimal control approach is to truncate the time horizon and use at each stage a decision based on lookahead of a small number of stages. The simplest possibility is a *one-step lookahead policy* whereby at state i we use the control $\bar{\mu}(i)$ that attains the minimum in the expression

$$\min_{u \in U(i)} \left[g(i, u) + \alpha \sum_{j=1}^n p_{ij}(u) \tilde{J}(j) \right],$$

where \tilde{J} is some approximation of the true optimal cost function J^* . Similarly, a *two-step lookahead policy* applies at state i , the control $\bar{\mu}(i)$ attaining the minimum in the preceding equation, where now \tilde{J} is obtained itself on the basis of a one-step lookahead approximation. In other words, for all states j that can be reached from i , we have

$$\tilde{J}(j) = \min_{u \in U(j)} \left[g(j, u) + \alpha \sum_{k=1}^n p_{jk}(u) \hat{J}(k) \right]$$

where \hat{J} is some approximation of J^* . Policies with lookahead of more than two stages are similarly defined. Important examples of one-step lookahead policies are the rollout algorithms, discussed in Section 6.4 of Vol. I, where \tilde{J} is the cost-to-go of a stationary policy, or a lower bound thereof. In this case, the one-step lookahead policy may be viewed as the result of a limited form of policy iteration, i.e., a single-step policy improvement.

In a variant of the method that aims at reducing the computation to obtain $\bar{\mu}(i)$, the minimization is done over a subset $\bar{U}(i) \subset U(i)$. Thus, the control $\bar{\mu}(i)$ used in this variant is one that attains the minimum in the expression

$$\min_{u \in \bar{U}(i)} \left[g(i, u) + \alpha \sum_{j=1}^n p_{ij}(u) \tilde{J}(j) \right]. \quad (1.53)$$

This is attractive for example when by using some heuristic or approximate optimization, we can identify a subset $\bar{U}(i)$ of promising controls, and to save computation, we restrict attention to this subset in the one-step lookahead minimization.

The following proposition gives some bounds for the performance of the policy obtained by one-step lookahead, which are related to those given in Section 6.3 of Vol. I.

Proposition 1.3.7: Let $\bar{\mu}$ be the one-step lookahead policy obtained by minimization in Eq. (1.53).

(a) Assume that for some scalar δ , we have

$$\hat{J} \leq \tilde{J} + \delta e, \quad (1.54)$$

where

$$\hat{J}(i) = \min_{u \in \bar{U}(i)} \left[g(i, u) + \alpha \sum_{j=1}^n p_{ij}(u) \tilde{J}(j) \right], \quad i = 1, \dots, n.$$

Then

$$J_{\bar{\mu}} \leq \hat{J} + \frac{\alpha \delta}{1 - \alpha} e \leq \tilde{J} + \frac{\delta}{1 - \alpha} e.$$

(b) Assume that $\bar{U}(i) = U(i)$ for all i and that for some scalar ϵ , we have

$$J^* - \epsilon e \leq \tilde{J} \leq J^* + \epsilon e,$$

where J^* is the optimal cost function. Then

$$J_{\bar{\mu}} \leq J^* + \frac{2\alpha\epsilon}{1 - \alpha} e.$$

Proof: (a) Assume first that $\delta = 0$. Then we have

$$\tilde{J} \geq \hat{J} \geq T\tilde{J} = T_{\bar{\mu}}\tilde{J},$$

from which by using the monotonicity of T , we obtain

$$\tilde{J} \geq \hat{J} \geq T_{\bar{\mu}}^k \tilde{J} \geq T_{\bar{\mu}}^{k+1} \tilde{J}, \quad k = 1, 2, \dots$$

By taking the limit as $k \rightarrow \infty$, we have $\tilde{J} \geq \hat{J} \geq J_{\bar{\mu}}$.

In the general case where $\delta \neq 0$, we have $\tilde{J} + \delta e \geq T\tilde{J}$, which by adding $\sum_{m=1}^k \alpha^m \delta e$ to both sides, yields for all k ,

$$\tilde{J} + \sum_{m=0}^{k-1} \alpha^m \delta e + \alpha^k \delta e \geq T\tilde{J} + \sum_{m=1}^k \alpha^m \delta e = T \left(\tilde{J} + \sum_{m=0}^{k-1} \alpha^m \delta e \right).$$

By taking the limit as $k \rightarrow \infty$, we see that $J^+ \geq T J^+$, where

$$J^+ = \tilde{J} + \frac{\delta}{1-\alpha} e.$$

Since the one-step lookahead policy does not change if \tilde{J} is replaced by J^+ , by applying the special case shown already where $\delta = 0$, with \tilde{J} replaced by J^+ , we obtain

$$J^+ \geq \hat{J}^+ \geq J_{\bar{\mu}},$$

where \hat{J}^+ is given by

$$\begin{aligned}\hat{J}^+(i) &= \min_{u \in \bar{U}(i)} \left[g(i, u) + \alpha \sum_{j=1}^n p_{ij}(u) J^+(j) \right] \\ &= \min_{u \in \bar{U}(i)} \left[g(i, u) + \alpha \sum_{j=1}^n p_{ij}(u) \left(\tilde{J}(j) + \frac{\delta}{1-\alpha} \right) \right] \\ &= \hat{J}(i) + \frac{\alpha\delta}{1-\alpha} e.\end{aligned}$$

This completes the proof.

(b) By applying T to both sides of the inequality $\tilde{J} \leq J^* + \epsilon e$ and by using the inequality $J^* \leq \tilde{J} + \epsilon e$, we obtain

$$T\tilde{J} \leq TJ^* + \alpha\epsilon e = J^* + \alpha\epsilon e \leq \tilde{J} + (1+\alpha)\epsilon e.$$

Since $\bar{U}(i) = U(i)$ for all i , we have $\hat{J} = T\tilde{J}$, so we obtain $\hat{J} \leq J^* + \alpha\epsilon e$ as well as

$$\hat{J} \leq \tilde{J} + (1+\alpha)\epsilon e.$$

By applying part (a) with $\delta = (1+\alpha)\epsilon$ and by using the relation $\hat{J} \leq J^* + \alpha\epsilon e$ just shown, we have

$$J_{\bar{\mu}} \leq \hat{J} + \frac{\alpha\delta}{1-\alpha} e = \hat{J} + \frac{\alpha(1+\alpha)\epsilon}{1-\alpha} e \leq J^* + \alpha\epsilon e + \frac{\alpha(1+\alpha)\epsilon}{1-\alpha} e = J^* + \frac{2\alpha\epsilon}{1-\alpha} e.$$

Q.E.D.

Part (a) of the above proposition gives a bound on the performance of the one-step lookahead policy $\bar{\mu}$, which is readily computable when δ is known, since $\hat{J}(i)$ is computed on-line during calculation of the one-step lookahead control at state i . Part (b) says that if the one-step lookahead approximation \tilde{J} is within ϵ of the optimal, the performance of the one-step lookahead policy is within $2\alpha\epsilon/(1-\alpha)$ of the optimal. Unfortunately, this is not very reassuring when α is close to 1, in which case the error bound is very large relative to ϵ . Nonetheless, an example given by Bertsekas and

Tsitsiklis [BeT96], p. 263, shows that the error bound of part (b) is tight in the sense that for any $\alpha < 1$, there is a problem with just two states where the error bound holds with equality.

An example of application of the preceding proposition is a rollout scheme with multiple heuristics. In particular, let μ_1, \dots, μ_M be stationary policies such that

$$\mu_1(i), \dots, \mu_M(i) \in \overline{U}(i), \quad i = 1, \dots, n,$$

and let

$$\tilde{J}(i) = \min\{J_{\mu_1(i)}, \dots, J_{\mu_M(i)}\}, \quad i = 1, \dots, n.$$

Then, for all $i = 1, \dots, n$, and $m = 1, \dots, M$, we have

$$\begin{aligned} \hat{J}(i) &= \min_{u \in \overline{U}(i)} \left[g(i, u) + \alpha \sum_{j=1}^n p_{ij}(u) \tilde{J}(j) \right] \\ &\leq \min_{u \in \overline{U}(i)} \left[g(i, u) + \alpha \sum_{j=1}^n p_{ij}(u) \tilde{J}_{\mu_m}(j) \right] \\ &\leq J_{\mu_m}(i), \end{aligned}$$

from which, by taking minimum of the right-hand side over m , it follows that

$$\hat{J}(i) \leq \tilde{J}(i), \quad i = 1, \dots, n.$$

Using Prop. 1.3.7(a) with $\delta = 0$, we see that the rollout policy $\overline{\mu}$, obtained by using \overline{J} as one-step lookahead approximation satisfies

$$J_{\overline{\mu}}(i) \leq \min\{J_{\mu_1(i)}, \dots, J_{\mu_M(i)}\}, \quad i = 1, \dots, n,$$

i.e., it is an improved policy over each of the policies μ_1, \dots, μ_M .

1.4 THE ROLE OF CONTRACTION MAPPINGS

Two key structural properties in DP models are responsible for most of the mathematical results one can prove about them. The first is the *monotonicity property* of the mappings T and T_μ (cf. Lemma 1.1.1). This property is fundamental for total cost infinite horizon problems. For example, it forms the basis for the results on positive and negative DP models to be shown in Chapter 3.

When the cost per stage is bounded and there is discounting, however, we have another property that strengthens the effects of monotonicity: the mappings T and T_μ are *contraction mappings*. In this section, we explain the meaning and implications of this property.

Generally, given a real vector space Y with a norm $\|\cdot\|$ (i.e., a real-valued function satisfying $\|y\| \geq 0$ for all $y \in Y$, $\|y\| = 0$ if and only if $y = 0$, and $\|y + z\| \leq \|y\| + \|z\|$ for all $y, z \in Y$), a function $F : Y \mapsto Y$ is said to be a *contraction mapping* if for some $\rho \in (0, 1)$, we have

$$\|F(y) - F(z)\| \leq \rho \|y - z\|, \quad \text{for all } y, z \in Y.$$

The scalar ρ is called the *modulus of contraction* of F . The space Y is said to be *complete* under the norm $\|\cdot\|$ if every Cauchy sequence $\{y_k\} \subset Y$ is convergent.[†] When Y is complete, an important property of a contraction mapping $F : Y \mapsto Y$ is that it has a unique fixed point, i.e., the equation $y = F(y)$ has a unique solution y^* , called the *fixed point of F* . Furthermore, the sequence $\{y_k\}$ generated by the iteration

$$y_{k+1} = F(y_k)$$

converges to y^* , starting from an arbitrary initial point y_0 . We will shortly prove this property in a specialized setting; our method of proof, however, applies to the more general case as well.

1.4.1 Sup-Norm Contractions

We will focus on contraction mappings within a specialized context that is particularly important in DP. Let S be a set (typically the state space in a DP context), and let $v : S \mapsto \mathbb{R}$ be a positive-valued function,

$$v(s) > 0, \quad \text{for all } s \in S.$$

Let $B(S)$ denote the set of all functions $J : S \mapsto \mathbb{R}$ such that $J(s)/v(s)$ is bounded over $s \in S$. We define a norm on $B(S)$, called the *weighted sup-norm*, by

$$\|J\| = \max_{s \in S} \frac{|J(s)|}{v(s)}.$$

(The maximum in the above equation need not be attained. We are still following the convention that “max” denotes the least upper bound of the

[†] In this section we will use some introductory material from real analysis; we refer to Luenberger [Lue69], Liusternik and Sobolev [LiS61], Royden [Roy88], Rudin [Rud76], who give alternative treatments aimed at a variety of audiences. A sequence $\{y_k\} \subset Y$ is said to be a *Cauchy sequence* if $\|y_m - y_n\| \rightarrow 0$ as $m, n \rightarrow \infty$, i.e., given any $\epsilon > 0$, there exists N such that $\|y_m - y_n\| \leq \epsilon$ for all $m, n \geq N$. Note that a Cauchy sequence is always bounded. Also, a Cauchy sequence of real numbers is convergent, so the real line is a complete space and so is every finite-dimensional space. On the other hand, an infinite dimensional space may not be complete under some norms, while it may be complete under other norms.

relevant set, regardless of whether it is attained.) It is easily verified that $\|\cdot\|$ thus defined has the required properties for being a norm. Furthermore, $B(S)$ is complete under this norm.[†]

For a mapping $F : B(S) \mapsto B(S)$ and a function $J \in B(S)$, we denote by FJ the function in $B(S)$ obtained by applying F to J , and for $k > 1$, we denote by $F^k J$ the function obtained by applying F to J successively k times. The case where S is countable (or, as a special case, finite) is frequently encountered in DP. The following proposition provides some useful criteria for verifying the contraction property of mappings that are either linear or are obtained via a parametric minimization of other contraction mappings.

Proposition 1.4.1: Let $S = \{1, 2, \dots\}$.

(a) Let $F : B(S) \mapsto B(S)$ be a linear mapping of the form

[†] To see this, take a Cauchy sequence $\{J_k\} \subset B(S)$, and note that $\|J_m - J_n\| \rightarrow 0$ as $m, n \rightarrow \infty$ implies that for all $s \in S$, $\{J_k(s)\}$ is a Cauchy sequence of real numbers, so it converges to some $J^*(s)$. We will show that $J^* \in B(S)$ and that $\|J_k - J^*\| \rightarrow 0$. To this end, it will be sufficient to show that given any $\epsilon > 0$, there exists a K such that

$$|J_k(s) - J^*(s)|/v(s) \leq \epsilon$$

for all $s \in S$ and $k \geq K$; this will imply that

$$\max_{s \in S} |J^*(s)|/v(s) \leq \epsilon + \|J_k\|,$$

so that $J^* \in B(S)$, and will also imply that $\|J_k - J^*\| \leq \epsilon$, so that $\|J_k - J^*\| \rightarrow 0$. Assume the contrary, i.e., that there exists an $\epsilon > 0$ and a subsequence $\{s_{m_1}, s_{m_2}, \dots\} \subset S$ such that $m_i < m_{i+1}$ and

$$\epsilon < |J_{m_i}(s_{m_i}) - J^*(s_{m_i})|/v(s_{m_i}), \quad \text{for all } i \geq 1.$$

The right-hand side above is less or equal to

$$|J_{m_i}(s_{m_i}) - J_n(s_{m_i})|/v(s_{m_i}) + |J_n(s_{m_i}) - J^*(s_{m_i})|/v(s_{m_i}), \text{ for all } n \geq 1, i \geq 1.$$

The first term in the above sum is less than $\epsilon/2$ for i and n larger than some threshold; fixing i and letting n be sufficiently large, the second term can also be made less than $\epsilon/2$, so the sum is made less than ϵ - a contradiction.

$$(FJ)(i) = b(i) + \sum_{j \in S} a(i, j) J(j), \quad i = 1, 2, \dots,$$

where $b(i)$ and $a(i, j)$ are some scalars. Then F is a contraction with modulus ρ if

$$\frac{\sum_{j \in S} |a(i, j)| v(j)}{v(i)} \leq \rho, \quad i = 1, 2, \dots.$$

(b) Let $F : B(S) \mapsto B(S)$ be a mapping of the form

$$(FJ)(i) = \min_{\mu \in M} (F_\mu J)(i), \quad i = 1, 2, \dots,$$

where M is parameter set, and for each $\mu \in M$, F_μ is a contraction mapping from $B(S)$ to $B(S)$ with modulus ρ . Then F is a contraction mapping with modulus ρ .

Proof: (a) For any $J, J' \in B(S)$, we have

$$\begin{aligned} \|FJ - FJ'\| &= \max_{i \in S} \left| \frac{\sum_{j \in S} a(i, j) (J(j) - J'(j))}{v(i)} \right| \\ &\leq \max_{i \in S} \frac{\sum_{j \in S} |a(i, j)| v(j) (|J(j) - J'(j)| / v(j))}{v(i)} \\ &\leq \max_{i \in S} \frac{\sum_{j \in S} |a(i, j)| v(j)}{v(i)} \|J - J'\| \\ &\leq \rho \|J - J'\|, \end{aligned}$$

where the last inequality follows from the hypothesis.

(b) Since F_μ is a contraction of modulus ρ , we have for any $J, J' \in B(S)$, $(F_\mu J)(i)/v(i) \leq (F_\mu J')(i)/v(i) + \rho \|J - J'\|$, so by taking the minimum over $\mu \in M$, we obtain

$$\frac{(FJ)(i)}{v(i)} \leq \frac{(FJ')(i)}{v(i)} + \rho \|J - J'\|.$$

Using also the companion inequality with the roles of J and J' reversed, we obtain

$$\frac{|(FJ)(i) - (FJ')(i)|}{v(i)} \leq \rho \|J - J'\|, \quad i = 1, 2, \dots,$$

and by maximizing over i , the contraction property of F is proved. **Q.E.D.**

The preceding proposition assumes that $FJ \in B(S)$ for all $J \in B(S)$. The following proposition provides conditions, particularly relevant to the DP context, which imply this assumption.

Proposition 1.4.2: Let $S = \{1, 2, \dots\}$, let M be a parameter set, and for each $\mu \in M$, let F_μ be a linear mapping of the form

$$(F_\mu J)(i) = b(i, \mu) + \sum_{j \in S} a(i, j, \mu) J(j), \quad i = 1, 2, \dots$$

- (a) We have $F_\mu J \in B(S)$ for all $J \in B(S)$ provided $b_\mu \in B(S)$ and $V_\mu \in B(S)$, where

$$b_\mu = \{b(1, \mu), b(2, \mu), \dots\}, \quad V_\mu = \{V(1, \mu), V(2, \mu), \dots\},$$

with

$$V(i, \mu) = \sum_{j \in S} |a(i, j, \mu)| v(j), \quad i = 1, 2, \dots$$

- (b) Consider the mapping F

$$(FJ)(i) = \min_{\mu \in M} (F_\mu J)(i), \quad i = 1, 2, \dots$$

We have $FJ \in B(S)$ for all $J \in B(S)$, provided $b \in B(S)$ and $V \in B(S)$, where

$$b = \{b(1), b(2), \dots\}, \quad V = \{V(1), V(2), \dots\},$$

with $b(i) = \max_{\mu \in M} b(i, \mu)$ and $V(i) = \max_{\mu \in M} V(i, \mu)$.

Proof: (a) For all $\mu \in M$, $J \in B(S)$ and $i \in S$, we have

$$\begin{aligned} (F_\mu J)(i) &\leq |b(i, \mu)| + \sum_{j \in S} |a(i, j, \mu)| |J(j)/v(j)| v(j) \\ &\leq |b(i, \mu)| + \|J\| \sum_{j \in S} |a(i, j, \mu)| v(j) \\ &= |b(i, \mu)| + \|J\| V(i, \mu), \end{aligned}$$

and similarly $(F_\mu J)(i) \geq -|b(i, \mu)| - \|J\| V(i, \mu)$. Thus

$$|(F_\mu J)(i)| \leq |b(i, \mu)| + \|J\| V(i, \mu), \quad i = 1, 2, \dots$$

By dividing this inequality with $v(i)$ and by taking the maximum over $i \in S$, we obtain

$$\|F_\mu J\| \leq \|b_\mu\| + \|J\| \|V_\mu\| < \infty.$$

(b) By doing the same as in (a), but after first taking the minimum of $(F_\mu J)(i)$ over μ , we obtain

$$\|FJ\| \leq \|b\| + \|J\| \|V\| < \infty.$$

Q.E.D.

We now prove the central result regarding contraction mappings, specialized to the case of $B(S)$.

Proposition 1.4.3: (Contraction Mapping Fixed-Point Theorem) If $F : B(S) \mapsto B(S)$ is a contraction mapping with modulus $\rho \in (0, 1)$, then there exists a unique $J^* \in B(S)$ such that

$$J^* = FJ^*.$$

Furthermore, if J is any function in $B(S)$, then $\{F^k J\}$ converges to J^* and we have

$$\|F^k J - J^*\| \leq \rho^k \|J - J^*\|, \quad k = 1, 2, \dots$$

Proof: Fix some $J \in B(S)$ and consider the sequence $\{J_k\}$ generated by $J_{k+1} = FJ_k$ starting with $J_0 = J$. By the contraction property of F ,

$$\|J_{k+1} - J_k\| \leq \rho \|J_k - J_{k-1}\|, \quad k = 1, 2, \dots,$$

which implies that

$$\|J_{k+1} - J_k\| \leq \rho^k \|J_1 - J_0\|, \quad k = 1, 2, \dots$$

It follows that for every $k \geq 0$ and $m \geq 1$, we have

$$\begin{aligned} \|J_{k+m} - J_k\| &\leq \sum_{i=1}^m \|J_{k+i} - J_{k+i-1}\| \\ &\leq \rho^k (1 + \rho + \dots + \rho^{m-1}) \|J_1 - J_0\| \\ &\leq \frac{\rho^k}{1 - \rho} \|J_1 - J_0\|. \end{aligned}$$

Therefore, $\{J_k\}$ is a Cauchy sequence and must converge to a limit $J^* \in B(S)$, since $B(S)$ is complete. We have for all $k \geq 1$,

$$\|FJ^* - J^*\| \leq \|FJ^* - J_k\| + \|J_k - J^*\| \leq \rho \|J^* - J_{k-1}\| + \|J_k - J^*\|$$

and since J_k converges to J^* , we obtain $FJ^* = J^*$. Thus, the limit J^* of J_k is a fixed point of F . It is a unique fixed point because if \tilde{J} were another fixed point, we would have

$$\|J^* - \tilde{J}\| = \|FJ^* - F\tilde{J}\| \leq \rho \|J^* - \tilde{J}\|,$$

which implies that $J^* = \tilde{J}$.

To show the last part, note that

$$\|F^k J - J^*\| = \|F^k J - FJ^*\| \leq \rho \|F^{k-1} J - J^*\|.$$

Repeating this process for a total of k times, we obtain the desired result.
Q.E.D.

Consider now the mappings T and T_μ associated with the discounted cost problem with bounded cost per stage [cf. Eqs. (1.4) and (1.5)]. Proposition 1.2.4 and Cor. 1.2.4.1 show that T and T_μ are contraction mappings ($\rho = \alpha$) with respect to the “unweighted” sup-norm, where $v(s) \equiv 1$. Their unique fixed points are J^* (the optimal cost function) and J_μ , respectively. Furthermore, the convergence of the value iteration method to J^* follows from the contraction mapping theorem. Note also that, by Prop. 1.3.2, the mapping F corresponding to the Gauss-Seidel variant of the value iteration method is also a contraction mapping with $\rho = \alpha$, and the convergence result of Prop. 1.3.2 is again a special case of the contraction mapping theorem. In Chapter 2, we will see some examples of DP problems where the underlying DP mapping T is not a contraction with respect to the (unweighted) sup-norm, but is instead a contraction with respect to a suitable weighted sup-norm.

1.4.2 m -Stage Sup-Norm Contractions

In some DP contexts, the mappings T and T_μ are not contraction mappings, but become contractions when iterated a finite number of times. In this case, one may use a slightly different version of the contraction mapping fixed point theorem, which we now present.

Let us say that a function $F : B(S) \mapsto B(S)$ is an *m -stage contraction mapping* if there exists a positive integer m and some $\rho < 1$ such that

$$\|F^m J - F^m J'\| \leq \rho \|J - J'\|, \quad \text{for all } J, J' \in B(S),$$

where F^m denotes the composition of F with itself m times. Thus, F is an m -stage contraction if F^m is a contraction. Again, the scalar ρ is called

the modulus of contraction. We have the following generalization of Prop. 1.4.3.

Proposition 1.4.4: (*m*-Stage Contraction Mapping Fixed-Point Theorem) If $F : B(S) \mapsto B(S)$ is an m -stage contraction mapping with modulus $\rho \in (0, 1)$, then there exists a unique $J^* \in B(S)$ such that

$$J^* = FJ^*.$$

Furthermore, if J is any function in $B(S)$, then $\{F^k J\}$ converges to J^* .

Proof: Since F^m maps $B(S)$ into $B(S)$ and is a contraction mapping, by Prop. 1.4.3, it has a unique fixed point in $B(S)$, denoted J^* . Applying F to both sides of the relation $J^* = F^m J^*$, we see that FJ^* is also a fixed point of F^m , so by the uniqueness of the fixed point, we have $J^* = FJ^*$. Therefore J^* is a fixed point of F . If F had another fixed point, say \tilde{J} , then we would have $\tilde{J} = F^m \tilde{J}$, which by the uniqueness of the fixed point of F^m implies that $\tilde{J} = J^*$. Thus, J^* is the unique fixed point of F .

To show the convergence of $\{F^k J\}$, note that by Prop. 1.4.3, we have for all $J \in B(S)$,

$$\lim_{k \rightarrow \infty} \|F^{mk} J - J^*\| \rightarrow 0.$$

Using $F^i J$ in place of J , we obtain

$$\lim_{k \rightarrow \infty} \|F^{mk+i} J - J^*\| \rightarrow 0, \quad i = 0, 1, \dots, m-1,$$

which proves the desired result. Q.E.D.

An interesting discounted problem that cannot be analyzed with the theory of Sections 1.1-1.3, but can be addressed with m -stage contraction mapping theory, is given in the next section.

1.4.3 Discounted Problems – Unbounded Cost per Stage

We have considered so far in this chapter discounted problems with a possibly infinite state space, but also a bounded cost per stage. This latter assumption has been essential for the DP mapping T to be a contraction with respect to the (unweighted) sup-norm. On the other hand the boundedness assumption on the cost per stage is often restrictive. For example, in problems involving queues or inventory facilities with infinite storage capacity, it is natural for the cost per stage to increase to infinity with the system occupancy. It turns out that for many discounted problems involving a countable state space there is a method of analysis that relies on weighted sup-norm contractions, as we now proceed to discuss.

Let us consider a problem where the state space is $S = \{1, 2, \dots\}$, the discount factor is $\alpha \in (0, 1)$, the transition probabilities are denoted $p_{ij}(u)$ for $i, j \in S$ and $u \in U(i)$, and the expected cost per stage is denoted by $g(i, u)$, $i \in S$, $u \in U(i)$. The constraint set $U(i)$ may be infinite/arbitrary. We introduce a positive sequence $v = \{v_0, v_1, \dots\}$ and the weighted supremum norm

$$\|J\| = \max_{i \in S} \frac{|J(i)|}{v_i} \quad (1.55)$$

in the space $B(S)$ of sequences $\{J(1), J(2), \dots\}$ such that $\|J\| < \infty$. The following assumption will allow the use of Props. 1.4.1 and 1.4.2 for the purpose of showing that the DP mappings T and T_μ are m -stage contraction mappings. We assume the following.

Assumption 1.4.1:

- (a) The sequence $G = \{G(1), G(2), \dots\}$, where

$$G(i) = \max_{u \in U(i)} |g(i, u)|, \quad i = 1, 2, \dots,$$

belongs to $B(S)$.

- (b) The sequence $V = \{V(1), V(2), \dots\}$, where

$$V(i) = \max_{u \in U(i)} \sum_{j \in S} p_{ij}(u) v_j, \quad i = 1, 2, \dots,$$

belongs to $B(S)$.

- (c) There exists an integer $m \geq 1$ and a scalar $\rho \in (0, 1)$ such that for every policy π , we have

$$\alpha^m \frac{\sum_{j \in S} P(x_m = j \mid x_0 = i, \pi) v_j}{v_i} \leq \rho, \quad i = 1, 2, \dots$$

Assumption 1.4.1(a) is satisfied if the absolute expected cost per stage as a function of the state i grows proportionally to v_i . In particular, it is satisfied when

$$v_i = \max \left\{ 1, \max_{u \in U(i)} \overline{|g(i, u)|} \right\}, \quad i = 1, 2, \dots$$

Assumption 1.4.1(b) is a boundedness assumption on the ratio $V(i)/v_i$, i.e., the maximum over u of the expected value of the ratio v_j/v_i . Assumption

1.4.1(c) is satisfied if the expression

$$\frac{\sum_{j \in S} P(x_m = j \mid x_0 = i, \pi) v_j}{v_i}$$

is uniformly bounded over all i , m , and π by some scalar $B > 0$, since then it is sufficient to take m large enough so that $\alpha^m B \leq \rho$. This expression is the expected value of v_j/v_i , m stages after reaching state i while using policy π .

Example 1.4.1

Let

$$v_i = i, \quad i = 1, 2, \dots$$

Then Assumption 2.5.1(a) is satisfied if the maximum expected absolute cost per stage at state i grows no faster than linearly with i . Assumption 2.5.1(b) states that the maximum expected next state following state i ,

$$\max_{u \in U(i)} E\{j \mid i, u\},$$

also grows no faster than linearly with i . Finally, Assumption 2.5.1(c) is satisfied if

$$\alpha^m \sum_{j \in S} P(x_m = j \mid x_0 = i, \pi) j \leq \rho i, \quad i = 1, 2, \dots$$

It requires that for all π , the expected value of the state obtained m stages after reaching state i is no more than $\alpha^{-m} \rho i$. In particular, if there is bounded upward expected change of the state starting at i , there exists m sufficiently large so that Assumption 2.5.1(c) is satisfied. Similar interpretations are possible for other choices of v_i , such as

$$v_i = i^t, \quad i = 1, 2, \dots,$$

for some positive integer t .

We now consider the DP mappings T_μ and T ,

$$(T_\mu J)(i) = g(i, \mu(i)) + \alpha \sum_{j \in S} p_{ij}(\mu(i)) J(j), \quad i = 0, 1, \dots,$$

$$(TJ)(i) = \min_{u \in U(i)} \left[g(i, u) + \alpha \sum_{j \in S} p_{ij}(u) J(j) \right], \quad i = 0, 1, \dots,$$

and show their contraction property.

Proposition 1.4.5: Under Assumption 1.4.1, the mappings T and T_μ map $B(S)$ into $B(S)$, and are m -stage contraction mappings with modulus ρ .

Proof: Assumptions 1.4.1(a) and 1.4.1(b), together with Prop. 1.4.2, show that if $J \in B(S)$, then $TJ \in B(S)$ and $T_\mu J \in B(S)$ for all μ . For any $J \in B(S)$, and any policy $\pi = \{\mu_0, \mu_1, \dots\}$, we have for all $i \in S$,

$$(T_{\mu_0} \cdots T_{\mu_{m-1}} J)(i) = b(i) + \alpha^m \sum_{j \in S} P(x_m = j \mid x_0 = i, \pi) J(j),$$

where $b(i)$ is the expected cost of the first m stages starting from state i and using policy π (with 0 terminal cost). Using Prop. 1.4.1(a) in conjunction with Assumption 1.4.1(c), it follows that $T_{\mu_0} \cdots T_{\mu_{m-1}}$ is a contraction of modulus ρ , and then using Prop. 1.4.1(b), it follows that the same is true for T^m . Q.E.D.

Note that, as an intermediate step, the preceding proof also shows that $T_{\mu_0} \cdots T_{\mu_{m-1}}$ is a contraction mapping with modulus ρ for all policies π . The m -stage contraction property of T and Prop. 1.4.4 can now be used to essentially replicate the analysis of Section 1.2, and to show the standard results:

- (a) The value iteration method $J_{k+1} = TJ_k$ converges to the unique solution J^* of Bellman's equation $J = TJ$.
- (b) The unique solution J^* of Bellman's equation is the optimal cost function of the problem.
- (c) A stationary policy μ is optimal if and only if $T_\mu J^* = TJ^*$.

The preceding analysis easily generalizes to the undiscounted case where $\alpha = 1$ (under some additional conditions), and the corresponding contraction property of T will be revisited in Section 2.5, in the context of stochastic shortest path problems with a countable number of states.

1.5 SCHEDULING AND MULTIARMED BANDIT PROBLEMS

In the problem of this section there are n projects (or activities) of which only one can be worked on at any time period. Each project i is characterized at time k by its state x_k^i . If project i is worked on at time k , one receives an expected reward $\alpha^k R^i(x_k^i)$, where $\alpha \in (0, 1)$ is a discount factor; the state x_k^i then evolves according to the equation

$$x_{k+1}^i = f^i(x_k^i, w_k^i), \quad \text{if } i \text{ is worked on at time } k,$$

where w_k^i is a random disturbance with probability distribution depending on x_k^i but not on prior disturbances. The states of all idle projects are unaffected; i.e.,

$$x_{k+1}^i = x_k^i, \quad \text{if } i \text{ is idle at time } k.$$

We assume perfect state information and that the reward functions $R^i(\cdot)$ are uniformly bounded above and below, so the problem comes under the discounted cost framework of Section 1.2.

We assume also that at any time k there is the option of permanently retiring from all projects, in which case a reward $\alpha^k M$ is received and no additional rewards are obtained in the future. The retirement reward M is given and provides a parameterization of the problem, which will prove very useful. Note that for M sufficiently small it is never optimal to retire, thereby allowing the possibility of modeling problems where retirement is not a real option.

The key characteristic of the problem is the independence of the projects manifested in our three basic assumptions:

1. States of idle projects remain fixed.
2. Rewards received depend only on the state of the project currently engaged.
3. Only one project can be worked on at a time.

The rich structure implied by these assumptions makes possible a powerful methodology. It turns out that optimal policies have the form of an *index rule*; that is, for each project i , there is a function $m^i(x^i)$ such that an optimal policy at time k is to

$$\begin{aligned} \text{retire} &\quad \text{if} \quad M > \max_j \{m^j(x_k^j)\}, \\ \text{work on project } i &\quad \text{if} \quad m^i(x_k^i) = \max_j \{m^j(x_k^j)\} \geq M. \end{aligned} \tag{1.56}$$

Thus $m^i(x_k^i)$ may be viewed as an index of profitability of operating the i th project, while M represents profitability of retirement at time k . The optimal policy is to exercise the option of maximum profitability.

The problem of this section has a colorful name. It is known as a *multiarmed bandit problem* after an early and somewhat specialized paradigm, whereby one is to select a sequence of plays on a slot machine that has several arms corresponding to different but unknown probability distributions of payoff. With each play the distribution of the selected arm is better identified, so at each play, the tradeoff is between playing arms with high expected payoff and exploring the winning potential of other arms.

Index of a Project

Let $J(x, M)$ denote the optimal reward attainable when the initial state is $x = (x^1, \dots, x^n)$ and the retirement reward is M . From Section 1.2 we

know that, for each M , $J(\cdot, M)$ is the unique bounded solution of Bellman's equation

$$J(x, M) = \max \left[M, \max_i L^i(x, M, J) \right], \quad \text{for all } x, \quad (1.57)$$

where L^i is defined by

$$L^i(x, M, J) = R^i(x^i) + \alpha E_{w^i} \{ J(x^1, \dots, x^{i-1}, f^i(x^i, w^i), x^{i+1}, \dots, x^n, M) \}. \quad (1.58)$$

The next proposition gives some useful properties of J .

Proposition 1.5.1: Let $B = \max_i \max_{x^i} |R^i(x^i)|$. For fixed x , the optimal reward function $J(x, M)$ has the following properties as a function of M :

- (a) $J(x, M)$ is convex and monotonically nondecreasing.
- (b) $J(x, M)$ is constant for $M \leq -B/(1 - \alpha)$.
- (c) $J(x, M) = M$ for all $M \geq B/(1 - \alpha)$.

Proof: Consider the value iteration method starting with the function

$$J_0(x, M) = \max[0, M].$$

Successive iterates are generated by

$$J_{k+1}(x, M) = \max \left[M, \max_i L^i(x, M, J_k) \right], \quad k = 0, 1, \dots, \quad (1.59)$$

and we know from Prop. 1.2.1 that

$$\lim_{k \rightarrow \infty} J_k(x, M) = J(x, M), \quad \text{for all } x, M.$$

We show inductively that $J_k(x, M)$ has the properties (a) to (c) stated in the proposition and, by taking the limit as $k \rightarrow \infty$, we establish the same properties for J . Clearly, $J_0(x, M)$ satisfies properties (a) to (c). Assume that $J_k(x, M)$ satisfies (a) to (c). Then from Eqs. (1.57) and (1.59) it follows that $J_{k+1}(x, M)$ is convex and monotonically nondecreasing in M , since the expectation and maximization operations preserve these properties. Verification of (b) and (c) is straightforward, and is left for the reader. Q.E.D.

Consider now a problem where there is only one project that can be worked on, say project i . The optimal reward function for this problem

is denoted $J^i(x^i, M)$ and has the properties indicated in Prop. 1.5.1. A typical form for $J^i(x^i, M)$, viewed as a function of M for fixed x^i , is shown in Fig. 1.5.1. Clearly, there is a minimal value $m^i(x^i)$ of M for which $J^i(x^i, M) = M$; i.e.,

$$m^i(x^i) = \min\{M \mid J^i(x^i, M) = M\}, \quad \text{for all } x^i. \quad (1.60)$$

The function $m^i(x^i)$ is called the *index function* (or simply index) of project i . It provides an indifference threshold at each state; i.e., $m^i(x^i)$ is the retirement reward for which we are indifferent between retiring and operating the project when at state x^i .

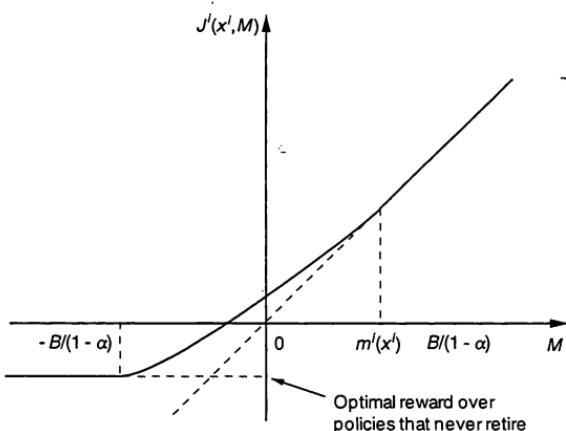


Figure 1.5.1 Form of the i th project reward function $J^i(x^i, M)$ for fixed x^i and definition of the index $m^i(x^i)$.

Our objective is to show the optimality of the index rule (1.56) for the index function defined by Eq. (1.60).

Project-by-Project Retirement Policies

Consider first a problem with a single project, say project i , and a fixed retirement reward M . Then by the definition (1.60) of the index, an optimal policy is to

$$\begin{aligned} &\text{retire project } i \text{ if } m^i(x^i) < M, \\ &\text{work on project } i \text{ if } m^i(x^i) \geq M. \end{aligned} \quad (1.61)$$

In other words, the project is operated continuously up to the time that its state falls into the *retirement set*

$$S^i = \{x^i \mid m^i(x^i) < M\}. \quad (1.62)$$

At that time the project is permanently retired.

Consider now the multiproject problem for fixed retirement reward M . Suppose that at some time we are at state $x = (x^1, \dots, x^n)$. Let us ask two questions:

1. Does it make sense to retire (from all projects) when there is still a project i with state x^i such that $m^i(x^i) > M$? The answer is negative. Retiring when $m^i(x^i) > M$ cannot be optimal, since if we operate project i exclusively up to the time that its state x^i falls within the retirement set S^i of Eq. (1.62) and then retire, we will gain a higher expected reward. [This follows from the definition (1.60) of the index and the nature of the optimal policy (1.61) for the single-project problem.]
2. Does it ever make sense to work on a project i with state in the retirement set S^i of Eq. (1.62)? Intuitively, the answer is negative; it seems unlikely that a project unattractive enough to be retired if it were the only choice would become attractive merely because of the availability of other projects that are independent in the sense assumed here.

We are led therefore to the conjecture that there is an optimal *project-by-project retirement (PPR) policy* that permanently retires projects in the same way as if they were the only project available. Thus at each time a PPR policy, when at state $x = (x^1, \dots, x^n)$,

$$\begin{array}{ll} \text{permanently retires project } i & \text{if } x^i \in S^i, \\ \text{works on some project} & \text{if } x^j \notin S^j \text{ for some } j, \end{array} \quad (1.63)$$

where S^i is the i th project retirement set of Eq. (1.62). Note that a PPR policy decides about retirement of projects but does not specify the project to be worked on out of those not yet retired.

The following proposition substantiates our conjecture. The proof is lengthy but quite simple.

Proposition 1.5.2: There exists an optimal PPR policy.

Proof: In view of Eqs. (1.57), and (1.63), existence of a PPR policy is equivalent to having, for all i ,

$$\max \left[M, \max_{j \neq i} L^j(x, M, J) \right] \geq L^i(x, M, J), \quad \text{for all } x \text{ with } x^i \in S^i, \quad (1.64)$$

$$M \leq L^i(x, M, J), \quad \text{for all } x \text{ with } x^i \notin S^i, \quad (1.65)$$

where L^i is given by

$$L^i(x, M, J) = R^i(x^i) + \alpha E_{w^i} \left\{ J(x^1, \dots, x^{i-1}, f^i(x^i, w^i), x^{i+1}, \dots, x^n, M) \right\}, \quad (1.66)$$

and $J(x, M)$ is the optimal reward function corresponding to x and M .

The i th single-project optimal reward function J^i clearly satisfies, for all x^i ,

$$J^i(x^i, M) \leq J(x^1, \dots, x^{i-1}, x^i, x^{i+1}, \dots, x^n, M), \quad (1.67)$$

since having the option of working at projects other than i cannot decrease the optimal reward. Furthermore, from the definition of the retirement set S^i [cf. Eq. (1.62)],

$$x^i \notin S^i, \quad \text{if } M \leq R^i(x^i) + \alpha E_{w^i} \left\{ J^i(f^i(x^i, w^i), M) \right\}. \quad (1.68)$$

Using Eqs. (1.66)-(1.68), we obtain Eq. (1.65).

It will suffice to show Eq. (1.64) for $i = 1$. Denote:

$\underline{x} = (x^2, \dots, x^n)$: The state of all projects other than project 1.

$J(\underline{x}, M)$: The optimal reward function for the problem resulting after project 1 is permanently retired.

$J(x^1, \underline{x}, M)$: The optimal reward function for the problem involving all projects and corresponding to state $x = (x^1, \underline{x})$.

We will show the following inequality for all $x = (x^1, \underline{x})$:

$$J(\underline{x}, M) \leq J(x^1, \underline{x}, M) \leq J(\underline{x}, M) + (J^1(x^1, M) - M). \quad (1.69)$$

In words this expresses the intuitively clear fact that at state (x^1, \underline{x}) one would be happy to retire project 1 permanently if one gets in return the maximum reward that can be obtained from project 1 in excess of the retirement reward M . We claim that to show Eq. (1.64) for $i = 1$, it will suffice to show Eq. (1.69). Indeed, when $x^1 \in S^1$, then $J^1(x^1, M) = M$, so from Eq. (1.69) we obtain $J(x^1, \underline{x}, M) = J(\underline{x}, M)$, which is in turn equivalent to Eq. (1.64) for $i = 1$.

We now turn to the proof of Eq. (1.69). Its left side is evident. To show the right side, we proceed by induction on the value iteration recursions

$$\begin{aligned} J_{k+1}(x^1, \underline{x}) &= \max \left[M, R^1(x^1) + \alpha E \left\{ J_k(f^1(x^1, w^1), \underline{x}) \right\}, \right. \\ &\quad \left. \max_{i \neq 1} [R^i(x^i) + \alpha E \{ J_k(x^1, F^i(\underline{x}, w^i)) \}] \right], \end{aligned} \quad (1.70)$$

$$J_{k+1}(\underline{x}) = \max \left[M, \max_{i \neq 1} [R^i(x^i) + \alpha E \{ J_k(F^i(\underline{x}, w^i)) \}] \right], \quad (1.71)$$

$$J_{k+1}^1(x^1) = \max [M, R^1(x^1) + \alpha E\{J_k^1(f^1(x^1, w^1))\}], \quad (1.72)$$

where, for all $i \neq 1$ and $\underline{x} = (x^2, \dots, x^n)$,

$$F^i(\underline{x}, w^i) = (x^2, \dots, x^{i-1}, f^i(x^i, w^i), x^{i+1}, \dots, x^n).$$

The initial conditions for the recursions (1.70)-(1.72) are

$$J_0(x^1, \underline{x}) = M, \quad \text{for all } (x^1, \underline{x}), \quad (1.73)$$

$$\underline{J}_0(\underline{x}) = M, \quad \text{for all } \underline{x}, \quad (1.74)$$

$$J_0^1(x^1) = M, \quad \text{for all } x^1. \quad (1.75)$$

We know that $J_k(x^1, \underline{x}) \rightarrow J(x^1, \underline{x}, M)$, $\underline{J}_k(\underline{x}) \rightarrow \underline{J}(\underline{x}, M)$, and $J_k^1(x^1) \rightarrow J^1(x^1, M)$, so to show Eq. (1.69) it will suffice to show that for all k and $x = (x^1, \underline{x})$ we have

$$J_k(x^1, \underline{x}) \leq \underline{J}_k(\underline{x}) + (J_k^1(x^1) - M). \quad (1.76)$$

In view of the definitions (1.73)-(1.75), we see that Eq. (1.76) holds for $k = 0$. Assume that it holds for some k . We will show that it holds for $k+1$. From Eqs. (1.70)-(1.72) and the induction hypothesis (1.76), we have

$$\begin{aligned} J_{k+1}(x^1, \underline{x}) &\leq \max \left[M, R^1(x^1) + \alpha E\{\underline{J}_k(\underline{x}) + J_k^1(f^1(x^1, w^1)) - M\}, \right. \\ &\quad \left. \max_{i \neq 1} [R^i(x^i) + \alpha E\{\underline{J}_k(F^i(\underline{x}, w^i)) + J_k^1(x^1) - M\}] \right]. \end{aligned}$$

Using the facts $\underline{J}_k(\underline{x}) \geq M$ and $J_k^1(x^1) \geq M$ [cf. Eqs. (1.70)-(1.72)], and the preceding equation, we see that

$$J_{k+1}(x^1, \underline{x}) \leq \max[\beta_1, \beta_2],$$

where

$$\beta_1 = \max \left[M, R^1(x^1) + \alpha E\{J_k^1(f^1(x^1, w^1))\} \right] + \alpha(\underline{J}_k(\underline{x}) - M),$$

$$\beta_2 = \max \left[M, \max_{i \neq 1} [R^i(x^i) + \alpha E\{\underline{J}_k(F^i(\underline{x}, w^i))\}] \right] + \alpha(J_k^1(x^1) - M).$$

Using Eqs. (1.71), (1.72), and the preceding equations, we see that

$$J_{k+1}(x^1, \underline{x}) \leq \max[J_{k+1}^1(x^1) + \widehat{\underline{J}_k(\underline{x})} - M, \underline{J}_{k+1}(\underline{x}) + J_k^1(x^1) - M]. \quad (1.77)$$

It can be seen from Eqs. (1.70)-(1.72) and (1.73)-(1.75) that $J_k^1(x^1) \leq J_{k+1}^1(x^1)$ and $\underline{J}_k(\underline{x}) \leq \underline{J}_{k+1}(\underline{x})$ for all k , x^1 , and \underline{x} , so from Eq. (1.77) we obtain that Eq. (1.76) holds for $k+1$. The induction is complete. **Q.E.D.**

As a first step towards showing optimality of the index rule, we use the preceding proposition to derive an expression for the partial derivative of $J(x, M)$ with respect of M .

Lemma 1.5.1: For fixed x , let K_M denote the retirement time under an optimal policy when the retirement reward is M . Then for all M for which $\partial J(x, M)/\partial M$ exists we have

$$\frac{\partial J(x, M)}{\partial M} = E\{\alpha^{K_M} \mid x_0 = x\}.$$

Proof: Fix x and M . Let π^* be an optimal policy and let K_M be the retirement time under π^* . If π^* is used for a problem with retirement reward $M + \epsilon$, we receive

$$E\{\text{reward prior to retirement}\} + (M + \epsilon)E\{\alpha^{K_M}\} = J(x, M) + \epsilon E\{\alpha^{K_M}\}.$$

The optimal reward $J(x, M + \epsilon)$ when the retirement reward is $M + \epsilon$ is no less than the preceding expression, so

$$J(x, M + \epsilon) \geq J(x, M) + \epsilon E\{\alpha^{K_M}\}.$$

Similarly, we obtain

$$J(x, M - \epsilon) \geq J(x, M) - \epsilon E\{\alpha^{K_M}\}.$$

For $\epsilon > 0$, these two relations yield

$$\frac{J(x, M) - J(x, M - \epsilon)}{\epsilon} \leq E\{\alpha^{K_M}\} \leq \frac{J(x, M + \epsilon) - J(x, M)}{\epsilon}.$$

The result follows by taking $\epsilon \rightarrow 0$. Q.E.D.

Note that the convexity of $J(x, \cdot)$ with respect to M (Prop. 1.5.1) implies that the derivative $\partial J(x, M)/\partial M$ exists almost everywhere with respect to Lebesgue measure (Rockafellar [Roc70]). Furthermore, it can be shown that $\partial J(x, M)/\partial M$ exists for all M for which the optimal policy is unique.

For a given M , initial state x , and optimal PPR policy, let T_i be the retirement time of project i if it were the only project available, let T be the retirement time for the multiproject problem. Both T_i and T take values that are either nonnegative or ∞ . The existence of an optimal PPR policy implies that we must have

$$T = T_1 + \cdots + T_n$$

and in addition $T_i, i = 1, \dots, n$, are independent random variables. Therefore,

$$E\{\alpha^T\} = E\{\alpha^{T_1+\dots+T_n}\} = \prod_{i=1}^n E\{\alpha^{T_i}\}.$$

Using Lemma 1.5.1, we obtain

$$\frac{\partial J(x, M)}{\partial M} = \prod_{i=1}^n \frac{\partial J^i(x^i, M)}{\partial M}. \quad (1.78)$$

Optimality of the Index Rule

We are now ready to show our main result.

Proposition 1.5.3: The index rule (1.56) is an optimal stationary policy.

Proof: Fix $x = (x^1, \dots, x^n)$, denote

$$m(x) = \max_j \{m^j(x^j)\},$$

and let i be such that

$$m^i(x^i) = \max_j \{m^j(x^j)\}.$$

If $m(x) < M$ the optimality of the index rule (1.56) at state x follows from the existence of an optimal PPR policy. If $m(x) \geq M$, we note that

$$J^i(x^i, M) = R^i(x^i) + \alpha E\{J^i(f^i(x^i, w^i), M)\}$$

and then use this relation together with Eq. (1.78) to write

$$\begin{aligned} \frac{\partial J(x, M)}{\partial M} &= \frac{\partial J^i(x^i, M)}{\partial M} \cdot \prod_{j \neq i} \frac{\partial J^j(x^j, M)}{\partial M} \\ &= \alpha \frac{\partial}{\partial M} E \left\{ J^i(f^i(x^i, w^i), M) \cdot \prod_{j \neq i} \frac{\partial J^j(x^j, M)}{\partial M} \right\} \\ &= \alpha E \left\{ \underbrace{\frac{\partial}{\partial M} J^i(f^i(x^i, w^i), M)}_{=} \cdot \prod_{j \neq i} \frac{\partial J^j(x^j, M)}{\partial M} \right\} \\ &= \alpha E \left\{ \frac{\partial}{\partial M} J(x^1, \dots, x^{i-1}, f^i(x^i, w^i), x^{i+1}, \dots, x^n, M) \right\} \\ &= \alpha \frac{\partial}{\partial M} E\{J(x^1, \dots, x^{i-1}, f^i(x^i, w^i), x^{i+1}, \dots, x^n, M)\}, \end{aligned}$$

and finally

$$\frac{\partial J(x, M)}{\partial M} = \frac{\partial}{\partial M} L^i(x, M, J),$$

where

$$L^i(x, M, J) = R^i(x^i) + \alpha E\{J(x^1, \dots, x^{i-1}, f^i(x^i, w^i), x^{i+1}, \dots, x^n, M)\}.$$

(The interchange of differentiation and expectation can be justified for almost all M ; see Bertsekas [Ber73a].) By the existence of an optimal PPR policy, we also have

$$J(x, m(x)) = L^i(x, m(x), J).$$

Therefore, the convex functions $J(x, M)$ and $L^i(x, M, J)$ viewed as functions of M for fixed x are equal for $M = m(x)$ and have equal derivative for almost all $M \leq m(x)$. It follows that for all $M \leq m(x)$ we have

$$J(x, M) = L^i(x, M, J).$$

This implies that the index rule (1.56) is optimal for all x with $m(x) \geq M$. Q.E.D.

Deteriorating and Improving Cases

It is evident that great simplification results from the optimality of the index rule (1.56), since optimization of a multiproject problem has been reduced to n separate single-project optimization problems. Nonetheless, solution of each of these single-project problems can be complicated. Under certain circumstances, however, the situation simplifies.

Suppose that for all i , x^i , and w^i that can occur with positive probability, we have either

$$m^i(x^i) \leq m^i(f^i(x^i, w^i)) \quad (1.79)$$

or

$$m^i(x^i) \geq m^i(f^i(x^i, w^i)). \quad (1.80)$$

Under Eq. (1.79) [or Eq. (1.80)] projects become more (less) profitable as they are worked on. We call these cases *improving* and *deteriorating*, respectively.

In the improving case the nature of the optimal policy is evident: either retire at the first period or else select a project with maximal index at the first period and continue engaging that project for all subsequent periods.

In the deteriorating case, note that Eq. (1.80) implies that if retirement is optimal when at state x^i then it is also optimal at each state $f^i(x^i, w^i)$. Therefore, for all x^i such that $M = m^i(x^i)$ we have, for all w^i ,

$$J^i(x^i, M) = M, \quad J^i(f^i(x^i, w^i), M) = M.$$

From Bellman's equation

$$J^i(x^i, M) = \max [M, R^i(x^i) + \alpha E\{J^i(f^i(x^i, w^i), M)\}]$$

we obtain

$$m^i(x^i) = R^i(x^i) + \alpha m^i(x^i)$$

or

$$m^i(x^i) = \frac{R^i(x^i)}{1 - \alpha}.$$

Thus the optimal policy in the deteriorating case is

retire if $M > \max_i \frac{R^i(x^i)}{1 - \alpha}$

engage the project i with maximal one-step reward $R^i(x^i)$ otherwise.

Example 1.5.1 (Treasure Hunting)

Consider a search problem involving N sites. Each site i may contain a treasure with expected value v_i . A search at site i costs c_i and reveals the treasure with probability β_i (assuming a treasure is there). Let P_i be the probability that there is a treasure at site i . We take P_i as the state of the project corresponding to searching site i . Then the corresponding one-step reward is

$$R^i(P_i) = \beta_i P_i v_i - c_i. \quad (1.81)$$

If a search at site i does not reveal the treasure, the probability P_i drops to

$$\bar{P}_i = \frac{P_i(1 - \beta_i)}{P_i(1 - \beta_i) + 1 - P_i},$$

as can be verified using Bayes' rule. If the search finds the treasure, the probability P_i drops to zero, since the treasure is removed from the site. Based on this and the fact that $R^i(P_i)$ is increasing with P_i [cf. Eq. (1.81)], it is seen that the deteriorating condition (1.80) holds. Therefore, it is optimal to search the site i for which the expression $R^i(P_i)$ of Eq. (1.81) is maximal, provided $\max_i R^i(P_i) > 0$, and to retire if $R^i(P_i) \leq 0$ for all i .

Suboptimal Policies for Multiarmed Bandit Problems

There are many types of problems that possess the basic character of multiarmed bandit problems, i.e., the selection of one out of several projects to work on at each time, but do not fully satisfy the assumptions of the present section. Here are some examples:

- (a) The state of the projects that are not worked on may not stay fixed, but rather may evolve according to some probabilistic mechanism.
- (b) There are time windows, which constrain the times at which each project may be worked on. The start time and the length of each time window may be known in advance or they may be generated according to a probabilistic mechanism as the system is operating.
- (c) There are precedence constraints, whereby some projects can be worked on only after some other projects have been worked on.

In situations such as the above an index policy may not be optimal, yet policies of the index type may form the basis for effective suboptimal control. One possibility is to use a (suboptimal) easily computable index policy as the base policy for a rollout algorithm, as described in Section 6.4 of Vol. I. The project indexes may be obtained by some heuristic or by solving a related multiarmed bandit problem that admits an optimal index policy. Examples of this type of rollout algorithm are given in the paper by Bertsekas and Castanon [BeC99].

An alternative possibility is to define a suboptimal policy by means of a cost function approximation. To illustrate this, consider the multiarmed bandit problem of this section, but with two differences:

- (1) When a project i is not worked on, its state changes according to

$$x_{k+1}^i = \bar{f}^i(x_k^i, \bar{w}_k^i),$$

where \bar{f}^i is a given function and \bar{w}_k^i is a random disturbance with distribution depending on x_k^i but not on prior disturbances. Furthermore, a reward $\bar{R}^i(x_k^i)$ is earned, where \bar{R}^i is a given function.

- (2) Retirement is not an option. (Alternatively, we could allow the possibility that no project is worked on at a given time. This would correspond to introducing an artificial project that earns no reward when worked on.)

For an example, let the projects correspond to machines only one of which can be maintained at each time. With maintenance, a machine goes to a “better” state where it earns a higher reward, while without maintenance the machine tends to drift to “worse” states where it tends to earn a lower reward. There may also be a machine-dependent cost for maintenance, which can be embedded in the function $R^i(x^i)$. The problem is in effect to

select at each time the project whose maintenance will contribute most to the total discounted reward earned by the system.

Suppose that the optimal reward function $J^*(x^1, \dots, x^n)$ is approximated by a separable function of the form $\sum_{i=1}^n \tilde{J}^i(x^i)$, where each \tilde{J}^i is a function corresponding to the contribution of the i th project to the total reward. The corresponding one-step lookahead policy selects the project i that maximizes

$$R^i(x^i) + \sum_{j \neq i} \bar{R}^j(x^j) + \alpha E\{\tilde{J}^i(f^i(x^i, w^i))\} + \alpha \sum_{j \neq i} E\left\{\tilde{J}^j(\bar{f}^j(x^j, \bar{w}^j))\right\},$$

which can also be written as

$$\begin{aligned} R^i(x^i) - \bar{R}^i(x^i) + \alpha E\left\{\tilde{J}^i(f^i(x^i, w^i) - \tilde{J}^i(\bar{f}^i(x^i, \bar{w}^i)))\right\} \\ + \sum_{j=1}^n \left\{\bar{R}^j(x^j) + \alpha E\{\tilde{J}^j(\bar{f}^j(x^j, \bar{w}^j))\}\right\}. \end{aligned}$$

Noting that the last term in the above expression does not depend on i , it follows that the one-step lookahead policy takes the form

$$\text{work on project } i \quad \text{if} \quad \tilde{m}^i(x^i) = \max_j \{\tilde{m}^j(x^j)\},$$

where for all i ,

$$\tilde{m}^i(x^i) = R^i(x^i) - \bar{R}^i(x^i) + \alpha E\{\tilde{J}^i(f^i(x^i, w^i)) - \tilde{J}^i(\bar{f}^i(x^i, \bar{w}^i))\}.$$

We may view $\tilde{m}^i(x^i)$ as an approximate index for project i , induced by the separable reward function approximation $\sum_{i=1}^n \tilde{J}^i(x^i)$.

An important question for the implementation of the above suboptimal index rule is the computation of the separable reward function terms $\tilde{J}^i(x^i)$. There are many possibilities here, and the best choice may depend strongly on the problem's structure. For example, one may obtain $\tilde{J}^i(x^i)$ by solving the corresponding single project problems using an algorithm such as value or policy iteration. It is interesting to note in this respect that adding a constant to the function $\tilde{J}^i(x^i)$ does not affect the value of the approximate index $\tilde{m}^i(x^i)$. Another possibility is to use a parameter approximation scheme, starting with a separable approximation architecture of the form $\sum_{i=1}^n \tilde{J}^i(x^i, r^i)$, where the r^i are vectors of “tunable” weights (cf. the discussion of Section 6.3.4 of Vol. I). The values of the r^i can be obtained by some form of heuristic search or by using the more systematic approximate dynamic programming methods to be discussed in Chapter 6.

1.6 NOTES, SOURCES, AND EXERCISES

Many authors have contributed to the analysis of the discounted problem with bounded cost per stage, most notably Shapley [Sha53], Bellman [Bel57], and Blackwell [Bla65]. For variations and extensions involving multiple criteria, weighted criteria, and constraints, see Feinberg and Shwartz [FeS94], Ghosh [Gho90], Ross [Ros89], and White and Kim [WhK80]. The mathematical issues relating to measurability concerns are analyzed extensively in Bertsekas and Shreve [BeS78], Dynkin and Yuskevich [DyY79], Hernandez-Lerma [Her89], and Hinderer [Hin70]. The lower semianalytic/universally measurable framework, described in Appendix A, was first proposed by Bertsekas and Shreve [BeS78].

The error bounds given in Section 1.3 and Exercise 1.9 are improvements on results of MacQueen [McQ66] (see Porteus [Por71], [Por75], Bertsekas [Ber76], and Porteus and Totten [PoT78]). The corresponding convergence rate was discussed by Morton [Mor71], and Morton and Wecker [MoW77]. The Gauss-Seidel method for discounted problems was proposed by Hastings [Has68]. An extensive discussion of the convergence aspects of the method and related background is given in Section 2.6 of Bertsekas and Tsitsiklis [BeT89]. The material on the generic rank-one correction, including the convergence analysis of Exercise 1.8 is due to Bertsekas [Ber95a], which also describes a multiple-rank correction method where the effect of several eigenvalues is nullified. Value iteration is particularly well-suited for parallel computation; see e.g., Archibald, McKinnon, and Thomas [AMT93], and Bertsekas and Tsitsiklis [BeT89].

Policy iteration for discounted problems was proposed by Bellman [Bel57]. The modified policy iteration algorithm was suggested and analyzed by van Nunen [Van76], and by Puterman and Shin [PuS78], [PuS82]; see also Puterman [Put94]. The convergence of the asynchronous policy iteration method (Prop. 1.3.5) was shown by Williams and Baird [WiB93]. The approximate policy iteration analysis (Prop. 1.3.6) is due to Bertsekas and Tsitsiklis [BeT96]. The relation between policy iteration and Newton's method (Exercise 1.10) was pointed out by Pollatschek and Avi-Itzhak [PoA69], and was further discussed by Puterman and Brumelle [Pub78].

The linear programming approach of Section 1.3.4 was proposed by D'Epenoux [D'Ep60]. There is a relation between policy iteration and the simplex method applied to solving the linear program associated with the discounted problem. In particular, it can be shown that the simplex method for linear programming with a block pivoting rule is mathematically equivalent to the policy iteration algorithm; see e.g., Kallenberg [Kal83], [Kal94a], [Kal94b], and Puterman [Put94], and the references quoted there. Approximation methods using basis functions and linear programming were proposed by Schweitzer and Seidman [ScS85], and have been further developed by de Farias and Van Roy [DFV03], [DFV04], [DeF04]. For an application

to pricing of network services, see Paschalidis and Tsitsiklis [PaT00].

The error bound of Prop. 1.3.7(a) for one-step lookahead policies is new. The error bound of Prop. 1.3.7(b) has been proved in several sources, including Williams and Baird [WiB93], and Tsitsiklis and Van Roy [TsV96]. Chapter 6 of Vol. I and the book by Bertsekas and Tsitsiklis [BeT96] contain a great deal of discussion on limited lookahead policies and rollout algorithms. A survey with emphasis on rollout algorithms and their connection with model predictive control is given by the author in [Ber05a]. Rollout algorithms for constrained DP problems are discussed in Bertsekas [Ber05a], [Ber05b].

A complexity analysis of finite-state infinite horizon problems is given by Papadimitriou and Tsitsiklis [PaT87]. Discretization methods that approximate infinite state space systems with finite-state Markov chains, are discussed by Bertsekas [Ber75], Fox [Fox71], Haurie and L’Ecuyer [HaL86], Whitt [Whi78], [Whi79], and White [Whi80a]. For related multigrid approximation methods and associated complexity analysis, see Chow and Tsitsiklis [ChT89], [ChT91]. A different approach to deal with infinite state spaces, which is based on randomization, has been introduced by Rust [Rus97]; see also Rust [Rus95].

The role of contraction mappings in discounted problems was first recognized and exploited by Shapley [Sha53], which considered two-player dynamic games. Abstract DP models and the implications of monotonicity and contraction have been explored in detail in Denardo [Den67], Bertsekas [Ber77], Bertsekas and Shreve [BeS78], and Verd'u and Poor [VeP84], [VeP87]. Countable-state discounted problems with unbounded cost per stage (cf. Section 1.4.4) were discussed by Harrison [Har72], Lippman [Lip73], [Lip75a], van Nunen [Van76], Wessels [Wes77], van Nunen and Wessels [VaW78], and Cavazos-Cadena [Cav86].

The index rule solution of the multiarmed bandit problem is due to Gittins [Git79], and Gittins and Jones [GiJ74]. Subsequent contributions include Whittle [Whi80b], Kelly [Kel81], and Whittle [Whi81], [Whi82]. The proof given here is due to Tsitsiklis [Tsi86]. Alternative proofs and further analysis are given in Kumar [Kum85], Varaiya, Walrand, and Buyukkoc [VWB85], Kumar and Varaiya [KuV86], Nain, Tsoucas, and Walrand [NTW89], Weber [Web93], Bertsimas and Nino-Mora [BeN93], Tsitsiklis [Tsi94a], and Bertsimas, Paschalidis, and Tsitsiklis [BPT94a], [BPT94b].

Finally, we note that even though our analysis in this chapter requires a countable disturbance space, it may still serve as the starting point of analysis of problems with uncountable disturbance space. This can be done by reducing such problems to deterministic problems having as state space a set of probability measures. The basic idea of this reduction is demonstrated in Exercise 1.13. This line of analysis was adopted by Bertsekas and Shreve [BeS78] (Chapter 9) for the resolution of measurability questions in infinite horizon stochastic control problems.

E X E R C I S E S

1.1

Write a computer program and compute iteratively the vector J_μ satisfying

$$J_\mu = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + \alpha \begin{bmatrix} 3/4 & 1/4 & 0 \\ 1/4 & 3/4 - \epsilon & \epsilon \\ 0 & \epsilon & 1 - \epsilon \end{bmatrix} J_\mu.$$

Do your computations for all combinations of $\alpha = 0.9$ and $\alpha = 0.999$, and $\epsilon = 0.5$ and $\epsilon = 0.001$. Try value iteration with and without error bounds. Discuss your results.

1.2

The purpose of this problem is to show that shortest path problems with a discount factor make little sense. Suppose that we have a graph with a nonnegative length a_{ij} for each arc (i, j) . The cost of a path (i_0, i_1, \dots, i_m) is $\sum_{k=0}^{m-1} \alpha^k a_{i_k i_{k+1}}$, where α is a discount factor from $(0, 1)$. Consider the problem of finding a path of minimum cost that connects two given nodes. Show that this problem need not have a solution.

1.3

Consider a problem similar to that of Section 1.1 except that when we are at state x_k , there is a probability β , where $0 < \beta < 1$, that the next state x_{k+1} will be determined according to $x_{k+1} = f(x_k, u_k, w_k)$ and a probability $(1 - \beta)$ that the system will move to a termination state, where it stays permanently thereafter at no cost. Show that even if $\alpha = 1$, the problem can be put into the discounted cost framework.

1.4

Consider a problem similar to that of Section 1.2 except that the discount factor α depends on the current state x_k , the control u_k , and the disturbance w_k ; i.e., the cost function has the form

$$J_\pi(x_0) = \lim_{N \rightarrow \infty} \underset{\substack{w_k \\ k=0,1,\dots}}{E} \left\{ \sum_{k=0}^{N-1} \alpha_{\pi,k} g(x_k, \mu_k(x_k), w_k) \right\},$$

where

$$\alpha_{\pi,k} = \alpha(x_0, \mu_0(x_0), w_0) \alpha(x_1, \mu_1(x_1), w_1) \cdots \alpha(x_k, \mu_k(x_k), w_k),$$

with $\alpha(x, u, w)$ a given function satisfying

$$\begin{aligned} 0 &\leq \min\{\alpha(x, u, w) \mid x \in S, u \in C, w \in D\} \\ &\leq \max\{\alpha(x, u, w) \mid x \in S, u \in C, w \in D\} \\ &< 1. \end{aligned}$$

Argue that the results and algorithms of Sections 1.2 and 1.3 have direct counterparts for such problems.

1.5 (Column Reduction [Por75]) www

The purpose of this problem is to provide a transformation of a certain type of discounted problem into another discounted problem with smaller discount factor. Consider the n -state discounted problem under the assumptions of Section 1.3. The cost per stage is $g(i, u)$, the discount factor is α , and the transition probabilities are $p_{ij}(u)$. For each $j = 1, \dots, n$, let

$$m_j = \min_{i=1, \dots, n} \min_{u \in U(i)} p_{ij}(u).$$

For all i, j , and u , let

$$\tilde{p}_{ij}(u) = \frac{p_{ij}(u) - m_j}{1 - \sum_{k=1}^n m_k},$$

assuming that $\sum_{k=1}^n m_k < 1$.

(a) Show that $\tilde{p}_{ij}(u)$ are transition probabilities.

(b) Consider the discounted problem with cost per stage $g(i, u)$, discount factor

$$\alpha \left(1 - \sum_{j=1}^n m_j \right),$$

and transition probabilities $\tilde{p}_{ij}(u)$. Show that this problem has the same optimal policies as the original, and that its optimal cost vector J' satisfies

$$J^* = J' + \frac{\alpha \sum_{j=1}^n m_j J'(j)}{1 - \alpha} e,$$

where J^* is the optimal cost vector of the original problem and e is the unit vector.

1.6

Let $\bar{J} : S \mapsto \mathbb{R}$ be any bounded function on S and consider the value iteration method of Section 1.3 with a starting function $J : S \mapsto \mathbb{R}$ of the form

$$J(x) = \bar{J}(x) + r, \quad x \in S,$$

where r is some scalar. Show that the bounds $(T^k J)(x) + \underline{c}_k$ and $(T^k J)(x) + \bar{c}_k$ of Prop. 1.3.1 are independent of the scalar r for all $x \in S$. Show also that if S consists of a single state \tilde{x} (i.e., $S = \{\tilde{x}\}$), then

$$(TJ)(\tilde{x}) + \underline{c}_1 = (TJ)(\tilde{x}) + \bar{c}_1 = J^*(\tilde{x}).$$

1.7 (Jacobi Version of Value Iteration) [www](#)

Consider the problem of Section 1.3 and the version of the value iteration method that starts with an arbitrary function $J : S \mapsto \mathbb{R}$ and generates recursively FJ, F^2J, \dots , where F is the mapping given by

$$(FJ)(i) = \min_{u \in U(i)} \frac{g(i, u) + \alpha \sum_{j \neq i} p_{ij}(u) J(j)}{1 - \alpha p_{ii}(u)}.$$

Show that $(F^k J)(i) \rightarrow J^*(i)$ as $k \rightarrow \infty$ and provide a rate of convergence estimate that is at least as favorable as the one for the ordinary method (cf. Prop. 1.2.3). Show that F is the DP mapping for an equivalent DP problem where there is 0 probability of self-transition at every state.

1.8 (Convergence Properties of Rank-One Correction [Ber95a])

Consider the solution of the system $J = FJ$, where $F : \mathbb{R}^n \mapsto \mathbb{R}^n$ is the mapping

$$FJ = h + QJ,$$

h is a given vector in \mathbb{R}^n , and Q is an $n \times n$ matrix. Consider the generic rank-one correction iteration $J := MJ$, where $M : \mathbb{R}^n \mapsto \mathbb{R}^n$ is the mapping

$$MJ = FJ + \gamma z,$$

and

$$z = Qd, \quad \gamma = \frac{(d - z)'(FJ - J)}{\|d - z\|^2}.$$

- (a) Show that any solution J^* of the system $J = FJ$ satisfies $J^* = MJ^*$.
- (b) Verify that the value iteration method that uses the error bounds in the manner of Eq. (1.21) is a special case of the iteration $J := MJ$ with d equal to the unit vector.
- (c) Assume that d is an eigenvector of Q , let λ be the corresponding eigenvalue, and let $\lambda_1, \dots, \lambda_{n-1}$ be the remaining eigenvalues. Show that MJ can be written as

$$MJ = h + RJ,$$

where h is some vector in \mathbb{R}^n and

$$R = Q - \frac{\lambda}{(1 - \lambda)\|d\|^2} dd'(I - Q).$$

Show also that $Rd = 0$ and that for all k and J ,

$$R^k = RQ^{k-1}, \quad M^k J = M(F^{k-1}J).$$

Furthermore, the eigenvalues of R are $0, \lambda_1, \dots, \lambda_{n-1}$. (This last statement requires a somewhat complicated proof; see [Ber95a].)

- (d) Let d be as in part (c), and suppose that e_1, \dots, e_{n-1} are eigenvectors corresponding to $\lambda_1, \dots, \lambda_{n-1}$. Suppose that a vector J can be written as

$$J = J^* + \xi e + \sum_{i=1}^{n-1} \xi_i e_i,$$

where J^* is a solution of the system. Show that, for all $k > 1$,

$$M^k J = J^* + \sum_{i=1}^{n-1} \xi_i \lambda_i^{k-1} R e_i,$$

so that if λ is a dominant eigenvalue and $\lambda_1, \dots, \lambda_{n-1}$ lie within the unit circle, $M^k J$ converges to J^* at a rate governed by the subdominant eigenvalue. Note: This result can be generalized for the case where Q does not have a full set of linearly independent eigenvectors, and for the case where F is modified through multiple-rank corrections [Ber95a].

1.9 (Generalized Error Bounds [Ber76]) www

Let S be a set and $B(S)$ be the set of all bounded real-valued functions on S . Let $T : B(S) \mapsto B(S)$ be a mapping with the following two properties:

- (1) $TJ \leq TJ'$ for all $J, J' \in B(S)$ with $J \leq J'$.
- (2) For every scalar $r \neq 0$ and all $x \in S$,

$$\alpha_1 \leq \frac{(T(J+r\epsilon))(x) - (TJ)(x)}{r} \leq \alpha_2,$$

where α_1, α_2 are two scalars with $0 \leq \alpha_1 \leq \alpha_2 < 1$.

- (a) Show that T is a contraction mapping on $B(S)$, and hence for every $J \in B(S)$ we have

$$\lim_{k \rightarrow \infty} (T^k J)(x) = J^*(x), \quad x \in S,$$

where J^* is the unique fixed point of T in $B(S)$.

- (b) Show that for all $J \in B(S)$, $x \in S$, and $k = 1, 2, \dots$,

$$\begin{aligned} (T^k J)(x) + \underline{c}_k &\leq (T^{k+1} J)(x) + \underline{c}_{k+1} \leq J^*(x) \leq (T^{k+1} J)(x) + \bar{c}_{k+1} \\ &\leq (T^k J)(x) + \bar{c}_k, \end{aligned}$$

where for all k

$$\underline{c}_k = \min \left\{ \frac{\alpha_1}{1 - \alpha_1} \min_{x \in S} [(T^k J)(x) - (T^{k-1} J)(x)], \right. \\ \left. \frac{\alpha_2}{1 - \alpha_2} \min_{x \in S} [(T^k J)(x) - (T^{k-1} J)(x)] \right\}, \quad (1.82)$$

$$\bar{c}_k = \max \left\{ \frac{\alpha_1}{1 - \alpha_1} \max_{x \in S} [(T^k J)(x) - (T^{k-1} J)(x)], \frac{\alpha_2}{1 - \alpha_2} \max_{x \in S} [(T^k J)(x) - (T^{k-1} J)(x)] \right\}. \quad (1.83)$$

A geometric interpretation of these relations for the case where S consists of a single element is provided in Fig. 1.6.1.

- (c) Consider the following algorithm:

$$J_k(x) = (TJ_{k-1})(x) + \gamma_k, \quad x \in S,$$

where J_0 is any function in $B(S)$, γ_k is any scalar in the range $[\underline{c}_k, \bar{c}_k]$, and \underline{c}_k and \bar{c}_k are given by Eqs. (6.1) and (6.2) with $(T^k J)(x) - (T^{k-1} J)(x)$ replaced by $(TJ_{k-1})(x) - J_{k-1}(x)$. Show that for all k ,

$$\max_{x \in S} |J_k(x) - J^*(x)| \leq \alpha_2^k \max_{x \in S} |J_0(x) - J^*(x)|.$$

- (d) Let $J \in \mathbb{R}^n$ and consider the equation $J = TJ$, where

$$TJ = h + MJ$$

and the vector $h \in \mathbb{R}^n$ and the matrix M are given. Let s_i be the i th row sum of M , i.e.,

$$s_i = \sum_{j=1}^n m_{ij},$$

and let $\alpha_1 = \min_i s_i$, $\alpha_2 = \max_i s_i$. Show that if the elements m_{ij} of M are all nonnegative and $\alpha_2 < 1$, then the conclusions of parts (a) and (b) hold.

- (e) [Por75] Consider the Gauss-Seidel method for solving the system $J = g + \alpha P J$, where $0 < \alpha < 1$ and P is a transition probability matrix. Use part (d) to obtain suitable error bounds.

1.10 (Policy Iteration and Newton's Method)

The purpose of this problem is to demonstrate a relation between policy iteration and Newton's method for solving nonlinear equations. Consider an equation of the form $F(J) = 0$, where $F: \mathbb{R}^n \mapsto \mathbb{R}^n$. Given a vector $J_k \in \mathbb{R}^n$, Newton's method determines J_{k+1} by solving the linear system of equations

$$F(J_k) + \frac{\partial F(J_k)}{\partial J} (J_{k+1} - J_k) = 0,$$

where $\partial F(J_k)/\partial J$ is the Jacobian matrix of F evaluated at J_k .

- (a) Consider the discounted finite-state problem of Section 1.3 and define

$$F(J) = TJ - J.$$

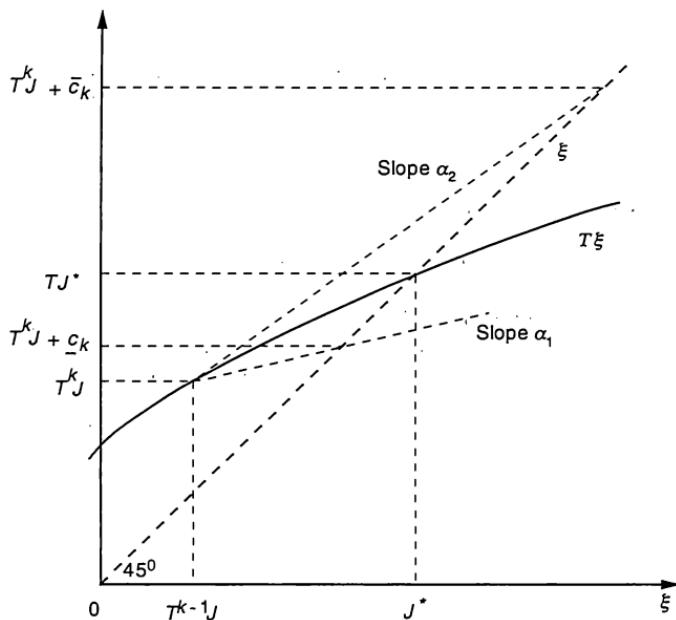


Figure 1.6.1 Graphical interpretation of the error bounds of Exercise 1.9.

Show that if there is a unique μ such that

$$T_\mu J = TJ,$$

then the Jacobian matrix of F at J is

$$\frac{\partial F(J)}{\partial J} = \alpha P_\mu - I,$$

where I is the $n \times n$ identity.

- (b) Show that the policy iteration algorithm can be identified with Newton's method for solving $F(J) = 0$ (assuming it gives a unique policy at each step).

1.11 (Minimax Problems)

Provide analogs of the results and algorithms of Sections 1.2 and 1.3 for the minimax problem where the cost is

$$J_\pi(x_0) = \lim_{N \rightarrow \infty} \max_{w_k \in W(x_k, \mu_k(x_k))} \sum_{k=0,1,\dots}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k),$$

g is bounded, x_k is generated by $x_{k+1} = f(x_k, \mu_k(x_k), w_k)$, and $W(x, u)$ is a given nonempty subset of D for each $(x, u) \in S \times C$. (Compare with Exercise 1.5 in Chapter 1 of Vol. I.)

1.12 (Data Transformations [Sch72]) [www](#)

A finite-state problem where the discount factor at each stage depends on the state can be transformed into a problem with state-independent discount factors. To see this, consider the following set of equations in the variables $J(i)$:

$$J(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^n m_{ij}(u) J(j) \right], \quad i = 1, \dots, n, \quad (1.84)$$

where we assume that for all $i, u \in U(i)$, and j , $m_{ij}(u) \geq 0$ and

$$M_i(u) = \sum_{j=1}^n m_{ij}(u) < 1.$$

Let

$$\alpha = \max_{i=1, \dots, n} \left\{ \frac{M_i(u) - m_{ii}(u)}{1 - m_{ii}(u)} \right\},$$

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases}$$

and define, for all i and j ,

$$\bar{g}(i, u) = \frac{g(i, u)(1 - \alpha)}{1 - M_i(u)},$$

$$\bar{m}_{ij}(u) = \delta_{ij} + \frac{(1 - \alpha)(m_{ij}(u) - \delta_{ij})}{1 - M_i(u)}.$$

Show that, for all i and j ,

$$\sum_{j=1}^n \bar{m}_{ij}(u) = \alpha < 1, \quad \bar{m}_{ij}(u) \geq 0,$$

and that a solution $\{J(i) \mid i = 1, \dots, n\}$ of Eq. (1.84) is also a solution of the equations

$$J(i) = \min_{u \in U(i)} \left[\bar{g}(i, u) + \alpha \sum_{j=1}^n \bar{p}_{ij}(u) J(j) \right], \quad i = 1, \dots, n,$$

where $\bar{p}_{ij}(u)$ are the transitions probabilities defined by

$$\bar{p}_{ij}(u) = \frac{\bar{m}_{ij}(u)}{\alpha}.$$

1.13 (Stochastic to Deterministic Problem Transformation)

Under the assumptions and notation of Section 1.3, consider the controlled system

$$p_{k+1} = p_k P_{\mu_k}, \quad k = 0, 1, \dots,$$

where p_k is a probability distribution over S viewed as a row vector, and P_{μ_k} is the transition probability matrix corresponding to the control function μ_k . The state is p_k and the control is μ_k . Consider also the cost function

$$\lim_{N \rightarrow \infty} \sum_{k=0}^{N-1} \alpha^k p_k g_{\mu_k}.$$

Show that the optimal cost and an optimal policy for the deterministic problem involving the above system and cost function yield the optimal cost and an optimal policy for the discounted cost problem of Section 1.3.

1.14 (Threshold Policies and Policy Iteration)

- (a) Consider the machine replacement example of Section 1.2, and assume that the condition (1.10) holds. Let us define a *threshold* policy to be a stationary policy that replaces if and only if the state is greater than or equal to some fixed state i . Suppose that we start the policy iteration algorithm using a threshold policy. Show that all the subsequently generated policies will be threshold policies, so that the algorithm will terminate after at most n iterations.
- (b) Prove the result of part (a) for the asset selling example of Vol. I, Section 7.3. Here, a threshold policy is a stationary policy that sells the asset if the offer is higher than a certain fixed number.

1.15 (Convergence of Modified Policy Iteration)

Let $\{J_k\}$ and $\{\mu_k\}$ be the sequences generated by the modified policy iteration algorithm. Show that $\{J_k\}$ converges to J^* . Furthermore, there exists an integer \bar{k} such that for all $k \geq \bar{k}$, μ^k is optimal. *Hint:* Complete the details of the following argument. Let r be a scalar such that the vector \bar{J}_0 , defined by $\bar{J}_0 = J_0 + re$, satisfies $T\bar{J}_0 \leq \bar{J}_0$. [Any scalar r such that $\max_i [(TJ_0)(i) - J_0(i)] \leq (1-\alpha)r$ has this property.] Define for all k , $\bar{J}_{k+1} = T_{\mu_k}^{m_k} \bar{J}_k$. Then, it can be seen by induction that for all k and $m = 0, 1, \dots, m_k$, the vectors $T_{\mu_k}^m J_k$ and $T_{\mu_k}^m \bar{J}_k$ differ by the multiple of the unit vector $r\alpha^{m_0+\dots+m_{k-1}+m}e$. It follows that if J_0 is replaced by \bar{J}_0 as the starting vector in the algorithm, the same sequence of policies $\{\mu_k\}$ will be obtained. Use Prop. 1.3.5.

1.16

Assume that we have two gold mines, Anaconda and Bonanza, and a gold-mining machine. Let x_A and x_B be the current amounts of gold in Anaconda and Bonanza, respectively. When the machine is used in Anaconda (or Bonanza), there is a probability p_A (or p_B , respectively) that $r_A x_A$ (or $r_B x_B$, respectively) of the gold will be mined without damaging the machine, and a probability $1 - p_A$ (or $1 - p_B$, respectively) that the machine will be damaged beyond repair and no gold will be mined. We assume that $0 < r_A < 1$ and $0 < r_B < 1$.

- (a) Assume that $p_A = p_B = p$, where $0 < p < 1$. Find the mine selection policy that maximizes the expected amount of gold mined before the machine breaks down. *Hint:* This problem can be viewed as a discounted multiarmed bandit problem with a discount factor p .
- (b) Assume that $p_A < 1$ and $p_B = 1$. Argue that the optimal expected amount of gold mined has the form $J^*(x_A, x_B) = \tilde{J}_A(x_A) + x_B$, where $\tilde{J}_A(x_A)$ is the optimal expected amount of gold mined if mining is restricted just to Anaconda. Show that there is no policy that attains the optimal amount $J^*(x_A, x_B)$.

1.17 (The Tax Problem [VWB85])

This problem is similar to the multiarmed bandit problem. The only difference is that, if we engage project i at period k , we pay a tax $\alpha^k C^j(x^j)$ for every other project j [for a total of $\alpha^k \sum_{j \neq i} C^j(x^j)$], instead of earning a reward $\alpha^k R^i(x^i)$. The objective is to find a project selection policy that minimizes the total tax paid. Show that the problem can be converted into a bandit problem with reward function for project i equal to

$$R^i(x^i) = C^i(x^i) - \alpha E\{C^i(f^i(x^i, w^i))\}.$$

1.18 (The Restart Problem [KaV87])

The purpose of this exercise is to show that the index of a project in the multi-armed bandit context can be calculated by solving an associated infinite horizon discounted cost problem. In what follows we consider a single project with reward function $R(x)$, a fixed initial state x_0 , and the calculation of the value of index $m(x_0)$ for that state. Consider the problem where at state x_k and time k there are two options: (1) Continue, which brings reward $\alpha^k R(x_k)$ and moves the project to state $x_{k+1} = f(x_k, w)$, or (2) restart the project, which moves the state to x_0 , brings reward $\alpha^k R(x_0)$, and moves the project to state $x_{k+1} = f(x_0, w)$. Show that the optimal reward functions of this problem and of the bandit problem with $M = m(x_0)$ are identical, and therefore the optimal reward for both problems when starting at x_0 equals $m(x_0)$. *Hint:* Show that Bellman's equation for both problems takes the form

$$J(x) = \max[R(x_0) + \alpha E\{J(f(x_0, w))\}, R(x) + \alpha E\{J(f(x, w))\}].$$

1.19 (λ -Policy Iteration [BeI96], [BeT96])

This exercise describes an approach whereby the discount factor is effectively reduced in order to accelerate the policy evaluation step in the policy iteration method. It is based on the notion of *temporal differences* (TD for short), which will be discussed further in the context of the simulation-based approximate dynamic programming methods of Chapter 6.

Consider a policy-iteration like algorithm (called λ -policy iteration) that maintains a cost-policy vector pair (J, μ) . In the typical iteration, given (J, μ) , we obtain the next policy $\bar{\mu}$ by the standard policy improvement step

$$T_{\bar{\mu}} J = TJ. \quad (1.85)$$

To calculate the next cost vector \bar{J} , we define the TD associated with each transition (i, j) under $\bar{\mu}$:

$$d(i, j) = g(i, \bar{\mu}(i), j) + \alpha J(j) - J(i). \quad (1.86)$$

We also consider a parameter λ from the range $[0, 1]$. We view $d(i, j)$ as the one-stage cost of policy μ^{k+1} for an $\alpha\lambda$ -discounted DP problem with the transition probabilities $p_{ij}(\mu^{k+1}(i))$ of the original problem, and we calculate the corresponding cost-to-go vector. This vector, denoted by Δ , has components given by

$$\Delta(i) = \sum_{k=0}^{\infty} E\{(\alpha\lambda)^k d(i_k, i_{k+1}) \mid i_0 = i\}, \quad i = 1, \dots, n. \quad (1.87)$$

The vector \bar{J} is then obtained by

$$\bar{J} = J + \Delta. \quad (1.88)$$

- (a) Show that for $\lambda = 1$, the method is equivalent to ordinary policy iteration, and that for $\lambda = 0$, the method is equivalent to ordinary value iteration.
- (b) Let λ be chosen from $[0, 1)$. Consider a modified policy iteration approach whereby, given J and $\bar{\mu}$, we evaluate $J_{\bar{\mu}}$ approximately using M value iterations starting with J , and let \hat{J}_M be the resulting function; i.e.,

$$\hat{J}_M = \alpha^M J(i_M) + \sum_{k=0}^{M-1} E\left\{ \alpha^k g(i_k, \mu^{k+1}(i_k), i_{k+1}) \mid i_0 = i \right\}.$$

Show that \bar{J} is equal to the *expected* value of \hat{J}_M when the number M of value iterations is randomly chosen from a geometric distribution with mean $E\{M\} = 1/(1 - \lambda)$.

- (c) Assume that $\lambda \in [0, 1)$, and let (J_k, μ^k) be the sequence generated by the λ -policy iteration algorithm. Then J_k converges to J^* and for all k greater than some index \bar{k} , we have

$$\max_{i=1, \dots, n} |J_{k+1}(i) - J^*(i)| \leq \frac{\alpha(1 - \lambda)}{1 - \alpha\lambda} \max_{i=1, \dots, n} |J_k(i) - J^*(i)|.$$

1.20 (Distributed Asynchronous DP [Ber82a], [BeT89])

The value iteration method is well suited for distributed (or parallel) computation since the iteration

$$J(i) := (TJ)(i)$$

can be executed in parallel for all states i . Consider the finite-state discounted problem of Section 1.3, and assume that the above iteration is executed *asynchronously* at a different processor i for each state i . By this we mean that the i th processor holds a vector J^i and updates the i th component of that vector at *arbitrary* times with an iteration of the form

$$J^i(i) := (TJ^i)(i),$$

and at *arbitrary* times transmits the results of the latest computation to other processors m who then update $J^m(i)$ according to

$$J^m(i) := J^i(i).$$

Assume that all processors never stop computing and transmitting the results of their computation to the other processors. Show that the estimates J_t^i of the optimal cost function available at each processor i at time t converge to the optimal solution function J^* as $t \rightarrow \infty$. *Hint:* Let \bar{J} and \underline{J} be two functions such that $\underline{J} \leq T\underline{J}$ and $T\bar{J} \leq \bar{J}$, and suppose that for all initial estimates J_0^i of the processors, we have $\underline{J} \leq J_0^i \leq \bar{J}$. Show that the estimates J_t^i of the processors at time t satisfy $\underline{J} \leq J_t^i \leq \bar{J}$ for all $t \geq 0$, and $T\underline{J} \leq J_t^i \leq T\bar{J}$ for t sufficiently large.

Stochastic Shortest Path Problems

Contents

2.1. Problem Formulation	p. 94
2.2. Bellman's Equation	p. 97
2.3. Value Iteration	p. 105
2.4. Policy Iteration	p. 108
2.5. Countable State Problems	p. 112
2.6. Notes, Sources, and Exercises	p. 114

In this chapter we consider a stochastic version of the shortest path problem. An introductory analysis of this problem was given in Section 7.2 of Vol. I. The analysis of this chapter is more sophisticated and uses weaker assumptions, which are patterned after the ones for the deterministic shortest path problem, given in Chapter 2 of Vol. I.

2.1 PROBLEM FORMULATION

For the purpose of orientation into the problem of this chapter, let us take the following view of the deterministic shortest path problem of Chapter 2 of Vol. I: We have a graph with nodes $1, 2, \dots, n, t$, where t is a special state called the *destination* or the *termination state*, and we want to choose for each node $i \neq t$, a successor node $\mu(i)$ so that $(i, \mu(i))$ is an arc, and the path formed by a sequence of successor nodes starting at any node j terminates at t and has minimum sum of arc lengths over all paths that start at j and terminate at t .

The stochastic shortest path problem (SSP problem for short) is a generalization whereby at each node i , we must select a probability distribution over all possible successor nodes j out of a given set of probability distributions $p_{ij}(u)$ parameterized by a control $u \in U(i)$. For a given selection of distributions and for a given origin node, the path traversed as well as its length are now random, but we wish that the path leads to the destination t with probability 1 and has minimum expected length. Note that if every feasible probability distribution assigns probability 1 to a single successor node, we obtain the deterministic shortest path problem.

We formulate this problem as the special case of the total cost infinite horizon problem where:

- (a) There is no discounting ($\alpha = 1$).
- (b) The state space is $S = \{1, 2, \dots, n, t\}$ with transition probabilities denoted by

$$p_{ij}(u) = P(x_{k+1} = j \mid x_k = i, u_k = u), \quad i, j \in S, u \in U(i).$$

Furthermore, the destination t is absorbing, i.e., for all $u \in U(t)$,

$$p_{tt}(u) = 1.$$

- (c) The control constraint set $U(i)$ is a finite set for all i .
- (d) A cost $g(i, u)$ is incurred when control $u \in U(i)$ is selected. Furthermore, the destination is *cost-free*, i.e., $g(t, u) = 0$ for all $u \in U(t)$.

Note that as in Section 1.3, we assume that the cost per stage does not depend on the successor state. This amounts to using expected cost

per stage in all calculations. In particular, if the cost of applying control u at state i and moving to state j is $\tilde{g}(i, u, j)$, we use as cost per stage the expected cost

$$g(i, u) = \sum_{j=1, \dots, n, t} p_{ij}(u) \tilde{g}(i, u, j).$$

Since the destination t is cost-free and absorbing, the cost starting from t is zero for every policy. Accordingly, for all cost functions, we ignore the component that corresponds to t , and define the mappings T and T_μ on functions J with components $J(1), \dots, J(n)$ by

$$(TJ)(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u) J(j) \right], \quad i = 1, \dots, n,$$

$$(T_\mu J)(i) = g(i, \mu(i)) + \sum_{j=1}^n p_{ij}(\mu(i)) J(j), \quad i = 1, \dots, n.$$

These mappings parallel the ones introduced in Section 1.3 for the discounted problem. The difference is that here we have no discounting ($\alpha = 1$), but for the states i and controls u for which $p_{it}(u) > 0$, we have

$$\sum_{j=1}^n p_{ij}(u) = 1 - p_{it}(u) < 1.$$

As in Section 1.3, for any stationary policy μ , we use the compact notation

$$P_\mu = \begin{bmatrix} p_{11}(\mu(1)) & \cdots & p_{1n}(\mu(1)) \\ \vdots & \ddots & \vdots \\ p_{n1}(\mu(n)) & \cdots & p_{nn}(\mu(n)) \end{bmatrix},$$

and

$$g_\mu = \begin{bmatrix} g(1, \mu(1)) \\ \vdots \\ g(n, \mu(n)) \end{bmatrix}.$$

We can then write in vector notation

$$T_\mu J = g_\mu + P_\mu J.$$

In terms of this notation, the cost function of a policy $\pi = \{\mu_0, \mu_1, \dots\}$ can be written as

$$J_\pi = \limsup_{N \rightarrow \infty} T_{\mu_0} \cdots T_{\mu_{N-1}} J_0 = \limsup_{N \rightarrow \infty} \left(g_{\mu_0} + \sum_{k=1}^{N-1} P_{\mu_0} \cdots P_{\mu_{k-1}} g_{\mu_k} \right),$$

where J_0 denotes the zero vector. The cost function of a stationary policy μ can be written as

$$J_\mu = \limsup_{N \rightarrow \infty} T_\mu^N J_0 = \limsup_{N \rightarrow \infty} \sum_{k=0}^{N-1} P_\mu^k g_\mu.$$

The SSP problem was discussed in Section 7.2 of Vol. I, under the assumption that all policies lead to the destination with probability 1, regardless of the initial state. In order to analyze the problem under weaker conditions, we introduce the notion of a proper policy.

Definition 2.1.1: A stationary policy μ is said to be *proper* if, when using this policy, there is positive probability that the destination will be reached after at most n stages, regardless of the initial state; i.e., if

$$\rho_\mu = \max_{i=1,\dots,n} P\{x_n \neq t \mid x_0 = i, \mu\} < 1. \quad (2.1)$$

A stationary policy that is not proper is said to be *improper*.

With a little thought, it can be seen that μ is proper if and only if in the Markov chain corresponding to μ , each state i is connected to the destination with a path of positive probability transitions. Note from the definition (2.1) that

$$\begin{aligned} P\{x_{2n} \neq t \mid x_0 = i, \mu\} &= P\{x_{2n} \neq t \mid x_n \neq t, x_0 = i, \mu\} \\ &\quad \times P\{x_n \neq t \mid x_0 = i, \mu\} \\ &\leq \rho_\mu^2. \end{aligned}$$

More generally, by repeating the preceding argument, we see that for a proper policy μ , the probability of not reaching the destination after k stages satisfies

$$P\{x_k \neq t \mid x_0 = i, \mu\} \leq \rho_\mu^{\lfloor k/n \rfloor}, \quad i = 1, \dots, n. \quad (2.2)$$

Thus the destination will eventually be reached with probability 1 under a proper policy. Furthermore, the limit defining the associated total cost vector J_μ will exist and be finite, since the expected cost incurred in the k th period is bounded in absolute value by

$$\rho_\mu^{\lfloor k/n \rfloor} \max_{i=1,\dots,n} |g(i, \mu(i))|,$$

so that

$$|J_\mu(i)| \leq \lim_{N \rightarrow \infty} \sum_{k=0}^{N-1} \rho_\mu^{\lfloor k/n \rfloor} \max_{i=1,\dots,n} |g(i, \mu(i))| < \infty.$$

Note that under a proper policy, the cost structure is similar to the one for discounted problems, the main difference being that the effective discount factor depends on the current state and stage, but builds up to at least ρ_μ per n stages.

With the exception of Section 2.5, where we will deal with the case of a countably infinite state space, we assume the following:

Assumption 2.1.1: There exists at least one proper policy.

Assumption 2.1.2: For every improper policy μ , the corresponding cost $J_\mu(i)$ is ∞ for at least one state i ; i.e., some component of the sum $\sum_{k=0}^{N-1} P_\mu^k g_\mu$ diverges to ∞ as $N \rightarrow \infty$.

In the case of a deterministic shortest path problem, Assumption 2.1.1 is satisfied if and only if every node is connected to the destination with a path, while Assumption 2.1.2 is satisfied if and only if each cycle that does not contain the destination has positive length. A simple condition that implies Assumption 2.1.2 is that the cost $g(i, u)$ is strictly positive for all $i \neq t$ and $u \in U(i)$. Another important case where Assumptions 2.1.1 and 2.1.2 are satisfied is when *all* policies are proper, i.e., when termination is inevitable under all stationary policies (this was assumed in Section 7.2 of Vol. I).

2.2 BELLMAN'S EQUATION

We will now develop the main analytical results for SSP problems under Assumptions 2.1.1 and 2.1.2. These results are almost as strong as those for discounted problems with bounded cost per stage. In particular, we show that:

- (a) The optimal cost vector is the unique solution of Bellman's equation $J^* = TJ^*$.
- (b) The value iteration method converges to the optimal cost vector J^* for an arbitrary starting vector.
- (c) A stationary policy μ is optimal if and only if $T_\mu J^* = TJ^*$.

- (d) The policy iteration algorithm yields an optimal proper policy starting from an arbitrary proper policy.

The following proposition provides some basic preliminary results:

Proposition 2.2.1:

- (a) For a proper policy μ , the associated cost vector J_μ satisfies

$$\lim_{k \rightarrow \infty} (T_\mu^k J)(i) = J_\mu(i), \quad i = 1, \dots, n,$$

for every vector J . Furthermore,

$$J_\mu = T_\mu J_\mu,$$

and J_μ is the unique solution of this equation.

- (b) A stationary policy μ satisfying for some vector J ,

$$J(i) \geq (T_\mu J)(i), \quad i = 1, \dots, n,$$

is proper.

Proof: (a) Using an induction argument, we have for all $J \in \Re^n$ and $k \geq 1$

$$T_\mu^k J = P_\mu^k J + \sum_{m=0}^{k-1} P_\mu^m g_\mu. \quad (2.3)$$

Equation (2.2) implies that for all $J \in \Re^n$, we have

$$\lim_{k \rightarrow \infty} P_\mu^k J = 0,$$

so that

$$\lim_{k \rightarrow \infty} T_\mu^k J = \lim_{k \rightarrow \infty} \sum_{m=0}^{k-1} P_\mu^m g_\mu = J_\mu,$$

where the limit above can be shown to exist using Eq. (2.2).

Also we have by definition

$$T_\mu^{k+1} J = g_\mu + P_\mu T_\mu^k J,$$

and by taking the limit as $k \rightarrow \infty$, we obtain

$$J_\mu = g_\mu + P_\mu J_\mu,$$

which is equivalent to $J_\mu = T_\mu J_\mu$.

Finally, to show uniqueness, note that if $J = T_\mu J$, then we have $J = T_\mu^k J$ for all k , so that $J = \lim_{k \rightarrow \infty} T_\mu^k J = J_\mu$.

(b) The hypothesis $J \geq T_\mu J$, the monotonicity of T_μ , and Eq. (2.3) imply that

$$J \geq T_\mu^k J = P_\mu^k J + \sum_{m=0}^{k-1} P_\mu^m g_\mu, \quad k = 1, 2, \dots$$

If μ were not proper, by Assumption 2.1.2, some component of the sum in the right-hand side of the above relation would diverge to ∞ as $k \rightarrow \infty$, which is a contradiction. **Q.E.D.**

The following proposition is the main result of this section, and provides analogs to the main results for discounted cost problems (Props. 1.2.1-1.2.3).

Proposition 2.2.2:

(a) The optimal cost vector J^* satisfies Bellman's equation

$$J^* = TJ^*.$$

Furthermore, J^* is the unique solution of this equation.

(b) We have

$$\lim_{k \rightarrow \infty} (T^k J)(i) = J^*(i), \quad i = 1, \dots, n,$$

for every vector J .

(c) A stationary policy μ is optimal if and only if

$$T_\mu J^* = TJ^*.$$

Proof: (a), (b) We first show that T has at most one fixed point. Indeed, if J and J' are two fixed points, then we select μ and μ' such that $J = TJ = T_\mu J$ and $J' = TJ' = T_{\mu'} J'$; this is possible because the control constraint set is finite. By Prop. 2.2.1(b), we have that μ and μ' are proper, and Prop. 2.2.1(a) implies that $J = J_\mu$ and $J' = J_{\mu'}$. We have $J = T^k J \leq T_{\mu'}^k J$ for all $k \geq 1$, and by Prop. 2.2.1(a), we obtain $J \leq \lim_{k \rightarrow \infty} T_{\mu'}^k J = J_{\mu'} = J'$. Similarly, $J' \leq J$, showing that $J = J'$ and that T has at most one fixed point.

We next show that T has at least one fixed point. Let μ be a proper policy (there exists one by Assumption 2.1.1). Choose μ' such that

$$T_{\mu'} J_\mu = T J_\mu.$$

Then we have $J_\mu = T_\mu J_\mu \geq T_{\mu'} J_\mu$. By Prop. 2.2.1(b), μ' is proper, and using the monotonicity of $T_{\mu'}$ and Prop. 2.2.1(a), we obtain

$$J_\mu \geq \lim_{k \rightarrow \infty} T_{\mu'}^k J_\mu = J'_\mu. \quad (2.4)$$

Continuing in the same manner, we construct a sequence $\{\mu^k\}$ such that each μ^k is proper and

$$J_{\mu^k} \geq T J_{\mu^k} \geq J_{\mu^{k+1}}, \quad k = 0, 1, \dots \quad (2.5)$$

Since the set of proper policies is finite, some policy μ must be repeated within the sequence $\{\mu^k\}$, and by Eq. (2.5), we have

$$J_\mu = T J_\mu.$$

Thus J_μ is a fixed point of T , and in view of the uniqueness property shown earlier, J_μ is the unique fixed point of T .

Next we show that the unique fixed point of T is equal to the optimal cost vector J^* , and that $T^k J \rightarrow J^*$ for all J . The construction of the preceding paragraph provides a proper μ such that $T J_\mu = J_\mu$. We will show that $T^k J \rightarrow J_\mu$ for all J and that $J_\mu = J^*$. Let $e = (1, 1, \dots, 1)$, let $\delta > 0$ be some scalar, and let \hat{J} be the vector satisfying

$$T_\mu \hat{J} = \hat{J} - \delta e.$$

There is a unique such vector because the equation $\hat{J} = T_\mu \hat{J} + \delta e$ can be written as $\hat{J} = g_\mu + \delta e + P_\mu \hat{J}$, so \hat{J} is the cost vector corresponding to μ for g_μ replaced by $g_\mu + \delta e$. Since μ is proper, by Prop. 2.2.1(a), \hat{J} is unique. Furthermore, we have $J_\mu \leq \hat{J}$, which implies that

$$J_\mu = T J_\mu \leq T \hat{J} \leq T_\mu \hat{J} = \hat{J} - \delta e \leq \hat{J}.$$

Using the monotonicity of T and the preceding relation, we obtain

$$J_\mu = T^k J_\mu \leq T^k \hat{J} \leq T^{k-1} \hat{J} \leq \hat{J}, \quad k \geq 1.$$

Hence, $T^k \hat{J}$ converges to some vector \tilde{J} , and we have

$$T \tilde{J} = T \left(\lim_{k \rightarrow \infty} T^k \hat{J} \right).$$

The mapping T can be seen to be continuous, so we can interchange T with the limit in the preceding relation, thereby obtaining $\tilde{J} = T\tilde{J}$. By the uniqueness of the fixed point of T shown earlier, we must have $\tilde{J} = J_\mu$. It is also seen that

$$J_\mu - \delta e = TJ_\mu - \delta e \leq T(J_\mu - \delta e) \leq TJ_\mu = J_\mu.$$

Thus, $T^k(J_\mu - \delta e)$ is monotonically increasing and bounded above. As earlier, it follows that $\lim_{k \rightarrow \infty} T^k(J_\mu - \delta e) = J_\mu$. For any J , we can find $\delta > 0$ such that

$$J_\mu - \delta e \leq J \leq \hat{J}.$$

By the monotonicity of T , we then have

$$T^k(J_\mu - \delta e) \leq T^k J \leq T^k \hat{J}, \quad k \geq 1,$$

and since $\lim_{k \rightarrow \infty} T^k(J_\mu - \delta e) = \lim_{k \rightarrow \infty} T^k \hat{J} = J_\mu$, it follows that

$$\lim_{k \rightarrow \infty} T^k J = J_\mu.$$

To show that $J_\mu = J^*$, take any policy $\pi = \{\mu_0, \mu_1, \dots\}$. We have

$$T_{\mu_0} \cdots T_{\mu_{k-1}} J_0 \geq T^k J_0,$$

where J_0 is the zero vector. Taking the \limsup of both sides as $k \rightarrow \infty$ in the preceding inequality, we obtain

$$J_\pi \geq J_\mu,$$

so μ is an optimal stationary policy and $J_\mu = J^*$.

(c) If μ is optimal, then $J_\mu = J^*$ and, by Assumptions 2.1.1 and 2.1.2, μ is proper, so by Prop. 2.2.1(a),

$$T_\mu J^* = T_\mu J_\mu = J_\mu = J^* = TJ^*.$$

Conversely, if $J^* = TJ^* = T_\mu J^*$, it follows from Prop. 2.2.1(b) that μ is proper, and by using Prop. 2.2.1(a), we obtain $J^* = J_\mu$. Therefore, μ is optimal. Q.E.D.

The results of Prop. 2.2.2 can also be proved (with minor changes) assuming, in place of Assumption 2.1.2, that $g(i, u) \geq 0$ for all i and $u \in U(i)$, and that there exists an optimal proper policy; see Exercise 2.12.

Underlying Contractions

We mentioned in Section 1.4 that the strong results we derived for discounted problems in Chapter 1 owe their validity to the contraction property of the mapping T . Despite the similarity of Prop. 2.2.2 with the corresponding discounted cost results of Section 1.2, the mapping T of this section need not be a contraction mapping with respect to any norm; see Exercise 2.13 for a counterexample. On the other hand, if we strengthen Assumptions 2.1.1 and 2.1.2, to assume that all stationary policies are proper, it turns out that T is a contraction mapping with respect to a *weighted sup-norm*. This is the subject of the following proposition.

Proposition 2.2.3: Assume that all stationary policies are proper. Then, there exists a vector v with positive components such that T and T_μ , for all stationary policies μ , are contraction mappings with respect to the weighted sup-norm

$$\|J\|_v = \max_{i=1,\dots,n} \frac{J(i)}{v_i}.$$

Proof: We first define the vector v as the solution of a certain DP problem, and then show that it has the required property. Consider a new SSP problem where the transition probabilities are the same as in the original, but the transition costs are all equal to -1 (except at the termination state t , where the self-transition cost is 0). Let $\hat{J}(i)$ be the optimal cost-to-go from state i in this new problem. By Prop. 2.2.2, we have for all $i = 1, \dots, n$, and stationary policies μ ,

$$\hat{J}(i) = -1 + \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u) \hat{J}(j) \leq -1 + \sum_{j=1}^n p_{ij}(\mu(i)) \hat{J}(j). \quad (2.6)$$

Define

$$v_i = -\hat{J}(i), \quad i = 1, \dots, n.$$

Then for all i , we have $v_i \geq 1$, and for all stationary policies μ , we have from Eq. (2.6),

$$\sum_{j=1}^n p_{ij}(\mu(i)) v_j \leq v_i - 1 \leq \rho v_i, \quad i = 1, \dots, n, \quad (2.7)$$

where ρ is defined by

$$\rho = \max_{i=1,\dots,n} \frac{v_i - 1}{v_i} < 1.$$

The contraction property of T_μ now follows from Eq. (2.7) and Prop. 1.4.1(a). The contraction property of T follows from Prop. 1.4.1(b) and the fact $TJ = \min_\mu T_\mu J$. Q.E.D.

Using the preceding proposition and under the (somewhat restrictive) assumption that all stationary policies are proper, one can obtain the most powerful analytical and algorithmic results for SSP problems: essentially all the results derived in Chapter 1 for discounted problems with bounded cost per state. These results were given in Section 7.2 of Vol. I, and it is now clear that their analysis draws their validity from the contraction mapping structure established in Prop. 2.2.3.

We note a generalization of the contraction property of Prop. 2.2.3 for SSP problems with a termination state t and a countable number of nontermination states, denoted $1, 2, \dots$. Let v_i be the maximum (over all policies) expected number of stages up to termination, starting from state i . Then if v_i is finite and bounded over i , the mappings T and T_μ are contraction mappings with respect to the weighted sup-norm $\|\cdot\|_v$, with modulus of contraction

$$\rho = \max_{i=1,2,\dots} \frac{v_i - 1}{v_i}.$$

The proof is similar, but somewhat more sophisticated than the proof of Prop. 2.2.3, and is outlined in Exercise 2.15.

Compact Control Constraint Sets

It turns out that the finiteness assumption on the control constraint $U(i)$ can be weakened. It is sufficient that, for each i , $U(i)$ be a compact subset of a Euclidean space, and that $p_{ij}(u)$ and $g(i, u)$ be continuous in u over $U(i)$, for all i and j . Under these compactness and continuity assumptions, and also Assumptions 2.1.1 and 2.1.2, Prop. 2.2.2 holds as stated. The proof is similar to the one given above, but is technically much more complex. It can be found in Bertsekas and Tsitsiklis [BeT91b].

Pathologies of Stochastic Shortest Path Problems

We now give two examples that illustrate the sensitivity of our results to seemingly minor changes in our assumptions.

Example 2.2.1 (The Blackmailer's Dilemma [Whi82])

This example shows that the assumption of a finite or compact control constraint set cannot be easily relaxed. Here, there are two states, state 1 and the destination state t . At state 1, we can choose a control u with $0 < u \leq 1$, while incurring a cost $-u$; we then move to state t with probability u^2 , and stay in state 1 with probability $1 - u^2$. Note that every stationary policy is proper in the sense that it leads to the destination with probability 1.

We may regard u as a demand made by a blackmailer, and state 1 as the situation where the victim complies. State t is the situation where the victim refuses to yield to the blackmailer's demand. The problem then can be seen as one whereby the blackmailer tries to maximize his total gain by balancing his desire for increased demands with keeping his victim compliant.

If controls were chosen from a *finite* subset of the interval $(0, 1]$, the problem would come under the framework of this section. The optimal cost would then be finite, and there would exist an optimal stationary policy. It turns out, however, that *without the finiteness restriction the optimal cost starting at state 1 is $-\infty$ and there exists no optimal stationary policy*. Indeed, for any stationary policy μ with $\mu(1) = u$, we have

$$J_\mu(1) = -u + (1 - u^2)J_\mu(1)$$

from which

$$J_\mu(1) = -\frac{1}{u}.$$

Since u can be taken arbitrarily close to 0, it follows that $J^*(1) = -\infty$, but there is no stationary policy that achieves the optimal cost. Note also that this situation would not change if the constraint set were $u \in [0, 1]$ (i.e., $u = 0$ were an allowable control), although in this case the stationary policy that applies $\mu(1) = 0$ is improper and its corresponding cost vector is zero, thus violating Assumption 2.1.2. Furthermore, it can be shown that Bellman's equation, which is

$$J^*(1) = (TJ^*)(1) = \min_{u \in (0, 1]} [-u + (1 - u^2)J^*(1)],$$

has no (real number) solution. Indeed, the equation cannot have a solution with $J^*(1) \geq 0$, since then $u^* = 1$ attains the minimum leading to a contradiction, and it cannot have a solution with $J^*(1) < 0$, since then the minimizing value of u is

$$u^* = \min \left[1, -\frac{1}{2J^*(1)} \right],$$

and by substitution, we have

$$J^*(1) = (TJ^*)(1) = \begin{cases} -1 & \text{if } J^*(1) \geq -1/2, \\ J^*(1) + \frac{1}{4J^*(1)} & \text{if } J^*(1) \leq -1/2, \end{cases}$$

a contradiction.[†]

Another interesting fact about this problem is that there is an optimal *nonstationary* policy π . This is the policy $\pi = \{\mu_0, \mu_1, \dots\}$ that applies $\mu_k(1) = \gamma/(k+1)$ at time k and state 1, where γ is a scalar in the interval

[†] The lack of existence of a real-number solution to Bellman's equation also follows from general results to be given in Chapter 3 on total cost problems with nonpositive stage costs [see Prop. 3.1.2(b)], which implies that if $J^*(i) = -\infty$ for some i , then Bellman's equation has no solution in real numbers; it has a solution but for some states, the solution will be $-\infty$.

(0, 1/2). We leave it for the reader to verify that $J_\pi(1) = -\infty$. What happens with the policy π is that the blackmailer requests diminishing amounts over time, which nonetheless add to ∞ . However, the probability of the victim's refusal diminishes at a much faster rate over time, and as a result, the probability of the victim remaining compliant forever is strictly positive, leading to an infinite total expected payoff to the blackmailer.

Example 2.2.2 (Pure Stopping Problems)

This example illustrates why we need to assume that all improper policies have infinite cost for at least some initial state (Assumption 2.1.2). Consider an optimal stopping problem where a state-dependent cost is incurred only when invoking a stopping action that drives the system to the destination; all costs are zero prior to stopping. Eventual stopping is a requirement here, so to properly formulate such a stopping problem as a total cost infinite horizon problem, it is essential to make the stopping costs negative (by adding a negative constant to all stopping costs if necessary), providing an incentive to stop. We then come under the framework of this section but with Assumption 2.1.2 violated because the improper policy that never stops does not yield infinite cost for any starting state. Unfortunately, this seemingly small relaxation of our assumptions invalidates our results as shown by the example of Fig. 2.2.1, where T has multiple fixed points. This example is in effect a deterministic shortest path problem involving a cycle with zero length, and there is a (nonoptimal) improper policy that yields finite cost for all initial states (rather than infinite cost for some initial state).

2.3 VALUE ITERATION

All the methods developed in connection with the discounted cost problem in Section 1.3, have SSP analogs. In this section and the next we will focus on value and policy iteration, respectively, (the linear programming approach of Section 1.3.4 has a straightforward extension, since by using Prop. 2.2.2, it can be seen that J^* is the largest solution of the system of inequalities $J \leq TJ$). We will discuss in some detail certain SSP problems with special structure. It turns out that by exploiting this structure, we can improve the convergence properties of some of the methods. For example, in deterministic shortest path problems, value iteration terminates finitely (Section 2.1 of Vol. I), whereas this does not happen for any significant class of discounted cost problems.

As shown by Prop. 2.2.2(b), value iteration converges to the optimal cost function for SSP problems. Furthermore, several of the enhancements and variations of value iteration for discounted problems have SSP analogs. In particular, there are error bounds similar to the ones of Prop. 1.3.1 (although not quite as powerful; see Section 7.2 of Vol. I). It can also

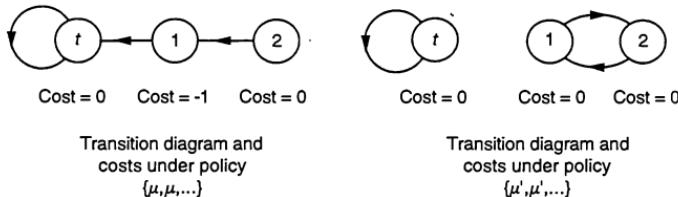


Figure 2.2.1 Example where Prop. 2.2.2 fails to hold when Assumption 2.1.2 is violated. There are two stationary policies, μ and μ' , with transition probabilities and costs as shown. The equation $J = TJ$ is given by

$$J(1) = \min\{-1, J(2)\},$$

$$J(2) = J(1),$$

and is satisfied by any J of the form

$$J(1) = \delta, \quad J(2) = \delta,$$

with $\delta \leq -1$. Here the proper policy μ is optimal and the corresponding optimal cost vector is

$$J(1) = -1, \quad J(2) = -1.$$

The difficulty is that the improper policy μ' has finite (zero) cost for all initial states.

be shown that the Gauss-Seidel version of the method works and that its rate of convergence is typically faster than that of the ordinary method (Exercise 2.6). Furthermore, the rank-one correction method described in Section 1.3.1 is straightforward and effective, as long as there is some separation between the dominant and the subdominant eigenvalue moduli.

Finite Termination of Value Iteration

Generally, the value iteration method requires an infinite number of iterations in SSP problems. However, under special circumstances, the method can terminate finitely. A prominent example is the case of a deterministic shortest path problem, but there are other more general circumstances where termination occurs. In particular, let us assume that the transition probability graph corresponding to some optimal stationary policy μ^* is acyclic. By this we mean that there are no cycles in the graph that has as nodes the states $1, \dots, n, t$, and has an arc (i, j) for each pair of states i and j such that $p_{ij}(\mu^*(i)) > 0$. Implicit in this assumption is that there are no positive self-transition probabilities $p_{ii}(\mu^*(i))$ for $i \neq t$, but it turns out that under Assumptions 2.1.1 and 2.1.2, a SSP problem with such self-transitions can be converted into another SSP problem where $p_{ii}(u) = 0$.

for all $i \neq t$ and $u \in U(i)$. In particular, it can be shown (Exercise 2.8) that the modified SSP problem that has costs

$$\tilde{g}(i, u) = \frac{g(i, u)}{1 - p_{ii}(u)}, \quad i = 1, \dots, n,$$

in place of $g(i, u)$, and transition probabilities

$$\tilde{p}_{ij}(u) = \begin{cases} 0 & \text{if } j = i, \\ \frac{p_{ij}(u)}{1 - p_{ii}(u)} & \text{if } j \neq i, \end{cases} \quad i = 1, \dots, n,$$

instead of $p_{ij}(u)$ is equivalent to the original in the sense that it has the same optimal costs and policies.

We claim that, under the preceding acyclicity assumption, the value iteration method will yield J^* after at most n iterations when started from the vector J given by

$$J(i) = \infty, \quad i = 1, \dots, n. \quad (2.8)$$

To show this, consider the sets of states S_0, S_1, \dots , defined by

$$S_0 = \{t\},$$

$$S_{k+1} = \{i \mid p_{ij}(\mu^*(i)) = 0 \text{ for all } j \notin \cup_{m=0}^k S_m\}, \quad k = 0, 1, \dots,$$

and let $S_{\bar{k}}$ be the last of these sets that is nonempty. Then in view of our acyclicity assumption, we have

$$\cup_{m=0}^{\bar{k}} S_m = \{1, \dots, n, t\}.$$

Let us show by induction that, starting from the vector J of Eq. (2.8), the value iteration method will yield for $k = 0, 1, \dots, \bar{k}$,

$$(T^k J)(i) = J^*(i), \quad \text{for all } i \in \cup_{m=0}^k S_m, i \neq t.$$

Indeed, this is so for $k = 0$. Assume that $(T^k J)(i) = J^*(i)$ if $i \in \cup_{m=0}^k S_m$. Then, by the monotonicity of T , we have for all i ,

$$J^*(i) \leq (T^{k+1} J)(i),$$

while we have by the induction hypothesis, the definition of the sets S_k , and the optimality of μ^* ,

$$\begin{aligned} (T^{k+1} J)(i) &\leq g(i, \mu^*(i)) + \sum_{j \in \cup_{m=0}^k S_m} p_{ij}(\mu^*(i)) J^*(j) \\ &= J^*(i), \quad \text{for all } i \in \cup_{m=0}^{k+1} S_m, i \neq t. \end{aligned}$$

The last two relations complete the induction.

Thus, we have shown that if there is some optimal stationary policy μ^* with an associated transition probability graph that is acyclic, at the k th iteration, the value iteration method, will find the optimal costs of the states in the set S_k . In particular, all optimal costs will be found after \bar{k} iterations.

Consistently Improving Policies

The properties of value iteration can be further improved if there is an optimal policy μ^* under which from a given state, we can only go to a state of lower cost; i.e., for all i , we have

$$p_{ij}(\mu^*(i)) > 0 \quad \Rightarrow \quad J^*(i) > J^*(j).$$

We call such a policy *consistently improving*.

A case where a consistently improving policy exists arises in deterministic shortest path problems when all the arc lengths are positive. Another important case arises in continuous-space shortest path problems; see Tsitsiklis [Tsi95] and Exercise 2.10.

The transition probability graph corresponding to a consistently improving policy is seen to be acyclic, so when such a policy exists, by the preceding discussion, the value iteration method terminates finitely. However, a stronger property can be proved. As discussed in Chapter 2 of Vol. I, for shortest path problems with positive arc lengths, one can use Dijkstra's algorithm. This is the label correcting method, which removes from the OPEN list a node with minimum label at each iteration and requires just one iteration per node. A similar property holds for SSP problems if there is a consistently improving policy: if one removes from the OPEN list a state j with minimum cost estimate $J(j)$, the Gauss-Seidel version of the value iteration method requires just one iteration per state; see Exercise 2.11.

For problems where a consistently improving policy exists, it is also appropriate to use straightforward adaptations of the label correcting shortest path methods discussed in Section 2.3.1 of Vol. I. In particular, one may approximate the policy of removing from the OPEN list a minimum cost state by using the SLF and LLL strategies (see Polymenakos, Bertsekas, and Tsitsiklis [PBT98]).

2.4 POLICY ITERATION

The policy iteration algorithm is based on the construction used in the proof of Prop. 2.2.2 to show that T has a fixed point. In the typical iteration, given a proper policy μ and the corresponding cost vector J_μ , one obtains a new proper policy $\bar{\mu}$ satisfying $T_{\bar{\mu}}J_\mu = TJ_\mu$. It was shown in Eq. (2.4) that $J_{\bar{\mu}} \leq J_\mu$. It can be seen also that strict inequality $J_{\bar{\mu}}(i) < J_\mu(i)$ holds for at least one state i , if μ is nonoptimal; otherwise we would have $J_\mu = TJ_\mu$ and by Prop. 2.2.2(c), μ would be optimal. Therefore, the new policy is strictly better if the current policy is nonoptimal. Since the number of proper policies is finite, the policy iteration algorithm terminates after a finite number of iterations with an optimal proper policy.

It is possible to execute approximately the policy evaluation step of policy iteration, using a finite number of value iterations, as in the discounted case. Here we start with some vector J_0 . For all k , a stationary policy μ^k is defined from J_k according to $T_{\mu^k} J_k = TJ_k$, the cost J_{μ^k} is approximately evaluated by $m_k - 1$ additional value iterations, yielding the vector J_{k+1} , which is used in turn to define μ^{k+1} . The proof of Prop. 1.3.5 can be essentially repeated to show that $J_k \rightarrow J^*$, assuming that the initial vector J_0 satisfies $TJ_0 \leq J_0$. This assumption is essential, unless all stationary policies are proper, in which case T is a contraction mapping (cf. Prop. 2.2.3).

Approximate Policy Iteration

Let us consider an approximate policy iteration algorithm that generates a sequence of stationary policies $\{\mu^k\}$ and a corresponding sequence of approximate cost vectors $\{J_k\}$ satisfying

$$\max_{i=1,\dots,n} |J_k(i) - J_{\mu^k}(i)| \leq \delta, \quad k = 0, 1, \dots \quad (2.9)$$

and

$$\max_{i=1,\dots,n} |(T_{\mu^{k+1}} J_k)(i) - (TJ_k)(i)| \leq \epsilon, \quad k = 0, 1, \dots \quad (2.10)$$

where δ and ϵ are some positive scalars, and μ^0 is some proper policy. One difficulty with such an algorithm is that, even if the current policy μ^k is proper, the next policy μ^{k+1} may not be proper. In this case, we have $J_{\mu^{k+1}}(i) = \infty$ for some i , and the method breaks down. Note, however, that for a sufficiently small ϵ , Eq. (2.10) implies that $T_{\mu^{k+1}} J_k = TJ_k$, so by Prop. 2.2.1(b), μ^{k+1} will be proper. In any case, we will analyze the method under the assumption that all generated policies are proper. The following proposition parallels Prop. 1.3.6. It provides an estimate of the difference $J_{\mu^k} - J^*$ in terms of the scalar

$$\rho = \max_{\substack{i=1,\dots,n \\ \mu: \text{proper}}} P\{x_n \neq t \mid x_0 = i, \mu\}.$$

Note that for every proper policy μ and state i , we have $P\{x_n \neq t \mid x_0 = i, \mu\} < 1$ by the definition of a proper policy, and since the number of proper policies is finite, we have $\rho < 1$.

Proposition 2.4.1: Assume that the stationary policies μ^k generated by the approximate policy iteration algorithm are all proper. Then

$$\limsup_{k \rightarrow \infty} \max_{i=1,\dots,n} (J_{\mu^k}(i) - J^*(i)) \leq \frac{n(1 - \rho + n)(\epsilon + 2\delta)}{(1 - \rho)^2}. \quad (2.11)$$

Proof: The proof is similar to the one of Prop. 1.3.6. We modify the arguments in order to use the relations $T_\mu(J+re) \leq T_\mu J + re$ and $P_\mu^n e \leq \rho e$, which hold for all proper policies μ and positive scalars r . We use Eqs. (2.9) and (2.10) to obtain for all k

$$T_{\mu^{k+1}} J_{\mu^k} \leq T J_{\mu^k} + (\epsilon + 2\delta)e \leq T_{\mu^k} J_{\mu^k} + (\epsilon + 2\delta)e. \quad (2.12)$$

From Eq. (2.12) and the equation $T_{\mu^k} J_{\mu^k} = J_{\mu^k}$, we have

$$T_{\mu^{k+1}} J_{\mu^k} \leq J_{\mu^k} + (\epsilon + 2\delta)e.$$

By subtracting from this relation the equation $T_{\mu^{k+1}} J_{\mu^{k+1}} = J_{\mu^{k+1}}$, we obtain

$$T_{\mu^{k+1}} J_{\mu^k} - T_{\mu^{k+1}} J_{\mu^{k+1}} \leq J_{\mu^k} - J_{\mu^{k+1}} + (\epsilon + 2\delta)e.$$

This relation can be written as

$$J_{\mu^{k+1}} - J_{\mu^k} \leq P_{\mu^{k+1}}(J_{\mu^{k+1}} - J_{\mu^k}) + (\epsilon + 2\delta)e, \quad (2.13)$$

where $P_{\mu^{k+1}}$ is the transition probability matrix corresponding to μ^{k+1} . Let

$$\xi_k = \max \left[0, \max_{i=1, \dots, n} (J_{\mu^{k+1}}(i) - J_{\mu^k}(i)) \right].$$

Then Eq. (2.13) yields

$$\xi_k e \leq \xi_k P_{\mu^{k+1}} e + (\epsilon + 2\delta)e.$$

By multiplying this relation with $P_{\mu^{k+1}}$ and by adding $(\epsilon + 2\delta)e$, we obtain

$$\xi_k e \leq \xi_k P_{\mu^{k+1}} e + (\epsilon + 2\delta)e \leq \xi_k P_{\mu^{k+1}}^2 e + 2(\epsilon + 2\delta)e.$$

By repeating this process for a total of $n - 1$ times, we have

$$\xi_k e \leq \xi_k P_{\mu^{k+1}}^n e + n(\epsilon + 2\delta)e \leq \rho \xi_k e + n(\epsilon + 2\delta)e.$$

Thus,

$$\xi_k \leq \frac{n(\epsilon + 2\delta)}{1 - \rho}. \quad (2.14)$$

Let μ^* be an optimal stationary policy. From Eq. (2.12), we have

$$\begin{aligned} T_{\mu^{k+1}} J_{\mu^k} &\leq T_{\mu^*} J_{\mu^k} + (\epsilon + 2\delta)e \\ &= T_{\mu^*} J_{\mu^k} - T_{\mu^*} J^* + J^* + (\epsilon + 2\delta)e \\ &= P_{\mu^*}(J_{\mu^k} - J^*) + J^* + (\epsilon + 2\delta)e. \end{aligned}$$

We also have

$$T_{\mu^{k+1}} J_{\mu^k} = J_{\mu^{k+1}} + T_{\mu^{k+1}} J_{\mu^k} - T_{\mu^{k+1}} J_{\mu^{k+1}} = J_{\mu^{k+1}} + P_{\mu^{k+1}}(J_{\mu^k} - J_{\mu^{k+1}}).$$

By subtracting the last two relations, and by using the definition of ξ_k and Eq. (2.14), we obtain

$$\begin{aligned} J_{\mu^{k+1}} - J^* &\leq P_{\mu^*}(J_{\mu^k} - J^*) + P_{\mu^{k+1}}(J_{\mu^{k+1}} - J_{\mu^k}) + (\epsilon + 2\delta)e \\ &\leq P_{\mu^*}(J_{\mu^k} - J^*) + \xi_k P_{\mu^{k+1}}e + (\epsilon + 2\delta)e \\ &\leq P_{\mu^*}(J_{\mu^k} - J^*) + \xi_k e + (\epsilon + 2\delta)e \\ &\leq P_{\mu^*}(J_{\mu^k} - J^*) + \frac{(1 - \rho + n)(\epsilon + 2\delta)}{1 - \rho}e. \end{aligned} \quad (2.15)$$

Let

$$\zeta_k = \max_{i=1,\dots,n} (J_{\mu^k}(i) - J^*(i)).$$

Then Eq. (2.15) yields, for all k ,

$$\zeta_{k+1}e \leq \zeta_k P_{\mu^*}e + \frac{(1 - \rho + n)(\epsilon + 2\delta)}{1 - \rho}e.$$

By multiplying this relation with P_{μ^*} and by adding $(1 - \rho + n)(\epsilon + 2\delta)e/(1 - \rho)$, we obtain

$$\zeta_{k+2}e \leq \zeta_{k+1}P_{\mu^*}e + \frac{(1 - \rho + n)(\epsilon + 2\delta)}{1 - \rho}e \leq \zeta_k P_{\mu^*}^2e + \frac{2(1 - \rho + n)(\epsilon + 2\delta)}{1 - \rho}e.$$

By repeating this process for a total of $n - 1$ times, we have

$$\zeta_{k+n}e \leq \zeta_k P_{\mu^*}^n e + \frac{n(1 - \rho + n)(\epsilon + 2\delta)}{1 - \rho}e \leq \rho \zeta_k e + \frac{n(1 - \rho + n)(\epsilon + 2\delta)}{1 - \rho}e.$$

By taking the limit superior as $k \rightarrow \infty$, we obtain

$$(1 - \rho) \limsup_{k \rightarrow \infty} \zeta_k \leq \frac{n(1 - \rho + n)(\epsilon + 2\delta)}{1 - \rho},$$

which was to be proved. **Q.E.D.**

The error bound (2.11) uses the worst-case estimate of the number of stages required to reach t with positive probability, which is n . We can strengthen the error bound if we have a better estimate. In particular, for all $m \geq 1$, let

$$\rho_m = \max_{\substack{i=1,\dots,n \\ \mu: \text{proper}}} P\{x_m \neq t \mid x_0 = i, \mu\},$$

and let \bar{m} be the minimal m for which $\rho_m < 1$. Then the proof of Prop. 2.4.1 can be adapted to show that

$$\limsup_{k \rightarrow \infty} \max_{i=1,\dots,n} (J_{\mu^k}(i) - J^*(i)) \leq \frac{\bar{m}(1 - \rho_{\bar{m}} + \bar{m})(\epsilon + 2\delta)}{(1 - \rho_{\bar{m}})^2}.$$

2.5 COUNTABLE STATE PROBLEMS

We will now consider an extension of the SSP problem where the number of states is countable. In particular, we assume that the state space is the set

$$S = \{1, 2, \dots\},$$

plus an absorbing destination state t . The transition probabilities are denoted $p_{ij}(u)$ for $i, j \in S \cup \{t\}$ and $u \in U(i)$, and the expected cost per stage is denoted by $g(i, u)$, $i \in S$, $u \in U(i)$. The constraint set $U(i)$ may be infinite/arbitrary. We follow the formalism of Section 1.4.3 that deals with discounted problems with a countable state space and possibly unbounded costs per stage. We introduce a positive sequence $v = \{v_1, v_2, \dots\}$, and the weighted sup-norm

$$\|J\| = \max_{i \in S} \frac{|J(i)|}{v_i}$$

in the space $B(S)$ of sequences $\{J(1), J(2), \dots\}$ such that $\|J\| < \infty$. The following assumption parallels Assumption 1.4.1 in Section 1.4.3.

Assumption 2.5.1:

- (a) We have $G = \{G(1), G(2), \dots\} \in B(S)$, where

$$G(i) = \max_{u \in U(i)} |g(i, u)|, \quad i = 1, 2, \dots$$

- (b) We have $V = \{V(1), V(2), \dots\} \in B(S)$, where

$$V(i) = \max_{u \in U(i)} \sum_{j \in S} p_{ij}(u) v_j, \quad i = 1, 2, \dots$$

- (c) There exists an integer $m \geq 1$ and a scalar $\rho \in (0, 1)$ such that for every policy π , we have

$$\frac{\sum_{j \in S} P(x_m = j \mid x_0 = i, \pi) v_j}{v_i} \leq \rho, \quad i = 1, 2, \dots$$

As an example consider the case

$$v_i \equiv 1.$$

Then Assumption 2.5.1(a) is equivalent to the expected cost per stage being uniformly bounded over i and u . Assumption 2.5.1(b) is automatically

satisfied, since $V(i) \equiv 1$. Finally, Assumption 2.5.1(c) is satisfied if there exists an integer $m \geq 1$ such that the probability of reaching the destination t within m stages is bounded away from 0 uniformly over all initial states i and policies π . Thus, in the case where $v_i \equiv 1$, Assumption 2.5.1 parallels (and generalizes to the countable state space case) the assumption that was used in Section 7.2 of Vol. I (which is equivalent to all stationary policies being proper; see Exercise 2.3).

Note, however, that with choices of v_i other than $v_i \equiv 1$, Assumption 2.5.1 applies to considerably more general problems, involving unbounded costs per stage. This is illustrated in the following example.

Example 2.5.1

Let

$$v_i = i, \quad i = 1, 2, \dots$$

Then Assumption 2.5.1(a) is satisfied if the maximum expected absolute cost per stage at state i grows no faster than linearly with i . Assumption 2.5.1(b) states that the expected next state following state i ,

$$E\{j \mid i, u\},$$

also grows no faster than linearly with i for all $u \in U(i)$. Finally, Assumption 2.5.1(c) is satisfied if for some m , and all i and π ,

$$\sum_{j \in S} P(x_m = j \mid x_0 = i, \pi) j \leq \rho i.$$

This means that for all π , the expected value of the state reached m stages following state i is no more than a fraction ρ of i , i.e., every m stages, there is at least a $(1 - \rho)i$ downward expected change of the state starting at i .

We now consider the DP mappings T_μ and T :

$$(T_\mu J)(i) = g(i, \mu(i)) + \sum_{j \in S} p_{ij}(\mu(i)) J(j), \quad i = 1, 2, \dots,$$

$$(TJ)(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j \in S} p_{ij}(u) J(j) \right], \quad i = 0, 1, \dots$$

A nearly verbatim repetition of the proof of Prop. 1.4.5 of Section 1.4.3 yields the following.

Proposition 2.5.1: Under Assumption 2.5.1, the mappings T and T_μ map $B(S)$ into $B(S)$, and are m -stage contraction mappings with modulus ρ .

The m -stage contraction property and the contraction mapping theorem for m -stage contractions (cf. Prop. 1.4.3) can be used to show the standard DP results:

- (a) The value iteration method $J_{k+1} = TJ_k$ converges to the unique solution J^* of Bellman's equation $J = TJ$.
- (b) The unique solution J^* of Bellman's equation is the optimal cost function of the problem.
- (c) A stationary policy μ is optimal if and only if $T_\mu J^* = TJ^*$.

We finally note a connection of the analysis of this section with the analysis of Section 2.2, where we showed that when the number of states is finite and all stationary policies are proper, the mappings T_μ and T are weighted (one-step) contraction mappings. Indeed, the proof of this fact (Prop. 2.2.3) constructs a set of weights v_i for which Assumption 2.5.1 is satisfied with $m = 1$ [cf. Eq. (2.7)]; see also Exercise 2.15.

2.6 NOTES, SOURCES, AND EXERCISES

The analysis of the SSP problem of Section 2.1 is due to Bertsekas and Tsitsiklis [BeT89], [BeT91b]. The latter reference proves the results shown here under a more general compactness assumption on $U(i)$, and continuity assumption on $g(i, u)$ and $p_{ij}(u)$. The first formulation of SSP problems was given by Eaton and Zadeh [EaZ62], under the assumption $g(i, u) > 0$ for all $i = 1, \dots, n$ and $u \in U(i)$. The proof of the weighted sup-norm contraction property of the mapping T is from Bertsekas and Tsitsiklis [BeT96] (Prop. 2.2), and was also independently given by Littman [Lit96] (earlier proofs were given in [BeT89], p. 325, and Tseng [Tse90]).

Finitely terminating value iteration algorithms have been developed for several types of SSP problems (see Nguyen and Pallottino [NgP86], Polychronopoulos and Tsitsiklis [PoT96], Psarafitis and Tsitsiklis [PsT93], Tsitsiklis [Tsi95], Bertsekas, Guerriero, and Musmanno [BGM95], Polymenakos, Bertsekas, and Tsitsiklis [PBT98]). The use of a Dijkstra-like algorithm for continuous space shortest path problems involving a consistently improving policy was first proposed by Tsitsiklis [Tsi95] (see Exercise 2.10). Tsitsiklis' algorithm was rediscovered later, under the name “fast marching method,” by Sethian [Set99a], [Set99b], who discusses several other related methods and applications, as well as by Helmsen et al. [HPC96]. A Dijkstra-like algorithm was also proposed for another class of problems involving a consistently improving policy by Nguyen and Pallottino [NgP86]. The Dijkstra-like algorithm of Exercise 2.11 is new in the general form given here. The error bound on the performance of approximate policy iteration (Prop. 2.4.1) is due to Bertsekas and Tsitsiklis [BeT96]. Computational methods for problems with a countably infinite

state space are given by Hinderer and Waldmann [HiW05]. Two-player dynamic game versions of the SSP problem were discussed by Pollatschek and Avi-Itzhak [PoA69]; see also the survey by Raghavan and Filar [RaF91], and the book by Filar and Vrieze [FiV96]. An analysis which is closer to the spirit of the present chapter is given by Patek and Bertsekas [PaB99].

Generalized forms of stochastic shortest path problems, which involve an infinite (uncountable) number of states are analyzed by Pliska [Pli78], Hernandez-Lerma et al. [HCP99], and James and Collins [JaC06].

E X E R C I S E S

2.1

Suppose that you want to travel from a start point S to a destination point D in minimum average time. There are two options:

- (1) Use a direct route that requires a time units.
- (2) Take a potential shortcut that requires b time units to go to an intermediate point I . From I you can either go to the destination D in c time units or return to the start (this will take an additional b time units). You will find out the value of c once you reach the intermediate point I . What you know a priori is that c has one of the m values c_1, \dots, c_m with corresponding probabilities p_1, \dots, p_m . Consider two cases: (i) The value of c is constant over time, and (ii) The value of c changes each time you return to the start independently of the value at the previous time periods.
 - (a) Formulate the problem as an SSP problem. Write Bellman's equation and characterize the optimal stationary policies as best as you can in terms of the given problem data. Solve the problem for the case $a = 2$, $b = 1$, $c_1 = 0$, $c_2 = 5$, $p_1 = 0.5$, $p_2 = 0.5$.
 - (b) Formulate as an SSP problem the variation where once you reach the intermediate point I , you can wait there. Each d time units the value of c changes to one of the values c_1, \dots, c_m with probabilities p_1, \dots, p_m , independently of its earlier values. Each time the value of c changes, you have the option of waiting for an extra d units, returning to the start, or going to the destination. Characterize the optimal stationary policies as best as you can.

2.2

A gambler engages in a game of successive coin flipping over an infinite horizon. He wins one dollar each time heads comes up, and loses $m > 0$ dollars each

time two successive tails come up (so the sequence TTTT loses $3m$ dollars). The gambler at each time period either flips a fair coin or else cheats by flipping a two-headed coin. In the latter case, however, he gets caught with probability $p > 0$ before he flips the coin, the game terminates, and the gambler keeps his earnings thus far. The gambler wishes to maximize his expected earnings.

- (a) View this as an SSP problem and identify all proper and all improper policies.
- (b) Identify a critical value \bar{m} such that if $m > \bar{m}$, then all improper policies give an infinite cost for some initial state.
- (c) Assume that $m > \bar{m}$, and show that it is then optimal to try to cheat if the last flip was tails and to play fair otherwise.
- (d) Show that if $m < \bar{m}$ it is optimal to always play fair.

2.3

Consider an SSP problem where all stationary policies are proper. Show that for every policy π there exists an $m > 0$ such that

$$P(x_m = t \mid x_0 = i, \pi) > 0$$

for all $i = 1, \dots, n$. *Abbreviated Proof:* Consider another SSP problem, which is the same as the original except that $g(i, u, j) = -1$ for all $i = 1, \dots, n$, $j = 1, \dots, n$, t , and $u \in U(i)$. By Prop. 2.2.2, the optimal cost $J^*(i)$ is finite for all i . Conclude that for every policy π and state i , there must exist a path that leads from i to t and contains no more than $\max_{s=1, \dots, n} |J^*(s)|$ (positive probability) transitions.

2.4

Consider the SSP problem, and assume that $g(i, u) \leq 0$ for all i and $u \in U(i)$. Show that either the optimal cost is $-\infty$ for some initial state, or else, under every policy, the system eventually enters with probability 1 a set of cost-free states and never leaves that set thereafter.

2.5

Consider the SSP problem, and assume that there exists at least one proper policy. Proposition 2.2.2 implies that if, for each improper policy μ , we have $J_\mu(i) = \infty$ for at least one state i , then there is no improper policy μ' such that $J_{\mu'}(j) = -\infty$ for at least one state j . Give an alternative proof of this fact that does not use Prop. 2.2.2. Hint: Suppose that there exists an improper policy μ' such that $J_{\mu'}(j) = -\infty$ for at least one state j . Combine this policy with a proper policy to produce another improper policy μ'' for which $J_{\mu''}(i) < \infty$ for all i .

2.6 (Gauss-Seidel Method for Stochastic Shortest Paths)

Show that the Gauss-Seidel version of the value iteration method for SSP converges under the same assumptions as the ordinary method (Assumptions 2.1.1 and 2.1.2). *Hint:* Consider two functions \underline{J} and \bar{J} that differ by a constant from J^* at all states except the destination, and are such that $\underline{J} \leq T\underline{J}$ and $T\bar{J} \leq \bar{J}$.

2.7 (Sequential Space Decomposition) www

Consider the SSP problem, and suppose that there is a finite sequence of subsets of states S_1, S_2, \dots, S_M such that each of the states $i = 1, \dots, n$ belongs to one and only one of these subsets, and the following property holds:

For all $m = 1, \dots, M$ and states $i \in S_m$, the successor state j is either the termination state t or else belongs to one of the subsets S_m, S_{m-1}, \dots, S_1 for all choices of the control $u \in U(i)$.

- (a) Show that the solution of this problem decomposes into the solution of M SSP problems, each involving the states in a subset S_m plus a termination state.
- (b) Show also that a finite horizon problem with N stages can be viewed as an SSP problem with the property given above.

2.8 www

Consider an SSP problem under Assumptions 2.1.1 and 2.1.2. Assuming $p_{ii}(u) < 1$ for all $i \neq t$ and $u \in U(i)$, consider another SSP problem that has transition probabilities

$$\tilde{p}_{ij}(u) = \begin{cases} 0 & \text{if } j = i, \\ \frac{p_{ij}(u)}{1-p_{ii}(u)} & \text{if } j \neq i, \end{cases} \quad i = 1, \dots, n,$$

and costs

$$\tilde{g}(i, u) = \frac{g(i, u)}{1 - p_{ii}(u)}.$$

- (a) Show that the two problems are equivalent in that they have the same optimal costs and policies. How would you deal with the case where $p_{ii}(u) = 1$ for some $i \neq t$ and $u \in U(i)$?
- (b) Interpret $\tilde{g}(i, u)$ as an average cost incurred between arrival to state i and transition to a state $j \neq i$.

2.9 (Simplifications for Uncontrollable State Components)

Consider an SSP problem under Assumptions 2.1.1 and 2.1.2, where the state is a composite (i, y) of two components i and y , and the evolution of the main

component i can be directly affected by the control u , but the evolution of the other component y cannot (cf. Section 1.4 of Vol. I). In particular, we assume that given the state (i, y) and the control u , the next state (j, z) is determined as follows: first j is generated according to transition probabilities $p_{ij}(u, y)$, and then z is generated according to conditional probabilities $p(z | j)$ that depend on the main component j of the new state. We also assume that the cost per stage is $g(i, y, u, j)$ and does not depend on the second component z of the next state (j, z) . For functions $\hat{J}(i)$, $i = 1, \dots, n$, consider the mapping

$$(\hat{T}\hat{J})(i) = \sum_y p(y | i) \left(\min_{u \in U(i, y)} \sum_j p_{ij}(u, y) (g(i, y, u, j) + \hat{J}(j)) \right),$$

and the corresponding mapping of a stationary policy μ ,

$$(\hat{T}_\mu \hat{J})(i) = \sum_y p(y | i) \sum_j p_{ij}(\mu(i, y), y) (g(i, y, \mu(i, y), j) + \hat{J}(j)).$$

- (a) Show that $\hat{J} = \hat{T}\hat{J}$ is a form of Bellman's equation and can be used to characterize the optimal stationary policies. *Hint:* Given $J(i, y)$, define $\hat{J}(i) = \sum_y p(y | i) J(i, y)$.
- (b) Show the validity of a modified value iteration algorithm that starts with an arbitrary function \hat{J} and sequentially produces $\hat{T}\hat{J}$, $\hat{T}^2\hat{J}$, ...
- (c) Show the validity of a modified policy iteration algorithm whose typical iteration, given the current policy $\mu^k(i, y)$, consists of two steps: (1) The policy evaluation step, which computes the unique function \hat{J}_{μ^k} that solves the linear system of equations $\hat{J}_{\mu^k} = \hat{T}_{\mu^k} \hat{J}_{\mu^k}$. (2) The policy improvement step, which computes the improved policy $\mu^{k+1}(i, y)$ from the equation $\hat{T}_{\mu^{k+1}} \hat{J}_{\mu^k} = \hat{T} \hat{J}_{\mu^k}$.

2.10 (Discretized Shortest Path Problems [Tsi95])

Suppose that the states are the grid points of a grid on the plane. The set of neighbors of each grid point x is denoted $U(x)$ and includes between two and four grid points. At each grid point x , we have two options:

- (1) Choose two neighbors $x^+, x^- \in U(x)$ and a probability $p \in [0, 1]$, pay a cost $g(x)\sqrt{p^2 + (1-p)^2}$, and move to x^+ or to x^- with probability p or $1-p$, respectively. Here g is a function such that $g(x) > 0$ for all x .
- (2) Stop and pay a cost $t(x)$.

Show that there exists a consistently improving optimal policy for this problem. *Note:* This problem can be used to model discretized versions of deterministic continuous space 2-dimensional shortest path problems. (Compare also with Exercise 6.11 in Chapter 6 of Vol. I.)

2.11 (Dijkstra's Algorithm and Consistently Improving Policies)

Consider the SSP problem under Assumptions 2.1.1 and 2.1.2, and assume that there exists a consistently improving optimal stationary policy.

- (a) Show that the transition probability graph of this policy is acyclic.
- (b) Consider the following algorithm, which maintains two subsets of states P and L , and a function J defined on the state space. (To relate the algorithm with Dijkstra's method of Section 2.3.1 of Vol. I, associate J with the node labels, L with the OPEN list, and P with the subset of nodes that have already exited the OPEN list.) Initially, $P = \emptyset$, $L = \{t\}$, and

$$J(i) = \begin{cases} \infty & \text{if } i = 1, \dots, n, \\ 0 & \text{if } i = t. \end{cases}$$

At the typical iteration, select a state j^* from L such that

$$j^* = \arg \min_{j \in L} J(j).$$

(If L is empty the algorithm terminates.) Remove j^* from L and place it in P . In addition, for all $i \notin P$ such that there exists a $u \in U(i)$ with $p_{ij^*}(u) > 0$, and

$$p_{ij}(u) = 0 \quad \text{for all } j \notin P,$$

define

$$\hat{U}(i) = \{u \in U(i) \mid p_{ij^*}(u) > 0 \text{ and } p_{ij}(u) = 0 \text{ for all } j \notin P\},$$

set

$$J(i) := \min \left[J(i), \min_{u \in \hat{U}(i)} \left[g(i, u) + \sum_{j \in P} p_{ij}(u) J(j) \right] \right],$$

and place i in L if it is not already there. Show that the algorithm is well defined in the sense that $\hat{U}(i)$ is nonempty and the set L does not become empty until all states are in P . Furthermore, each state j is removed from L once, and at the time it is removed, we have $J(j) = J^*(j)$.

2.12 (Alternative Assumptions for Prop. 2.2.2)

Consider a variation of Assumption 2.1.2, whereby we assume that $g(i, u) \geq 0$ for all i and $u \in U(i)$, and that there exists a proper policy. Prove the assertions of Prop. 2.2.2, except that, in part (a), uniqueness of the solution of Bellman's equation should be shown within the set $\mathbb{R}^+ = \{J \mid J \geq 0\}$ (rather than within \mathbb{R}^n), and the vector J in part (b) must belong to \mathbb{R}^+ . Hint: Proposition 2.2.1 is not valid, so a somewhat different proof is needed. Complete the details of the following argument. The assumptions guarantee that J^* is finite and $J^* \in \mathbb{R}^+$. [We have $J^* \geq 0$ because $g(i, u) \geq 0$, and $J^*(i) < \infty$ because a proper policy exists.] The idea now is to show that $J^* \geq TJ^*$, and then to choose μ such that

$T_\mu J^* = TJ^*$ and show that μ is optimal. Let $\pi = \{\mu_0, \mu_1, \dots\}$ be a policy. We have for all i ,

$$J_\pi(i) = g(i, \mu_0(i)) + \sum_{j=1}^n p_{ij}(\mu_0(i)) J_{\pi_1}(j)$$

where π_1 is the policy $\{\mu_1, \mu_2, \dots\}$. Since $J_{\pi_1} \geq J^*$, we obtain

$$J_\pi(i) \geq g(i, \mu_0(i)) + \sum_{j=1}^n p_{ij}(\mu_0(i)) J^*(j) = (T_{\mu_0} J^*)(i) \geq (TJ^*)(i).$$

Taking the infimum over π in the preceding equation, we obtain

$$J^* \geq TJ^*. \quad (2.16)$$

Let μ be such that $T_\mu J^* = TJ^*$. From Eq. (2.16), we have $J^* \geq T_\mu J^*$, and using the monotonicity of T_μ , we obtain

$$J^* \geq T_\mu J^* \geq T_\mu^N J^* = P_\mu^N J^* + \sum_{k=0}^{N-1} P_\mu^k g_\mu \geq \sum_{k=0}^{N-1} P_\mu^k g_\mu, \quad N \geq 1.$$

By taking limit superior as $N \rightarrow \infty$, we obtain $J^* \geq J_\mu$. Therefore, μ is an optimal policy, and $J^* = J_\mu$. Since μ was selected so that $T_\mu J^* = TJ^*$, we obtain, using $J^* = J_\mu$ and $J_\mu = T_\mu J_\mu$, that $J^* = TJ^*$. For the rest of the proof, use the vector δe similar to the proof of Prop. 2.2.2.

2.13 (A Contraction Counterexample)

Consider an SSP problem with a single state 1, in addition to the termination state t . At state 1 there are two controls u and u' . Under u the cost is 1 and the system remains in state 1 for one more stage; under u' the cost is 2 and the system moves to t . Show that Assumptions 2.1.1 and 2.1.2 are satisfied, but T is not a contraction mapping with respect to any norm.

2.14 (Multistage Lookahead Policy Iteration)

- (a) Consider the SSP problem under Assumptions 2.1.1 and 2.1.2. Let μ be a stationary policy, let J be a function such that $TJ \leq J \leq J_\mu$ ($J = J_\mu$ is one possibility), and let $\{\bar{\mu}_0, \bar{\mu}_1, \dots, \bar{\mu}_{N-1}\}$ be an optimal policy for the N -stage problem with terminal cost function J , i.e.

$$T_{\bar{\mu}_k} T^{N-k-1} J = T^{N-k} J, \quad k = 0, 1, \dots, N-1.$$

- (a) Show that

$$J_{\bar{\mu}_k} \leq J_\mu, \quad \text{for all } k = 0, 1, \dots, N-1.$$

Hint: First show that $T^{k+1} J \leq T^k J \leq J$ for all k , and then show that the hypothesis $T_{\bar{\mu}_k} T^{N-k-1} J = T^{N-k} J$ implies that $J_{\bar{\mu}_k} \leq T^{N-k-1} J$.

- (b) Use part (a) to show the validity of the multistage policy iteration algorithm discussed in Section 2.3.3.

2.15 (Contractions for Countable Number of States)

Consider the countable space SSP problem of Section 2.5. Let v_i be the maximum (over all policies) expected number of stages up to termination, starting from state i . Assume that v_i is finite and bounded over i . Show that the mappings T and T_μ are contraction mappings with respect to the weighted sup-norm $\|\cdot\|_v$, with modulus of contraction

$$\rho = \max_{i=1,2,\dots} \frac{v_i - 1}{v_i}.$$

Hint: Consider the problem of maximizing the expected time to termination starting from an initial state i . Show that v_i , the optimal value of this problem, satisfies Bellman's equation

$$v_i = 1 + \max_{u \in U(i)} \sum_{j=1}^{\infty} p_{ij}(u) v_j, \quad i = 1, 2, \dots,$$

[this can be shown by using value iteration arguments, and is also a special case of general results shown in Chapter 3 (see Prop. 3.1.1)]. Conclude that for all stationary policies μ , we have

$$1 + \sum_{j=1}^{\infty} p_{ij}(\mu(i)) v_j \leq v_i, \quad i = 1, 2, \dots,$$

so that

$$\frac{\sum_{j=1}^{\infty} p_{ij}(\mu(i)) v_j}{v_i} \leq \rho, \quad i = 1, 2, \dots$$

Use Prop. 1.4.1.

2.16

A certain activity is performed for an infinite sequence of days by a single machine. At the beginning of each day, we must decide whether the existing machine will be used, or be replaced by a new machine at cost C . On the k th day of its operation, the machine will produce revenue r_k , and it will break down at the end of the k th day with probability p_k , in which case it will have to be replaced by a new machine. Assume that there exists $\epsilon > 0$ such that $p_k \geq \epsilon$ for all k . Show that the mappings T and T_μ for this problem are contraction mappings.

Hint: Use the result of the preceding exercise.

Undiscounted Problems ¹**Contents**

3.1. Unbounded Costs per Stage	p. 124
3.2. Linear Systems and Quadratic Cost	p. 140
3.3. Inventory Control	p. 142
3.4. Optimal Stopping	p. 145
3.5. Optimal Gambling Strategies	p. 150
3.6. Nonstationary and Periodic Problems	p. 157
3.7. Notes, Sources, and Exercises	p. 162

In this chapter we consider total cost infinite horizon problems where we allow costs per stage that are unbounded above or below. Also, the discount factor α does not have to be less than one. The complications resulting are substantial, and the analysis required is considerably more sophisticated than the one given thus far. We also consider applications of the theory to important classes of problems. The exercise section touches on several related topics.

3.1 UNBOUNDED COSTS PER STATE

In this section we consider the total cost infinite horizon problem of Section 1.1 under one of the following two assumptions.

Assumption P: (Positivity) The cost per stage g satisfies

$$0 \leq g(x, u, w), \quad \text{for all } (x, u, w) \in S \times C \times D. \quad (3.1)$$

Assumption N: (Negativity) The cost per stage g satisfies

$$g(x, u, w) \leq 0, \quad \text{for all } (x, u, w) \in S \times C \times D. \quad (3.2)$$

Somewhat paradoxically, problems corresponding to Assumption P are sometimes referred to in the research literature as *negative DP problems*. This choice of name is due to historical reasons. It was introduced in an early paper by Strauch [Str66], where the problem of maximizing the infinite sum of negative rewards per stage was considered. Similarly, problems corresponding to Assumption N are sometimes referred to as *positive DP problems* (Blackwell [Bla65], Strauch [Str66]). Assumption N arises in problems where there is a nonnegative reward per stage and the total expected reward is to be *maximized*.

Note that when $\alpha < 1$ and g is either bounded above or below, we may add a suitable scalar to g in order to satisfy Eq. (3.1) or Eq. (3.2), respectively. An optimal policy will not be affected by this change since, because of the discount factor, the addition of a constant r to g merely adds $(1 - \alpha)^{-1}r$ to the cost of every policy.

One complication arising from unbounded costs per stage is that, for some initial states x_0 and some genuinely interesting admissible policies

$\pi = \{\mu_0, \mu_1, \dots\}$, the cost $J_\pi(x_0)$ may be ∞ (in the case of Assumption P) or $-\infty$ (in the case of Assumption N). Here is an example:

Example 3.1.1

Consider the scalar system

$$x_{k+1} = \beta x_k + u_k, \quad k = 0, 1, \dots,$$

where $x_k \in \mathbb{R}$ and $u_k \in \mathbb{R}$, for all k , and β is a positive scalar. The control constraint is $|u_k| \leq 1$, and the cost is

$$J_\pi(x_0) = \lim_{N \rightarrow \infty} \sum_{k=0}^{N-1} \alpha^k |x_k|.$$

Consider the policy $\bar{\pi} = \{\bar{\mu}, \bar{\mu}, \dots\}$, where $\bar{\mu}(x) = 0$ for all $x \in \mathbb{R}$. Then

$$J_{\bar{\pi}}(x_0) = \lim_{N \rightarrow \infty} \sum_{k=0}^{N-1} \alpha^k \beta^k |x_0|,$$

and hence

$$J_{\bar{\pi}}(x_0) = \begin{cases} 0 & \text{if } x_0 = 0 \\ \infty & \text{if } x_0 \neq 0 \end{cases} \quad \text{if } \alpha\beta \geq 1,$$

while

$$J_{\bar{\pi}}(x_0) = \frac{|x_0|}{1 - \alpha\beta} \quad \text{if } \alpha\beta < 1.$$

Note a peculiarity here: if $\beta > 1$ the state x_k diverges to ∞ or to $-\infty$, but if the discount factor is sufficiently small ($\alpha < 1/\beta$), the cost $J_{\bar{\pi}}(x_0)$ is finite.

It is also possible to verify that when $\beta > 1$ and $\alpha\beta \geq 1$ the optimal cost $J^*(x_0)$ is equal to ∞ for $|x_0| \geq 1/(\beta-1)$ and is finite for $|x_0| < 1/(\beta-1)$. What happens here is that when $\beta > 1$ the system is unstable, and in view of the restriction $|u_k| \leq 1$ on the control, it may not be possible to force the state near zero once it has reached sufficiently large magnitude.

The preceding example shows that there is not much that can be done about the possibility of the cost function being infinite for some policies. To cope with this situation, we conduct our analysis with the notational understanding that the costs $J_\pi(x_0)$ and $J^*(x_0)$ may be ∞ (or $-\infty$) under Assumption P (or N, respectively) for some initial states x_0 and policies π . In other words, we consider $J_\pi(\cdot)$ and $J^*(\cdot)$ to be extended real-valued functions. In fact, the entire subsequent analysis is valid even if the cost $g(x, u, w)$ is ∞ or $-\infty$ for some (x, u, w) , as long as Assumption P or Assumption N holds.

The line of analysis of this section is fundamentally different from the one of the discounted problem of Section 1.2. For the latter problem, the analysis was based on ignoring the “tails” of the cost sequences. In

this section, the tails of the cost sequences may not be small, and for this reason, the control is much more focused on affecting the long-term behavior of the state. For example, let $\alpha = 1$, and assume that the stage cost at all states is nonzero except for a cost-free and absorbing termination state. Then, a primary task of control under Assumption P (or Assumption N) is roughly to bring the state of the system to the termination state or to a region where the cost per stage is nearly zero as *quickly* as possible (as *late* as possible, respectively). Note the difference in control objective between Assumptions P and N. It accounts for some strikingly different results under the two assumptions.

Main Results – Bellman’s Equation

We now present results that characterize the optimal cost function J^* , as well as optimal stationary policies. We also give conditions under which value iteration converges to the optimal cost function J^* . In the proofs we will often need to interchange expectation and limit in various relations. This interchange is valid under the assumptions of the following theorem.

Monotone Convergence Theorem: Let $P = (p_1, p_2, \dots)$ be a probability distribution over $S = \{1, 2, \dots\}$. Let $\{h_N\}$ be a sequence of extended real-valued functions on S such that for all $i \in S$ and $N = 1, 2, \dots$,

$$0 \leq h_N(i) \leq h_{N+1}(i).$$

Let $h : S \mapsto [0, \infty]$ be the limit function

$$h(i) = \lim_{N \rightarrow \infty} h_N(i).$$

Then

$$\lim_{N \rightarrow \infty} \sum_{i=1}^{\infty} p_i h_N(i) = \sum_{i=1}^{\infty} p_i \lim_{N \rightarrow \infty} h_N(i) = \sum_{i=1}^{\infty} p_i h(i).$$

Proof: We have

$$\sum_{i=1}^{\infty} p_i h_N(i) \leq \underbrace{\sum_{i=1}^{\infty} p_i h(i)}_{\text{by assumption}}.$$

By taking the limit, we obtain

$$\lim_{N \rightarrow \infty} \sum_{i=1}^{\infty} p_i h_N(i) \leq \sum_{i=1}^{\infty} p_i h(i),$$

so there remains to prove the reverse inequality. For every integer $M \geq 1$, we have

$$\lim_{N \rightarrow \infty} \sum_{i=1}^{\infty} p_i h_N(i) \geq \lim_{N \rightarrow \infty} \sum_{i=1}^M p_i h_N(i) = \sum_{i=1}^M p_i h(i),$$

and by taking the limit as $M \rightarrow \infty$ the reverse inequality follows. Q.E.D.

Similar to all the infinite horizon problems considered so far, the optimal cost function satisfies Bellman's equation.

Proposition 3.1.1: (Bellman's Equation) Under either Assumption P or N the optimal cost function J^* satisfies

$$J^*(x) = \min_{u \in U(x)} E_w \{ g(x, u, w) + \alpha J^*(f(x, u, w)) \}, \quad x \in S$$

or, equivalently,

$$J^* = TJ^*.$$

Proof: For any admissible policy $\pi = \{\mu_0, \mu_1, \dots\}$, consider the cost $J_\pi(x)$ corresponding to π when the initial state is x . We have

$$J_\pi(x) = E_w \{ g(x, \mu_0(x), w) + V_\pi(f(x, \mu_0(x), w)) \}, \quad (3.3)$$

where, for all $x_1 \in S$,

$$V_\pi(x_1) = \lim_{N \rightarrow \infty} E_{w_k} \left\{ \sum_{k=1}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}.$$

Thus, $V_\pi(x_1)$ is the cost from stage 1 to infinity using π when the initial state is x_1 . We clearly have

$$V_\pi(x_1) \geq \alpha J^*(x_1), \quad \text{for all } x_1 \in S.$$

Hence, from Eq. (3.3),

$$\begin{aligned} J_\pi(x) &\geq E_w \{ g(x, \mu_0(x), w) + \alpha J^*(f(x, \mu_0(x), w)) \} \\ &\geq \min_{u \in U(x)} E_w \{ g(x, u, w) + \alpha J^*(f(x, u, w)) \}. \end{aligned}$$

Taking the minimum over all admissible policies, we obtain

$$\begin{aligned} \min_{\pi} J_\pi(x) &= J^*(x) \\ &\geq \min_{u \in U(x)} E_w \{ g(x, u, w) + \alpha J^*(f(x, u, w)) \} \\ &= (TJ^*)(x). \end{aligned}$$

Thus there remains to prove that the reverse inequality also holds. We prove this separately for Assumption N and for Assumption P.

Assume P. The following proof of $J^* \leq TJ^*$ under this assumption would be considerably simplified if we knew that there exists a μ such that $T_\mu J^* = TJ^*$. Since in general such a μ need not exist, we introduce a positive sequence $\{\epsilon_k\}$, and we choose an admissible policy $\pi = \{\mu_0, \mu_1, \dots\}$ such that

$$(T_{\mu_k} J^*)(x) \leq (TJ^*)(x) + \epsilon_k, \quad x \in S, \quad k = 0, 1, \dots$$

Such a choice is possible because under P, we have $0 \leq J^*(x)$ for all x . By using the inequality $TJ^* \leq J^*$ shown earlier, we obtain

$$(T_{\mu_k} J^*)(x) \leq J^*(x) + \epsilon_k, \quad x \in S, \quad k = 0, 1, \dots$$

Applying $T_{\mu_{k-1}}$ to both sides of this relation, we have

$$\begin{aligned} (T_{\mu_{k-1}} T_{\mu_k} J^*)(x) &\leq (T_{\mu_{k-1}} J^*)(x) + \alpha \epsilon_k \\ &\leq (TJ^*)(x) + \epsilon_{k-1} + \alpha \epsilon_k \\ &\leq J^*(x) + \epsilon_{k-1} + \alpha \epsilon_k. \end{aligned}$$

Continuing this process, we obtain

$$(T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} J^*)(x) \leq (TJ^*)(x) + \sum_{i=0}^k \alpha^i \epsilon_i.$$

By taking the limit as $k \rightarrow \infty$ and noting that

$$J^*(x) \leq J_\pi(x) = \lim_{k \rightarrow \infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} J_0)(x) \leq \lim_{k \rightarrow \infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} J^*)(x),$$

where J_0 is the zero function, it follows that

$$J^*(x) \leq J_\pi(x) \leq (TJ^*)(x) + \sum_{i=0}^{\infty} \alpha^i \epsilon_i, \quad x \in S.$$

Since the sequence $\{\epsilon_k\}$ is arbitrary, we can take $\sum_{i=0}^{\infty} \alpha^i \epsilon_i$ as small as desired, and we obtain $J^*(x) \leq (TJ^*)(x)$ for all $x \in S$. Combining this with the inequality $J^*(x) \geq (TJ^*)(x)$ shown earlier, the result follows (under Assumption P).

Assume N and let J_N be the optimal cost function for the corresponding N-stage problem

$$J_N(x_0) = \min_{\pi} E \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}.$$

We first show that

$$J^*(x) = \lim_{N \rightarrow \infty} J_N(x), \quad x \in S. \quad (3.4)$$

Indeed, in view of Assumption N, we have $J^* \leq J_N$ for all N , so

$$J^*(x) \leq \lim_{N \rightarrow \infty} J_N(x), \quad x \in S. \quad (3.5)$$

Also, for all $\pi = \{\mu_0, \mu_1, \dots\}$, we have

$$E \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\} \geq J_N(x_0),$$

and by taking the limit as $N \rightarrow \infty$,

$$J_\pi(x) \geq \lim_{N \rightarrow \infty} J_N(x), \quad x \in S.$$

Taking the minimum over π , we obtain $J^*(x) \geq \lim_{N \rightarrow \infty} J_N(x)$, and combining this relation with Eq. (3.5), we obtain Eq. (3.4).

For every admissible μ , we have

$$T_\mu J_N \geq J_{N+1},$$

and by taking the limit as $N \rightarrow \infty$, and using the monotone convergence theorem and Eq. (3.4), we obtain

$$T_\mu J^* \geq J^*.$$

Taking the minimum over μ , we obtain $TJ^* \geq J^*$, which combined with the inequality $J^* \geq TJ^*$ shown earlier, proves the result under Assumption N. Q.E.D.

Similar to Cor. 1.2.2.1, we have:

Corollary 3.1.1.1: Let μ be a stationary policy. Then under Assumption P or N, we have

$$J_\mu(x) = E_w \{g(x, \mu(x), w) + \alpha J_\mu(f(x, \mu(x), w))\}, \quad x \in S$$

or, equivalently,

$$J_\mu = T_\mu J_\mu. \quad (3.6)$$

Contrary to discounted problems with bounded cost per stage, the optimal cost function J^* under Assumption P or N need not be the unique solution of Bellman's equation. Consider the following example.

Example 3.1.2

Let $S = [0, \infty)$ (or $S = (-\infty, 0]$) and

$$g(x, u, w) = 0, \quad f(x, u, w) = \frac{x}{\alpha}.$$

Then for every β , the function J given by $J(x) = \beta x$ for all $x \in S$, is a solution of Bellman's equation, so T has an infinite number of fixed points. Note, however, that there is a unique fixed point within the class of bounded functions, the zero function $J_0(x) \equiv 0$, which is the optimal cost function for this problem. More generally, it can be shown by using the following Prop. 3.1.2 that if $\alpha < 1$ and there exists a bounded function that is a fixed point of T , then that function must be equal to the optimal cost function J^* (see Exercise 3.5). When $\alpha = 1$, Bellman's equation may have an infinity of solutions even within the class of bounded functions. This is because if $\alpha = 1$ and $J(\cdot)$ is any solution, then for any scalar r , $J(\cdot) + r$ is also a solution.

The optimal cost function J^* , however, has the property that it is the smallest (under Assumption P) or largest (under Assumption N) fixed point of T in the sense described in the following proposition.

Proposition 3.1.2:

- (a) Under Assumption P, if $\tilde{J} : S \mapsto (-\infty, \infty]$ satisfies $\tilde{J} \geq T\tilde{J}$ and either \tilde{J} is bounded below and $\alpha < 1$, or $\tilde{J} \geq 0$, then $\tilde{J} \geq J^*$.
- (b) Under Assumption N, if $\tilde{J} : S \mapsto [-\infty, \infty)$ satisfies $\tilde{J} \leq T\tilde{J}$ and either \tilde{J} is bounded above and $\alpha < 1$, or $\tilde{J} \leq 0$, then $\tilde{J} \leq J^*$.

Proof: (a) Under Assumption P, let r be a scalar such that $\tilde{J}(x) + r \geq 0$ for all $x \in S$ and if $\alpha \geq 1$ let $r = 0$. For any sequence $\{\epsilon_k\}$ with $\epsilon_k > 0$, let $\tilde{\pi} = \{\tilde{\mu}_0, \tilde{\mu}_1, \dots\}$ be an admissible policy such that, for every $x \in S$ and k ,

$$\mathop{\mathbb{E}}_w \{g(x, \mu_k(x), w) + \alpha \tilde{J}(f(x, \mu_k(x), w))\} \leq (T\tilde{J})(x) + \epsilon_k. \quad (3.7)$$

Such a policy exists since $(T\tilde{J})(x) > -\infty$ for all $x \in S$. We have for any initial state $x_0 \in S$,

$$\begin{aligned} J^*(x_0) &= \min_{\pi} \lim_{N \rightarrow \infty} E \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\} \\ &\leq \min_{\pi} \liminf_{N \rightarrow \infty} E \left\{ \alpha^N (\tilde{J}(x_N) + r) + \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\} \\ &\leq \liminf_{N \rightarrow \infty} E \left\{ \alpha^N (\tilde{J}(x_N) + r) + \sum_{k=0}^{N-1} \alpha^k g(x_k, \tilde{\mu}_k(x_k), w_k) \right\}. \end{aligned}$$

Using Eq. (3.7) and the assumption $\tilde{J} \geq T\tilde{J}$, we obtain

$$\begin{aligned}
& E \left\{ \alpha^N \tilde{J}(x_N) + \sum_{k=0}^{N-1} \alpha^k g(x_k, \tilde{\mu}(x_k), w_k) \right\} \\
&= E \left\{ \alpha^N \tilde{J}(f(x_{N-1}, \tilde{\mu}_{N-1}(x_{N-1}), w_{N-1})) + \sum_{k=0}^{N-1} \alpha^k g(x_k, \tilde{\mu}_k(x_k), w_k) \right\} \\
&\leq E \left\{ \alpha^{N-1} \tilde{J}(x_{N-1}) + \sum_{k=0}^{N-2} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\} + \alpha^{N-1} \epsilon_{N-1} \\
&\leq E \left\{ \alpha^{N-2} \tilde{J}(x_{N-2}) + \sum_{k=0}^{N-3} \alpha^k g(x_k, \tilde{\mu}_k(x_k), w_k) \right\} + \alpha^{N-2} \epsilon_{N-2} \\
&\quad + \alpha^{N-1} \epsilon_{N-1} \\
&\leq \tilde{J}(x_0) + \sum_{k=0}^{N-1} \alpha^k \epsilon_k.
\end{aligned}$$

Combining these inequalities, we obtain

$$J^*(x_0) \leq \tilde{J}(x_0) + \lim_{N \rightarrow \infty} \left(\alpha^N r + \sum_{k=0}^{N-1} \alpha^k \epsilon_k \right).$$

Since $\{\epsilon_k\}$ is an arbitrary positive sequence, we may select $\{\epsilon_k\}$ so that $\lim_{N \rightarrow \infty} \sum_{k=0}^{N-1} \alpha^k \epsilon_k$ is arbitrarily close to zero, and the result follows.

(b) Under Assumption N, let r be a scalar such that $\tilde{J}(x) + r \leq 0$ for all $x \in S$, and if $\alpha \geq 1$, let $r = 0$. We have for every initial state $x_0 \in S$,

$$\begin{aligned}
J^*(x_0) &= \min_{\pi} \lim_{N \rightarrow \infty} E \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\} \\
&\geq \min_{\pi} \limsup_{N \rightarrow \infty} E \left\{ \alpha^N (\tilde{J}(x_N) + r) + \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\} \\
&\geq \limsup_{N \rightarrow \infty} \min_{\pi} E \left\{ \alpha^N (\tilde{J}(x_N) + r) + \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\},
\end{aligned} \tag{3.8}$$

where the last inequality follows from the fact that for any sequence $\{h_N(\xi)\}$ of functions of a parameter ξ we have

$$\min_{\xi} \limsup_{N \rightarrow \infty} h_N(\xi) \geq \limsup_{N \rightarrow \infty} \min_{\xi} h_N(\xi).$$

This inequality follows by writing

$$h_N(\xi) \geq \min_{\xi} h_N(\xi)$$

and by subsequently taking the \limsup of both sides and the minimum over ξ of the left-hand side.

Now we have, by using the assumption $\tilde{J} \leq T\tilde{J}$,

$$\begin{aligned} \min_{\pi} E & \left\{ \alpha^N \tilde{J}(x_N) + \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\} \\ & = \min_{\pi} E \left\{ \alpha^{N-1} \min_{u_{N-1} \in U(x_{N-1})} E_{w_{N-1}} \{ g(x_{N-1}, u_{N-1}, w_{N-1}) \right. \\ & \quad \left. + \alpha \tilde{J}(f(x_{N-1}, u_{N-1}, w_{N-1})) \} \right. \\ & \quad \left. + \sum_{k=0}^{N-2} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\} \\ & \geq \min_{\pi} E \left\{ \alpha^{N-1} \tilde{J}(x_{N-1}) + \sum_{k=0}^{N-2} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\} \\ & \quad \vdots \\ & \geq \tilde{J}(x_0). \end{aligned}$$

Using this relation in Eq. (3.8), we obtain

$$J^*(x_0) \geq \tilde{J}(x_0) + \lim_{N \rightarrow \infty} \alpha^N r = \tilde{J}(x_0).$$

Q.E.D.

As before, we have the following corollary:

Corollary 3.1.2.1: Let μ be an admissible stationary policy.

- (a) Under Assumption P, if $\tilde{J} : S \mapsto (-\infty, \infty]$ satisfies $\tilde{J} \geq T_\mu \tilde{J}$ and either \tilde{J} is bounded below and $\alpha < 1$, or $\tilde{J} \geq 0$, then $\tilde{J} \geq J_\mu$.
- (b) Under Assumption N, if $\tilde{J} : S \mapsto [-\infty, \infty)$ satisfies $\tilde{J} \leq T_\mu \tilde{J}$ and either \tilde{J} is bounded above and $\alpha < 1$, or $\tilde{J} \leq 0$, then $\tilde{J} \leq J_\mu$.

Conditions for Optimality of a Stationary Policy

Under Assumption P, we have the same optimality condition as for discounted problems with bounded cost per stage.

Proposition 3.1.3: (Necessary and Sufficient Condition for Optimality under P) Let Assumption P hold. A stationary policy μ is optimal if and only if

$$TJ^* = T_\mu J^*.$$

Proof: If $TJ^* = T_\mu J^*$, Bellman's equation ($J^* = TJ^*$) implies that $J^* = T_\mu J^*$. From Cor. 3.1.2.1(a) we then obtain $J^* \geq J_\mu$, showing that μ is optimal. Conversely, if $J^* = J_\mu$, we have using Cor. 3.1.1.1, $TJ^* = J^* = J_\mu = T_\mu J_\mu = T_\mu J^*$. Q.E.D.

Note that when $U(x)$ is a finite set for every $x \in S$, the above proposition implies the existence of an optimal stationary policy under Assumption P. This need not be true under Assumption N (see the subsequent Example 3.4.4).

Unfortunately, the sufficiency part of the above proposition need not be true under Assumption N; i.e., we may have $TJ^* = T_\mu J^*$ while μ is not optimal. This is illustrated in the following example.

Example 3.1.3

Let $S = C = (-\infty, 0]$, $U(x) = C$ for all $x \in S$, and

$$g(x, u, w) = f(x, u, w) = u,$$

for all $(x, u, w) \in S \times C \times D$. Then $J^*(x) = -\infty$ for all $x \in S$, and every stationary policy μ satisfies the condition of the preceding proposition. On the other hand, when $\mu(x) = 0$ for all $x \in S$, we have $J_\mu(x) = 0$ for all $x \in S$, and hence μ is not optimal.

Under Assumption N, we have a different characterization of an optimal stationary policy.

Proposition 3.1.4: (Necessary and Sufficient Condition for Optimality under N) Let Assumption N hold. A stationary policy μ is optimal if and only if

$$TJ_\mu = T_\mu J_\mu. \quad (3.9)$$

Proof: If $TJ_\mu = T_\mu J_\mu$, then from Cor. 3.1.1.1 we have $J_\mu = T_\mu J_\mu$, so that J_μ is a fixed point of T . Then by Prop. 3.1.2, we have $J_\mu \leq J^*$, which

implies that μ is optimal. Conversely, if $J_\mu = J^*$, then $T_\mu J_\mu = J_\mu = J^* = TJ^* = TJ_\mu$. Q.E.D.

The interpretation of the preceding optimality condition is that persistently using μ is optimal if and only if this performs at least as well as using any $\bar{\mu}$ at the first stage and using μ thereafter. Under Assumption P this condition is not sufficient to guarantee optimality of the stationary policy μ , as the following example shows.

Example 3.1.4

Let $S = (-\infty, \infty)$, $U(x) = (0, 1]$ for all $x \in S$,

$$g(x, u, w) = |x|, \quad f(x, u, w) = \alpha^{-1}ux,$$

for all $(x, u, w) \in S \times C \times D$. Let $\mu(x) = 1$ for all $x \in S$. Then $J_\mu(x) = \infty$ if $x \neq 0$ and $J_\mu(0) = 0$. Furthermore, we have $J_\mu = T_\mu J_\mu = TJ_\mu$, as the reader can easily verify. It can also be verified that $J^*(x) = |x|$, and hence the stationary policy μ is not optimal.

The Value Iteration Method

We now turn to the question whether the DP algorithm converges to the optimal cost function J^* . Let J_0 be the zero function on S ,

$$J_0(x) = 0, \quad x \in S.$$

Then under Assumption P, we have

$$J_0 \leq TJ_0 \leq T^2J_0 \leq \cdots \leq T^k J_0 \leq \cdots,$$

while under Assumption N, we have

$$J_0 \geq TJ_0 \geq T^2J_0 \geq \cdots \geq T^k J_0 \geq \cdots$$

In either case the limit function

$$J_\infty(x) = \lim_{k \rightarrow \infty} (T^k J_0)(x), \quad x \in S,$$

is well defined, provided we allow the possibility that J_∞ can take the value ∞ (under Assumption P) or $-\infty$ (under Assumption N). The question is whether the value iteration method is valid in the sense

$$J_\infty = J^*.$$

This question is, of course, of computational interest, but it is also of analytical interest since, if we know that $J^* = \lim_{k \rightarrow \infty} T^k J_0$, we can infer properties of the unknown function J^* from properties of the k -stage optimal cost functions $T^k J_0$, which are defined in a concrete algorithmic manner.

We will show that $J_\infty = J^*$ under Assumption N. It turns out, however, that under Assumption P, we may have $J_\infty \neq J^*$ (see Exercise 3.1). We will later provide easily verifiable conditions guaranteeing that $J_\infty = J^*$ under Assumption P. We have the following proposition.

Proposition 3.1.5:

- (a) Let Assumption P hold and assume that

$$J_\infty(x) = (T J_\infty)(x), \quad x \in S.$$

Then if $J : S \mapsto \mathbb{R}$ is any bounded function and $\alpha < 1$, or otherwise if $J_0 \leq J \leq J^*$, we have

$$\lim_{k \rightarrow \infty} (T^k J)(x) = J^*(x), \quad x \in S. \quad (3.10)$$

- (b) Let Assumption N hold. Then if $J : S \mapsto \mathbb{R}$ is any bounded function and $\alpha < 1$, or otherwise if $J^* \leq J \leq J_0$, we have

$$\lim_{k \rightarrow \infty} (T^k J)(x) = J^*(x), \quad x \in S.$$

Proof: (a) Since under Assumption P, we have

$$J_0 \leq T J_0 \leq \cdots \leq T^k J_0 \leq \cdots \leq J^*,$$

it follows that $\lim_{k \rightarrow \infty} T^k J_0 = J_\infty \leq J^*$. Since J_∞ is also a fixed point of T by assumption, we obtain from Prop. 3.1.2(a) that $J^* \leq J_\infty$. It follows that

$$J_\infty = J^*,$$

and hence Eq. (3.10) is proved for the case $J = J_0$.

For the case where $\alpha < 1$ and J is bounded, let r be a scalar such that

$$J_0 - re \leq J \leq J_0 + re.$$

Applying T^k to this relation, we obtain

$$T^k J_0 - \alpha^k re \leq T^k J \leq T^k J_0 + \alpha^k re.$$

Since $T^k J_0$ converges to J^* , as shown earlier, this relation implies that $T^k J$ converges also to J^* .

In the case where $J_0 \leq J \leq J^*$, we have by applying T^k

$$T^k J_0 \leq T^k J \leq J^*, \quad k = 0, 1, \dots$$

Since $T^k J_0$ converges to J^* , so does $T^k J$.

(b) It was shown earlier [cf. Eq. (3.4)] that under Assumption N, we have

$$J_\infty(x) = \lim_{k \rightarrow \infty} (T^k J_0)(x) = J^*(x).$$

The proof from this point is identical to that for part (a). Q.E.D.

We now derive conditions guaranteeing that $J_\infty = TJ_\infty$ holds under Assumption P, which by Prop. 3.1.5 implies that $J_\infty = J^*$. We prove two propositions. The first admits an easy proof but requires a finiteness assumption on the control constraint set. The second is harder to prove but requires a weaker compactness assumption.

Proposition 3.1.6: Let Assumption P hold and assume that the control constraint set is finite for every $x \in S$. Then

$$J_\infty = TJ_\infty = J^*.$$

Proof: As shown in the proof of Prop. 3.1.5(a), we have for all k , $T^k J_0 \leq J_\infty \leq J^*$. Applying T to this relation, we obtain

$$\begin{aligned} (T^{k+1} J_0)(x) &= \min_{u \in U(x)} E_w \{g(x, u, w) + \alpha(T^k J_0)(f(x, u, w))\} \\ &\leq (TJ_\infty)(x), \end{aligned} \tag{3.11}$$

and by taking the limit as $k \rightarrow \infty$, it follows that

$$J_\infty \leq TJ_\infty.$$

Suppose that there existed a state $\tilde{x} \in S$ such that

$$J_\infty(\tilde{x}) < (TJ_\infty)(\tilde{x}). \tag{3.12}$$

Let u_k minimize in Eq. (3.11) when $x = \tilde{x}$. Since $U(\tilde{x})$ is finite, there must exist some $\tilde{u} \in U(\tilde{x})$ such that $u_k = \tilde{u}$ for all k in some infinite subset K of the positive integers. By Eq. (3.11) we have for all $k \in K$

$$\begin{aligned} (T^{k+1} J_0)(\tilde{x}) &= E_w \{g(\tilde{x}, \tilde{u}, w) + \alpha(T^k J_0)(f(\tilde{x}, \tilde{u}, w))\} \\ &\leq (TJ_\infty)(\tilde{x}). \end{aligned}$$

Taking the limit as $k \rightarrow \infty$, $k \in K$, we obtain

$$\begin{aligned} J_\infty(\tilde{x}) &= \mathbb{E}_w \{ g(\tilde{x}, \tilde{u}, w) + \alpha J_\infty(f(\tilde{x}, \tilde{u}, w)) \} \\ &\geq (TJ_\infty)(\tilde{x}) \\ &= \min_{u \in U(\tilde{x})} \mathbb{E}_w \{ g(\tilde{x}, u, w) + \alpha J_\infty(f(\tilde{x}, u, w)) \}. \end{aligned}$$

This contradicts Eq. (3.12), so we have $J_\infty(\tilde{x}) = (TJ_\infty)(\tilde{x})$. Q.E.D.

The following proposition strengthens Prop. 3.1.6 in that it requires a compactness rather than a finiteness assumption. We recall (see Appendix A of Vol. I) that a subset X of the n -dimensional Euclidean space \mathbb{R}^n is said to be *compact* if every sequence $\{x_k\}$ with $x_k \in X$ contains a subsequence $\{x_{k_j}\}_{j \in \mathbb{N}}$ that converges to a point $x \in X$. Equivalently, X is compact if and only if it is closed and bounded. The empty set is (trivially) considered compact. Given any collection of compact sets, their intersection is a compact set (possibly empty). Given a sequence of nonempty compact sets $X_1, X_2, \dots, X_k, \dots$ such that

$$X_1 \supset X_2 \supset \cdots \supset X_k \supset X_{k+1} \supset \cdots$$

their intersection $\cap_{k=1}^\infty X_k$ is both nonempty and compact. In view of this fact, it follows that if $f : \mathbb{R}^n \mapsto [-\infty, \infty]$ is a function such that the set

$$F_\lambda = \{x \in \mathbb{R}^n \mid f(x) \leq \lambda\} \quad (3.13)$$

is compact for every $\lambda \in \mathbb{R}$, then there exists a vector x^* minimizing f ; i.e., there exists an $x^* \in \mathbb{R}^n$ such that

$$f(x^*) = \min_{x \in \mathbb{R}^n} f(x).$$

To see this, take a sequence $\{\lambda_k\}$ such that $\lambda_k \rightarrow \min_{x \in \mathbb{R}^n} f(x)$ and $\lambda_k \geq \lambda_{k+1}$ for all k . If $\min_{x \in \mathbb{R}^n} f(x) < \infty$, such a sequence exists and the sets

$$F_{\lambda_k} = \{x \in \mathbb{R}^n \mid f(x) \leq \lambda_k\}$$

are nonempty and compact. Furthermore, $F_{\lambda_k} \supset F_{\lambda_{k+1}}$ for all k , and hence the intersection $\cap_{k=1}^\infty F_{\lambda_k}$ is also nonempty and compact. Let x^* be any vector in $\cap_{k=1}^\infty F_{\lambda_k}$. Then

$$f(x^*) \leq \lambda_k, \quad k = 1, 2, \dots,$$

and taking the limit as $k \rightarrow \infty$, we obtain $f(x^*) \leq \min_{x \in \mathbb{R}^n} f(x)$, proving that x^* minimizes $f(x)$. The most common case where we can guarantee

that the set F_λ of Eq. (3.13) is compact for all λ is when f is continuous and $f(x) \rightarrow \infty$ as $\|x\| \rightarrow \infty$.

Proposition 3.1.7: Let Assumption P hold, and assume that the sets

$$U_k(x, \lambda) = \left\{ u \in U(x) \mid \underset{w}{E} \{ g(x, u, w) + \alpha(T^k J_0)(f(x, u, w)) \} \leq \lambda \right\} \quad (3.14)$$

are compact subsets of a Euclidean space for every $x \in S$, $\lambda \in \mathfrak{R}$, and for all k greater than some integer \bar{k} . Then

$$J_\infty = TJ_\infty = J^*. \quad (3.15)$$

Furthermore, there exists a stationary optimal policy.

Proof: As in Prop. 3.1.6, we have $J_\infty \leq TJ_\infty$. Suppose that there existed a state $\tilde{x} \in S$ such that

$$J_\infty(\tilde{x}) < (TJ_\infty)(\tilde{x}). \quad (3.16)$$

Clearly, we must have $J_\infty(\tilde{x}) < \infty$. For every $k \geq \bar{k}$, consider the sets

$$\begin{aligned} U_k(\tilde{x}, J_\infty(\tilde{x})) \\ = \left\{ u \in U(\tilde{x}) \mid \underset{w}{E} \{ g(\tilde{x}, u, w) + \alpha(T^k J_0)(f(\tilde{x}, u, w)) \} \leq J_\infty(\tilde{x}) \right\}. \end{aligned}$$

Let also u_k be a point attaining the minimum in

$$(T^{k+1} J_0)(\tilde{x}) = \min_{u \in U(\tilde{x})} \underset{w}{E} \{ g(\tilde{x}, u, w) + \alpha(T^k J_0)(f(\tilde{x}, u, w)) \};$$

i.e., u_k is such that

$$(T^{k+1} J_0)(\tilde{x}) = \underset{w}{E} \{ g(\tilde{x}, u_k, w) + \alpha(T^k J_0)(f(\tilde{x}, u_k, w)) \}.$$

Such minimizing points u_k exist by our compactness assumption. For every $k \geq \bar{k}$, consider the sequence $\{u_i\}_{i=k}^\infty$. Since $T^k J_0 \leq T^{k+1} J_0 \leq \dots \leq J_\infty$, it follows that

$$\begin{aligned} & \underset{w}{E} \{ g(\tilde{x}, u_i, w) + \alpha(T^k J_0)(\tilde{f}(\tilde{x}, u_i, w)) \} \\ & \leq \underset{w}{E} \{ g(\tilde{x}, u_i, w) + \alpha(T^i J_0)(f(\tilde{x}, u_i, w)) \} \\ & \leq J_\infty(\tilde{x}), \quad i \geq k. \end{aligned}$$

Therefore $\{u_i\}_{i=k}^{\infty} \subset U_k(\tilde{x}, J_{\infty}(\tilde{x}))$, and since $U_k(\tilde{x}, J_{\infty}(\tilde{x}))$ is compact, all the limit points of $\{u_i\}_{i=k}^{\infty}$ belong to $U_k(\tilde{x}, J_{\infty}(\tilde{x}))$ and at least one such limit point exists. Hence the same is true of the limit points of the whole sequence $\{u_i\}_{i=\bar{k}}^{\infty}$. It follows that if \tilde{u} is a limit point of $\{u_i\}_{i=\bar{k}}^{\infty}$ then

$$\tilde{u} \in \cap_{k=\bar{k}}^{\infty} U_k(\tilde{x}, J_{\infty}(\tilde{x})).$$

This implies by Eq. (3.14) that for all $k \geq \bar{k}$

$$J_{\infty}(\tilde{x}) \geq E_w \{g(\tilde{x}, \tilde{u}, w) + \alpha(T^k J_0)(f(\tilde{x}, \tilde{u}, w))\} \geq (T^{k+1} J_0)(\tilde{x}).$$

Taking the limit as $k \rightarrow \infty$, we obtain

$$J_{\infty}(\tilde{x}) = E_w \{g(\tilde{x}, \tilde{u}, w) + \alpha J_{\infty}(f(\tilde{x}, \tilde{u}, w))\}.$$

Since the right-hand side is greater than or equal to $(T J_{\infty})(\tilde{x})$, Eq. (3.16) is contradicted. Hence $J_{\infty} = T J_{\infty}$ and Eq. (3.15) is proved in view of Prop. 3.1.5(a).

To show that there exists an optimal stationary policy, observe that Eq. (3.15) and the last relation imply that \tilde{u} attains the minimum in

$$J^*(\tilde{x}) = \min_{u \in U(\tilde{x})} E_w \{g(\tilde{x}, u, w) + \alpha J^*(f(\tilde{x}, u, w))\}$$

for a state $\tilde{x} \in S$ with $J^*(\tilde{x}) < \infty$. For states $\tilde{x} \in S$ such that $J^*(\tilde{x}) = \infty$, every $u \in U(\tilde{x})$ attains the preceding minimum. Hence by Prop. 3.1.3(a) an optimal stationary policy exists. **Q.E.D.**

The reader may verify by inspection of the preceding proof that if $\mu_k(\tilde{x})$, $k = 0, 1, \dots$, attains the minimum in the relation

$$(T^{k+1} J_0)(\tilde{x}) = \min_{u \in U(\tilde{x})} E_w \{g(\tilde{x}, u, w) + \alpha(T^k J_0)(f(\tilde{x}, u, w))\},$$

then if $\mu^*(\tilde{x})$ is a limit point of $\{\mu_k(\tilde{x})\}$, for every $\tilde{x} \in S$, the stationary policy μ^* is optimal. Furthermore, $\{\mu_k(\tilde{x})\}$ has at least one limit point for every $\tilde{x} \in S$ for which $J^*(\tilde{x}) < \infty$. Thus the value iteration method under the assumptions of either Prop. 3.1.6 or Prop. 3.1.7 yields in the limit not only the optimal cost function J^* but also an optimal stationary policy.

Other Computational Methods

Unfortunately, policy iteration is not a valid procedure under either P or N in the absence of further conditions. If μ and $\bar{\mu}$ are stationary policies such that $T_{\bar{\mu}} J_{\mu} = T J_{\mu}$, then it can be shown that under Assumption P we have

$$J_{\bar{\mu}}(x) \leq J_{\mu}(x), \quad x \in S. \quad (3.17)$$

To see this, note that $T_{\bar{\mu}} J_{\mu} = T J_{\mu} \leq T_{\mu} J_{\mu} = J_{\mu}$ from which we obtain $\lim_{N \rightarrow \infty} T_{\bar{\mu}}^N J_{\mu} \leq J_{\mu}$. Since $J_{\bar{\mu}} = \lim_{N \rightarrow \infty} T_{\bar{\mu}}^N J_0$ and $J_0 \leq J_{\mu}$, we obtain $J_{\bar{\mu}} \leq J_{\mu}$. However, $J_{\bar{\mu}} \leq J_{\mu}$ by itself is not sufficient to guarantee the validity of policy iteration. For example, it is not clear that strict inequality holds in Eq. (3.17) for at least one state $x \in S$ when μ is not optimal. The difficulty here is that the equality $J_{\mu} = T J_{\mu}$ does not imply that μ is optimal, and additional conditions are needed to guarantee the validity of policy iteration. However, for special cases such conditions can be verified (see for example Section 3.2 and Exercise 3.16).

It is possible to devise a computational method based on mathematical programming when S , C , and D are finite sets by making use of Prop. 3.1.2. Under N and $\alpha = 1$, the corresponding (linear) program is (compare with Section 1.3.4)

$$\begin{aligned} & \text{maximize} \quad \sum_{i=1}^n \lambda_i \\ & \text{subject to} \quad \lambda_i \leq g(i, u) + \sum_{j=1}^n p_{ij}(u) \lambda_j, \quad i = 1, 2, \dots, n, \quad u \in U(i). \end{aligned}$$

When $\alpha = 1$ and Assumption P holds, the corresponding program takes the form

$$\begin{aligned} & \text{minimize} \quad \sum_{i=1}^n \lambda_i \\ & \text{subject to} \quad \lambda_i \geq \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u) \lambda_j \right], \quad i = 1, \dots, n, \end{aligned}$$

but unfortunately this program is not linear or even convex.

3.2 LINEAR SYSTEMS AND QUADRATIC COST

Consider the case of the linear system

$$x_{k+1} = Ax_k + Bu_k + w_k, \quad k = 0, 1, \dots,$$

where $x_k \in \mathbb{R}^n$, $u_k \in \mathbb{R}^m$ for all k , and the matrices A , B are known. As in Sections 4.1 and 5.2 of Vol. I, we assume that the random disturbances w_k are independent with zero mean and finite second moments. The cost function is quadratic and has the form

$$J_{\pi}(x_0) = \lim_{N \rightarrow \infty} \underset{k=0,1,\dots,N-1}{E}_{w_k} \left\{ \sum_{k=0}^{N-1} \alpha^k (x'_k Q x_k + \mu_k(x_k)' R \mu_k(x_k)) \right\},$$

where $\alpha \in (0, 1)$, Q is a positive semidefinite symmetric $n \times n$ matrix, and R is a positive definite symmetric $m \times m$ matrix. Clearly, Assumption P of Section 3.1 holds.

Our approach will be to use the DP algorithm to obtain the functions TJ_0, T^2J_0, \dots , as well as the pointwise limit function $J_\infty = \lim_{k \rightarrow \infty} T^k J_0$. Subsequently, we show that J_∞ satisfies $J_\infty = TJ_\infty$ and hence, by Prop. 3.1.5(a), $J_\infty = J^*$. The optimal policy is then obtained from the optimal cost function J^* by minimizing in Bellman's equation (cf. Prop. 3.1.3).

As in Section 4.1 of Vol. I, we have

$$J_0(x) = 0, \quad x \in \mathbb{R}^n,$$

$$(TJ_0)(x) = \min_u [x'Qx + u'Ru] = x'Qx, \quad x \in \mathbb{R}^n,$$

$$\begin{aligned} (T^2J_0)(x) &= \min_u E\{x'Qx + u'Ru + \alpha(Ax + Bu + w)'Q(Ax + Bu + w)\} \\ &= x'K_1x + \alpha E\{w'Qw\}, \quad x \in \mathbb{R}^n, \end{aligned}$$

$$(T^{k+1}J_0)(x) = x'K_kx + \sum_{m=0}^{k-1} \alpha^{k-m} E\{w'K_mw\}, \quad x \in \mathbb{R}^n, \quad k = 1, 2, \dots,$$

where the matrices K_0, K_1, K_2, \dots are given recursively by

$$K_0 = Q,$$

$$K_{k+1} = A'(\alpha K_k - \alpha^2 K_k B (\alpha B' K_k B + R)^{-1} B' K_k) A + Q, \quad k = 0, 1, \dots$$

By defining $\tilde{R} = R/\alpha$ and $\tilde{A} = \sqrt{\alpha}A$, the preceding equation may be written as

$$K_{k+1} = \tilde{A}'(K_k - K_k B (\tilde{B}' K_k \tilde{B} + \tilde{R})^{-1} \tilde{B}' K_k) \tilde{A} + Q,$$

and is of the form considered in Section 4.1 of Vol. I. By using the result shown there, we have that the generated matrix sequence $\{K_k\}$ converges to a positive definite symmetric matrix K ,

$$K_k \rightarrow K,$$

provided the pairs (\tilde{A}, B) and (\tilde{A}, C) , where $Q = C'C$, are controllable and observable, respectively. Since $\tilde{A} = \sqrt{\alpha}A$, controllability and observability of (A, B) or (A, C) are clearly equivalent to controllability and observability of (\tilde{A}, B) or (\tilde{A}, C) , respectively. The matrix K is the unique solution of the equation

$$K = A'(\alpha K - \alpha^2 K B (\alpha B' K B + R)^{-1} B' K) A + Q \quad (3.18)$$

Because $K_k \rightarrow K$, it can also be seen that the limit

$$c = \lim_{k \rightarrow \infty} \sum_{m=0}^{k-1} \alpha^{k-m} E\{w' K_m w\}$$

is well defined, and in fact

$$c = \frac{\alpha}{1 - \alpha} E\{w' K w\}. \quad (3.19)$$

Thus, in conclusion, if the pairs (A, B) and (A, C) are controllable and observable, respectively, the limit of the functions $T^k J_0$ is given by

$$J_\infty(x) = \lim_{k \rightarrow \infty} (T^k J_0)(x) = x' K x + c. \quad (3.20)$$

Using Eqs. (3.18)-(3.20), it can be verified by straightforward calculation that for all $x \in S$

$$J_\infty(x) = (T J_\infty)(x) = \min_u [x' Q x + u' R u + \alpha E\{J_\infty(Ax + Bu + w)\}] \quad (3.21)$$

and hence, by Prop. 3.1.5(a), $J_\infty = J^*$. Another way to prove that $J_\infty = T J_\infty$ is to show that the assumption of Prop. 3.1.7, is satisfied; i.e., the sets

$$U_k(x, \lambda) = \{u \mid E\{x' Q x + u' R u + \alpha(T^k J_0)(Ax + Bu + w)\} \leq \lambda\}$$

are compact for all k and scalars λ . This can be verified using the fact that $T^k J_0$ is a positive semidefinite quadratic function and R is positive definite. The optimal stationary policy μ^* , obtained by minimization in Eq. (3.21), has the form

$$\mu^*(x) = -\alpha(\alpha B' K B + R)^{-1} B' K A x, \quad x \in \mathbb{R}^n.$$

This policy is attractive for practical implementation since it is linear and stationary. A number of generalized versions of the problem of this section, including the case of imperfect state information, are treated in the exercises. Interestingly, the problem can be solved by policy iteration (see Exercise 3.16), even though, as discussed in Section 3.1, policy iteration is not valid in general under Assumption P.

3.3 INVENTORY CONTROL

Let us consider a discounted, infinite horizon version of the inventory control problem of Section 4.2 in Vol. I. Inventory stock evolves according to the equation

$$x_{k+1} = x_k + u_k - w_k, \quad k = 0, 1, \dots$$

We assume that the successive demands w_k are independent and bounded, and have identical probability distributions. We also assume for simplicity that there is no fixed cost. The case of a nonzero fixed cost can be treated similarly. The cost function is

$$J_\pi(x_0) = \lim_{N \rightarrow \infty} E_{w_k \sim \mu_{x_k}} \left\{ \sum_{k=0}^{N-1} \alpha^k (c\mu_k(x_k) + H(x_k + \mu(x_k) - w_k)) \right\},$$

where

$$H(y) = p \max(0, -y) + h \max(0, y).$$

The DP algorithm is given by

$$J_0(x) = 0,$$

$$(T^{k+1} J_0)(x) = \min_{0 \leq u} E \{ c u + H(x + u - w) + \alpha (T^k J_0)(x + u - w) \}.$$

We first show that the optimal cost is finite for all initial states:

$$J^*(x_0) = \min_{\pi} J_\pi(x_0) < \infty, \quad \text{for all } x_0 \in S.$$

Indeed, consider the policy $\tilde{\pi} = \{\tilde{\mu}, \tilde{\mu}, \dots\}$, where $\tilde{\mu}$ is defined by

$$\tilde{\mu}(x) = \begin{cases} 0 & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases}$$

Since w_k is nonnegative and bounded, it follows that the inventory stock x_k when the policy $\tilde{\pi}$ is used satisfies

$$-w_{k-1} \leq x_k \leq \max(0, x_0), \quad k = 1, 2, \dots,$$

and is bounded. Hence $\tilde{\mu}(x_k)$ is also bounded. It follows that the cost per stage incurred when $\tilde{\pi}$ is used is bounded, and in view of the presence of the discount factor we have

$$J_{\tilde{\pi}}(x_0) < \infty, \quad x_0 \in S.$$

Since $J^* \leq J_{\tilde{\pi}}$, the finiteness of the optimal cost follows.

Next we observe that, under the assumption $c < p$, the functions $T^k J_0$ are real-valued and convex. Indeed, we have

$$J_0 \leq T J_0 \leq \dots \leq T^k J_0 \leq \dots \leq J^*,$$

which implies that $T^k J_0$ is real-valued. Convexity follows by induction as shown in Section 4.2 of Vol. I.

Consider now the sets

$$U_k(x, \lambda) = \{u \geq 0 \mid E\{cu + H(x+u-w) + \alpha(T^k J_0)(x_u - w)\} \leq \lambda\}. \quad (3.22)$$

These sets are bounded since the expected value within the braces above tends to ∞ as $u \rightarrow \infty$. Also, the sets $U_k(x, \lambda)$ are closed since the expected value in Eq. (3.22) is a continuous function of u [recall that $T^k J_0$ is a real-valued convex and hence continuous function]. Thus we may invoke Prop. 3.1.7 and assert that

$$\lim_{k \rightarrow \infty} (T^k J_0)(x) = J^*(x), \quad x \in S.$$

It follows from the convexity of the functions $T^k J_0$ that the limit function J^* is a real-valued convex function. Furthermore, an optimal stationary policy μ^* can be obtained by minimizing in the right-hand side of Bellman's equation

$$J^*(x) = \min_{u \geq 0} E\{cu + H(x+u-w) + \alpha J^*(x+u-w)\}.$$

We have

$$\mu^*(x) = \begin{cases} S^* - x & \text{if } x \leq S^*, \\ 0 & \text{otherwise,} \end{cases}$$

where S^* is a minimizing point of

$$G^*(y) = cy + L(y) + \alpha E\{J^*(y-w)\},$$

with

$$L(y) = E\{H(y-w)\}.$$

It can be seen that if $p > c$, we have $\lim_{|y| \rightarrow \infty} G^*(y) = \infty$, so that such a minimizing point exists. Furthermore, by using the observation made near the end of Section 3.1, it follows that a minimizing point S^* of $G^*(y)$ may be obtained as a limit point of a sequence $\{S_k\}$, where for each k the scalar S_k minimizes

$$G_k(y) = cy + L(y) + \alpha E\{(T^k J_0)(y-w)\}$$

and is obtained by means of the value iteration method.

It turns out that the critical level S^* has a simple characterization. It can be shown that S^* minimizes over y the expression $(1-\alpha)cy + L(y)$, and it can be essentially obtained in closed form (see Exercise 3.18, and Heyman and Sobel [HeS84], Ch. 2).

In the case where there is a positive fixed cost ($K > 0$), the same line of argument may be used. Similarly, we prove that J^* is a real-valued K -convex function. A separate argument is necessary to prove that J^* is also continuous (this is intuitively clear and is left for the reader). Once K -convexity and continuity of J^* are established, the optimality of a stationary (s^*, S^*) policy follows from the equation

$$J^*(x) = \min_{u \geq 0} E\{C(u) + H(x+u-w) + \alpha J^*(x+u-w)\},$$

where $C(u) = K + cu$ if $u > 0$ and $C(0) = 0$.

3.4 OPTIMAL STOPPING

Consider an infinite horizon version of the stopping problems of Section 4.4 of Vol. I. At each state x , we must choose between two actions: pay a cost $s(x)$ and *stop* with no further cost incurred, or pay a cost $c(x)$ and *continue* the process according to the system equation

$$x_{k+1} = f_c(x_k, w_k), \quad k = 0, 1, \dots \quad (3.23)$$

The objective is to find the optimal stopping policy that minimizes the total expected cost over an infinite number of stages. It is assumed that the input disturbances w_k have the same probability distribution for all k , which depends only on the current state x_k .

This problem may be viewed as a special case of the stochastic shortest path problem of Section 2.1, but here we will not assume that the state space is finite and that only proper policies can be optimal, as we did in Section 2.1. Instead we will rely on the general theory of unbounded cost problems developed in Section 3.1.

To put the problem within the framework of the total cost infinite horizon problem, we introduce an additional state t (termination state) and we complete the system equation (3.23) as in Section 4.4 of Vol. I by letting

$$x_{k+1} = t, \quad \text{if } u_k = \text{stop or } x_k = t.$$

Once the system reaches the termination state, it remains there permanently at no cost.

We first assume that

$$s(x) \geq 0, \quad c(x) \geq 0, \quad \text{for all } x \in S, \quad (3.24)$$

thus coming under the framework of Assumption P of Section 3.1. The case corresponding to Assumption N, where $s(x) \leq 0$ and $c(x) \leq 0$ for all $x \in S$ will be considered later. Actually, whenever there exists an $\epsilon > 0$ such that $c(x) \geq \epsilon$ for all $x \in S$, the results to be obtained under the assumption (3.24) apply also to the case where $s(x)$ is bounded below by some scalar rather than bounded by zero. The reason is that, if $c(x)$ is assumed to be greater than $\epsilon > 0$ for all $x \in S$, any policy that will not stop within a finite expected number of stages results in infinite cost and can be excluded from consideration. As a result, if we reformulate the problem and add a constant r to $s(x)$ so that $s(x) + r \geq 0$ for all $x \in S$, the optimal cost $J^*(x)$ will merely be increased by r , while optimal policies will remain unaffected.

The mapping T that defines the DP algorithm takes the form

$$(TJ)(x) = \begin{cases} \min[s(x), c(x) + E\{J(f_c(x, w))\}] & \text{if } x \neq t, \\ 0 & \text{if } x = t, \end{cases} \quad (3.25)$$

where $s(x)$ is the cost of the stopping action, and $c(x) + E\{J(f_c(x, w))\}$ is the cost of the continuation action. Since the control space has only two elements, by Prop. 3.1.6, we have

$$\lim_{k \rightarrow \infty} (T^k J_0)(x) = J^*(x), \quad x \in S,$$

where J_0 is the zero function [$J_0(x) = 0$, for all $x \in S$]. By Prop. 3.1.3, there exists a stationary optimal policy given by

$$\begin{aligned} \text{stop} & \quad \text{if } s(x) < c(x) + E\{J^*(f_c(x, w))\}, \\ \text{continue} & \quad \text{if } s(x) \geq c(x) + E\{J^*(f_c(x, w))\}. \end{aligned}$$

Let us denote by S^* the optimal stopping set (which may be empty)

$$S^* = \{x \in S \mid s(x) < c(x) + E\{J^*(f_c(x, w))\}\}.$$

Consider also the sets

$$S_k = \{x \in S \mid s(x) < c(x) + E\{(T^k J_0)(f_c(x, w))\}\}$$

that determine the optimal policy for finite horizon versions of the stopping problem. Since we have

$$J_0 \leq T J_0 \leq \cdots \leq T^k J_0 \leq \cdots \leq J^*,$$

it follows that

$$S_1 \subset S_2 \subset \cdots \subset S_k \subset \cdots \subset S^*$$

and therefore $\cup_{k=1}^{\infty} S_k \subset S^*$. Also, if $\tilde{x} \notin \cup_{k=1}^{\infty} S_k$, then we have

$$s(\tilde{x}) \geq c(\tilde{x}) + E\{(T^k J_0)(f_c(\tilde{x}, w))\}, \quad k = 0, 1, \dots$$

By taking the limit as $k \rightarrow \infty$, and by using the monotone convergence theorem and the fact $T^k J_0 \rightarrow J^*$, we obtain

$$s(\tilde{x}) \geq c(\tilde{x}) + E\{J^*(f_c(\tilde{x}, w))\},$$

from which $\tilde{x} \notin S^*$. Hence

$$S^* = \cup_{k=1}^{\infty} S_k.$$

In other words, the *optimal stopping set* S^* for the infinite horizon problem is equal to the union of all the finite horizon stopping sets S_k .

Consider now, as in Section 4.4 of Vol. I, the one-step-to-go stopping set

$$\tilde{S}_1 = \{x \in S \mid s(x) \leq c(x) + E\{t(f_c(x, w))\}\} \tag{3.26}$$

and assume that \tilde{S}_1 is *absorbing* in the sense

$$f_c(x, w) \in \tilde{S}_1, \quad \text{for all } x \in \tilde{S}_1, \quad w \in D. \tag{3.27}$$

Then, as in Section 4.4 of Vol. I, it follows that the one-step lookahead policy

$$\text{stop if and only if } x \in \tilde{S}_1$$

is optimal. We now provide some examples.

Example 3.4.1 (Asset Selling)

Consider the version of the asset selling example of Sections 4.4 and 7.3 of Vol. I, where the rate of interest r is zero and there is instead a maintenance cost $c > 0$ per period for which the house remains unsold. Furthermore, past offers can be accepted at any future time. We have the following optimality equation:

$$J^*(x) = \max[x, -c + E\{\max(x, w)\}].$$

In this case we consider maximization of total expected reward, the continuation cost is strictly negative, and the stopping reward x is positive. Hence the assumption (3.24) is not satisfied. If, however, we assume that x takes values in a bounded interval $[0, M]$, where M is an upper bound on the possible values of offers, our analysis is still applicable [cf. the discussion following Eq. (3.24)]. Consider the one-step-to-go stopping set given by

$$\tilde{S}_1 = \{x \mid x \geq -c + E\{\max(x, w)\}\}.$$

After a calculation similar to the one given in Section 4.4 of Vol. I, we see that

$$\tilde{S}_1 = \{x \mid x \geq \bar{a}\},$$

where \bar{a} is the scalar satisfying

$$\bar{a} = P(\bar{a})\bar{a} + \int_{\bar{a}}^{\infty} w \, dP(w) - c.$$

Clearly, \tilde{S}_1 is absorbing in the sense of Eq. (3.27), and therefore the one-step lookahead policy, which accepts the first offer that is greater or equal to \bar{a} is optimal.

Example 3.4.2 (Sequential Hypothesis Testing)

Consider the hypothesis testing problem of Example 5.4.4 of Vol. I for the case where the number of possible observations is unlimited. Here the states are x^0 and x^1 (true distribution of the observations is f_0 and f_1 , respectively). The set S is the interval $[0, 1]$ and corresponds to the sufficient statistic

$$p_k = P(x_k = x^0 \mid z_0, z_1, \dots, z_k).$$

To each $p \in [0, 1]$ we may assign the stopping cost

$$s(p) = \min[(1-p)L_0, pL_1],$$

i.e., the cost associated with optimal choice between the distributions f_0 and f_1 . The mapping T of Eq. (3.25) takes the form

$$(TJ)(p) = \min \left[(1-p)L_0, pL_1, c + E_z \left\{ J \left(\frac{pf_0(z)}{pf_0(z) + (1-p)f_1(z)} \right) \right\} \right]$$

for all $p \in [0, 1]$, where the expectation over z is taken with respect to the probability distribution

$$P(z) = pf_0(z) + (1-p)f_1(z), \quad z \in Z.$$

The optimal cost function J^* satisfies Bellman's equation

$$J^*(p) = \min \left[(1-p)L_0, pL_1, c + \underset{z}{E} \left\{ J^* \left(\frac{pf_0(z)}{pf_0(z) + (1-p)f_1(z)} \right) \right\} \right]$$

and is obtained in the limit through the equation

$$J^*(p) = \lim_{k \rightarrow \infty} (T^k J_0)(p), \quad p \in [0, 1],$$

where J_0 is the zero function on $[0, 1]$.

Now consider the functions $T^k J_0$, $k = 0, 1, \dots$. It is clear that

$$J_0 \leq TJ_0 \leq \dots \leq T^k J_0 \leq \dots \leq \min[(1-p)L_0, pL_1].$$

Furthermore, in view of the analysis of Section 5.5 of Vol. I, we have that the function $T^k J_0$ is concave on $[0, 1]$ for all k . Hence the pointwise limit function J^* is also concave on $[0, 1]$. In addition, Bellman's equation implies that

$$J^*(0) = J^*(1) = 0,$$

$$J^*(p) \leq \min[(1-p)L_0, pL_1].$$

Using the reasoning illustrated in Fig. 3.4.1 it follows that [provided $c < L_0 L_1 / (L_0 + L_1)$] there exist two scalars $\bar{\alpha}$, $\bar{\beta}$ with $0 < \bar{\beta} \leq \bar{\alpha} < 1$, that determine an optimal stationary policy of the form

$$\begin{aligned} \text{accept } f_0 &\quad \text{if } p \leq \bar{\alpha}, \\ \text{accept } f_1 &\quad \text{if } p \leq \bar{\beta}, \\ \text{continue the observations} &\quad \text{if } \bar{\beta} < p < \bar{\alpha}. \end{aligned}$$

In view of the optimality of the preceding stationary policy, the sequential probability ratio test described in Section 5.5 of Vol. I is justified when the number of possible observations is infinite.

The Case of Negative Transition-Costs

We now consider the stopping problem under Assumption N, i.e.,

$$s(x) \leq 0, \quad c(x) \leq 0, \quad \text{for all } x \in S.$$

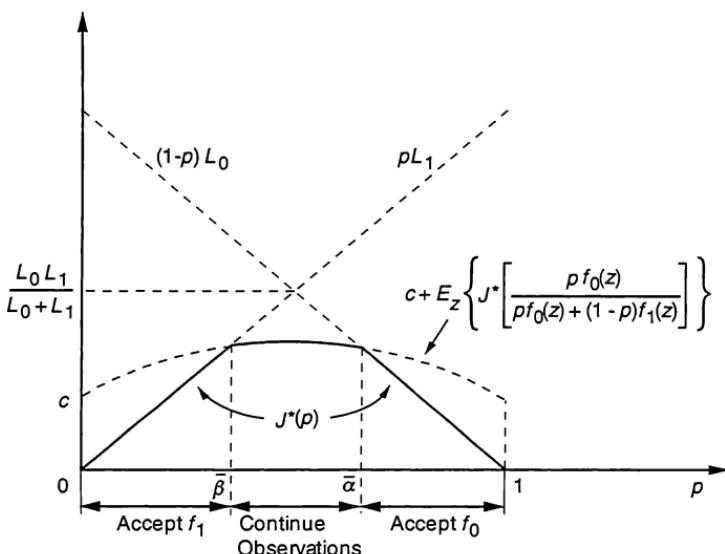


Figure 3.4.1 Derivation of the sequential probability ratio test.

Under these circumstances there is no penalty for continuing operation of the system (although by not stopping at a given state, a favorable opportunity may be missed). The mapping T is given by

$$(TJ)(x) = \min[s(x), c(x) + E\{J(f_c(x, w))\}].$$

The optimal cost function J^* satisfies

$$J^*(x) \leq s(x), \quad x \in S,$$

and by using Props. 3.1.1 and 3.1.5(b), we have

$$J^* = TJ^*, \quad J^* = \lim_{k \rightarrow \infty} T^k J_0 = \lim_{k \rightarrow \infty} T^k s,$$

where J_0 is the zero function. It can also be seen that if the one-step-to-go stopping set \tilde{S}_1 is *absorbing* [cf. Eq. (3.27)], a one-step lookahead policy is optimal.

Example 3.4.3 (The Rational Burglar)

This example was considered at the end of Section 4.4 of Vol. I where it was shown that a one-step lookahead policy is optimal for any finite horizon length. The optimality equation is

$$J^*(x) = \max[x, (1-p)E\{J^*(x+w)\}].$$

The problem is equivalent to a minimization problem where

$$s(x) = -x, \quad c(x) = 0,$$

so Assumption N holds. From the preceding analysis, we have that $T^k s \rightarrow J^*$ and that a one-step lookahead policy is optimal if the one-step stopping set is absorbing [cf. Eqs. (3.26) and (3.27)]. It can be shown (see the analysis of Section 4.4 of Vol. I) that this condition holds, so the finite horizon optimal policy whereby the burglar retires when his accumulated earnings reach or exceed $(1-p)\bar{w}/p$ is optimal for an infinite horizon as well.

Example 3.4.4 (A Problem with no Optimal Policy)

This is a deterministic stopping problem where Assumption N holds, and an optimal policy does not exist, even though only two controls are available at each state (stop and continue). The states are the positive integers, and continuation from state i leads to state $i + 1$ with certainty and no cost, i.e., $S = \{1, 2, \dots\}$, $c(i) = 0$, and $f_c(i, w) = i + 1$ for all $i \in S$ and $w \in D$. The stopping cost is $s(i) = -1 + (1/i)$ for all $i \in S$, so that there is an incentive to delay stopping at every state. We have $J^*(i) = -1$ for all i , and the optimal cost -1 can be approached arbitrarily closely by postponing the stopping action for a sufficiently long time. However, there does not exist an optimal policy that attains the optimal cost.

3.5 OPTIMAL GAMBLING STRATEGIES

A gambler enters a certain game played as follows. The gambler may stake at any time k any amount $u_k \geq 0$ that does not exceed his current fortune x_k (defined to be his initial capital plus his gain or minus his loss thus far). He wins his stake back and as much more with probability p and he loses his stake with probability $(1-p)$. Thus the gambler's fortune evolves according to the equation

$$x_{k+1} = x_k + w_k u_k, \quad k = 0, 1, \dots, \quad (3.28)$$

where $w_k = 1$ with probability p and $w_k = -1$ with probability $(1-p)$. Several games, such as playing red and black in roulette, fit this description.

The gambler enters the game with an initial capital x_0 , and his goal is to increase his fortune up to a level X . He continues gambling until he either reaches his goal or loses his entire initial capital, at which point he leaves the game. The problem is to determine the optimal gambling strategy for maximizing the probability of reaching his goal. By a gambling strategy, we mean a rule that specifies what the stake should be at time k when the gambler's fortune is x_k , for every x_k with $0 < x_k < X$.

The problem may be cast within the total cost, infinite horizon framework, where we consider maximization in place of minimization. Let us assume for convenience that fortunes are normalized so that $X = 1$. The state space is the set $[0, 1] \cup \{t\}$, where t is a termination state to which the system moves with certainty from both states 0 and 1 with corresponding rewards 0 and 1. When $x_k \neq 0, x_k \neq 1$, the system evolves according to Eq. (3.28). The control constraint set is specified by

$$0 \leq u_k \leq x_k, \quad 0 \leq u_k \leq 1 - x_k.$$

The reward per stage when $x_k \neq 0$ and $x_k \neq 1$ is zero. Under these circumstances the probability of reaching the goal is equal to the total expected reward. Assumption N holds since our problem is equivalent to a problem of minimizing expected total cost with nonpositive costs per stage.

The mapping T defining the DP algorithm takes the form

$$(TJ)(x) = \begin{cases} \max_{\substack{0 \leq u \leq x \\ 0 \leq u \leq 1-x}} [pJ(x+u) + (1-p)J(x-u)] & \text{if } x \in (0, 1), \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x = 1, \end{cases}$$

for any function $J : [0, 1] \mapsto [0, \infty]$.

Consider now the case where

$$0 < p < \frac{1}{2},$$

i.e., the game is unfair to the gambler. A discretized version of the case where $1/2 \leq p < 1$ is considered in Exercise 3.21. When $0 < p < 1/2$, it is intuitively clear that if the gambler follows a very conservative strategy and stakes a very small amount at each time, he is all but certain to lose his capital. For example, if the gambler adopts a strategy of betting $1/n$ at each time, then it may be shown (see Exercise 3.21 or Ash [Ash70], p. 182) that his probability of attaining the target fortune of 1 starting with an initial capital i/n , $0 < i < n$, is given by

$$\left(\left(\frac{1-p}{p} \right)^i - 1 \right) \left(\left(\frac{1-p}{p} \right)^n - 1 \right)^{-1}$$

If $0 < p < 1/2$, n tends to infinity, and i/n tends to a constant, the above probability tends to zero, thus indicating that placing consistently small bets is a bad strategy.

[↑] We are thus led to a policy that places large bets and, in particular, the *bold strategy* whereby the gambler stakes at each time k his entire fortune x_k or just enough to reach his goal, whichever is least. In other words, the bold strategy is the stationary policy μ^* given by

$$\mu^*(x) = \begin{cases} x & \text{if } 0 < x \leq 1/2, \\ 1-x & \text{if } 1/2 \leq x < 1. \end{cases}$$

We will prove that the bold strategy is indeed an optimal policy. To this end it is sufficient to show that for every initial fortune $x \in [0, 1]$ the value of the reward function $J_{\mu^*}(x)$ corresponding to the bold strategy μ^* satisfies the sufficiency condition (cf. Prop. 3.1.4)

$$TJ_{\mu^*} = J_{\mu^*},$$

or equivalently

$$J_{\mu^*}(0) = 0, \quad J_{\mu^*}(1) = 1,$$

$$J_{\mu^*}(x) \geq pJ_{\mu^*}(x+u) + (1-p)J_{\mu^*}(x-u),$$

for all $x \in (0, 1)$ and $u \in [0, x] \cap [0, 1-x]$.

By using the definition of the bold strategy, Bellman's equation

$$J_{\mu^*} = T_{\mu^*}J_{\mu^*},$$

is written as

$$J_{\mu^*}(0) = 0, \quad J_{\mu^*}(1) = 1, \tag{3.29}$$

$$J_{\mu^*}(x) = \begin{cases} pJ_{\mu^*}(2x) & \text{if } 0 < x \leq 1/2, \\ p + (1-p)J_{\mu^*}(2x-1) & \text{if } 1/2 \leq x < 1. \end{cases} \tag{3.30}$$

The following lemma shows that J_{μ^*} is uniquely defined from these relations.

Lemma 3.5.1: For every p , with $0 < p \leq 1/2$, there is only one bounded function on $[0, 1]$ satisfying Eqs. (3.29) and (3.30), the function J_{μ^*} . Furthermore, J_{μ^*} is continuous and strictly increasing on $[0, 1]$.

Proof: Suppose that there existed two bounded functions $J_1 : [0, 1] \mapsto \mathbb{R}$ and $J_2 : [0, 1] \mapsto \mathbb{R}$ such that $J_i(0) = 0$, $J_i(1) = 1$, $i = 1, 2$, and

$$J_i(x) = \begin{cases} pJ_i(2x) & \text{if } 0 < x \leq 1/2, \\ p + (1-p)J_i(2x-1) & \text{if } 1/2 \leq x < 1, \end{cases} \quad i = 1, 2.$$

Then we have

$$J_1(2x) - J_2(2x) = \frac{J_1(x) - J_2(x)}{p}, \quad \text{if } 0 \leq x \leq 1/2, \tag{3.31}$$

$$J_1(2x-1) - J_2(2x-1) = \frac{J_1(x) - J_2(x)}{1-p}, \quad \text{if } 1/2 \leq x \leq 1. \tag{3.32}$$

Let z be any real number with $0 \leq z \leq 1$. Define

$$\begin{aligned} z_1 &= \begin{cases} 2z & \text{if } 0 \leq z \leq 1/2, \\ 2z - 1 & \text{if } 1/2 < z \leq 1, \end{cases} \\ &\quad \vdots \\ z_k &= \begin{cases} 2z_{k-1} & \text{if } 0 \leq z_{k-1} \leq 1/2, \\ 2z_{k-1} - 1 & \text{if } 1/2 < z_{k-1} \leq 1, \end{cases} \end{aligned}$$

for $k = 1, 2, \dots$. Then from Eqs. (3.31) and (3.32) it follows (using $p \leq 1/2$) that

$$|J_1(z_k) - J_1(z_k)| \geq \frac{|J_1(z) - J_2(z)|}{(1-p)^k}, \quad k = 1, 2, \dots$$

Since $J_1(z_k) - J_2(z_k)$ is bounded, it follows that $J_1(z) - J_2(z) = 0$, for otherwise the right side of the inequality would tend to ∞ . Since $z \in [0, 1]$ is arbitrary, we obtain $J_1 = J_2$. Hence J_{μ^*} is the unique bounded function on $[0, 1]$ satisfying Eqs. (3.29) and (3.30).

To show that J_{μ^*} is strictly increasing and continuous, we consider the mapping T_{μ^*} , which operates on functions $J : [0, 1] \mapsto [0, 1]$ and is defined by

$$\begin{aligned} (T_{\mu^*} J)(x) &= \begin{cases} pJ(2x) + (1-p)J(0) & \text{if } 0 < x \leq 1/2, \\ pJ(1) + (1-p)J(2x-1) & \text{if } 1/2 \leq x < 1, \end{cases} \\ (T_{\mu^*} J)(0) &= 0, \quad (T_{\mu^*} J)(1) = 1. \end{aligned} \tag{3.33}$$

Consider the functions $J_0, T_{\mu^*}^* J_0, \dots, T_{\mu^*}^k J_0, \dots$, where J_0 is the zero function [$J_0(x) = 0$ for all $x \in [0, 1]$]. We have

$$J_{\mu^*}(x) = \lim_{k \rightarrow \infty} (T_{\mu^*}^k J_0)(x), \quad x \in [0, 1]. \tag{3.34}$$

Furthermore, the functions $T_{\mu^*}^k J_0$ can be shown to be monotonically nondecreasing in the interval $[0, 1]$. Hence, by Eq. (3.34), J_{μ^*} is also monotonically nondecreasing.

Consider now for $n = 0, 1, \dots$ the sets

$$S_n = \{x \in [0, 1] \mid x = k2^{-n}, k = \text{nonnegative integer}\}.$$

It is straightforward to verify that

$$(T_{\mu^*}^m J_0)(x) = (T_{\mu^*}^n J_0)(x), \quad x \in S_{n-1}, \quad m \geq n \geq 1.$$

As a result of this equality and Eq. (3.34),

$$J_{\mu^*}(x) = (T_{\mu^*}^n J_0)(x), \quad x \in S_{n-1}, \quad n \geq 1. \tag{3.35}$$

A further fact that may be verified by using induction and Eqs. (3.33) and (3.35) is that for any nonnegative integers k, n for which $0 \leq k2^{-n} < (k+1)2^{-n} \leq 1$, we have

$$p^n \leq J_{\mu^*}((k+1)2^{-n}) - J_{\mu^*}(k2^{-n}) \leq (1-p)^n. \quad (3.36)(5.11)$$

Since any number in $[0, 1]$ can be approximated arbitrarily closely from above and below by numbers of the form $k2^{-n}$, and since J_{μ^*} has been shown to be monotonically nondecreasing, it follows from Eq. (3.36) that J_{μ^*} is continuous and strictly increasing. Q.E.D.

We are now in a position to prove the following proposition.

Proposition 3.5.1: The bold strategy is an optimal stationary gambling policy.

Proof: We will prove the sufficiency condition

$$J_{\mu^*}(x) \geq pJ_{\mu^*}(x+u) + (1-p)J_{\mu^*}(x-u), \quad x \in [0, 1], \quad u \in [0, 1] \cap [0, 1-x]. \quad (3.37)$$

In view of the continuity of J_{μ^*} established in the previous lemma, it is sufficient to establish Eq. (3.37) for all $x \in [0, 1]$ and $u \in [0, x] \cap [0, 1-x]$ that belong to the union $\cup_{n=0}^{\infty} S_n$ of the sets S_n defined by

$$S_n = \{z \in [0, 1] \mid z = k2^{-n}, k = \text{nonnegative integer}\}.$$

We will use induction. By using the fact that $J_{\mu^*}(0) = 0$, $J_{\mu^*}(1/2) = p$, and $J_{\mu^*}(1) = 1$, we can show that Eq. (3.37) holds for all x and u in S_0 and S_1 . Assume that Eq. (3.37) holds for all $x, u \in S_n$. We will show that it holds for all $x, u \in S_{n+1}$.

For any $x, u \in S_{n+1}$ with $u \in [0, x] \cap [0, 1-x]$, there are four possibilities:

1. $x + u \leq 1/2$,
2. $x - u \geq 1/2$,
3. $x - u \leq x \leq 1/2 \leq x + u$,
4. $x - u \leq 1/2 \leq x \leq x + u$,

We will prove Eq. (3.37) for each of these cases.

Case 1. If $x, u \in S_{n+1}$, then $2x \in S_n$, and $2u \in S_n$, and by the induction hypothesis

$$J_{\mu^*}(2x) - pJ_{\mu^*}(2x + 2u) - (1-p)J_{\mu^*}(2x - 2u) \geq 0. \quad (3.38)$$

If $x + u \leq 1/2$, then by Eq. (3.30)

$$\begin{aligned} J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u) \\ = p(J_{\mu^*}(2x) - pJ_{\mu^*}(2x+2u) - (1-p)J_{\mu^*}(2x-2u)) \end{aligned}$$

and using Eq. (3.38), the desired relation Eq. (3.37) is proved for the case under consideration.

Case 2. If $x, u \in S_{n+1}$, then $(2x-1) \in S_n$ and $2u \in S_n$, and by the induction hypothesis

$$J_{\mu^*}(2x-1) - pJ_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u-1) \geq 0.$$

If $x - u \geq 1/2$, then by Eq. (3.30)

$$\begin{aligned} J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u) \\ = p + (1-p)J_{\mu^*}(2x-1) - p(p + (1-p)J_{\mu^*}(2x+2u-1)) \\ - (1-p)(p + (1-p)J_{\mu^*}(2x-2u-1)) \\ = (1-p)(J_{\mu^*}(2x-1) - pJ_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u-1)) \\ \geq 0, \end{aligned}$$

and Eq. (3.37) follows from the preceding relations.

Case 3. Using Eq. (3.30), we have

$$\begin{aligned} J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u) \\ = pJ_{\mu^*}(2x) - p(p + (1-p)J_{\mu^*}(2x+2u-1)) - p(1-p)J_{\mu^*}(2x-2u) \\ = p(J_{\mu^*}(2x) - p - (1-p)J_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u)). \end{aligned}$$

Now we must have $x \geq \frac{1}{4}$, for otherwise $u < \frac{1}{4}$ and $x+u < 1/2$. Hence $2x \geq 1/2$ and the sequence of equalities can be continued as follows:

$$\begin{aligned} J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u) \\ = p(p + (1-p)J_{\mu^*}(4x-1) - p \\ - (1-p)J_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u)) \\ = p(1-p)(J_{\mu^*}(4x-1) - J_{\mu^*}(2x+2u-1) - J_{\mu^*}(2x-2u)) \\ = (1-p)(J_{\mu^*}(2x-1/2) - pJ_{\mu^*}(2x+2u-1) - pJ_{\mu^*}(2x-2u)). \end{aligned}$$

Since $p \leq (1-p)$, the last expression is greater than or equal to both

$$(1-p)(J_{\mu^*}(2x-1/2) - pJ_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u))$$

and

$$(1-p)(J_{\mu^*}(2x-1/2) - (1-p)J_{\mu^*}(2x+2u-1) - pJ_{\mu^*}(2x-2u)).$$

Now for $x, u \in S_{n+1}$, and $n \geq 1$, we have $(2x - 1/2) \in S_n$ and $(2u - 1/2) \in S_n$ if $(2u - 1/2) \in [0, 1]$, and $(1/2 - 2u) \in S_n$ if $(1/2 - 2u) \in [0, 1]$. By the induction hypothesis, the first or the second of the preceding expressions is nonnegative, depending on whether $2x + 2u - 1 \geq 2x - 1/2$ or $2x - 2u \geq 2x - 1/2$ (i.e., $u \geq \frac{1}{4}$ or $u \leq \frac{1}{4}$). Hence Eq. (3.37) is proved for case 3.

Case 4. The proof resembles the one for case 3. Using Eq. (3.30), we have

$$\begin{aligned} J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u) \\ = p + (1-p)J_{\mu^*}(2x-1) - p(p + (1-p)J_{\mu^*}(2x+2u-1)) \\ - (1-p)pJ_{\mu^*}(2x-2u) \\ = p(1-p) \\ + (1-p)(J_{\mu^*}(2x-1) - pJ_{\mu^*}(2x+2u-1) - pJ_{\mu^*}(2x-2u)). \end{aligned}$$

We must have $x \leq \frac{3}{4}$ for otherwise $u < \frac{1}{4}$ and $x-u > \frac{1}{2}$. Hence $0 \leq 2x-1 \leq 1/2 \leq 2x-2u \leq 1$, and using Eq. (3.30) we have

$$(1-p)J_{\mu^*}(2x-1) = (1-p)pJ_{\mu^*}(4x-2) = p(J_{\mu^*}(2x-1/2) - p).$$

Using the preceding relations, we obtain

$$\begin{aligned} J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u) \\ = p(1-p) + p(J_{\mu^*}(2x-1/2) - p) - p(1-p)J_{\mu^*}(2x+2u-1) \\ - p(1-p)J_{\mu^*}(2x-2u) \\ = p((1-2p) + J_{\mu^*}(2x-1/2) - (1-p)J_{\mu^*}(2x+2u-1) \\ - (1-p)J_{\mu^*}(2x-2u)). \end{aligned}$$

These relations are equal to both

$$\begin{aligned} p((1-2p)(1 - J_{\mu^*}(2x+2u-1)) \\ + J_{\mu^*}(x-1/2) - pJ_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u)) \end{aligned}$$

and

$$\begin{aligned} p((1-2p)(1 - J_{\mu^*}(2x-2u)) \\ + J_{\mu^*}(2x-1/2) - (1-p)J_{\mu^*}(2x+2u-1) - pJ_{\mu^*}(2x-2u)). \end{aligned}$$

Since $0 \leq J_{\mu^*}(2x+2u-1) \leq 1$ and $0 \leq J_{\mu^*}(2x-2u) \leq 1$, these expressions are greater than or equal to both

$$p(J_{\mu^*}(2x-1/2) - pJ_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u))$$

and

$$p(J_{\mu^*}(2x-1/2) - (1-p)J_{\mu^*}(2x+2u-1) - pJ_{\mu^*}(2x-2u))$$

and the result follows as in case 3. Q.E.D.

We note that the bold strategy is not the unique optimal stationary gambling strategy. For a characterization of all optimal strategies, see Dubins and Savage [DuS65], p. 90. Several other gambling problems where strategies of the bold type are optimal are described in Dubins and Savage [DuS65], Chapters 5 and 6.

3.6 NONSTATIONARY AND PERIODIC PROBLEMS

The standing assumption so far in this book has been that the problem involves a stationary system and a stationary cost per stage (except for the presence of the discount factor). Problems with nonstationary system or cost per stage arise occasionally in practice or in theoretical studies and are thus of some interest. It turns out that such problems can be converted to stationary ones by a simple reformulation. We can then obtain results analogous to those obtained earlier for stationary problems.

Consider a nonstationary system of the form

$$x_{k+1} = f_k(x_k, u_k, w_k), \quad k = 0, 1, \dots,$$

and a cost function of the form

$$J_\pi(x_0) = \lim_{N \rightarrow \infty} E_{w_k} \left\{ \sum_{k=0}^{N-1} \alpha^k g_k(x_k, \mu_k(x_k), w_k) \right\}.$$

In these equations, for each k , x_k belongs to a space S_k , u_k belongs to a space C_k and satisfies $u_k \in U_k(x_k)$ for all $x_k \in S_k$, and w_k belongs to a countable space D_k . The sets S_k , C_k , $U_k(x_k)$, D_k may differ from one stage to the next. The random disturbances w_k are characterized by probabilities $P_k(\cdot | x_k, u_k)$, which depend on x_k and u_k as well as the time index k . The set of admissible policies Π is the set of all sequences $\pi = \{\mu_0, \mu_1, \dots\}$ with $\mu_k : S_k \mapsto C_k$ and $\mu_k(x_k) \in U_k(x_k)$ for all $x_k \in S_k$ and $k = 0, 1, \dots$. The functions $g_k : S_k \times C_k \times D_k \mapsto \mathbb{R}$ are given and are assumed to satisfy one of the following three assumptions:

Assumption D': We have $\alpha < 1$, and the functions g_k satisfy, for all $k = 0, 1, \dots$,

$$|g_k(x_k, u_k, w_k)| \leq M, \quad \text{for all } (x_k, u_k, w_k) \in S_k \times C_k \times D_k,$$

where M is some scalar.

Assumption P': The functions g_k satisfy, for all $k = 0, 1, \dots$,

$$0 \leq g_k(x_k, u_k, w_k), \quad \text{for all } (x_k, u_k, w_k) \in S_k \times C_k \times D_k.$$

Assumption N': The functions g_k satisfy, for all $k = 0, 1, \dots$,

$$g_k(x_k, u_k, w_k) \leq 0, \quad \text{for all } (x_k, u_k, w_k) \in S_k \times C_k \times D_k.$$

We will refer to the problem formulated as the *nonstationary problem* (NSP for short). We can get an idea on how the NSP can be converted to a stationary problem by considering the special case where the state space is the same for each stage (i.e., $S_k = S$ for all k). We consider an augmented state

$$\tilde{x} = (x, k),$$

where $x \in S$, and k is the time index. The new state space is $\tilde{S} = S \times K$, where K denotes the set of nonnegative integers. The augmented system evolves according to

$$(x, k) \rightarrow (f_k(x, u_k, w_k), k + 1), \quad (x, k) \in \tilde{S}.$$

Similarly, we can define a cost per stage as

$$\tilde{g}((x, k), u_k, w_k) = g_k(x, u_k, w_k), \quad (x, k) \in \tilde{S}.$$

It is evident that the problem corresponding to the augmented system is stationary. If we restrict attention to initial states $\tilde{x}_0 \in S \times \{0\}$, it can be seen that this stationary problem is equivalent to the NSP.

Let us now consider the more general case. To simplify notation, we will assume that the state spaces S_i , $i = 0, 1, \dots$, the control spaces C_i , $i = 0, 1, \dots$, and the disturbance spaces D_i , $i = 0, 1, \dots$, are all mutually disjoint. This assumption does not involve a loss of generality since, if necessary, we may relabel the elements of S_i , C_i , and D_i without affecting the structure of the problem. Define now a new state space S , a new control space C , and a new (countable) disturbance space D by

$$S = \cup_{i=0}^{\infty} S_i, \quad C = \cup_{i=0}^{\infty} C_i, \quad D = \cup_{i=0}^{\infty} D_i.$$

Introduce a new (stationary) system

$$\tilde{x}_{k+1} = f(\tilde{x}_k, \tilde{u}_k, \tilde{w}_k), \quad k = 0, 1, \dots,$$

where $\tilde{x}_k \in S$, $\tilde{u}_k \in C$, $\tilde{w}_k \in D$, and the system function $f : S \times C \times D \mapsto S$ is defined by

$$f(\tilde{x}, \tilde{u}, \tilde{w}) = f_i(\tilde{w}, \tilde{u}, \tilde{w}), \quad \text{if } \tilde{x} \in S_i, \quad \tilde{u} \in C_i, \quad \tilde{w} \in D_i, \quad i = 0, 1, \dots$$

For triplets $(\tilde{x}, \tilde{u}, \tilde{w})$, where for some $i = 0, 1, \dots$, we have $\tilde{x} \in S_i$, but $\tilde{u} \notin C_i$ or $\tilde{w} \notin D_i$, the definition of f is immaterial; any definition is adequate for

our purposes in view of the control constraints to be introduced. The control constraint is taken to be $\tilde{u} \in U(\tilde{x})$ for all $\tilde{x} \in S$, where $U(\cdot)$ is defined by

$$U(\tilde{x}) = U_i(\tilde{x}), \quad \text{if } \tilde{x} \in S_i, \quad i = 0, 1, \dots$$

The disturbance \tilde{w} is characterized by probabilities $P(\tilde{w} | \tilde{x}, \tilde{u})$ such that

$$P(\tilde{w} \in D_i | \tilde{x} \in S_i, \tilde{u} \in C_i) = 1, \quad i = 0, 1, \dots,$$

$$P(\tilde{w} \notin D_i | \tilde{x} \in S_i, \tilde{u} \in C_i) = 0, \quad i = 0, 1, \dots$$

Furthermore, for any $w_i \in D_i$, $x_i \in S_i$, $u_i \in C_i$, $i = 0, 1, \dots$, we have

$$P(w_i | x_i, u_i) = P_i(w_i | x_i, u_i).$$

We also introduce a new cost function

$$\tilde{J}_{\tilde{\pi}}(\tilde{x}_0) = \lim_{N \rightarrow \infty} E_{\substack{w_k \\ k=0,1,\dots,N-1}} \left\{ \sum_{k=0}^{N-1} \alpha^k g(\tilde{x}_k, \mu_k(\tilde{x}_k), \tilde{w}_k) \right\},$$

where the (stationary) cost per stage $g : S \times C \times D \mapsto \mathbb{R}$ is defined for all $i = 0, 1, \dots$ by

$$g(\tilde{x}, \tilde{u}, \tilde{w}) = g_i(\tilde{x}, \tilde{u}, \tilde{w}), \quad \text{if } \tilde{x} \in S_i, \quad \tilde{u} \in C_i, \quad \tilde{w} \in D_i.$$

For triplets $(\tilde{x}, \tilde{u}, \tilde{w})$, where for some $i = 0, 1, \dots$, we have $\tilde{x} \in S_i$ but $\tilde{u} \notin C_i$ or $\tilde{w} \notin D_i$, any definition of g is adequate provided $|g(\tilde{x}, \tilde{u}, \tilde{w})| \leq M$ for all $(\tilde{x}, \tilde{u}, \tilde{w})$ when Assumption D' holds, $0 \leq g(\tilde{x}, \tilde{u}, \tilde{w})$ when P' holds, and $g(\tilde{x}, \tilde{u}, \tilde{w}) \leq 0$ when N' holds. The set of admissible policies $\tilde{\Pi}$ for the new problem consists of all sequences $\tilde{\pi} = \{\tilde{\mu}_0, \tilde{\mu}_1, \dots\}$, where $\tilde{\mu}_k : S \mapsto C$ and $\tilde{\mu}_k(\tilde{x}) \in U(\tilde{x})$ for all $\tilde{x} \in S$ and $k = 0, 1, \dots$

The construction given defines a problem that clearly fits the framework of the infinite horizon total cost problem. We will refer to this problem as the *stationary problem* (SP for short).

It is important to understand the nature of the intimate connection between the NSP and the SP formulated here. Let $\pi = \{\mu_0, \mu_1, \dots\}$ be an admissible policy for the NSP. Also, let $\tilde{\pi} = \{\tilde{\mu}_0, \tilde{\mu}_1, \dots\}$ be an admissible policy for the SP such that

$$\tilde{\mu}_i(\tilde{x}) = \mu_i(\tilde{x}), \quad \text{if } \tilde{x} \in S_i, \quad i = 0, 1, \dots \tag{3.39}$$

Let $x_0 \in S_0$ be the initial state for the NSP and consider the same initial state for the SP (i.e., $\tilde{x}_0 = x_0 \in S_0$). Then the sequence of states $\{\tilde{x}_i\}$ generated in the SP will satisfy $\tilde{x}_i \in S_i$, $i = 0, 1, \dots$, with probability 1 (i.e., the system will move from the set S_0 to the set S_1 , then to S_2 , etc., just as in the NSP). Furthermore, the probabilistic law of generation of

states and costs is identical in the NSP and the SP. As a result, it is easy to see that for any admissible policies π and $\tilde{\pi}$ satisfying Eq. (3.39) and initial states x_0, \tilde{x}_0 satisfying $x_0 = \tilde{x}_0 \in S_0$, the sequence of generated states in the NSP and the SP is the same ($x_i = \tilde{x}_i$, for all i) provided the generated disturbances w_i and \tilde{w}_i are also the same for all i ($w_i = \tilde{w}_i$, for all i). Furthermore, if π and $\tilde{\pi}$ satisfy Eq. (3.39), we have $J_\pi(x_0) = \tilde{J}_\pi(\tilde{x}_0)$ if $x_0 = \tilde{x}_0 \in S_0$. Let us also consider the optimal cost functions for the NSP and the SP:

$$J^*(x_0) = \min_{\pi \in \Pi} J_\pi(x_0), \quad x_0 \in S_0,$$

$$\tilde{J}^*(\tilde{x}_0) = \min_{\tilde{\pi} \in \tilde{\Pi}} J_{\tilde{\pi}}(\tilde{x}_0), \quad \tilde{x}_0 \in S_0.$$

Then it follows from the construction of the SP that

$$\tilde{J}^*(\tilde{x}_0) = \tilde{J}^*(\tilde{x}_0, i), \quad \text{if } \tilde{x}_0 \in S_i, \quad i = 0, 1, \dots,$$

where, for all $i = 0, 1, \dots$,

$$\tilde{J}^*(\tilde{x}_0, i) = \min_{\pi \in \Pi} \lim_{N \rightarrow \infty} \sum_{k=0,1,\dots,N-1}^E \left\{ \sum_{k=i}^{N-1} \alpha^{k-i} g_k(x_k, \mu_k(x_k), w_k) \right\}, \quad (3.40)$$

if $\tilde{x}_0 = x_i \in S_i$. Note that in this equation, the right-hand side is defined in terms of the data of the NSP. As a special case of this equation, we obtain

$$\tilde{J}^*(\tilde{x}_0) = \tilde{J}^*(\tilde{x}_0, 0) = J^*(x_0), \quad \text{if } \tilde{x}_0 = x_0 \in S_0.$$

Thus the optimal cost function J^* of the NSP can be obtained from the optimal cost function \tilde{J}^* of the SP. Furthermore, if $\tilde{\pi}^* = \{\tilde{\mu}_0^*, \tilde{\mu}_1^*, \dots\}$ is an optimal policy for the SP, then the policy $\pi^* = \{\mu_0^*, \mu_1^*, \dots\}$ defined by

$$\mu_i^*(x_i) = \tilde{\mu}_i^*(x_i), \quad \text{for all } x_i \in S_i, \quad i = 0, 1, \dots, \quad (3.41)$$

is an optimal policy for the NSP. Thus optimal policies for the SP yield optimal policies for the NSP via Eq. (3.41). Another point to be noted is that if Assumption D' (P', N') is satisfied for the NSP, then Assumption D (P, N) introduced earlier in this chapter is satisfied for the SP.

These observations show that one may analyze the NSP by means of the SP. Every result given in the preceding sections when applied to the SP yields a corresponding result for the NSP. We will just provide the form of the optimality equation for the NSP in the following proposition.

Proposition 3.6.1: Under Assumption D' (P', N'), there holds

$$J^*(x_0) = \tilde{J}^*(x_0, 0), \quad x_0 \in S_0,$$

where for all $i = 0, 1, \dots$, the functions $\tilde{J}^*(\cdot, i)$ map S_i into $\Re ([0, \infty], [-\infty, 0])$, are given by Eq. (3.40), and satisfy for all $x_i \in S_i$ and $i = 0, 1, \dots$,

$$\tilde{J}^*(x_i, i) = \min_{u_i \in U_i(x_i)} E \{ g_i(x_i, u_i, w_i) + \alpha \tilde{J}^*(f_i(x_i, u_i, w_i), i+1) \}. \quad (3.42)$$

Under Assumption D' the functions $\tilde{J}^*(\cdot, i)$, $i = 0, 1, \dots$, are the unique bounded solutions of the set of equations Eq. (3.42). Furthermore, under Assumption D' or P', if $\mu_i^*(x_i) \in U_i(x_i)$ attains the minimum in Eq. (3.42) for all $x_i \in S_i$ and i , then the policy $\pi^* = \{\mu_0^*, \mu_1^*, \dots\}$ is optimal for the NSP.

Periodic Problems

Assume within the framework of the NSP that there exists an integer $p \geq 2$ (called the *period*) such that for all integers i and j with $|i - j| = mp$, $m = 1, 2, \dots$, we have

$$S_i = S_j, \quad C_i = C_j, \quad D_i = D_j, \quad U_i(\cdot) = U_j(\cdot),$$

$$f_i = f_j, \quad g_i = g_j, \quad P_i(\cdot | x, j) = P_j(\cdot | x, u), \quad (x, u) \in S_i \times C_i.$$

We assume that the spaces S_i , C_i , D_i , $i = 0, 1, \dots, p-1$, are mutually disjoint. We define new state, control, and disturbance spaces by

$$S = \cup_{i=0}^{p-1} S_i, \quad C = \cup_{i=0}^{p-1} C_i, \quad D = \cup_{i=0}^{p-1} D_i.$$

The optimality equation for the equivalent stationary problem reduces to the system of p equations

$$\tilde{J}^*(x_0, 0) = \min_{u_0 \in U_0(x_0)} E \{ g_0(x_0, u_0, w_0) + \alpha \tilde{J}^*(f_0(x_0, u_0, w_0), 1) \},$$

$$\tilde{J}^*(x_1, 1) = \min_{u_1 \in U_1(x_1)} E \{ g_1(x_1, u_1, w_1) + \alpha \tilde{J}^*(f_1(x_1, u_1, w_1), 2) \},$$

$$\begin{aligned} \tilde{J}^*(x_{p-1}, p-1) = & \min_{u_{p-1} \in U_{p-1}(x_{p-1})} E \{ g_{p-1}(x_{p-1}, u_{p-1}, w_{p-1}) \\ & + \alpha \tilde{J}^*(f_{p-1}(x_{p-1}, u_{p-1}, w_{p-1}), 0) \}. \end{aligned}$$

These equations may be used to obtain (under Assumption D' or P') a periodic policy of the form $\{\mu_0^*, \dots, \mu_{p-1}^*, \mu_0^*, \dots, \mu_{p-1}^*, \dots\}$ whenever the minimum of the right-hand side is attained for all x_i , $i = 0, 1, \dots, p - 1$. When all spaces involved are finite, an optimal policy may be found by means of the algorithms of Section 1.3, appropriately adapted to the corresponding SP.

3.7 NOTES, SOURCES, AND EXERCISES

Undiscounted problems and discounted problems with unbounded cost per stage were first analyzed systematically by Dubins and Savage [DuS65], Blackwell [Bla65], and Strauch [Str66]. The monograph by Bertsekas and Shreve [BeS78] provides an extensive treatment, which also resolves the associated measurability questions. Sufficient conditions for convergence of the value iteration method under Assumption P (cf. Props. 3.1.6 and 3.1.7) were derived independently in Bertsekas [Ber77] and Schal [Sch75]. The former reference also derives necessary conditions for convergence. Problems involving convexity assumptions are analyzed in Bertsekas [Ber73b].

We have bypassed a number of complex theoretical issues relating to stationary policies that historically have played an important role in the development of the subject of this chapter. The main question is to what extent is it possible to restrict attention to stationary policies. Much theoretical work has been done on this question (Bertsekas and Shreve [BeS79], Blackwell [Bla65], Blackwell [Bla70], Dubins and Savage [DuS65], Feinberg [Fei78], [Fei92a], [Fei92b], Ornstein [Orn69]), and some aspects are still open. Suppose, for example, that we are given an $\epsilon > 0$. One issue is whether there exists an ϵ -optimal stationary policy, i.e., a stationary policy μ such that

$$J_\mu(x) \leq J^*(x) + \epsilon, \quad \text{for all } x \in S \text{ with } J^*(x) > -\infty,$$

$$J_\mu(x) \geq -\frac{1}{\epsilon}, \quad \text{for all } x \in S \text{ with } J^*(x) = -\infty.$$

The answer is positive under any one of the following conditions:

1. Assumption P holds and $\alpha < 1$ (see Exercise 3.8).
2. Assumption N holds, S is a finite set, $\alpha = 1$, and $J^*(x) > -\infty$ for all $x \in S$ (see Exercise 3.11 or Blackwell [Bla65], [Bla70], and Ornstein [Orn69]).
3. Assumption N holds, S is a countable set, $\alpha = 1$, and the problem is deterministic (see Bertsekas and Shreve [BeS79]).

The answer can be negative under any one of the following conditions:

1. Assumption P holds and $\alpha = 1$ (see Exercise 3.8).

2. Assumption N holds and $\alpha < 1$ (see Exercise 3.11 or Bertsekas and Shreve [BeS79]).

The existence of an ϵ -optimal stationary policy for stochastic shortest path problems with a finite state space, but under somewhat different assumptions than the ones of Section 2.1 is established by Feinberg [Fei92b].

Another issue is whether one can confine the search for an optimal policy within the class of stationary policies, i.e., whether there exists an optimal stationary policy when there exists an optimal policy for each initial state. This is true under Assumption P (see Exercise 3.9). It is also true (but very hard to prove) under Assumption N if $J^*(x) > -\infty$ for all $x \in S$, $\alpha = 1$, and the disturbance space D is countable (Blackwell [Bla70], Dubins and Savage [DuS65], Ornstein [Orn69]). Simple two-state examples can be constructed showing that the result fails to hold if $\alpha = 1$ and $J^*(x) = -\infty$ for some state x (see Exercises 3.10 and 3.25). However, these examples rely on the presence of a stochastic element in the problem. If the problem is deterministic, stronger results are available; one can find an optimal stationary policy if there exists an optimal policy at each initial state and either $\alpha = 1$ or $\alpha < 1$ and $J^*(x) > -\infty$ for all $x \in S$. These results also require a difficult proof (Bertsekas and Shreve [BeS79]).

The gambling problem and its solution are taken from Dubins and Savage [DuS65]. A surprising property of the optimal reward function J^* for this problem has been shown by Billingsley [Bil83]: J^* is almost everywhere differentiable with derivative zero, yet it is strictly increasing, taking values that range from 0 to 1.

E X E R C I S E S

3.1

Let $S = [0, \infty)$ and $C = U(x) = (0, \infty)$ be the state and control spaces, respectively, let the system equation be

$$x_{k+1} = \left(\frac{2}{\alpha} \right) x_k + u_k, \quad k = 0, 1, \dots,$$

where $\alpha \in (0, 2)$, and let

$$g(x_k, u_k) = x_k + u_k$$

be the cost per stage. Show that for this deterministic problem, Assumption P holds and that $J^*(x) = \infty$ for all $x \in S$, but $(T^k J_0)(0) = 0$ for all k [J_0 is the zero function, $J_0(x) = 0$, for all $x \in S$].

3.2

Let Assumption P hold and consider the finite-state case $S = D = \{1, 2, \dots, n\}$, $\alpha = 1$, $x_{k+1} = w_k$. The mapping T is represented as

$$(TJ)(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u) J(j) \right], \quad i = 1, \dots, n,$$

where $p_{ij}(u)$ denotes the transition probability that the next state will be j when the current state is i and control u is applied. Assume that the sets $U(i)$ are compact subsets of \mathbb{R}^m for all i , and that $p_{ij}(u)$ and $g(i, u)$ are continuous on $U(i)$ for all i and j . Show that $\lim_{k \rightarrow \infty} (T^k J_0)(i) = J^*(i)$, where $J_0(i) = 0$ for all $i = 1, \dots, n$. Show also that there exists an optimal stationary policy.

3.3

Consider a deterministic problem involving a linear system

$$x_{k+1} = Ax_k + Bu_k, \quad k = 0, 1, \dots,$$

where the pair (A, B) is controllable and $x_k \in \mathbb{R}^n$, $u_k \in \mathbb{R}^m$. Assume no constraints on the control and a cost per stage g satisfying

$$0 \leq g(x, u), \quad (x, u) \in \mathbb{R}^n \times \mathbb{R}^m.$$

Assume furthermore that g is continuous in x and u , and that $g(x_n, u_n) \rightarrow \infty$ if $\{x_n\}$ is bounded and $\|u_n\| \rightarrow \infty$.

- (a) Show that for a discount factor $\alpha < 1$, the optimal cost satisfies $0 \leq J^*(x) < \infty$, for all $x \in \mathbb{R}^n$. Furthermore, there exists an optimal stationary policy and

$$\lim_{k \rightarrow \infty} (T^k J_0)(x) = J^*(x), \quad x \in \mathbb{R}^n.$$

- (b) Show that the same is true, except perhaps for $J^*(x) < \infty$, when the system is of the form $x_{k+1} = f(x_k, u_k)$, with $f : \mathbb{R}^n \times \mathbb{R}^m \mapsto \mathbb{R}^n$ being a continuous function.
- (c) Prove the same results assuming that the control is constrained to lie in a compact set $U \in \mathbb{R}^m$ [$U(x) = U$ for all x] in place of the assumption $g(x_n, u_n) \rightarrow \infty$ if $\{x_n\}$ is bounded and $\|u_n\| \rightarrow \infty$. Hint: Show that $T^k J_0$ is real valued and continuous for every k , and use Prop. 3.1.7.

3.4

Under Assumption P, let μ be such that for all $x \in S$, $\mu(x) \in U(x)$ and

$$(T_\mu J^*)(x) \leq (TJ^*)(x) + \epsilon,$$

where ϵ is some positive scalar. Show that, if $\alpha < 1$,

$$J_\mu(x) \leq J^*(x) + \frac{\epsilon}{1 - \alpha}, \quad x \in S.$$

Hint: Show that $(T_\mu^k J^*)(x) \leq J^*(x) + \sum_{i=0}^{k-1} \alpha^i \epsilon$.

3.5

Under Assumption P or N, show that if $\alpha < 1$ and $J' : S \mapsto \mathbb{R}$ is a bounded function satisfying $J' = T J'$, then $J' = J^*$. Hint: Under P, let r be a scalar such that $J^* + r\epsilon \geq J'$. Argue that $J^* \geq J'$ and use Prop. 3.1.2(a).

3.6

We want to find a scalar sequence $\{u_0, u_1, \dots\}$ that satisfies $\sum_{k=0}^{\infty} u_k \leq c$, $u_k \geq 0$, for all k , and maximizes $\sum_{k=0}^{\infty} g(u_k)$, where $c > 0$ and $g(u) \geq 0$ for all $u \geq 0$, $g(0) = 0$. Assume that g is monotonically nondecreasing on $[0, \infty)$. Show that the optimal value of the problem is $J^*(c)$, where J^* is a monotonically nondecreasing function on $[0, \infty)$ satisfying $J^*(0) = 0$ and

$$J^*(x) = \max_{0 \leq u \leq x} \{g(u) + J^*(x - u)\}, \quad x \in [0, \infty).$$

3.7

Let Assumption P hold and assume that $\pi^* = \{\mu_0^*, \mu_1^*, \dots\} \in \Pi$ satisfies $J^* = T_{\mu_k^*} J^*$ for all k . Show that π^* is optimal, i.e., $J_{\pi^*} = J^*$.

3.8

Under Assumption P, show that, given $\epsilon > 0$, there exists a policy $\pi_\epsilon \in \Pi$ such that $J_{\pi_\epsilon}(x) \leq J^*(x) + \epsilon$ for all $x \in S$, and that for $\alpha < 1$ the policy π_ϵ can be taken stationary. Give an example where $\alpha = 1$ and for each stationary policy π we have $J_\pi(x) = \infty$, while $J^*(x) = 0$ for all x . Hint: See the proof of Prop. 3.1.1.

3.9

Under Assumption P, show that if there exists an optimal policy (a policy $\pi^* \in \Pi$ such that $J_{\pi^*} = J^*$), then there exists an optimal stationary policy.

3.10

Use the following counterexample to show that the result of Exercise 3.9 may fail to hold under Assumption N if $J^*(x) = -\infty$ for some $x \in S$. Let $S = D = \{0, 1\}$, $f(x, u, w) = w$, $g(x, u, w) = u$, $U(0) = (-\infty, 0]$, $U(1) = \{0\}$, $p(w = 0 | x = 0, u) = \frac{1}{2}$, and $p(w = 1 | x = 1, u) = 1$. Show that $J^*(0) = -\infty$, $J^*(1) = 0$ and that the admissible nonstationary policy $\{\mu_0^*, \mu_1^*, \dots\}$ with $\mu_k^*(0) = -(2/\alpha)^k$ is optimal. Show that every stationary policy μ satisfies $J_\mu(0) = (2/(2-\alpha))\mu(0)$, $J_\mu(1) = 0$ (see [Bla70], [DuS65], and [Orn69] for related analysis).

3.11

Show that the result of Exercise 3.8 holds under Assumption N if S is a finite set, $\alpha = 1$, and $J^*(x) > -\infty$ for all $x \in S$. Construct a counterexample to show that the result can fail to hold if S is countable and $\alpha < 1$ [even if $J^*(x) > -\infty$ for all $x \in S$]. Hint: Consider an integer N such that the N -stage optimal cost J_N satisfies

$$J_N(x) \leq J^*(x) + \epsilon, \quad x \in S.$$

For a counterexample, see [BeS79].

3.12 (Deterministic Linear-Quadratic Problems)

Consider the deterministic linear-quadratic problem involving the system

$$x_{k+1} = Ax_k + Bu_k$$

and the cost

$$J_\pi(x_0) = \sum_{k=0}^{\infty} (x'_k Q x_k + \mu_k(x_k)' R \mu_k(x_k)).$$

We assume that R is positive definite symmetric, Q is of the form $C'C$, and the pairs (A, B) , (A, C) are controllable and observable, respectively. Use the theory of Section 4.1 of Vol. I to show that the stationary policy μ^* with

$$\mu^*(x) = -(B'KB + R)^{-1}B'KAx$$

is optimal, where K is the unique positive semidefinite symmetric solution of the algebraic Riccati equation (cf. Section 4.1 of Vol. I):

$$K = A'(K - KB(B'KB + R)^{-1}B'K)A + Q.$$

Provide a similar result under an appropriate controllability assumption for the case of a periodic deterministic linear system and a periodic quadratic cost (cf. Section 3.6).

3.13

Consider the linear-quadratic problem of Section 3.2 with the only difference that the disturbances w_k have zero mean, but their covariance matrices are non-stationary and uniformly bounded over k . Show that the optimal control law remains unchanged.

3.14 (Periodic Linear-Quadratic Problems)

Consider the linear system

$$x_{k+1} = A_k x_k + B_k u_k + w_k, \quad k = 0, 1, \dots,$$

and the quadratic cost

$$J_\pi(x_0) = \lim_{N \rightarrow \infty} \frac{E}{w_k} \left\{ \sum_{k=0}^{N-1} \alpha^k (x'_k Q_k x_k + u'_k R_k u_k) \right\},$$

where the matrices have appropriate dimensions, Q_k and R_k are positive semidefinite and positive definite symmetric, respectively, for all k , and $0 < \alpha < 1$. Assume that the system and cost are periodic with period p (cf. Section 3.6), that the controls are unconstrained, and that the disturbances are independent, and have zero mean and finite covariance. Assume further that the following (controllability) condition is in effect.

For any state \bar{x}_0 , there exists a finite sequence of controls $\{\bar{u}_0, \bar{u}_1, \dots, \bar{u}_r\}$ such that $\bar{x}_{r+1} = 0$, where \bar{x}_{r+1} is generated by

$$\bar{x}_{k+1} = A_k \bar{x}_k + B_k \bar{u}_k, \quad k = 0, 1, \dots, r.$$

Show that there is an optimal periodic policy π^* of the form

$$\pi^* = \{\mu_0^*, \mu_1^*, \dots, \mu_{p-1}^*, \mu_0^*, \mu_1^*, \dots, \mu_{p-1}^*, \dots\},$$

where $\mu_0^*, \mu_1^*, \dots, \mu_{p-1}^*$ are given by

$$\mu_i^*(x) = -\alpha(\alpha B'_i K_{i+1} B_i + R_i)^{-1} B'_i K_{i+1} A_i x, \quad i = 0, \dots, p-2,$$

$$\mu_{p-1}^*(x) = -\alpha(\alpha B'_{p-1} K_0 B_{p-1} + R_{p-1})^{-1} B'_{p-1} K_0 A_{p-1} x,$$

and the matrices K_0, K_1, \dots, K_{p-1} satisfy the coupled set of p algebraic Riccati equations given for $i = 0, 1, \dots, p-1$ by

$$K_i = A'_i (\alpha K_{i+1} - \alpha^2 K_{i+1} B_i (\alpha B'_i K_{i+1} B_i + R_i)^{-1} B'_i K_{i+1} A_i) + Q_i,$$

with

$$K_p = K_0.$$

3.15 (Linear-Quadratic Problems – Imperfect State Information)

Consider the linear-quadratic problem of Section 3.2 with the difference that the controller, instead of having perfect state information, has access to measurements of the form

$$z_k = Cx_k + v_k, \quad k = 0, 1, \dots$$

As in Section 5.2 of Vol. I, the disturbances v_k are independent and have identical statistics, zero mean, and finite covariance matrix. Assume that for every admissible policy π the matrices

$$E\{(x_k - E\{x_k | I_k\})(x_k - E\{x_k | I_k\})' | \pi\}$$

are uniformly bounded over k , where I_k is the information vector defined in Section 5.2 of Vol. I. Show that the stationary policy μ^* given by

$$\mu^*(I_k) = -\alpha(\alpha B'KB + R)^{-1}B'KAE\{x_k | I_k\}, \quad \text{for all } I_k, k = 0, 1, \dots$$

is optimal. Show also that the same is true if w_k and v_k are nonstationary with zero mean and covariance matrices that are uniformly bounded over k . Hint: Combine the theory of Sections 5.2 of Vol. I and 3.2.

3.16 (Policy Iteration for Linear-Quadratic Problems [Kle68])

Consider the problem of Section 3.2 and let L_0 be an $m \times n$ matrix such that the matrix $(A + BL_0)$ has eigenvalues strictly within the unit circle.

- (a) Show that the cost corresponding to the stationary policy μ_0 , where $\mu_0(x) = L_0x$ is of the form

$$J_{\mu_0}(x) = x'K_0x + \text{constant},$$

where K_0 is a positive semidefinite symmetric matrix satisfying the (linear) equation

$$K_0 = \alpha(A + BL_0)'K_0(A + BL_0) + Q + L_0'R'L_0.$$

- (b) Let $\mu_1(x)$ attain the minimum for each x in the expression

$$\min_u \{u'R'u + \alpha(Ax + Bu)'K_0(Ax + bu)\}.$$

Show that for all x we have

$$J_{\mu_1}(x) = x'K_1x + \text{constant} \leq J_{\mu_0}(x),$$

where K_1 is some positive semidefinite symmetric matrix.

- (c) Show that the policy iteration process described in parts (a) and (b) yields a sequence $\{K_k\}$ such that

$$K_k \rightarrow K,$$

where K is the optimal cost matrix of the problem.

3.17 (Periodic Inventory Control Problems)

In the inventory control problem of Section 3.3, consider the case where the statistics of the demands w_k , the prices c_k , and the holding and the shortage costs are periodic with period p . Show that there exists an optimal periodic policy of the form $\pi^* = \{\mu_0^*, \dots, \mu_{p-1}^*, \mu_0^*, \dots, \mu_{p-1}^*, \dots\}$,

$$\mu_i^*(x) = \begin{cases} S_i^* - x & \text{if } x \leq S_i^*, \\ 0 & \text{if otherwise,} \end{cases} \quad i = 0, 1, \dots, p-1,$$

where S_0^*, \dots, S_{p-1}^* are appropriate scalars.

3.18 [HeS84]

Show that the critical level S^* for the inventory problem with zero fixed cost of Section 3.3 minimizes $(1-\alpha)cy + L(y)$ over y . Hint: Show that the cost can be expressed as

$$J_\pi(x_0) = E \left\{ \sum_{k=0}^{\infty} \alpha^k ((1-\alpha)cy_k + L(y_k)) + \frac{c\alpha}{1-\alpha} E\{w\} - cx_0 \right\},$$

where $y_k = x_k + \mu_k(x_k)$.

3.19

Consider a machine that may break down and can be repaired. When it operates over a time unit, it costs -1 (i.e., it produces a benefit of 1 unit), and it may break down with probability 0.1 . When it is in the breakdown mode, it may be repaired with an effort u . The probability of making it operative over one time unit is then u , and the cost is Cu^2 . Determine the optimal repair effort over an infinite time horizon with discount factor $\alpha < 1$.

3.20

Let z_0, z_1, \dots be a sequence of independent and identically distributed random variables taking values on a finite set Z . We know that the probability distribution of the z_k 's is one out of n distributions f_1, \dots, f_n , and we are trying to decide which distribution is the correct one. At each time k after observing z_1, \dots, z_k , we may either stop the observations and accept one of the n distributions as correct, or take another observation at a cost $c > 0$. The cost for accepting f_i given that f_j is correct is L_{ij} , $i, j = 1, \dots, n$. We assume $L_{ij} > 0$ for $i \neq j$, $L_{ii} = 0$, $i = 1, \dots, n$. The a priori distribution of f_1, \dots, f_n is denoted

$$P_0 = \{p_0^1, p_0^2, \dots, p_0^n\}, \quad p_0^i \geq 0, \quad \sum_{i=1}^n p_0^i = 1.$$

Show that the optimal cost $J^*(P_0)$ is a concave function of P_0 . Characterize the optimal acceptance regions and show how they can be obtained in the limit by means of a value iteration method.

3.21 (Gambling Strategies for Favorable Games)

A gambler plays a game such as the one of Section 3.5, but where the probability of winning p satisfies $1/2 \leq p < 1$. His objective is to reach a final fortune n , where n is an integer with $n \geq 2$. His initial fortune is an integer i with $0 < i < n$, and his stake at time k can take only integer values u_k satisfying $0 \leq u_k \leq x_k$, $0 \leq u_k \leq n - x_k$, where x_k is his fortune at time k . Show that the strategy that always stakes one unit is optimal [i.e., $\mu^*(x) = 1$ for all integers x with $0 < x < n$ is optimal]. Hint: Show that if $p \in (1/2, 1)$,

$$J_{\mu^*}(i) = \left[\left(\frac{1-p}{p} \right)^i - 1 \right] \left[\left(\frac{1-p}{p} \right)^n - 1 \right]^{-1}, \quad 0 \leq i \leq n,$$

and if $p = 1/2$,

$$J_{\mu^*}(i) = \frac{i}{n}, \quad 0 \leq i \leq n,$$

(or see [Ash70], p. 182, for a proof). Then use the sufficiency condition of Prop. 3.1.4.

3.22 [Sch81]

Consider a network of n queues whereby a customer at queue i upon completion of service is routed to queue j with probability p_{ij} , and exits the network with probability $1 - \sum_j p_{ij}$. For each queue i denote:

r_i : the external customer arrival rate,

$\frac{1}{\mu_i}$: the average customer service time,

λ_i : the customer departure rate,

a_i : the total customer arrival rate (sum of external rate and departure rates from upstream queues weighted by the corresponding probabilities).

We have

$$a_i = r_i + \sum_{j=1}^n \lambda_j p_{ji}, \quad \text{for all } i,$$

and we assume that any portion of the arrival rate a_i in excess of the service rate μ_i is lost; so the departure rate at queue i satisfies

$$\lambda_i = \min[\mu_i, a_i] = \min \left[\mu_i, r_i + \sum_{j=1}^n \lambda_j p_{ji} \right].$$

Assume that $r_i > 0$ for at least one i , and that for every queue i_1 with $r_{i_1} > 0$, there is a queue i with $1 - \sum_j p_{ij} > 0$, and a sequence i_1, i_2, \dots, i_k , i such that $p_{i_1 i_2} > 0, \dots, p_{i_k i} > 0$. Show that the departure rates λ_i satisfying the preceding equations are unique and can be found by value iteration or policy iteration. Hint: This problem does not quite fit our framework because we may have $\sum_j p_{ji} > 1$ for some i . However, it is possible to carry out an analysis based on m -stage contraction mappings.

3.23 (Infinite Time Reachability [Ber71], [Ber72])

Consider the stationary system

$$x_{k+1} = f(x_k, u_k, w_k), \quad k = 0, 1, \dots,$$

where the disturbance space D is an arbitrary (not necessarily countable) set. The disturbances w_k can take values in a subset $W(x_k, u_k)$ of D that may depend on x_k and u_k . This problem deals with the following question: Given a nonempty subset X of the state space S , under what conditions does there exist an admissible policy that keeps the state of the (closed-loop) system

$$x_{k+1} = f(x_k, \mu_k(x_k), w_k) \quad (3.43)$$

in the set X for all k and all possible values $w_k \in W(x_k, \mu_k(x_k))$, i.e.,

$$x_k \in X, \quad \text{for all } w_k \in W(x_k, \mu_k(x_k)), \quad k = 0, 1, \dots \quad (3.44)$$

The set X is said to be *infinitely reachable* if there exists an admissible policy $\{\mu_0, \mu_1, \dots\}$ and *some* initial state $x_0 \in X$ for which the above relations are satisfied. It is said to be *strongly reachable* if there exists an admissible policy $\{\mu_0, \mu_1, \dots\}$ such that for *all* initial states $x_0 \in X$ the above relations are satisfied.

Consider the function R mapping any subset Z of the state space S into a subset $R(Z)$ of S defined by

$$R(Z) = \{x \mid \text{for some } u \in U(x), f(x, u, w) \in Z, \text{ for all } w \in W(x, u)\} \cap Z.$$

- (a) Show that the set X is strongly reachable if and only if $R(X) = X$.
- (b) Given X , consider the set X^* defined as follows: $x_0 \in X^*$ if and only if $x_0 \in X$ and there exists an admissible policy $\{\mu_0, \mu_1, \dots\}$ such that Eqs. (3.43) and (3.44) are satisfied when x_0 is taken as the initial state of the system. Show that a set X is infinitely reachable if and only if it contains a nonempty strongly reachable set. Furthermore, the largest such set is X^* in the sense that X^* is strongly reachable whenever nonempty, and if $\tilde{X} \subset X$ is another strongly reachable set, then $\tilde{X} \subset X^*$.
- (c) Show that if X is infinitely reachable, there exists an admissible stationary policy μ such that if the initial state x_0 belongs to X^* , then all subsequent states of the closed-loop system $x_{k+1} = f(x_k, \mu(x_k), w_k)$ are guaranteed to belong to X^* .
- (d) Given X , consider the sets $R^k(X)$, $k = 1, 2, \dots$, where $R^k(X)$ denotes the set obtained after k applications of the mapping R on X . Show that

$$X^* \subset \bigcap_{k=1}^{\infty} R^k(X).$$

- (e) Given X , consider for each $x \in X$ and $k = 1, 2, \dots$ the set

$$U_k(x) = \{u \mid f(x, u, w) \in R^k(X) \text{ for all } w \in W(x, u)\}.$$

Show that, if there exists an index \bar{k} such that for all $x \in X$ and $k \geq \bar{k}$ the set $U_k(x)$ is a compact subset of a Euclidean space, then $X^* = \bigcap_{k=1}^{\infty} R^k(X)$.

3.24 (Infinite Time Reachability for Linear Systems)

Consider the linear stationary system

$$x_{k+1} = Ax_k + Bu_k + Gw_k,$$

where $x_k \in \mathbb{R}^n$, $u_k \in \mathbb{R}^m$, and $w_k \in \mathbb{R}^r$, and the matrices A , B , and G are known and have appropriate dimensions. The matrix A is assumed invertible. The controls u_k and the disturbances w_k are restricted to take values in the ellipsoids $U = \{u \mid u'R u \leq 1\}$ and $W = \{w \mid w'Q w \leq 1\}$, respectively, where R and Q are positive definite symmetric matrices of appropriate dimensions. Show that in order for the ellipsoid $X = \{x \mid x'Kx \leq 1\}$, where K is a positive definite symmetric matrix, to be strongly reachable (in the terminology of Exercise 3.23), it is sufficient that for some positive definite symmetric matrix M and for some scalar $\beta \in (0, 1)$ we have

$$K = A' \left[(1 - \beta)K^{-1} - \frac{1 - \beta}{\beta}GQ^{-1}G' + BR^{-1}B' \right]^{-1} A + M,$$

$$K^{-1} - \frac{1}{\beta}GQ^{-1}G' : \text{positive definite.}$$

Show also that if the above relations are satisfied, the linear stationary policy μ^* , where $\mu^*(x) = Lx$ and

$$L = -(R + B'FB)^{-1}B'FA,$$

$$F = \left[(1 - \beta)K^{-1} - \frac{1 - \beta}{\beta}GQ^{-1}G' \right]^{-1},$$

achieves reachability of the ellipsoid $X = \{x \mid x'Kx \leq 1\}$. Furthermore, the matrix $(A + BL)$ has all its eigenvalues strictly within the unit circle. (For a proof together with a computational procedure for finding matrices K satisfying the above, see [Ber71] and [Ber72].)

3.25 (The Blackmailer's Dilemma)

Consider Example 2.1.1. Here, there are two states, state 1 and a termination state t . At state 1, we can choose a control u with $0 < u \leq 1$; we then move to state t at no cost with probability $p(u)$, and stay in state 1 at a cost $-u$ with probability $1 - p(u)$.

- (a) Let $p(u) = u^2$. For this case it was shown in Example 2.1.1 that the optimal costs are $J^*(1) = -\infty$ and $J^*(t) = 0$. Furthermore, it was shown that there is no optimal stationary policy, although there is an optimal nonstationary policy. Find the set of solutions to Bellman's equation and verify the result of Prop. 3.1.2(b).
- (b) Let $p(u) = u$. Find the set of solutions to Bellman's equation and use Prop. 3.1.2(b) to show that the optimal costs are $J^*(1) = -1$ and $J^*(t) = 0$. Show that there is no stationary optimal policy.

Average Cost per Stage Problems

Contents

4.1. Finite-Spaces Average Cost Models	p. 174
4.1.1. Relation with the Discounted Cost Problem	p. 178
4.1.2. Blackwell Optimal Policies	p. 184
4.1.3. Optimality Equations	p. 194
4.2. Conditions for Equal Average Cost for all Initial States	p. 198
4.3. Value Iteration	p. 204
4.3.1. Single-Chain Value Iteration	p. 207
4.3.2. Multi-Chain Value Iteration	p. 222
4.4. Policy Iteration	p. 229
4.4.1. Single-Chain Policy Iteration	p. 229
4.4.2. Multi-Chain Policy Iteration	p. 235
4.5 Linear Programming	p. 239
4.6. Infinite-Spaces Problems	p. 245
4.6.1. A Sufficient Condition for Optimality	p. 253
4.6.2. Finite State Space and Infinite Control Space	p. 255
4.6.3. Countable States – Vanishing Discount Approach	p. 264
4.6.4. Countable States – Contraction Approach	p. 267
4.6.5. Linear Systems with Quadratic Cost	p. 272
4.7. Notes, Sources, and Exercises	p. 274

The results of the preceding chapters apply mainly to problems where the optimal total expected cost is finite either because of discounting or because of a cost-free absorbing state that the system eventually enters. In many situations, however, discounting is inappropriate and there is no natural cost-free absorbing state. In such situations it is often meaningful to optimize the average cost per stage, to be defined shortly. In this chapter, we discuss this type of optimization, with an emphasis on the case of a finite-state Markov chain.

An introductory analysis of the problem of this chapter was given in Section 7.4 of Vol. I. That analysis was based on a connection between the average cost per stage and the stochastic shortest path problem. While this connection can be further enhanced (see Exercise 4.12), we develop here an alternative, more powerful line of analysis, which is based on a relation with the discounted cost problem. This relation allows us to use discounted cost results, derived in Sections 1.2 and 1.3, in order to motivate and prove results for the average cost problem.

In this chapter, we will also encounter a special feature of the average cost problem, which we have not seen in either Vol. I or the preceding chapters in Vol. II: the probabilistic structure of the system genuinely matters in DP-related analyses. In particular, features such as number of recurrent classes and periodicity of the associated Markov chains play an important role in both analysis and algorithms. As a result, the material of this chapter is far more closely intertwined with the theory of stochastic processes than we have seen so far. The theory of Markov chains as reviewed in Appendix D of Vol. I is for the most part adequate background for our development. Some additional facts, relating for example to periodic Markov chains, will be reviewed as needed; see also the probability text by Bertsekas and Tsitsiklis [BeT02].

4.1 FINITE-SPACES AVERAGE COST MODELS

Let us formulate the problem of this chapter for the case of finite state and control spaces. We adopt the Markov chain notation of Section 1.3. In particular, we denote the states by $1, \dots, n$. To each state i and control u there corresponds a set of transition probabilities $p_{ij}(u)$, $j = 1, \dots, n$. Each time the system is in state i and control u is applied, we incur an expected cost $g(i, u)$, and the system moves to state j with probability $p_{ij}(u)$. The objective is to minimize over all policies $\pi = \{\mu_0, \mu_1, \dots\}$ with $\mu_k(i) \in U(i)$, for all i and k , an average cost per stage starting from a given initial state x_0 . This is defined as

$$J_\pi(x_0) = \limsup_{N \rightarrow \infty} \frac{1}{N} E \left\{ \sum_{k=0}^{N-1} g(x_k, \mu_k(x_k)) \right\}. \quad (4.1)$$

The reason for using \limsup rather than \lim in this definition is that the \limsup is guaranteed to exist for all π and x_0 , but this is not necessarily true for the \lim . We will show later that we may use \lim to define the average cost of every stationary policy (see the subsequent Prop. 4.1.2). Furthermore, for finite-spaces problems, there exists an optimal stationary policy (see the subsequent Prop. 4.1.7). However, while our analysis will revolve around stationary policies almost exclusively, to maintain mathematical rigor, it is essential to define the average cost of nonstationary policies in terms of \limsup (see Exercise 4.3 for an example).

An alternative way to formulate the average cost problem is to introduce the “upper” and “lower” costs of a policy,

$$J_\pi^+(x_0) = \limsup_{N \rightarrow \infty} \frac{1}{N} E \left\{ \sum_{k=0}^{N-1} g(x_k, \mu_k(x_k)) \right\},$$

$$J_\pi^-(x_0) = \liminf_{N \rightarrow \infty} \frac{1}{N} E \left\{ \sum_{k=0}^{N-1} g(x_k, \mu_k(x_k)) \right\},$$

and through analysis, establish that for particular policies π of interest, we have $J_\pi^+(x_0) = J_\pi^-(x_0)$, in which case the common value of $J_\pi^+(x_0)$ and $J_\pi^-(x_0)$ can be viewed as the cost of π starting from x_0 .

As in earlier chapters, for a stationary policy μ , we denote by $J_\mu(x_0)$ the average cost of μ starting at x_0 , and we use the following shorthand notation:

$$g_\mu = \begin{bmatrix} g(1, \mu(1)) \\ \vdots \\ g(n, \mu(n)) \end{bmatrix}, \quad P_\mu = \begin{bmatrix} p_{11}(\mu(1)) & \dots & p_{1n}(\mu(1)) \\ & \ddots & \\ p_{n1}(\mu(n)) & \dots & p_{nn}(\mu(n)) \end{bmatrix}, \quad J_\mu = \begin{bmatrix} J_\mu(1) \\ \vdots \\ J_\mu(n) \end{bmatrix}.$$

An Overview of Results

While the material of this chapter does not rely on the average cost analysis of Section 7.4 in Vol. I, it is worth summarizing some of the salient features of that analysis. We assumed there that a special state, by convention state n , is recurrent in the Markov chain corresponding to each stationary policy. The idea was to consider a sequence of generated states, and to divide it into cycles marked by successive visits to the special state n . We then argued that each of the cycles can be viewed as a state trajectory of a corresponding stochastic shortest path problem with the termination state being essentially n .

More precisely, the states of this stochastic shortest path problem are $1, \dots, n$, plus an artificial termination state t to which we move from state i with transition probability equal to $p_{in}(u)$. The transition probabilities

from a state i to a state $j \neq n$ are the same as those of the original problem, while the transition probability from i to n is 0. The expected stage cost for each state-control pair (i, u) , is $g(i, u) - \lambda$, where λ is the optimal average cost per stage starting from the special state n .

We showed that this stochastic shortest path problem is essentially equivalent to the original average cost per stage problem, and that the corresponding Bellman's equations and optimal stationary policies essentially coincide. Based on this, we showed a number of results, which in summary are the following:

- (a) The optimal average cost is independent of the initial state.
- (b) Bellman's equation takes the form

$$\lambda + h(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u)h(j) \right], \quad i = 1, \dots, n, \quad (4.2)$$

where $h(n) = 0$, λ is the optimal average cost, and $h(i)$ has the interpretation of a relative or differential cost for each state i (it is the minimum, over all policies, of the difference between the expected cost to reach n from i for the first time and the cost that would be incurred if the cost per stage were equal to the average λ at all states).

- (c) There are versions of the value iteration, policy iteration, and linear programming methods that can be used for computational solution under reasonable conditions.

We will now provide the foundation for the more powerful analysis of this chapter, based on the connection between average cost and discounted problems. Here are the highlights of this connection:

- (1) The α -discounted optimal cost can be expressed as a series expansion in α , with the first and second terms in the series being the optimal average cost (now a vector with possibly unequal components) and a differential cost vector.
- (2) There exists a stationary policy that is α -discounted optimal simultaneously for all α sufficiently close to 1, and also optimal for the average cost problem. This is called a *Blackwell optimal policy*.
- (3) The α -series expansion and Blackwell optimal policies are used to develop a *pair* of coupled optimality equations as a substitute for Bellman's equation. When the optimal average cost is equal for all initial states, the equation pair reduces to Bellman's equation (4.2).

The connection between discounted and average cost problems is also useful beyond the realm of finite-spaces problems, as we will see in Section 4.6.3.

On Finite-State Markov Chains

The theory of finite-state Markov chains plays an important role in this chapter. For the purpose of easy reference, we summarize here the definitions and properties that we will be using (see also Appendix D of Vol. I).

Given a finite-state Markov chain with transition probability matrix P , we recall that a *recurrent class* is a set of states that communicate in the sense that from every state of the set, there is a probability 1 to eventually go to all other states of the set and a probability of 0 to ever go to any state outside the set. There are two kinds of states: *recurrent*, which are those that belong to some recurrent class (these are the states that after they are visited once, they will be visited an infinite number of times with probability 1), and *transient*, which are the ones that are not recurrent (these are the states that with probability 1 will be visited only a finite number of times, regardless of the initial state). The Markov chain, as well as P , are said to be *periodic* if there is a recurrent class whose states can be grouped in $d > 1$ disjoint subsets S_1, \dots, S_d such that all transitions from one subset lead to the next subset. More precisely,

$$\text{if } i \in S_k \text{ and } p_{ij} > 0, \quad \text{then } \begin{cases} j \in S_{k+1}, & \text{if } k = 1, \dots, d-1, \\ j \in S_1, & \text{if } k = d; \end{cases}$$

otherwise they are said to be *aperiodic* (see e.g., Bertsekas and Tsitsiklis [BeT02]). The eigenvalues of P lie within the unit circle of the complex plane (have modulus less or equal to 1), with 1 being an eigenvalue with corresponding eigenvector $e = (1, \dots, 1)'$. Note that P is aperiodic if and only if all its eigenvalues, except the eigenvalue 1, lie strictly within the unit circle.

It can be shown that P is aperiodic if and only if P^k converges as $k \rightarrow \infty$ to the matrix

$$P^* = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k.$$

The limit in the above definition always exists, as will be shown as part of the following Prop. 4.1.1. If the Markov chain consists of a single recurrent class, it can be shown that it is aperiodic if and only if for some k , all the components of the matrix P^k are positive.

We finally note the structure of the matrix P^* . Its ij th component $[P^*]_{ij}$ is the long-term relative frequency of visits to state j given that the initial state is i . Thus, if i is a recurrent state, then $[P^*]_{ij} > 0$ if and only if j is recurrent and belongs to the same recurrent class as i . Furthermore, $[P^*]_{ij} = 0$ for all i if and only if j is a transient state.

4.1.1 Relation with the Discounted Cost Problem

Let us consider the cost of a stationary policy μ for the corresponding α -discounted problem. It is given by

$$J_{\alpha,\mu} = \sum_{k=0}^{\infty} \alpha^k P_{\mu}^k g_{\mu} = \left(\sum_{k=0}^{\infty} \alpha^k P_{\mu}^k \right) g_{\mu} = (I - \alpha P_{\mu})^{-1} g_{\mu}, \quad \alpha \in (0, 1). \quad (4.3)$$

To get a sense of the relation with J_{μ} , the average cost of μ , we note that we can write

$$\begin{aligned} J_{\mu}(i) &= \limsup_{N \rightarrow \infty} \frac{1}{N} E \left\{ \sum_{k=0}^{N-1} g(x_k, \mu(x_k)) \right\} \\ &= \limsup_{N \rightarrow \infty} \lim_{\alpha \rightarrow 1} \frac{E \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu(x_k)) \right\}}{\sum_{k=0}^{N-1} \alpha^k}. \end{aligned}$$

Assuming that the order of the two limiting operations in the right-hand side above can be interchanged, we obtain

$$\begin{aligned} J_{\mu}(i) &= \lim_{\alpha \rightarrow 1} \limsup_{N \rightarrow \infty} \frac{E \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu(x_k)) \right\}}{\sum_{k=0}^{N-1} \alpha^k} \\ &= \lim_{\alpha \rightarrow 1} \frac{\lim_{N \rightarrow \infty} E \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu(x_k)) \right\}}{\lim_{N \rightarrow \infty} \sum_{k=0}^{N-1} \alpha^k} \\ &= \lim_{\alpha \rightarrow 1} (1 - \alpha) J_{\alpha,\mu}(i). \end{aligned}$$

The formal proof of the above relation will follow as a corollary to the next proposition, which also provides an estimate of the difference between J_{μ} and $(1 - \alpha)J_{\alpha,\mu}$ in the form

$$(1 - \alpha) J_{\alpha,\mu} = J_{\mu} + (1 - \alpha) h_{\mu} + O(|1 - \alpha|^2),$$

where h_{μ} is some vector and $O(|1 - \alpha|^2)$ is an α -dependent vector such that $\lim_{\alpha \rightarrow 1} O(|1 - \alpha|^2)/|1 - \alpha| = 0$ (see the subsequent Prop. 4.1.2). The following proposition provides some general results on transition probability matrices, which are fundamental background for this chapter. The proof does not provide strong insights into DP, and its detailed reading is not essential for understanding the developments of the chapter.

Proposition 4.1.1: For any stochastic matrix P and $\alpha \in (0, 1)$, there holds

$$(I - \alpha P)^{-1} = (1 - \alpha)^{-1} P^* + H + O(|1 - \alpha|), \quad (4.4)$$

where $O(|1 - \alpha|)$ is an α -dependent matrix such that

$$\lim_{\alpha \rightarrow 1} O(|1 - \alpha|) = 0,$$

and the matrices P^* and H are given by

$$P^* = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k, \quad (4.5)$$

$$H = (I - P + P^*)^{-1} - P^*. \quad (4.6)$$

[It will be shown as part of the proof that the limit in Eq. (4.5) and the inverse in Eq. (4.6) exist.] Furthermore, P^* and H satisfy the following equations:

$$P^* = PP^* = P^*P = P^*P^*, \quad (4.7)$$

$$P^*H = HP^* = 0, \quad (4.8)$$

$$P^* + H = I + PH. \quad (4.9)$$

Proof: From the matrix inversion formula that expresses each entry of the inverse as a ratio of two determinants (Cramer's rule), it is seen that the matrix $M(\alpha)$, given by

$$M(\alpha) = (1 - \alpha)(I - \alpha P)^{-1},$$

can be expressed as a matrix with elements that are either zero or fractions whose numerator and denominator are polynomials in α with no common divisor. The denominator polynomials of the nonzero elements of $M(\alpha)$ cannot have 1 as a root, since otherwise some elements of $M(\alpha)$ would tend to infinity as $\alpha \rightarrow 1$; this is not possible, because from Eq. (4.3) for any μ , we have $(1 - \alpha)^{-1} M(\alpha) g_\mu = (I - \alpha P)^{-1} g_\mu = J_{\alpha, \mu}$ and $|J_{\alpha, \mu}(j)| \leq (1 - \alpha)^{-1} \max_i |g_\mu(i)|$, implying that the absolute values of the coordinates of $M(\alpha) g_\mu$ are bounded by $\max_i |g_\mu(i)|$ for all $\alpha < 1$. Therefore, the (i, j) th element of $M(\alpha)$ is of the form

$$m_{ij}(\alpha) = \frac{\gamma(\alpha - \zeta_1) \cdots (\alpha - \zeta_p)}{(\alpha - \xi_1) \cdots (\alpha - \xi_q)}$$

where γ , ζ_i , $i = 1, \dots, p$, and ξ_i , $i = 1, \dots, q$, are scalars such that $\xi_i \neq 1$ for $i = 1, \dots, q$.

Define

$$P^* = \lim_{\alpha \rightarrow 1} M(\alpha), \quad (4.10)$$

and let H be the matrix having as (i, j) th element the 1st derivative of $-m_{ij}(\alpha)$ evaluated at $\alpha = 1$. By the 1st order Taylor expansion of the elements of $m_{ij}(\alpha)$ of $M(\alpha)$, we have for all α in a neighborhood of $\alpha = 1$

$$M(\alpha) = P^* + (1 - \alpha)H + O((1 - \alpha)^2), \quad (4.11)$$

where $O((1 - \alpha)^2)$ is an α -dependent matrix such that

$$\lim_{\alpha \rightarrow 1} \frac{O((1 - \alpha)^2)}{(1 - \alpha)} = 0.$$

Multiplying Eq. (4.11) with $(1 - \alpha)^{-1}$, we obtain the desired relation (4.4) [although, we have yet to show that P^* and H are also given by Eqs. (4.5) and (4.6), respectively].

We will now show that P^* as defined by Eq. (4.10), satisfies Eqs. (4.7), (4.6), (4.8), (4.9), and (4.5), in that order.

We have

$$(I - \alpha)(I - \alpha P)(I - \alpha P)^{-1} = (1 - \alpha)I \quad (4.12)$$

and by rearranging terms, we obtain

$$\alpha P(1 - \alpha)(I - \alpha P)^{-1} = (1 - \alpha)(I - \alpha P)^{-1} + (\alpha - 1)I.$$

By taking the limit as $\alpha \rightarrow 1$ and using the definition (4.10), it follows that

$$PP^* = P^*.$$

Also, by reversing the order of $(I - \alpha P)$ and $(I - \alpha P)^{-1}$ in Eqs. (4.12), it follows similarly that $P^*P = P^*$. From $PP^* = P^*$, we also obtain $(I - \alpha P)P^* = (1 - \alpha)P^*$ or $P^* = (1 - \alpha)(I - \alpha P)^{-1}P^*$, and by taking the limit as $\alpha \rightarrow 1$ and by using Eq. (4.10), we have $P^* = P^*P^*$. Thus Eq. (4.7) has been proved.

We have, using Eq. (4.7), $(P - P^*)^2 = P^2 - P^*$, and similarly

$$(P - P^*)^k = P^k - P^*, \quad k > 0.$$

Therefore,

$$\begin{aligned} (I - \alpha P)^{-1} - (1 - \alpha)^{-1}P^* &= \underbrace{\sum_{k=0}^{\infty} \alpha^k (P^k - P^*)}_{=} \\ &= I - P^* + \sum_{k=1}^{\infty} \alpha^k (P - P^*)^k \\ &= (I - \alpha(P - P^*))^{-1} - P^*. \end{aligned}$$

On the other hand, from Eq. (4.11), we have

$$\begin{aligned} H &= \lim_{\alpha \rightarrow 1} ((1 - \alpha)^{-1} M(\alpha) - (1 - \alpha)^{-1} P^*) \\ &= \lim_{\alpha \rightarrow 1} ((I - \alpha P)^{-1} - (1 - \alpha)^{-1} P^*). \end{aligned}$$

By combining the preceding two equations, we have

$$H + P^* = \lim_{k \rightarrow \infty} A_k, \quad (4.13)$$

where

$$A_k = (I - \alpha_k(P - P^*))^{-1}$$

and $\{\alpha_k\}$ is a sequence with $\alpha_k \uparrow 1$. Multiplying the left-hand side of Eq. (4.13) with $I - P + P^*$ and the right-hand side with the equal matrix $\lim_{k \rightarrow \infty} A_k^{-1}$, we obtain

$$(I - P + P^*)(H + P^*) = \lim_{k \rightarrow \infty} A_k^{-1} \lim_{k \rightarrow \infty} A_k = \lim_{k \rightarrow \infty} (A_k^{-1} A_k) = I,$$

where the second equality follows because the two limits on the left are known to exist. Thus we have

$$H + P^* = (I - P + P^*)^{-1},$$

which yields the desired formula (4.6) for H .

From Eq. (4.6), we obtain

$$(I - P + P^*)H = I - (I - P + P^*)P^*$$

or, using Eq. (4.7),

$$(I - P + P^*)H = I - P^*. \quad (4.14)$$

Premultiplying this relation by P^* and using Eq. (4.7), we obtain $P^*H = 0$, which is one part of Eq. (4.8). Equation (4.9) then follows from Eq. (4.14). Similarly, postmultiplying Eq. (4.14) by P^* and using Eq. (4.7), we obtain $(I - P + P^*)HP^* = 0$, which in view of the invertibility of $I - P + P^*$, implies that $HP^* = 0$, the remaining part of Eq. (4.8).

Multiplying Eq. (4.9) with P^k and using Eq. (4.7), we obtain

$$P^* + P^k H = P^k + P^{k+1} H, \quad k = 0, 1 \dots$$

Adding this relation over $k = 0, \dots, N - 1$, we have

$$NP^* + H = \sum_{k=0}^{N-1} P^k + P^N H.$$

Dividing by N , taking the limit as $N \rightarrow \infty$, and using the fact $P^N/N \rightarrow 0$ (since P and hence also P^N is a stochastic matrix) we obtain Eq. (4.5). Q.E.D.

Note that the matrix P^* of Eq. (4.5) can be used to express concisely the average cost vector J of any Markov chain with transition probability matrix P and cost vector g as a limit:

$$J = \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k g = \left(\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k \right) g,$$

and finally

$$J = P^* g.$$

To interpret this equation, note that we may view the i th row of P^* as a vector of steady-state occupancy frequencies corresponding to starting at state i ; i.e., the ij th element p_{ij}^* of P^* represents the long-term fraction of time that the Markov chain spends at state j given that it starts at state i . Thus the above equation gives the average cost $J(i)$, starting from state i , as the sum $\sum_{j=1}^n p_{ij}^* g_j$ of all the single-stage costs g_j weighted by the corresponding occupancy frequencies.

The following proposition relates the α -discounted and average costs corresponding to a stationary policy.

Proposition 4.1.2: (Laurent Series Expansion) For any stationary policy μ and $\alpha \in (0, 1)$, we have

$$J_{\alpha, \mu} = (1 - \alpha)^{-1} J_\mu + h_\mu + O(|1 - \alpha|), \quad (4.15)$$

where J_μ and h_μ are given by

$$J_\mu = P_\mu^* g_\mu, \quad h_\mu = H_\mu g_\mu, \quad (4.16)$$

with

$$P_\mu^* = \left(\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} P_\mu^k \right), \quad H_\mu = (I - P_\mu + P_\mu^*)^{-1} - P_\mu^*$$

[cf. Eqs. (4.5) and (4.6)]. Furthermore, we have

$$J_\mu = P_\mu J_\mu, \quad (4.17)$$

$$J_\mu + h_\mu = g_\mu + P_\mu h_\mu. \quad (4.18)$$

Proof: Equation (4.15) follows from Eqs. (4.3) and (4.4) with the identifications $F = P_\mu$, $P^* = P_\mu^*$, $H = H_\mu$, and $h_\mu = H_\mu g_\mu$. Equation (4.18) follows by multiplying Eq. (4.9) with g_μ and by using the same identifications. Equation (4.17) follows from $J_\mu = P_\mu^* g_\mu$ by multiplying with P_μ , and by using the fact $P_\mu^* = P_\mu P_\mu^*$ [cf. Eq. (4.7)]. **Q.E.D.**

Equation (4.15) will be referred to as the *Laurent series expansion* of the discounted cost of a stationary policy μ .[†] The vectors J_μ and h_μ in the Laurent series expansion are uniquely defined, and will be referred to as the *gain* and *bias* of μ , respectively.

We note two useful equations regarding the bias. The first is

$$P_\mu^* h_\mu = 0, \quad (4.19)$$

which follows from the definition $h_\mu = H_\mu g_\mu$ and the fact $P_\mu^* H_\mu = 0$ [cf. Eq. (4.8)]. The second is

$$h_\mu = \lim_{N \rightarrow \infty} \sum_{k=0}^N P_\mu^k (g_\mu - J_\mu), \quad (4.20)$$

which holds under the assumption

$$P_\mu^* = \lim_{N \rightarrow \infty} P_\mu^N. \quad (4.21)$$

[†] Equation (4.15) is sometimes referred to as the *truncated Laurent series expansion* to distinguish it from another more detailed version of the expansion, where the term $O(|1 - \alpha|)$ is explicitly defined as a power series involving α , g_μ , and P_μ . This power series expansion is given by

$$J_{\alpha, \mu} = (1 + \rho) \left(\rho^{-1} J_\mu + h_\mu + \sum_{k=1}^{\infty} \rho^k y_k \right),$$

where

$$\rho = \frac{1 - \alpha}{\alpha}, \quad y_k = (-1)^k H_\mu^{k+1} g_\mu, \quad k = 1, 2, \dots$$

and

$$H_\mu = (I - P_\mu + P_\mu^*)^{-1} - P_\mu^*$$

[cf. Eq. (4.6)]. Note that the above expansion is consistent with the truncated version given in Prop. 4.1.2 (the term ρh_μ can be lumped into the first term of the summation). The expansion is valid for $0 < \rho < |\nu|$, where ν is the nonzero eigenvalue of $I - P_\mu$ with the smallest modulus. We will not need this more detailed form of the expansion; it is useful in several topics of interest in average cost problem analysis, which are, however, beyond our scope.

(This assumption is in turn satisfied if P_μ corresponds to a Markov chain with no periodic recurrent classes; see e.g., [BeT02].) To show Eq. (4.20), we use the equation $g_\mu - J_\mu = h_\mu - P_\mu h_\mu$ [cf. Eq. (4.18)] to write

$$\sum_{k=0}^N P_\mu^k (g_\mu - J_\mu) = \sum_{k=0}^N P_\mu^k (h_\mu - P_\mu h_\mu) = h_\mu - P_\mu^{N+1} h_\mu,$$

then take the limit as $N \rightarrow \infty$, and use Eqs. (4.21) and (4.19). An interesting interpretation of Eq. (4.20) is that *the bias may be viewed as a relative cost*: it is the difference of the total cost of μ , and the total cost that would be incurred if the cost per stage were the average J_μ .

Unfortunately, the gain-bias pair (J_μ, h_μ) of μ cannot be determined by solving the system of equations (4.17) and (4.18) because this system has an infinite number of solutions; for example the pair $(J_\mu, h_\mu + \gamma e)$, where γ is a scalar and e is the unit vector (all components equal to 1), is a solution. We will address this issue and characterize the set of all solutions in the subsequent Prop. 4.1.9, and also when we discuss policy iteration in Section 4.4.

4.1.2 Blackwell Optimal Policies

The Laurent series expansion (Prop. 4.1.2) shows that J_μ , the average cost vector of a stationary policy μ , consists of three α -dependent terms:

$$J_\mu = (1 - \alpha) J_{\alpha, \mu} - (1 - \alpha) h_\mu + O(|1 - \alpha|^2), \quad (4.22)$$

of which the term $(1 - \alpha) J_{\alpha, \mu}$ tends to dominate for $\alpha \approx 1$, since the components of $J_{\alpha, \mu}$ ordinarily become infinite asymptotically as $\alpha \rightarrow 1$. It would thus appear that a policy minimizing $J_{\alpha, \mu}$ for all $\alpha \approx 1$ should also minimize the average cost J_μ . This motivates a special type of policy, which will provide a key conceptual link between average cost and discounted problems.

Definition 4.1.1: A stationary policy μ is said to be *Blackwell optimal* if it is simultaneously optimal for all the α -discounted problems with α in an interval $(\bar{\alpha}, 1)$, where $\bar{\alpha}$ is some scalar with $0 < \bar{\alpha} < 1$.

Note from Eq. (4.22) that for any stationary policy μ , we have

$$J_\mu = \lim_{\alpha \rightarrow 1} (1 - \alpha) J_{\alpha, \mu}.$$

Since a Blackwell optimal policy minimizes $J_{\alpha, \mu}$ over μ for all α sufficiently close to 1, it follows that a Blackwell optimal policy is optimal within the class of stationary policies.

We will show later that a Blackwell optimal policy is optimal over all policies, stationary or not (Prop. 4.1.7). There may exist stationary optimal policies that are not Blackwell optimal (see Exercise 4.4). However, a Blackwell optimal policy has some advantage over other types of average cost optimal policies: not only it minimizes the average cost per stage, as just mentioned, but also, by definition, it minimizes the α -discounted cost for $\alpha \approx 1$. Thus it optimizes not just the steady-state average performance, but also, to some extent, the *transient* performance of the system; see also the discussion of m -discount optimality in Section 4.7.

The following proposition establishes the existence of Blackwell optimal policies.

Proposition 4.1.3: There exists a Blackwell optimal policy.

Proof: From the relation

$$J_{\alpha,\mu} = (I - \alpha P_\mu)^{-1} g_\mu$$

[cf. Eq. (4.3)], and Cramer's rule for expressing the inverse of the matrix in terms of determinants, we know that, for each μ and state i , $J_{\alpha,\mu}(i)$ is a rational function of α , i.e., a ratio of two polynomials in α . Therefore, for any two stationary policies μ and μ' the graphs of $J_{\alpha,\mu}(i)$ and $J_{\alpha,\mu'}(i)$ either coincide or cross only a finite number of times in the interval $(0, 1)$. Since there are only a finite number of stationary policies, we conclude that for each state i there is a policy μ^i and a scalar $\bar{\alpha}_i \in (0, 1)$ such that μ^i is optimal for the α -discounted problem for $\alpha \in (\bar{\alpha}_i, 1)$ when the initial state is i . Then, for each i , $\mu^i(i)$ attains the minimum in Bellman's equation for the α -discounted problem

$$J_\alpha(i) = \min_{u \in U(i)} \left[g(i, u) + \alpha \sum_{j=1}^n p_{ij}(u) J_\alpha(j) \right],$$

for all α in the interval $(\max_i \bar{\alpha}_i, 1)$ (cf. Prop. 1.2.3). Consider the stationary policy defined for each i by $\mu^*(i) = \mu^i(i)$. Then for all i , $\mu^*(i)$ attains the minimum in Bellman's equation, so μ^* is optimal for the α -discounted problem for all $\alpha \in (\max_i \bar{\alpha}_i, 1)$ (cf. Prop. 1.2.3). Hence μ^* is Blackwell optimal. **Q.E.D.**

We will now use Blackwell optimal policies as an analytical vehicle, to derive the analog of Bellman's equation for average cost problems. In the process, we will also show that a Blackwell optimal policy is optimal over all policies. The following proposition provides the first step in this development.

Proposition 4.1.4:

- (a) All Blackwell optimal policies have the same gain and bias vectors, i.e., for any two Blackwell optimal μ and μ' ,

$$J_\mu = J_{\mu'}, \quad h_\mu = h_{\mu'},$$

where (J_μ, h_μ) and $(J_{\mu'}, h_{\mu'})$ correspond to μ and μ' , respectively, in the Laurent series expansion (cf. Prop. 4.1.2).

- (b) Let (J^*, h^*) be the gain-bias pair common to all Blackwell optimal policies as per part (a). We have

$$J^*(i) = \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u) J^*(j), \quad i = 1, \dots, n, \quad (4.23)$$

and

$$J^*(i) + h^*(i) = \min_{u \in \bar{U}(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u) h^*(j) \right], \quad i = 1, \dots, n, \quad (4.24)$$

where, for each i , $\bar{U}(i)$ is the set of controls attaining the minimum in Eq. (4.23). Furthermore, if μ^* is a Blackwell optimal policy, $\mu^*(i)$ attains the minimum in the right-hand sides of these two equations for all i .

Proof: (a) From Prop. 4.1.2, we have

$$J_{\alpha,\mu} = (1 - \alpha)^{-1} J_\mu + h_\mu + O(|1 - \alpha|),$$

$$J_{\alpha,\mu'} = (1 - \alpha)^{-1} J_{\mu'} + h_{\mu'} + O(|1 - \alpha|).$$

Since $J_{\alpha,\mu} = J_{\alpha,\mu'}$ for all α sufficiently close to 1, by taking the limit as $\alpha \rightarrow 1$ in the above equations, we obtain $J_\mu = J_{\mu'}$ and $h_\mu = h_{\mu'}$.

(b) Let μ^* be a Blackwell optimal policy. Since μ^* is optimal for the α -discounted problem for all α in an interval $(\bar{\alpha}, 1)$, we must have, for every stationary policy μ and $\alpha \in (\bar{\alpha}, 1)$,

$$g_{\mu^*} + \alpha P_{\mu^*} J_{\alpha,\mu^*} \leq g_\mu + \alpha P_\mu J_{\alpha,\mu^*}. \quad (4.25)$$

From Prop. 4.1.2, we have, for all $\alpha \in (\bar{\alpha}, 1)$,

$$J_{\alpha,\mu^*} = (1 - \alpha)^{-1} J^* + h^* + O(|1 - \alpha|).$$

Substituting this expression in Eq. (4.25), we obtain

$$0 \leq g_\mu - g_{\mu^*} + \alpha(P_\mu - P_{\mu^*})((1-\alpha)^{-1}J^* + h^* + O(|1-\alpha|)), \quad (4.26)$$

or equivalently, multiplying with $1-\alpha$,

$$0 \leq (1-\alpha)(g_\mu - g_{\mu^*}) + \alpha(P_\mu - P_{\mu^*})(J^* + (1-\alpha)h^* + O((1-\alpha)^2)).$$

By taking the limit as $\alpha \rightarrow 1$, we obtain $P_{\mu^*}J^* \leq P_\mu J^*$, which is equivalent to Eq. (4.23).

If μ is such that $P_{\mu^*}J^* = P_\mu J^*$, then from Eq. (4.26) we have

$$0 \leq g_\mu - g_{\mu^*} + \alpha(P_\mu - P_{\mu^*})(h^* + O(|1-\alpha|)).$$

By taking the limit as $\alpha \rightarrow 1$, we see that μ^* minimizes $g_\mu + P_\mu h^*$ over μ , and by using also the equation $J^* + h^* = g_{\mu^*} + P_{\mu^*}h^*$ (cf. Prop. 4.1.2), we obtain the desired relation (4.24). Q.E.D.

Some insight into the pair of equations satisfied by a Blackwell optimal policy [Eqs. (4.23) and (4.24)] may be obtained from the preceding proof. A Blackwell optimal policy first minimizes $P_\mu J^*$ over μ , which corresponds to the most significant $\alpha(1-\alpha)^{-1}$ -order term in Eq. (4.26), and among policies that minimize $P_\mu J^*$, it minimizes over μ the next most significant term $g_\mu + P_\mu h^*$.

The equations bear a resemblance to Bellman's equation that we encountered in Section 7.4 of Vol. I, but there is an important difference: the constraint set $\bar{U}(i)$ in the second equation depends on the outcome of the minimization in the first equation. In the special case where the average cost $J^*(i)$ of the Blackwell optimal policies is independent of the initial state i , the first equation is trivially satisfied, and we have $\bar{U}(i) = U(i)$ for all i . Then, the second equation becomes identical to Bellman's equation, as we have encountered it in Section 7.4 of Vol. I [cf. Eq. (4.2)]. In general, however, we may have $\bar{U}(i) \neq U(i)$. Here is an example.

Example 4.1.1

Consider an average cost problem with two states, 1 and 2, and two controls, 1 and 2. Control 1 keeps the system at the state where it is, at a cost of 1 or 2, when at state 1 or 2, respectively. Control 2 is available only at state 1, and it moves the system to state 2 at a cost of -10 (see Fig. 4.1.1).

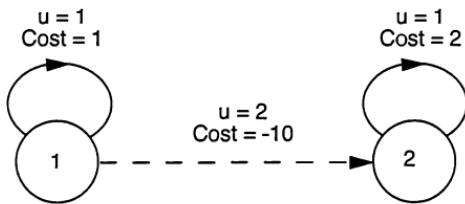


Figure 4.1.1 Transition probabilities for Example 4.1.1.

From the point of view of achieving a low short-term cost, there is an incentive to move from state 1 to state 2 by applying control 2, but this leads to the worse long-term cost of 2 per stage. Thus, there is a unique optimal policy μ , which is also Blackwell optimal: apply control 1 at both states, and attain an average cost of 1 and 2 starting from states 1 and 2, respectively. The transition probability matrix P_μ is the identity matrix,

$$P_\mu = I.$$

The associated average costs are

$$J^*(1) = 1, \quad J^*(2) = 2,$$

and the associated α -discounted costs are

$$J_{\alpha,\mu}(1) = (1 - \alpha)^{-1}, \quad J_{\alpha,\mu}(2) = 2(1 - \alpha)^{-1}.$$

The bias h^* corresponding to μ [cf. Eq. (4.15)] is

$$h^*(1) = h^*(2) = 0.$$

Equations (4.23) and (4.24) take the form

$$J^*(1) = \min\{J^*(1), J^*(2)\}, \quad J^*(2) = J^*(2),$$

$$J^*(1) + h^*(1) = 1 + h^*(1), \quad J^*(2) + h^*(2) = 2 + h^*(2),$$

and are clearly satisfied by J^* and h^* [we have $\bar{U}(1) = \bar{U}(2) = \{1\}$]. On the other hand, if we were to replace $\bar{U}(i)$ with $U(i)$ in Eq. (4.24), the corresponding equation would not be satisfied because

$$J^*(1) + h^*(1) \neq \min\{1 + h^*(1), -10 + h^*(2)\}.$$

The following proposition provides a mechanism for simplifying the coupled pair of equations (4.23) and (4.24), and leads to a proof of optimality of a Blackwell optimal policy. It introduces a “penalty” into the

second equation, which induces a preference for controls that satisfy the first equation.

Proposition 4.1.5: Let (J^*, h^*) satisfy the pair of equations

$$J^*(i) = \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u) J^*(j), \quad i = 1, \dots, n, \quad (4.27)$$

and

$$J^*(i) + h^*(i) = \min_{u \in \bar{U}(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u) h^*(j) \right], \quad i = 1, \dots, n, \quad (4.28)$$

where, for each i , $\bar{U}(i)$ is the set of controls attaining the minimum in Eq. (4.27). For any scalar γ , denote

$$h_\gamma = h^* + \gamma J^*.$$

Then there exists some $\bar{\gamma} \geq 0$ such that for all $\gamma \geq \bar{\gamma}$, we have

$$J^*(i) + h_\gamma(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u) h_\gamma(j) \right], \quad i = 1, \dots, n. \quad (4.29)$$

Furthermore, if a stationary policy μ is such that $\mu(i)$ attains the minimum in Eq. (4.28) for all i , then $\mu(i)$ attains the minimum in Eq. (4.29) for all i .

Proof: The idea of the proof is that the addition of the penalty

$$\gamma \sum_{j=1}^n p_{ij}(u) J^*(j) \quad (4.30)$$

to the expression minimized in Eq. (4.28) makes controls $u \notin \bar{U}(i)$ unattractive in the minimization of Eq. (4.29), for sufficiently large γ .

Let μ be such that for all i , $\mu(i)$ attains the minimum in Eq. (4.28). Then we have

$$J^*(i) + h^*(i) - \left(g(i, u) + \sum_{j=1}^n p_{ij}(u) h^*(j) \right) \leq 0, \quad i = 1, \dots, n, \quad u \in \bar{U}(i). \quad (4.31)$$

with equality holding when $u = \mu(i)$.

Using the definition $h_\gamma = h^* + \gamma J^*$, let us write

$$\begin{aligned} J^*(i) + h_\gamma(i) - \left(g(i, u) + \sum_{j=1}^n p_{ij}(u)h_\gamma(j) \right) &= J^*(i) + h^*(i) \\ &\quad - \left(g(i, u) + \sum_{j=1}^n p_{ij}(u)h^*(j) \right) + \gamma \left(J^*(i) - \sum_{j=1}^n p_{ij}(u)J^*(j) \right). \end{aligned} \quad (4.32)$$

For all i and $u \notin \overline{U}(i)$, we have

$$J^*(i) - \sum_{j=1}^n p_{ij}(u)J^*(j) < 0,$$

and from Eq. (4.32), it follows that there exists a positive scalar $\gamma(i, u)$ such that

$$J^*(i) + h_\gamma(i) - \left(g(i, u) + \sum_{j=1}^n p_{ij}(u)h_\gamma(j) \right) \leq 0, \quad \text{for all } \gamma \geq \gamma(i, u).$$

For $u \in \overline{U}(i)$, we have $J^*(i) - \sum_{j=1}^n p_{ij}(u)J^*(j) = 0$, so that using Eqs. (4.31) and (4.32), we obtain

$$J^*(i) + h_\gamma(i) - \left(g(i, u) + \sum_{j=1}^n p_{ij}(u)h_\gamma(j) \right) \leq 0, \quad \text{for all } \gamma \geq 0,$$

with equality when $u = \mu(i)$. By taking $\bar{\gamma}$ to be the maximum of $\gamma(i, u)$ over the set $\{(i, u) \mid u \notin \overline{U}(i)\}$ (or $\bar{\gamma} = 0$ if this set is empty), we obtain for all $\gamma \geq \bar{\gamma}$,

$$J^*(i) + h_\gamma(i) - \left(g(i, u) + \sum_{j=1}^n p_{ij}(u)h_\gamma(j) \right) \leq 0, \quad i = 1, \dots, n, \quad u \in U(i),$$

with equality when $u = \mu(i)$, which is the desired result. **Q.E.D.**

Note that by Prop. 4.1.4, the Blackwell optimal policies attain the minimum in Eq. (4.28), and their gain-bias pair (J^*, h^*) satisfies the pair of equations (4.27) and (4.28). Hence, by Prop. 4.1.5, they also attain the minimum in Eq. (4.29).

Example 4.1.1 (continued)

For this example, Eq. (4.29) has the form

$$J^*(1) + h^*(1) + \gamma J^*(1) = \min\{1 + h^*(1) + \gamma J^*(1), -10 + h^*(2) + \gamma J^*(2)\},$$

$$J^*(2) + h^*(2) + \gamma J^*(2) = 2 + h^*(2) + \gamma J^*(2),$$

and it is satisfied, for all $\gamma \geq 11$, by

$$J^* = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad h^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which are the gain and bias of the Blackwell optimal policy.

We can write Eq. (4.29) as

$$J^* + h_\gamma = Th_\gamma,$$

where, as in earlier chapters, the mapping T is defined by

$$(Th)(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u)h(j) \right], \quad i = 1, \dots, n.$$

For a stationary policy μ , we will also use the mapping T_μ , defined by

$$(T_\mu h)(i) = g(i, \mu(i)) + \sum_{j=1}^n p_{ij}(\mu(i))h(j), \quad i = 1, \dots, n.$$

Then, by Eqs. (4.17) and (4.18), we have

$$J_\mu = P_\mu J_\mu, \quad J_\mu + h_\mu = T_\mu h_\mu.$$

The following proposition provides the essence of the argument for optimality of a Blackwell optimal policy.

Proposition 4.1.6: Let J and h be n -dimensional vectors, and let $\pi = \{\mu_0, \mu_1, \dots\}$ be an admissible policy. Assume that for all k , we have

$$P_{\mu_k} J \geq J, \quad T_{\mu_k} h \geq J + h. \quad (4.33)$$

Then

$$J_\pi \geq J.$$

Furthermore, if equality holds in Eq. (4.33) for all k , then $J_\pi = J$.

Proof: For any μ , we have $T_\mu(J + h) = g_\mu + P_\mu(J + h) = P_\mu J + T_\mu h$, so from Eq. (4.33), it follows that for all k ,

$$T_{\mu_k}(J + h) \geq J + T_{\mu_k}h. \quad (4.34)$$

Let N be a positive integer. We have, from Eq. (4.33),

$$T_{\mu_{N-1}}h \geq J + h.$$

By applying $T_{\mu_{N-2}}$ to both sides of this relation, and by using the monotonicity of $T_{\mu_{N-2}}$ and Eqs. (4.33), (4.34), we see that

$$T_{\mu_{N-2}}T_{\mu_{N-1}}h \geq T_{\mu_{N-2}}(J + h) \geq J + T_{\mu_{N-2}}h \geq 2J + h.$$

By applying $T_{\mu_{N-3}}$ to both sides of this relation, and by continuing similarly, we obtain

$$T_{\mu_0}T_{\mu_1}\cdots T_{\mu_{N-1}}h \geq NJ + h, \quad (4.35)$$

with equality in the above relation if equality holds in Eq. (4.33) for all k .

As discussed in Section 1.1, $(T_{\mu_0}T_{\mu_1}\cdots T_{\mu_{N-1}}h)(i)$ is equal to the N -stage cost corresponding to initial state i , policy $\{\mu_0, \mu_1, \dots, \mu_{N-1}\}$, and terminal cost function h ; i.e.,

$$(T_{\mu_0}T_{\mu_1}\cdots T_{\mu_{N-1}}h)(i) = E \left\{ h(x_N) + \sum_{k=0}^{N-1} g(x_k, \mu_k(x_k)) \mid x_0 = i, \pi \right\}.$$

Using this relation in Eq. (4.35) and dividing by N , we obtain for all i

$$\begin{aligned} \frac{1}{N} E \left\{ h(x_N) \mid x_0 = i, \pi \right\} &+ \frac{1}{N} E \left\{ \sum_{k=0}^{N-1} g(x_k, \mu_k(x_k)) \mid x_0 = i, \pi \right\} \\ &\geq J(i) + \frac{1}{N} h(i). \end{aligned} \quad (4.36)$$

By taking \limsup as $N \rightarrow \infty$, we see that

$$J_\pi(i) \geq J(i), \quad i = 1, \dots, n.$$

If equality holds in Eq. (4.33) for all k , then all the preceding inequalities become equalities, so $J_\pi = J$. **Q.E.D.**

The preceding proposition provides a sufficient condition for optimality of a stationary policy. In particular, if μ is a stationary policy such that for some vectors J and h , we have

$$J = P_\mu J = \min_{\mu'} P_{\mu'} J, \quad J + h = T_\mu h = \min_{\mu'} T_{\mu'} h, \quad (4.37)$$

then Prop. 4.1.6 implies that μ is optimal and J is equal to the optimal average cost vector. In this way, we can show the optimality of a Blackwell optimal policy.

Proposition 4.1.7: A Blackwell optimal policy is optimal over all policies in the average cost problem.

Proof: Let μ be a Blackwell optimal policy, and denote

$$J = J_\mu, \quad h = h_\mu + \gamma J,$$

where (J_μ, h_μ) is the gain-bias pair corresponding to μ (cf. Prop. 4.1.2), and γ is a scalar that is sufficiently large so that $Th = J + h$ (cf. Prop. 4.1.5). Then, we have for all stationary policies μ' ,

$$T_{\mu'} h \geq J + h.$$

Since we also have, by Eq. (4.27), $P_{\mu'} J \geq J$ for all μ' , Prop. 4.1.6 implies that $J_\pi \geq J$ for all policies π . Q.E.D.

As an application of the preceding proposition, let us show that the optimal average cost is the same for all initial states under the condition that for some constants $L > 0$ and $\bar{\alpha} \in (0, 1)$, we have

$$|J_\alpha(i) - J_\alpha(j)| \leq L, \quad \text{for all } i, j = 1, \dots, n, \text{ and } \alpha \in (\bar{\alpha}, 1), \quad (4.38)$$

where J_α is the α -discounted optimal cost vector. Indeed, let μ be a Blackwell optimal policy. Then, for all $\alpha \in (\bar{\alpha}, 1)$ and i we have, by the Laurent series expansion (cf. Prop. 4.1.2),

$$J_\mu(i) = (1 - \alpha)J_\alpha(i) - (1 - \alpha)h_\mu(i) + O(|1 - \alpha|^2).$$

Writing this equation for states i and j , and subtracting, we obtain

$$|J_\mu(i) - J_\mu(j)| \leq (1 - \alpha)|J_\alpha(i) - J_\alpha(j)| + (1 - \alpha)|h_\mu(i) - h_\mu(j)| + O((1 - \alpha)^2).$$

Taking the limit as $\alpha \rightarrow 1$ and using the hypothesis (4.38), we see that $J_\mu(i) = J_\mu(j)$ for all i and j . Since by Prop. 4.1.7, μ is optimal for the average cost problem, we see that under the assumption (4.38), the optimal average cost is independent of the initial state. We note that while the condition (4.38) is of limited utility for finite-spaces problems, in Section 4.6.3 it will become the starting point of an important line of analysis for infinite-spaces average cost problems.

4.1.3 Optimality Equations

Let us consider the coupled pair of equations (4.27) and (4.28), repeated here for easy reference:

$$J^*(i) = \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u) J^*(j), \quad i = 1, \dots, n, \quad (4.39)$$

and

$$J^*(i) + h^*(i) = \min_{u \in \bar{U}(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u) h^*(j) \right], \quad i = 1, \dots, n, \quad (4.40)$$

where $\bar{U}(i)$ is the set of controls attaining the minimum in Eq. (4.39).

This pair of equations can be viewed as an analog to Bellman's equation, encountered in the preceding chapters for various types of total cost problems. Since the gain-bias pair (J^*, h^*) common to all Blackwell optimal policies satisfies these equations, and there exists a Blackwell optimal policy (cf. Prop. 4.1.3), the equations always have at least one solution. Conversely, any solution yields optimal stationary policies by minimization in the right-hand side, as shown in the following proposition.

Proposition 4.1.8: If J^* and h^* satisfy the pair of optimality equations (4.39) and (4.40), then J^* is equal to the optimal average cost vector. Furthermore, if $\mu^*(i)$ attains the minimum in Eqs. (4.39) and (4.40) for each i , then the stationary policy μ^* is optimal.

Proof: Using Prop. 4.1.5, we select γ such that $h_\gamma = h^* + \gamma J^*$ satisfies

$$J^*(i) + h_\gamma(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u) h_\gamma(j) \right], \quad i = 1, \dots, n. \quad (4.41)$$

Let $\mu^*(i)$ attain the minimum in Eq. (4.40) for each i , so that by Prop. 4.1.5, $\mu^*(i)$ attains the minimum in Eq. (4.41) for each i . By applying Prop. 4.1.6 with $J = J^*$ and $h = h_\gamma$, we have $J_\pi \geq J^*$, for every policy π . Applying Prop. 4.1.6 with $J = J^*$ and $\pi = \{\mu^*, \mu^*, \dots\}$, we obtain $J_{\mu^*} = J^*$. Thus,

$$J_\pi \geq J^* = J_{\mu^*}$$

for all π . Q.E.D.

In the important case where the optimal average cost is equal for all states $[J^*(i) = \lambda$ for some λ and all $i]$, Eq. (4.39) is automatically satisfied

and is superfluous, so we have $U(i) = \bar{U}(i)$ for all i . In this case, the pair of optimality equations is equivalent to the single equation

$$\lambda + h(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u)h(j) \right], \quad i = 1, \dots, n, \quad (4.42)$$

which is the one we encountered in our introductory treatment of average cost problems of Section 7.4, Vol. I.

In what follows, we will refer to Eq. (4.42) as *Bellman's equation* for the average cost problem, with the understanding that it holds only in the case where the optimal average cost is equal to a constant λ for all initial states. We can write this equation as $\lambda e + h = Th$, where

$$(Th)(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u)h(j) \right], \quad i = 1, \dots, n,$$

and $e = (1, \dots, 1)'$ is the unit vector. The vector h can be interpreted as a *differential* or *relative* cost vector, as discussed in Section 7.4 of Vol. I. Note that while at least one optimal stationary policy can be obtained by minimizing the right-hand side of Bellman's equation (including all Blackwell-optimal policies), not all optimal stationary policies can be obtained in this way. An example is given in Exercise 4.4 (see also Exercise 4.17); this is contrary to what happens in discounted problems (cf. Prop. 1.2.3). The intuitive reason is that optimal policies that are non-Blackwell optimal may have a larger bias than the bias of the Blackwell-optimal policies, which in turn is the one that solves Bellman's equation; see also the discussion of discount optimality in Section 4.7.

In the case of a stationary policy μ whose average cost is equal to a constant λ_μ for all initial states, Bellman's equation takes the form $\lambda_\mu e + h = T_\mu h$, where

$$(T_\mu h)(i) = g(i, \mu(i)) + \sum_{j=1}^n p_{ij}(\mu(i))h(j), \quad i = 1, \dots, n.$$

This is simply a restatement of Eq. (4.18) in Prop. 4.1.2, for the special case where the average cost of μ is the same for all initial states.

Given a stationary policy μ , we may consider a problem where the constraint set $U(i)$ is replaced by the set $\tilde{U}(i) = \{\mu(i)\}$; i.e., $\tilde{U}(i)$ contains a single element, the control $\mu(i)$. Then the optimality equations become the linear system of $2n$ equations with $2n$ unknowns for the gain-bias pair (J_μ, h_μ) , which was derived in Prop. 4.1.2. The following proposition characterizes the set of solutions of this system.

Proposition 4.1.9: Let μ be a stationary policy with gain-bias pair (J_μ, h_μ) . The set of solutions of the system of equations

$$J = P_\mu J, \quad (4.43)$$

$$J + h = g_\mu + P_\mu h \quad (4.44)$$

is the set of pairs of the form $(J_\mu, h_\mu + d)$ such that $d = P_\mu d$.

Proof: To simplify notation, we drop the subscript μ , and we denote the gain and bias of μ by \bar{J} and \bar{h} , to distinguish them from the generic vectors J and h . We will use the formulas of Prop. 4.1.1 for P^* and H :

$$P^* = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k, \quad H = (I - P + P^*)^{-1} - P^*.$$

Let (J, h) be a solution of the system (4.43), (4.44). By multiplying Eq. (4.44) by P^* , adding Eq. (4.43), and using the fact $P^*P = P^*$ [cf. Eq. (4.7)], we have

$$(I - P + P^*)J = P^*g.$$

From this, using also Eq. (4.6), and the facts $P^* = P^*P^*$ and $HP^* = 0$ [cf. Eqs. (4.7) and (4.8)], we obtain

$$J = (I - P + P^*)^{-1}P^*g = (H + P^*)P^*g = P^*g = \bar{J}.$$

Since we just showed that $J = \bar{J} = P^*g$, Eq. (4.44) is written as $P^*g + h = g + Ph$, or

$$(I - P)h = (I - P^*)g.$$

This equation, using the fact $I - P^* = (I - P)H$ [cf. Eq. (4.9)] and the fact $Hg = \bar{h}$ [cf. Eq. (4.16)], yields $(I - P)(h - \bar{h}) = 0$, i.e., the vector $d = h - \bar{h}$ satisfies $d = Pd$.

Conversely, by Prop. 4.1.2, the gain-bias pair (\bar{J}, \bar{h}) is a solution of the system (4.43), (4.44), and from the form of Eq. (4.44), it follows that $(\bar{J}, \bar{h} + d)$ is also a solution for every d with $d = Pd$. Q.E.D.

Modified Optimality Equations

Let us also note the result of Prop. 4.1.5, and the associated sufficient condition for optimality (4.37). They imply that if J^* and h satisfy the equations

$$J^*(i) = \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u) J^*(j), \quad i = 1, \dots, n, \quad (4.45)$$

$$J^*(i) + h(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u)h(j) \right], \quad i = 1, \dots, n, \quad (4.46)$$

and μ^* is a stationary policy such that $\mu^*(i)$ attains the minimum in Eqs. (4.45) and (4.46) for each i , then J^* is equal to the optimal average cost and μ^* is optimal.

We call Eqs. (4.45) and (4.46) the *modified optimality equations*. Note that if a pair (J^*, h) solves these equations, it does not follow that J^* is the optimal average cost (see the subsequent example). It is necessary also that there exists a stationary policy μ^* such that $\mu^*(i)$ simultaneously attains the minimum in Eqs. (4.45) and (4.46) for all i . Nonetheless, the modified optimality equations are often useful in various analyses. For example, Eq. (4.46) lies at the heart of the proof of optimality of a Blackwell optimal policy, given in Prop. 4.1.7. This equation will also be useful in the analysis of value iteration (see the subsequent Prop. 4.3.1).

The following example illustrates how the modified optimality equations can have a solution set that is very different from the one of the coupled pair of optimality equations (4.39), (4.40).

Example 4.1.1 (continued)

Consider again the deterministic 2-state, 2-control, average cost problem shown in Fig. 4.1.1. Recall that control 1 keeps the system at the state where it is, at a cost of 1 or 2, when at state 1 or 2, respectively. Control 2 is available only at state 1, and it moves the system to state 2 at a cost of -10.

Here, the modified optimality equations (4.45) and (4.46) take the form

$$J^*(1) = \min\{J^*(1), J^*(2)\}, \quad J^*(2) = J^*(2),$$

$$J^*(1) + h(1) = \min\{1 + h(1), -10 + h(2)\}, \quad J^*(2) + h(2) = 2 + h(2).$$

It is straightforward to verify that the solutions (J^*, h) are the ones that satisfy

$$J^*(1) = \min\{1, -10 + h(2) - h(1)\}, \quad J^*(2) = 2.$$

On the other hand, the coupled pair of Eqs. (4.39), (4.40) take the form

$$J^*(1) = \min\{J^*(1), J^*(2)\}, \quad J^*(2) = J^*(2),$$

$$J^*(1) + h^*(1) = 1 + h^*(1), \quad J^*(2) + h^*(2) = 2 + h^*(2).$$

Their solutions (J^*, h^*) are the ones where $J^*(1) = 1$ and $J^*(2) = 2$, and h^* is any vector. Thus the solutions of the two sets of optimality equations are quite different in this example, and in particular, the modified optimality equations have solutions (J^*, h) for which J^* is different from the optimal average cost, which is the vector $(1, 2)'$.

4.2 CONDITIONS FOR EQUAL AVERAGE COST FOR ALL INITIAL STATES

We will now specialize some of the preceding analysis to the case of equal optimal cost for each initial state, which is typical in practice for finite-state average cost problems, as discussed in Section 7.4 of Vol. I. As discussed in Section 4.1.3, the coupled pair of optimality equations reduces in this case to a single equation, Bellman's equation.

Proposition 4.2.1: If a scalar λ and a vector h satisfy

$$\lambda + h(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u)h(j) \right], \quad i = 1, \dots, n, \quad (4.47)$$

then λ is the optimal average cost $J^*(i)$ for all i ,

$$\lambda = \min_{\pi} J_{\pi}(i) = J^*(i), \quad i = 1, \dots, n.$$

Furthermore, if $\mu^*(i)$ attains the minimum in Eq. (4.47) for each i , the stationary policy μ^* is optimal, i.e., $J_{\mu^*}(i) = \lambda$ for all i .

Proof: This follows as a special case of Prop. 4.1.8. **Q.E.D.**

Specialized to a single stationary policy, Prop. 4.1.9 yields the following.

Proposition 4.2.2: Let μ be a stationary policy. If a scalar λ_{μ} and a vector h satisfy

$$\lambda_{\mu} + h(i) = g(i, \mu(i)) + \sum_{j=1}^n p_{ij}(\mu(i))h(j), \quad i = 1, \dots, n,$$

then $\lambda_{\mu} = J_{\mu}(i)$ for all i .

Weak Accessibility Condition — - -

We now provide a condition under which Bellman's equation has a solution, and by Prop. 4.2.1, the optimal cost is independent of the initial state. To understand this condition, consider two states i and j , and suppose

that under a policy π the system reaches j in finite expected number of transitions if started at i . Then the corresponding average costs satisfy $J^*(i) \leq J^*(j)$, since a possible option at initial state i , is to use π and then switch to an optimal policy upon reaching j , thereby achieving average cost $J^*(j)$ (starting from i), which must therefore be an upper bound to $J^*(i)$. This suggests that if every state is reachable from every other state using some policy, then the optimal average cost should be the same for all initial states. We will slightly generalize this condition as follows, by introducing the possibility of states that are transient under all policies.

Definition 4.2.1: We say that state i is accessible from state j if there exists a stationary policy μ and an integer k such that

$$P(x_k = j \mid x_0 = i, \mu) > 0.$$

Definition 4.2.2: We say that the *Weak Accessibility* (WA for short) condition holds if the set of states can be partitioned into two subsets S_t and S_c such that:

- (a) All states in S_t are transient under every stationary policy.
- (b) For every two states i and j in S_c , j is accessible from i .

We have the following proposition.

Proposition 4.2.3: Let the WA condition hold. Then the optimal average cost is the same for all initial states.

Proof: Let S_t and S_c be subsets of states satisfying the conditions of Definition 4.2.2, and consider an optimal stationary policy μ (at least one exists by Prop. 4.1.7). We first show that the average cost of μ is the same for all states in S_c . Assume the contrary, i.e., that both the set

$$M = \left\{ i \in S_c \mid J_\mu(i) = \max_{j=1,\dots,n} J_\mu(j) \right\}$$

and its complement in S_c , $\overline{M} = \{i \in S_c \mid i \notin M\}$, are nonempty. Take any states $i \in M$ and $j \in \overline{M}$, and a stationary policy μ' such that, for some k ,

$$P(x_k = j \mid x_0 = i, \mu') > 0,$$

and without loss of generality, let k be the minimal time index for which this inequality holds. Then there must exist states $m \in M$ and $\bar{m} \in \overline{M}$ such that

$$[P_{\mu'}]_{m\bar{m}} = P(x_{k+1} = \bar{m} \mid x_k = m, \mu') > 0.$$

It follows that the m th component of $P_{\mu'} J_{\mu}$ is strictly less than $\max_s J_{\mu}(s)$, which is equal to the m th component of J_{μ} . This contradicts the optimality equation (4.39), which implies that

$$J_{\mu} = P_{\mu} J_{\mu} \leq P_{\mu'} J_{\mu}.$$

Since the states in S_t are transient under μ , the system will move to a state in S_c within a finite expected number of transitions if started at any state in S_t . Thus the average cost of a state in S_t is equal to the common average cost of the states in S_c . Q.E.D.

Here is an example where the WA condition can be used.

Example 4.2.1 (Machine Replacement)

Consider a machine that can be in any one of n states, $1, \dots, n$. There is a cost $g(i)$ for operating for one time period the machine when it is in state i . The options at the start of each period are to (a) let the machine operate one more period in the state it currently is, or (b) repair the machine at a positive cost R and bring it to state 1 (corresponding to a machine in perfect condition). In the absence of repair, the transitions between states at the end of each time period are governed by given probabilities p_{ij} with $p_{ij} = 0$ for $j < i$. Once repaired, the machine is guaranteed to stay in state 1 for one period, and in subsequent periods, it may deteriorate to states $j \geq 1$ according to the transition probabilities p_{1j} . The problem is to find a policy that minimizes the average cost per stage. We have analyzed the discounted cost version of this problem in Example 1.2.1. It can be seen that the WA condition holds, so it follows that there exists a scalar λ and a vector h , such that for all i ,

$$\lambda + h(i) = \min \left[R + g(1) + h(1), g(i) + \sum_{j=1}^n p_{ij} h(j) \right],$$

while the policy that chooses the minimizing action above is average cost optimal.

Note that the WA condition depends solely on the transition probabilities of the problem, not on the transition costs. It is certainly possible that with a suitable choice of transition costs, the optimal average costs of all initial states are equal, even if the WA condition is violated. However, this will happen essentially by accident. To clarify this assertion, let us say that for a given set of transition probabilities, the optimal average costs of all initial states are *generically* equal if the set of cost vectors

$\{g(i, u) \mid i = 1, \dots, n, u \in U(i)\}$ that result in unequal optimal average costs has Lebesgue measure zero. Then it can be shown that the optimal average costs are generically equal if and only if the WA condition holds (see Tsitsiklis [Tsi06]).

Unichain Policies

We now consider a special type of stationary policy, called *unichain*, for which the corresponding Markov chain has a single recurrent class (and possibly some transient states). For a problem where there is only one control available at each state, the corresponding policy is unichain if and only if the WA condition holds. Hence by Prop. 4.2.3, for a unichain policy μ , the average cost is a common scalar λ_μ for all initial states,

$$J_\mu(i) = \lambda_\mu, \quad i = 1, \dots, n,$$

and Bellman's equation has the form

$$\lambda_\mu + h(i) = g(i, \mu(i)) + \sum_{j=1}^n p_{ij}(\mu(i))h(j), \quad i = 1, \dots, n.$$

This is a system of n linear equations with $n + 1$ unknowns, the scalars $\lambda_\mu, h(1), \dots, h(n)$, which has an infinite number of solutions, since by adding the same constant to all the components $h(1), \dots, h(n)$ of a solution, we obtain another solution. However, the system can be solved uniquely if we remove this degree of freedom and we fix a single component of h at some arbitrary value (0 for example). This is the subject of the following proposition.

Proposition 4.2.4: Let μ be a unichain policy, and let t be a fixed state. The system of the $n + 1$ linear equations

$$\lambda + h(i) = g(i, \mu(i)) + \sum_{j=1}^n p_{ij}(\mu(i))h(j), \quad i = 1, \dots, n, \quad (4.48)$$

$$h(t) = 0, \quad (4.49)$$

with the $n + 1$ unknowns $\lambda, h(1), \dots, h(n)$ has a unique solution.

Proof: This is Prop. 7.4.1(c) in Vol. I. For completeness, we (essentially) repeat the proof here. There is a quicker proof, which uses a standard fact from Markov chain theory: for a unichain μ , the set of vectors d satisfying $d = P_\mu d$ is the set of scalar multiples of the unit vector e . The result then follows from Prop. 4.1.9.

Assume first that t is a recurrent state of the Markov chain corresponding to μ . Then, in view of Eq. (4.49), we can write Eq. (4.48) as

$$h(i) = g(i, \mu(i)) - \lambda_\mu + \sum_{j=1, j \neq t}^n p_{ij}(\mu(i))h(j), \quad i = 1, \dots, n, i \neq t,$$

and is the same as Bellman's equation for a corresponding stochastic shortest path problem where t is the termination state, $g(i, \mu(i)) - \lambda_\mu$ is the expected stage cost at state i , and $h(i)$ is the average cost starting from i up to reaching t . By Prop. 2.1.2, this system has a unique solution, so $h(i)$ is uniquely defined by Eq. (4.48) for all $i \neq t$.

Assume next that t is a transient state. Then we choose another state \bar{t} that is recurrent and we introduce the transformation of variables $\bar{h}(i) = h(i) - h(\bar{t})$. The system of equations (4.48) and (4.49) can be written in terms of the variables λ and $\bar{h}(i)$ as

$$\bar{h}(i) = g(i, \mu(i)) - \lambda + \sum_{j=1, j \neq \bar{t}}^n p_{ij}(\mu(i))\bar{h}(j), \quad i = 1, \dots, n, i \neq \bar{t},$$

$$\bar{h}(\bar{t}) = 0,$$

so by the stochastic shortest path argument given earlier, it has a unique solution, implying that the solution of the system of equations (4.48) and (4.49) is also unique. **Q.E.D.**

Suppose now that all stationary policies are unichain. Then, any optimal stationary policy is unichain, and at least one is guaranteed to exist by Prop. 4.1.7. By the preceding proposition, this implies that the optimal average cost is independent of the initial state. An alternative way to see this is to combine Prop. 4.2.3 with the following proposition.

Proposition 4.2.5: If all stationary policies are unichain, the WA condition holds.

Proof: Assume the contrary, i.e., that there exist states i and j that are not transient under every stationary policy, and such that j is not accessible from i . Consider a stationary policy μ under which j is recurrent. Under μ and starting from i , the recurrent class of j is never reached, so some other recurrent class must be reached. Thus the Markov chain corresponding to μ has more than one recurrent classes, contradicting the unichain assumption. **Q.E.D.**

Note that in the machine replacement Example 4.2.1, for which the WA condition holds, not all policies are unichain. In particular, for the

stationary policy that replaces at every state except the worst state n (a poor but legitimate choice), the corresponding Markov chain has two recurrent classes, $\{1, 2, \dots, n-1\}$ and $\{n\}$ (assuming that $p_{1n} = 0$). Thus, the all-policies-unichain assumption is more restrictive than the WA condition.

Another interesting fact is that verifying the all-policies-unichain assumption for given transition probabilities is an NP-complete problem, as shown by Tsitsiklis [Tsi06]. By contrast one can verify the WA condition using simple polynomial-time graph algorithms.

Constructing Unichain Policies

While not all stationary policies are unichain under the WA condition, it is always possible to convert a stationary policy into one that is unichain, without affecting the average cost of any one chosen class of recurrent states. In particular, let μ be a stationary policy and let S be a set of states that forms a recurrent class under μ , with average cost λ_μ for each state in S .[†] We will redefine μ on the states outside S to construct a new unichain policy with recurrent class S , and average cost λ_μ for all states. The idea is to make the states not in S transient, so that they lead with probability 1 to the recurrent class S , thereby attaining the average cost λ_μ .

The construction starts with $S_0 = S$. At the k th step, given S_{k-1} , we stop if S_{k-1} is empty. Else, we define

$S_k = \{i \notin S_0 \cup \dots \cup S_{k-1} \mid p_{ij}(u) > 0 \text{ for some } u \in U(i) \text{ and } j \in S_{k-1}\}$, and for each $i \in S_k$, we redefine $\mu(i)$ to be equal to some $u \in U(i)$ with $p_{ij}(u) > 0$ for some $j \in S_{k-1}$. It can be seen that because of the WA condition, S_k will be nonempty unless $S_0 \cup \dots \cup S_{k-1}$ includes all states. Thus, this construction will redefine μ on all states not in S , and with the new definition of μ , these states will be transient, so the redefined policy is unichain, while the average cost of all the states will be λ_μ . Note that the construction requires a total of $O(n(nm))$ operations, where m is the maximum number of possible controls per state.

By using the above construction, we obtain the following proposition.

Proposition 4.2.6: If the WA condition holds, there exists an optimal stationary policy that is unichain.

Proof: Let μ^* be an optimal stationary policy, and let S be a set of states that forms a recurrent class under μ^* . Apply the construction just given. **Q.E.D.**

[†] One may identify all recurrent classes of the Markov chain corresponding to μ by using simple algorithms, for which we refer to the literature.

Note that the unichain optimal policy, guaranteed to exist by the preceding proposition, need not be Blackwell optimal. In fact, there may exist no unichain policy that is Blackwell optimal under the WA condition (see Exercise 4.4).

4.3 VALUE ITERATION

All the computational methods developed for discounted and stochastic shortest path problems (cf. Sections 1.3 and 2.2) have average cost per stage counterparts. However, the derivations of these methods are often intricate, and have no direct analogs in the discounted and stochastic shortest path context. In fact, the validity of these methods may depend on assumptions that relate to the structure of the underlying Markov chains, something that we have not encountered in earlier chapters.

Generally, the most important characteristic of an average cost problem is whether the Weak Accessibility (WA) condition holds, which is essentially equivalent to the optimal average cost $J^*(i)$ being generically independent of i (see the discussion in the preceding section). We will thus distinguish between two cases:

- (a) *The Single-Chain Case*, where the WA condition holds. Here $J^*(i)$ is independent of i , and there exists an optimal stationary policy that has a single recurrent class (plus possibly some transient states).
- (b) *The Multi-Chain Case*, where the WA condition does not hold. Here $J^*(i)$ is typically dependent on i , and optimal stationary policies typically have multiple recurrent classes.

We have already seen a major difference in the analysis of these two cases: in the single-chain case there is a single optimality equation, while in the multi-chain case there is a pair of coupled optimality equations (cf. Section 4.1.3). We will now discuss value iteration, placing primary emphasis to the more common single-chain case (under either the WA condition or some other assumption that implies the WA condition), but also addressing the multi-chain case. In subsequent sections, we will discuss policy iteration, and linear programming, for both the single-chain and the multi-chain cases.

The natural version of the value iteration method for the average cost problem is simply to generate successively the finite horizon optimal costs $T^k h, T^{2k} h, \dots$, starting with some initial vector h , where T is the DP mapping

$$Th = \min_{\mu} [\widehat{g_\mu} + P_\mu h].$$

It is then natural to speculate that the k -stage average cost “per stage” $(1/k)T^k h$ converges to the optimal average cost vector J^* as $k \rightarrow \infty$; this is in fact proved in Section 7.4 of Vol. I for the single-chain case.

We will now show that $(1/k)T^k h$ converges to J^* in the general case. The basic idea, captured in the following proposition, is to show that for all $k \geq 1$, we have

$$T^k \hat{h} = kJ^* + \hat{h}, \quad (4.50)$$

where J^* is the optimal average cost vector, and \hat{h} is a vector that satisfies the modified optimality equation $J^* + \hat{h} = T\hat{h}$, as per Prop. 4.1.5. We then argue that $T^k h - T^k \hat{h}$ is just the difference of the optimal values of k -stage cost functions that differ only in their terminal costs (h versus \hat{h}), so we have for all states i

$$\min_j [h(j) - \hat{h}(j)] \leq (T^k h)(i) - (T^k \hat{h})(i) \leq \max_j [h(j) - \hat{h}(j)],$$

and Eq. (4.50) shows that $(1/k)T^k h$ yields in the limit the optimal cost vector J^* .

Proposition 4.3.1: Let J^* be the optimal average cost vector, and let \hat{h} be a vector that satisfies the modified optimality equation $J^* + \hat{h} = T\hat{h}$, as per Prop. 4.1.5. Let also h be any vector in \mathbb{R}^n .

(a) For all k , we have

$$\begin{aligned} \min_{i=1,\dots,n} [h(i) - \hat{h}(i)] &\leq (T^k h)(i) - kJ^*(i) - \hat{h}(i) \\ &\leq \max_{i=1,\dots,n} [h(i) - \hat{h}(i)]. \end{aligned} \quad (4.51)$$

(b) For all k , we have

$$T^k \hat{h} = kJ^* + \hat{h}. \quad (4.52)$$

(c) The value iteration method yields J^* via

$$J^* = \lim_{k \rightarrow \infty} \frac{1}{k} T^k h. \quad (4.53)$$

Proof: (a) For any μ_0, \dots, μ_{k-1} , we have

$$(T_{\mu_0} \cdots T_{\mu_{k-1}})(h) = (T_{\mu_0} \cdots T_{\mu_{k-1}})(\hat{h}) + P_{\mu_0} \cdots P_{\mu_{k-1}}(h - \hat{h}). \quad (4.54)$$

Also, from the equation $J^* + \hat{h} = T\hat{h}$, we have

$$T_{\mu_{k-1}} \hat{h} \geq J^* + \hat{h},$$

and applying $T_{\mu_{k-2}}$ to both sides,

$$\begin{aligned} T_{\mu_{k-2}} T_{\mu_{k-1}} \hat{h} &\geq T_{\mu_{k-2}} (J^* + \hat{h}) \\ &= g_{\mu_{k-2}} + P_{\mu_{k-2}} J^* + P_{\mu_{k-2}} \hat{h} \\ &\geq J^* + T_{\mu_{k-2}} \hat{h} \\ &\geq 2J^* + \hat{h}, \end{aligned}$$

where the second inequality follows from the fact $P_\mu J^* \geq J^*$ for all μ (cf. Prop. 4.1.4), and the third inequality follows from the fact $T\hat{h} = J^* + \hat{h}$. Continuing similarly, we have

$$T_{\mu_0} \cdots T_{\mu_{k-1}} \hat{h} \geq kJ^* + \hat{h}. \quad (4.55)$$

Let μ^* be a Blackwell optimal policy. Then by Props. 4.1.4 and 4.1.5, we have $P_{\mu^*} J^* = J^*$ and $J^* + \hat{h} = T_{\mu^*} \hat{h}$, so if μ_0, \dots, μ_{k-1} are all equal to μ^* , equality holds in the preceding calculation, i.e.,

$$T_{\mu^*}^k \hat{h} = kJ^* + \hat{h}. \quad (4.56)$$

Applying Eq. (4.54) with $\mu_0 = \dots = \mu_{k-1} = \mu^*$, we obtain

$$T^k h \leq T_{\mu^*}^k h = T_{\mu^*}^k \hat{h} + P_{\mu^*}^k (h - \hat{h}) \leq T_{\mu^*}^k \hat{h} + \max_{i=1,\dots,n} [h(i) - \hat{h}(i)] e,$$

and by using Eq. (4.56), the right-hand side of Eq. (4.51) follows.

Also, for $m = 0, \dots, k-1$, let μ_m be such that $T_{\mu_m} T^m h = T^{m+1} h$. Then, combining Eqs. (4.54) and (4.55), we obtain

$$\begin{aligned} T^k h &= T_{\mu_0} \cdots T_{\mu_{k-1}} h \\ &= T_{\mu_0} \cdots T_{\mu_{k-1}} \hat{h} + P_{\mu_0} \cdots P_{\mu_{k-1}} (h - \hat{h}) \\ &\geq kJ^* + \hat{h} + P_{\mu_0} \cdots P_{\mu_{k-1}} (h - \hat{h}), \\ &\geq kJ^* + \hat{h} + \min_{i=1,\dots,n} [h(i) - \hat{h}(i)] e, \end{aligned}$$

which is the left-hand side of Eq. (4.51).

(b) Follows from Eq. (4.51) by setting $h = \hat{h}$.

(c) Divide both sides of Eq. (4.51) by k and take the limit as $k \rightarrow \infty$. Q.E.D.

While one may obtain J^* as the limit of $(1/k)T^k h$, this has two drawbacks. First, some of the components of $T^k h$ typically diverge to ∞ or $-\infty$, so direct calculation of $\lim_{k \rightarrow \infty} (1/k)T^k h$ is numerically impractical. Second, a corresponding differential cost vector will not be obtained. To address these issues, we will now distinguish between the single-chain and multi-chain cases.

4.3.1 Single-Chain Value Iteration

We may attempt to bypass the two difficulties with value iteration just mentioned by subtracting the same scalar δ^k from all values $(T^k h)(i)$, $i = 1, \dots, n$, so that the differences $(T^k h)(i) - \delta^k$ remain bounded. For this to work, we must assume that the optimal average cost is the same for all i ; otherwise the value iterates $(T^k h)(i)$ will grow at different rates for different i [cf. Prop. 4.3.1(b)].

We thus consider methods of the form

$$h^k = T^k h - \delta^k e,$$

where h is an arbitrary vector, and δ^k is some scalar satisfying

$$\min_{i=1, \dots, n} (T^k h)(i) \leq \delta^k \leq \max_{i=1, \dots, n} (T^k h)(i),$$

such as for example

$$\delta^k = (T^k h)(t),$$

where t is some fixed state. Then the differences

$$\max_i (T^k h)(i) - \min_i (T^k h)(i)$$

remain bounded as $k \rightarrow \infty$ by Prop. 4.3.1, so the vectors h^k also remain bounded, and we will see that with a proper choice of the scalar δ^k , $\{h^k\}$ converges to a differential cost vector.

Let us now restate the algorithm $h^k = T^k h - \delta^k e$ in a form that is suitable for iterative calculation. We have

$$h^{k+1} = T^{k+1} h - \delta^{k+1} e,$$

and since

$$T^{k+1} h = T(T^k h) = T(h^k + \delta^k e) = Th^k + \delta^k e,$$

we obtain

$$h^{k+1} = Th^k + (\delta^k - \delta^{k+1})e. \quad (4.57)$$

In the case where $\delta^k = (T^k h)(t)$ for some fixed state t , we have

$$\delta^{k+1} = (T^{k+1} h)(t) = (T(h^k + \delta^k e))(t) = (Th^k)(t) + \delta^k,$$

and the iteration (4.57) becomes

$$h^{k+1} = Th^k - (Th^k)(t)e. \quad (4.58)$$

We will henceforth restrict attention to the case where $\delta^k = (T^k h)(t)$, and call the corresponding algorithm (4.58) *relative value iteration*, since

the iterate h^k is equal to $T^k h - (T^k h)(t)e$ and may be viewed as a k -stage optimal cost vector *relative to state t*. The following results also apply to other versions of the algorithm (see Exercises 4.7 and 4.8). Note that relative value iteration, which generates h^k , is not really different than ordinary value iteration, which generates $T^k h$. The vectors generated by the two methods merely differ by a multiple of the unit vector, and the minimization problems involved in the corresponding iterations of the two methods are mathematically equivalent.

It can be seen that if the relative value iteration (4.58) converges to some vector h^* , then

$$(Th^*)(t)e + h^* = Th^*,$$

which by Prop. 4.2.1, implies that $(Th^*)(t)$ is the optimal average cost of all initial states, and h^* is an associated differential cost vector. Thus convergence can only be expected when the optimal average cost is independent of the initial state. However, it turns out that a stronger hypothesis is needed for convergence. The following example illustrates the reason.

Example 4.3.1

Consider value iteration for the case of a fixed stationary policy with transition matrix denoted by P and cost vector equal to 0. Then the method, starting with a vector h , generates

$$T^k h = P^k h.$$

While $\lim_{k \rightarrow \infty} (1/k)T^k h$ correctly yields the average cost vector, which is 0, the relative value iteration sequence

$$T^k h - (T^k h)(t)e$$

may not be convergent if P^k is not convergent, which will happen if P is periodic.

Indeed the relative value iteration

$$h^{k+1} = Ph^k - (Ph^k)(t)e$$

can be written as

$$h^{k+1} = Ph^k - ee'_t Ph^k = \hat{P}h^k.$$

where

$$\hat{P} = (I - ee'_t)P, \quad (4.59)$$

and e'_t is the row vector having all coordinates equal to 0 except for coordinate t which is equal to 1. The iteration will converge for all initial vectors h^0 if and only if all the eigenvalues of \hat{P} lie strictly within the unit circle. We have for any eigenvalue γ of P with corresponding eigenvector v ,

$$\hat{P}v = (I - ee'_t)Pv = \gamma(v - ee'_t v).$$

Therefore,

$$\hat{P}(v - ee_t'v) = \gamma(v - ee_t'v),$$

and it follows that each eigenvalue γ of P with corresponding eigenvector v , is also an eigenvalue of \hat{P} with corresponding eigenvector $(v - ee_t'v)$, as long as $v - ee_t'v \neq 0$. The eigenvalue $\gamma = 1$, with eigenvector $v = e$, fails the test $v - ee_t'v \neq 0$. However, if P has an eigenvalue $\gamma \neq 1$ that is *on* the unit circle, \hat{P} will have the same eigenvalue, and the iteration is not convergent. This occurs when P is periodic and has some nonunity eigenvalue on the unit circle. For example, suppose that

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

which has eigenvalues 1 and -1 . Then taking $t = 1$, the matrix \hat{P} of Eq. (4.59) is given by

$$\hat{P} = \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} [1 \ 0] \right) \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix},$$

and has eigenvalues 0 and -1 . As a result, because of the periodicity of P , the relative value iteration is not convergent, even though we are dealing with a single stationary policy that is unichain.

The following proposition shows convergence of the relative value iteration (4.58) under a technical condition that excludes situations such as the one of the preceding example. When there is only one control available per state, i.e., there is only one stationary policy μ , the condition of the following proposition requires that for some positive integer m , the matrix P_μ^m has at least one column all the components of which are positive. This can be shown to be equivalent to μ being unichain and the corresponding Markov chain being aperiodic (see e.g., Bertsekas and Tsitsiklis [BeT02]). However, we will later provide a variant of the relative value iteration (4.58), which converges under the condition that all stationary policies are unichain, regardless of whether the corresponding Markov chains are periodic or not (see Prop. 4.3.4).

Proposition 4.3.2: Assume that there exists a positive integer m such that for every admissible policy $\pi = \{\mu_0, \mu_1, \dots\}$, there exists an $\epsilon > 0$ and a state s such that

$$[P_{\mu_m} P_{\mu_{m-1}} \dots P_{\mu_1}]_{is} \geq \epsilon, \quad i = 1, \dots, n, \quad (4.60)$$

$$[P_{\mu_{m-1}} P_{\mu_{m-2}} \dots P_{\mu_0}]_{is} \geq \epsilon, \quad i = 1, \dots, n, \quad (4.61)$$

where $[\cdot]_{is}$ denotes the element of the i th row and s th column of the corresponding matrix. Fix a state t and consider the relative value iteration algorithm

$$h^{k+1}(i) = (Th^k)(i) - (Th^k)(t), \quad i = 1, \dots, n, \quad (4.62)$$

where h^0 is an arbitrary vector. Then the sequence $\{h^k\}$ converges to a vector h^* satisfying $(Th^*)(t)e + h^* = Th^*$, so that by Prop. 4.2.1, $(Th^*)(t)$ is equal to the optimal average cost of all initial states and h^* is an associated differential cost vector.

Proof: Denote

$$q^k(i) = h^{k+1}(i) - h^k(i), \quad i = 1, 2, \dots, n.$$

We will show that for all i and $k \geq m$ we have

$$\max_i q^k(i) - \min_i q^k(i) \leq (1 - \epsilon) \left(\max_i q^{k-m}(i) - \min_i q^{k-m}(i) \right), \quad (4.63)$$

where m and ϵ are as stated in the hypothesis. From this relation we then obtain, for some $B > 0$ and all k ,

$$\max_i q^k(i) - \min_i q^k(i) \leq B(1 - \epsilon)^{k/m}.$$

Since $q^k(t) = 0$, it follows that, for all i ,

$$|h^{k+1}(i) - h^k(i)| = |q^k(i)| \leq \max_j q^k(j) - \min_j q^k(j) \leq B(1 - \epsilon)^{k/m}.$$

Therefore, for every $r > 1$ and i we have

$$\begin{aligned} |h^{k+r}(i) - h^k(i)| &\leq \sum_{l=0}^{r-1} |h^{k+l+1}(i) - h^{k+l}(i)| \\ &\leq B(1 - \epsilon)^{k/m} \sum_{l=0}^{r-1} (1 - \epsilon)^{l/m} \\ &= \frac{B(1 - \epsilon)^{k/m} (1 - (1 - \epsilon)^{r/m})}{1 - (1 - \epsilon)^{1/m}}, \end{aligned} \quad (4.64)$$

so that $\{h^k(i)\}$ is a Cauchy sequence and converges to a limit $h^*(i)$. From Eq. (4.62) we see then that the equation $(Th^*)(t) + h^*(i) = (Th^*)(i)$ holds for all i . It will thus be sufficient to prove Eq. (4.63).

To prove Eq. (4.63), we denote by $\mu_k(i)$ the control that attains the minimum in the relation

$$(Th^k)(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u) h^k(j) \right], \quad (4.65)$$

for every k and i . Denote

$$\lambda_k = (Th^k)(t).$$

Then we have

$$h^{k+1} = g_{\mu_k} + P_{\mu_k} h^k - \lambda_k e \leq g_{\mu_{k-1}} + P_{\mu_{k-1}} h^k - \lambda_k e,$$

$$h^k = g_{\mu_{k-1}} + P_{\mu_{k-1}} h^{k-1} - \lambda_{k-1} e \leq g_{\mu_k} + P_{\mu_k} h^{k-1} - \lambda_{k-1} e,$$

where $e = (1, \dots, 1)'$ is the unit vector. From these relations, using the definition $q^k = h^{k+1} - h^k$, we obtain

$$P_{\mu_k} q^{k-1} + (\lambda_{k-1} - \lambda_k) e \leq q^k \leq P_{\mu_{k-1}} q^{k-1} + (\lambda_{k-1} - \lambda_k) e.$$

Since this relation holds for every $k \geq 1$, by iterating we obtain

$$\begin{aligned} P_{\mu_k} \dots P_{\mu_{k-m+1}} q^{k-m} + (\lambda_{k-m} - \lambda_k) e &\leq q^k \\ &\leq P_{\mu_{k-1}} \dots P_{\mu_{k-m}} q^{k-m} + (\lambda_{k-m} - \lambda_k) e. \end{aligned} \quad (4.66)$$

First, let us assume that the special state s corresponding to μ_{k-m}, \dots, μ_k as in Eqs. (4.60) and (4.61) is the fixed state t used in iteration (4.62); i.e.,

$$\begin{aligned} [P_{\mu_k} \dots P_{\mu_{k-m+1}}]_{it} &\geq \epsilon, \quad i = 1, \dots, n, \\ [P_{\mu_{k-1}} \dots P_{\mu_{k-m}}]_{it} &\geq \epsilon, \quad i = 1, \dots, n. \end{aligned} \quad (4.67)$$

The right-hand side of Eq. (4.66) yields

$$q^k(i) \leq \sum_{j=1}^n [P_{\mu_{k-1}} \dots P_{\mu_{k-m}}]_{ij} q^{k-m}(j) + \lambda_{k-m} - \lambda_k,$$

so using Eq. (4.67) and the fact $q^{k-m}(t) = 0$, we obtain

$$q^k(i) \leq (1 - \epsilon) \max_j q^{k-m}(j) + \lambda_{k-m} - \lambda_k, \quad i = 1, \dots, n,$$

implying that

$$\max_j q^k(j) \leq (1 - \epsilon) \max_j q^{k-m}(j) + \lambda_{k-m} - \lambda_k.$$

Similarly, from the left-hand side of Eq. (4.66) we obtain

$$\min_j q^k(j) \geq (1 - \epsilon) \min_j q^{k-m}(j) + \lambda_{k-m} - \lambda_k,$$

and by subtracting the last two relations, we obtain the desired Eq. (4.63).

When the special state s corresponding to μ_{k-m}, \dots, μ_k as in Eqs. (4.60) and (4.61) is not equal to t , we define a related iterative process

$$\tilde{h}^{k+1}(i) = (T\tilde{h}^k)(i) - (T\tilde{h}^k)(s), \quad i = 1, \dots, n, \quad (4.68)$$

$$\tilde{h}^0(i) = h^0(i), \quad i = 1, \dots, n.$$

Then, as earlier, we have

$$\max_i \tilde{q}^k(i) - \min_i \tilde{q}^k(i) \leq (1 - \epsilon) \left(\max_i \tilde{q}^{k-m}(i) - \min_i \tilde{q}^{k-m}(i) \right), \quad (4.69)$$

where

$$\tilde{q}^k = \tilde{h}^{k+1} - \tilde{h}^k.$$

It is straightforward to verify, using Eqs. (4.62) and (4.68), that for all i and k we have

$$h^k(i) = \tilde{h}^k(i) + (T\tilde{h}^{k-1})(s) - (T\tilde{h}^{k-1})(t).$$

Therefore, the coordinates of both h^k and q^k differ from the coordinates of \tilde{h}^k and \tilde{q}^k , respectively, by a constant. It follows that

$$\max_i q^k(i) - \min_i q^k(i) = \max_i \tilde{q}^k(i) - \min_i \tilde{q}^k(i),$$

and from Eq. (4.69) we obtain the desired Eq. (4.63). **Q.E.D.**

As a by-product of the preceding proof, we obtain a rate of convergence estimate. By taking the limit in Eq. (4.64) as $r \rightarrow \infty$, we have

$$\max_i |h^k(i) - h^*(i)| \leq \frac{B(1 - \epsilon)^{k/m}}{1 - (1 - \epsilon)^{1/m}}, \quad k = 0, 1, \dots,$$

so the bound on the error is reduced by $(1 - \epsilon)^{1/m}$ at each iteration. A sharper rate of convergence result can be obtained if we assume that there exists a unique optimal stationary policy μ^* . Then, it is possible to show that the minimum in Eq. (4.65) is attained by $\mu^*(i)$ for all i and all k after a certain index, so for such k , the relative value iteration takes the form $h^{k+1} = T_{\mu^*} h^k - (T_{\mu^*} h^k)(t)e$, and is governed by the largest eigenvalue modulus of the matrix

$$\hat{P}_{\mu^*} = (I - ee_t') P_{\mu^*}.$$

Error Bounds

Similar to discounted problems, the relative value iteration method can be strengthened by the calculation of monotonic error bounds.

Proposition 4.3.3: Under the assumption of Prop. 4.3.2, the iterates h^k of the relative value iteration method (4.62) satisfy

$$\underline{c}_k \leq \underline{c}_{k+1} \leq \lambda \leq \bar{c}_{k+1} \leq \bar{c}_k, \quad (4.70)$$

where λ is the optimal average cost of all initial states, and

$$\underline{c}_k = \min_i [(Th^k)(i) - h^k(i)],$$

$$\bar{c}_k = \max_i [(Th^k)(i) - h^k(i)].$$

Proof: Let $\mu_k(i)$ attain the minimum in

$$(Th^k)(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u) h^k(j) \right]$$

for each k and i . We have, using Eq. (4.62),

$$\begin{aligned} (Th^k)(i) &= g(i, \mu_k(i)) + \sum_{j=1}^n p_{ij}(\mu_k(i)) h^k(j) \\ &= g(i, \mu_k(i)) + \sum_{j=1}^n p_{ij}(\mu_k(i)) (Th^{k-1})(j) - (Th^{k-1})(i), \end{aligned}$$

and

$$h^k(i) \leq g(i, \mu_k(i)) + \sum_{j=1}^n p_{ij}(\mu_k(i)) h^{k-1}(j) - (Th^{k-1})(i).$$

Subtracting the last two relations, we obtain

$$(Th^k)(i) - h^k(i) \geq \sum_{j=1}^n p_{ij}(\mu_k(i)) ((Th^{k-1})(j) - h^{k-1}(j)),$$

and it follows that

$$\min_i [(Th^k)(i) - h^k(i)] \geq \min_i [(Th^{k-1})(i) - h^{k-1}(i)],$$

or equivalently

$$\underline{c}_{k-1} \leq \underline{c}_k.$$

A similar argument shows that

$$\bar{c}_k \leq \bar{c}_{k-1}.$$

By Prop. 4.3.2 we have $h^k(i) \rightarrow h^*(i)$ and $(Th^*)(i) - h^*(i) = \lambda$ for all i , so that $\underline{c}_k \rightarrow \lambda$. Since $\{\underline{c}_k\}$ is also nondecreasing, we must have $\underline{c}_k \leq \lambda$ for all k . Similarly, $\bar{c}_k \geq \lambda$ for all k . Q.E.D.

We now demonstrate the relative value iteration method and the error bounds (4.70) by means of an example.

Example 4.3.2

Consider an undiscounted version of the computational example of Section 1.3 (Example 1.3.1). We have

$$S = \{1, 2\}, \quad C = \{u^1, u^2\},$$

$$P(u^1) = \begin{bmatrix} p_{11}(u^1) & p_{12}(u^1) \\ p_{21}(u^1) & p_{22}(u^1) \end{bmatrix} = \begin{bmatrix} 3/4 & 1/4 \\ 3/4 & 1/4 \end{bmatrix},$$

$$P(u^2) = \begin{bmatrix} p_{11}(u^2) & p_{12}(u^2) \\ p_{21}(u^2) & p_{22}(u^2) \end{bmatrix} = \begin{bmatrix} 1/4 & 3/4 \\ 1/4 & 3/4 \end{bmatrix},$$

and

$$g(1, u^1) = 2, \quad g(1, u^2) = 0.5, \quad g(2, u^1) = 1, \quad g(2, u^2) = 3.$$

The mapping T has the form

$$(Th)(i) = \min \left\{ g(i, u^1) + \sum_{j=1}^2 p_{ij}(u^1)h(j), g(i, u^2) + \sum_{j=1}^2 p_{ij}(u^2)h(j) \right\}.$$

Letting $t = 1$ be the reference state, the relative value iteration (4.62) takes the form

$$\begin{aligned} h^{k+1}(1) &= 0 \\ h^{k+1}(2) &= \overline{(Th^k)(2)} - \overline{(Th^k)(1)}. \end{aligned}$$

The results of the computation starting with $h^0(1) = h^0(2) = 0$ are shown in the table of Fig. 4.3.1.

k	$h^k(1)$	$h^k(2)$	\underline{c}_k	\bar{c}_k
0	0	0		
1	0	0.500	0.625	0.875
2	0	0.250	0.687	0.812
3	0	0.375	0.719	0.781
4	0	0.312	0.734	0.765
5	0	0.344	0.742	0.758
6	0	0.328	0.746	0.754
7	0	0.336	0.748	0.752
8	0	0.332	0.749	0.751
9	0	0.334	0.749	0.750
10	0	0.333	0.750	0.750

Figure 4.3.1 Iterates and error bounds generated by the relative value iteration method for the problem of Example 4.3.2.

Other Versions of the Relative Value Iteration Method

As mentioned earlier, the relative value iteration method given in Prop. 4.3.2 may not converge under the unichain assumption; a stronger condition is necessary. We will now show that we can bypass this difficulty by modifying the problem without affecting either the optimal cost or the optimal policies, and by applying the relative value iteration method to the modified problem.

Let τ be any scalar with

$$0 < \tau < 1,$$

and consider the problem that results when each transition matrix P_μ corresponding to a stationary policy μ is replaced by

$$\tilde{P}_\mu = \tau P_\mu + (1 - \tau)I, \quad (4.71)$$

where I is the identity matrix. Note that \tilde{P}_μ is a transition probability matrix with the property that, at every state, a self-transition occurs with probability at least $(1 - \tau)$. This destroys any periodic character that P_μ may have, and makes \tilde{P}_μ aperiodic. For another view of the same point, note that each eigenvalue of \tilde{P}_μ is of the form $\tau\gamma + (1 - \tau)$, where γ is an eigenvalue of P_μ . Therefore, all eigenvalues $\gamma \neq 1$ of P_μ that lie on the unit circle are mapped into eigenvalues of \tilde{P}_μ strictly inside the unit circle.

Bellman's equation for the modified problem is

$$\tilde{\lambda}_\mu e + \tilde{h} = g_\mu + \tilde{P}_\mu \tilde{h} = g_\mu + (\tau P_\mu + (1 - \tau)I)\tilde{h},$$

which can be written as

$$\tilde{\lambda}_\mu e + \tau \tilde{h} = g_\mu + P_\mu(\tau \tilde{h}).$$

We observe that this equation is the same as Bellman's equation for the original problem,

$$\lambda_\mu e + h = g_\mu + P_\mu h,$$

with the identification

$$h = \tau \tilde{h}.$$

It follows from Prop. 4.2.2 that if the average cost per stage for the original problem is independent of i for every μ , then the same is true for the modified problem. Furthermore, the costs of all stationary policies, as well as the optimal cost, are equal for both the original and the modified problem.

Consider now the relative value iteration method (4.62) for the modified problem. A straightforward calculation shows that it takes the form

$$h^{k+1}(i) = (1 - \tau)h^k(i) + \min_{u \in U(i)} \left[g(i, u) + \tau \sum_{j=1}^n p_{ij}(u)h^k(j) \right] - \min_{u \in U(t)} \left[g(t, u) + \tau \sum_{j=1}^n p_{tj}(u)h^k(j) \right], \quad (4.72)$$

where t is some fixed state with $h^0(t) = 0$. Note that this iteration is as easy to execute as the original version. It is convergent, however, under weaker conditions than those required in Prop. 4.3.2.

Proposition 4.3.4: Assume that each stationary policy is unichain. Then, for $0 < \tau < 1$, the sequences $\{h^k(i)\}$ generated by the modified relative value iteration (4.72) satisfy

$$\lim_{k \rightarrow \infty} h^k(i) = \frac{h^*(i)}{\tau},$$

$$\lim_{k \rightarrow \infty} \min_{u \in U(i)} \left[g(t, u) + \tau \sum_{j=1}^n p_{tj}(u)h^k(j) \right] = \lambda,$$

where λ is the optimal average cost and h^* is a differential cost vector.

Proof: The proof consists of showing that the conditions of Prop. 4.3.2 hold for the modified problem involving the transition probability matrices \tilde{P}_μ of Eq. (4.71). The key fact is that if a component of a matrix \tilde{P}_μ is positive, it continues to be positive when this matrix is multiplied by a matrix $\tilde{P}_{\mu'}$, i.e., $[\tilde{P}_{\mu'} \tilde{P}_\mu]_{ij} > 0$ if $[\tilde{P}_\mu]_{ij} > 0$ or if $[\tilde{P}_{\mu'}]_{ij} > 0$, because of the $(1 - \tau)$ -multiple of the identity in the definition (4.71). Thus the positive components of a product $\tilde{P}_{\mu_m} \cdots \tilde{P}_{\mu_0}$ include the components that are positive in any product of a subset of matrices from the set $\tilde{P}_{\mu_m}, \dots, \tilde{P}_{\mu_0}$.

Let $m > nn_M$, where n is the number of states and n_M is the number of distinct stationary policies. Consider a set of control functions $\mu_0, \mu_1, \dots, \mu_m$, so at least one μ is repeated n times within the subset μ_1, \dots, μ_{m-1} . Then the positive components of $\tilde{P}_{\mu_m} \cdots \tilde{P}_{\mu_1}$ and $\tilde{P}_{\mu_{m-1}} \cdots \tilde{P}_{\mu_0}$ include the components of the matrix P_μ^n that are positive. Let s be a recurrent state under \tilde{P}_μ . Then, we have $[P_\mu^n]_{is} > 0$ for all i , so it follows that for some $\epsilon > 0$, the conditions

$$[\tilde{P}_{\mu_m} \cdots \tilde{P}_{\mu_1}]_{is} \geq \epsilon, \quad i = 1, \dots, n,$$

$$[\tilde{P}_{\mu_{m-1}} \cdots \tilde{P}_{\mu_0}]_{is} \geq \epsilon, \quad i = 1, \dots, n;$$

are satisfied. Q.E.D.

Note that since the modified value iteration method is nothing but the ordinary method applied to a modified problem, the error bounds of Prop. 4.3.3 apply in appropriately modified form.

We finally mention that relative value iteration can also be shown to converge in the manner of Prop. 4.3.2 under an alternative condition, which requires that each optimal stationary policy has an aperiodic transition matrix. We will show this with a more sophisticated analysis in the next section, where we will discuss value iteration in the context of multi-chain problems.

Contracting Value Iteration and the λ -SSP

Contrary to the value iteration methods for discounted and stochastic shortest path problems, relative value iteration does not involve a weighted sup-norm contraction or even a contraction of any kind. It is possible to develop a value iteration method that does involve a weighted sup-norm contraction by exploiting the connection with the stochastic shortest path problem that we developed in Section 7.4 of Vol. I. The advantage of such a method is that it inherits the robustness of contraction iterations. For example periodic transition matrices are not an issue in a contraction-based method. Furthermore, such a method admits a Gauss-Seidel as well as other related asynchronous variants, while there are no such variants of relative value iteration.

For our development, we need a somewhat restrictive assumption, namely that *all stationary policies are unichain, and state n is recurrent in the Markov chain corresponding to each stationary policy*. As in Section 7.4 of Vol. I, we consider the stochastic shortest path problem obtained by leaving unchanged all transition probabilities $p_{ij}(u)$ for $j \neq n$, by setting all transition probabilities $p_{in}(u)$ to 0, and by introducing an artificial termination state t to which we move from each state i with probability $p_{in}(u)$ (see Fig. 4.3.2). The one-stage cost is equal to $g(i, u) - \lambda$, where λ is a scalar parameter. We refer to this stochastic shortest path problem as the λ -SSP.

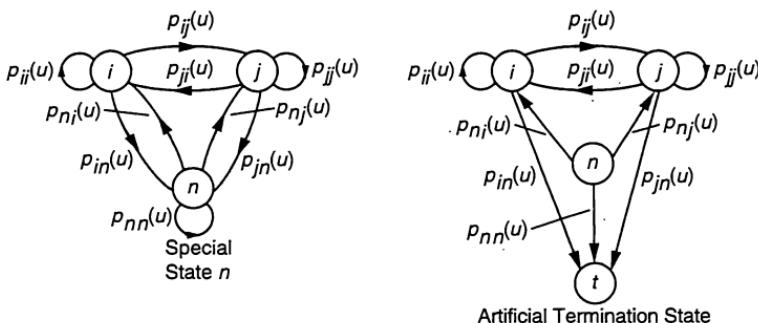


Figure 4.3.2 Transition probabilities for an average cost problem and its associated stochastic shortest path problem.

Let $h_{\mu, \lambda}(i)$ be the cost of a stationary policy μ for the λ -SSP, starting from state i ; i.e., $h_{\mu, \lambda}(i)$ is the total expected cost incurred starting from state i up to the first positive time that we reach the termination state n . Let $h_\lambda(i) = \min_\mu h_{\mu, \lambda}(i)$ be the corresponding optimal cost of the λ -SSP. Then the following can be shown (see Fig. 4.3.3):

- (a) For all μ , and all scalars λ and $\bar{\lambda}$, we have

$$h_{\mu, \lambda}(i) = h_{\mu, \bar{\lambda}}(i) + (\bar{\lambda} - \lambda)N_\mu(i), \quad i = 1, \dots, n,$$

where $N_\mu(i)$ is the expected value of the first positive time that s is reached under μ starting from state i . This is because the only difference between the λ -SSP and the $\bar{\lambda}$ -SSP is that the one-stage costs in the λ -SSP are offset from the one-stage costs in the $\bar{\lambda}$ -SSP by $\bar{\lambda} - \lambda$. Thus, in particular, for all scalars λ , we have

$$h_{\mu, \lambda}(i) = h_{\mu, \lambda_\mu}(i) + (\lambda_\mu - \lambda)N_\mu(i), \quad i = 1, \dots, n, \quad (4.73)$$

where λ_μ is the average cost per stage of μ . Furthermore, because $h_{\mu, \lambda_\mu}(n) = 0$ (compare with the analysis of Section 7.4 of Vol. I), we

have

$$h_{\mu, \lambda}(n) = (\lambda_\mu - \lambda)N_\mu(n). \quad (4.74)$$

(b) The functions

$$h_\lambda(i) = \min_{\mu} h_{\mu, \lambda}(i), \quad i = 1, \dots, n, \quad (4.75)$$

are concave, monotonically decreasing, and piecewise linear as functions of λ , and we have

$$h_\lambda(n) = 0 \quad \text{if and only if} \quad \lambda = \lambda^*. \quad (4.76)$$

Furthermore, the vector h_{λ^*} , together with λ^* , satisfies Bellman's equation $\lambda^*e + h_{\lambda^*} = Th_{\lambda^*}$.

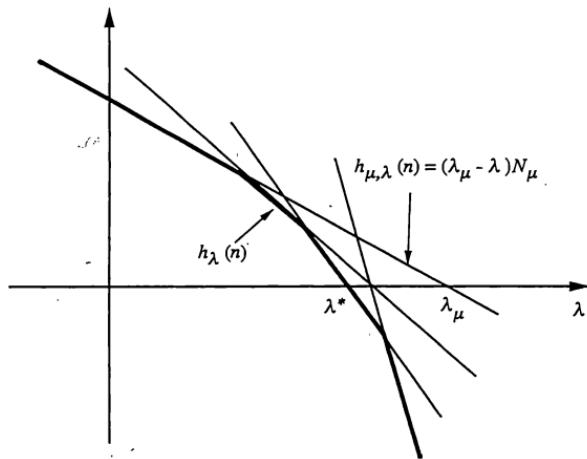


Figure 4.3.3: Relation of the costs of stationary policies in the average cost problem and the λ -SSP. Here, $h_{\mu, \lambda}$ is the cost-to-go vector of policy μ in the λ -SSP, while h_λ is the corresponding optimal cost-to-go vector; i.e., $h_\lambda(i) = \min_{\mu} h_{\mu, \lambda}(i)$ for all i . Note that $h_{\mu, \lambda}$ is linear in λ , cf. Eq. (4.73), and that h_λ is piecewise linear and concave as a function of λ . Furthermore, for the state n , we have

$$h_{\mu, \lambda}(n) = (\lambda_\mu - \lambda)N_\mu(n),$$

since $h_{\mu, \lambda_\mu}(n) = 0$ [cf. Eq. (4.74)].

From Fig. 4.3.3, it can be seen that λ^* can be obtained by a one-dimensional search procedure that brackets λ^* within a sequence of nested and diminishing intervals (see Exercise 4.11). This method is probably inefficient because it requires the (exact) solution of several λ -SSPs, corresponding to several different values of λ . An alternative method, which is

also inefficient because it requires the exact solution of several λ -SSPs, is to update λ by an iteration of the form

$$\lambda^{k+1} = \lambda^k + \gamma^k h_{\lambda^k}(n), \quad (4.77)$$

where γ^k is a positive stepsize parameter. This iteration is motivated by Fig. 4.3.3 where it is seen that $\lambda < \lambda^*$ (or $\lambda > \lambda^*$) if and only if $h_\lambda(n) > 0$ [or $h_\lambda(n) < 0$, respectively]. Indeed, it can be seen from Fig. 4.3.3 that the sequence $\{\lambda^k\}$ generated by Eq. (4.77) converges to λ^* provided the stepsize γ^k is the same for all iterations and does not exceed the threshold value $1/\max_\mu N_\mu(n)$. Such a stepsize is sufficiently small to guarantee that the difference $\lambda - \lambda^*$ does not change sign during the algorithm (4.77). Note that each λ -SSP can be solved by value iteration, which has the form

$$h^{k+1}(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^{n-1} p_{ij}(u) h^k(j) \right] - \lambda, \quad i = 1, \dots, n, \quad (4.78)$$

with λ kept fixed throughout the value iteration method.

A more efficient possibility is to change λ during the value iteration process (4.78) by using an iteration of the form (4.77), but with $h_{\lambda^k}(n)$ replaced by an approximation, the current value iterate $h^{k+1}(n)$. Such an algorithm may be viewed as a *value iteration algorithm for a slowly varying stochastic shortest path problem*. It has the form

$$h^{k+1}(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^{n-1} p_{ij}(u) h^k(j) \right] - \lambda^k, \quad i = 1, \dots, n, \quad (4.79)$$

$$\lambda^{k+1} = \lambda^k + \gamma^k h^{k+1}(n), \quad (4.80)$$

where γ^k is a positive stepsize.

The motivation for the method is that, under our recurrence assumption on state n under all stationary policies, value iteration for stochastic shortest path problems involves a contraction. In particular, according to Prop. 2.2.3, the mapping $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ with components given by

$$F_i(h) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^{n-1} p_{ij}(u) h(j) \right], \quad i = 1, \dots, n, \quad (4.81)$$

is a contraction with respect to some weighted sup-norm. Note here that while there is coupling between the iteration of h as per Eq. (4.79) and the iteration for λ as per Eq. (4.80), the latter iteration can be made much slower than the former through the use of the stepsize γ , so that

the weighted sup-norm contraction character of the iteration (4.79) is preserved. By contrast, the relative value iteration method does not involve a weighted sup-norm contraction.

Convergence of the method (4.79), (4.80) can be shown for a variety of rules for choosing the stepsize γ^k (see Bertsekas [Ber98]). One possibility is to keep γ^k equal to a sufficiently small constant. The main tradeoff is that if γ^k is chosen constant but very small, or diminishing at the rate of say $1/k$ (as is common in many stochastic iterative algorithms), then λ changes slowly relative to h , and iteration (4.80) essentially becomes identical to iteration (4.77) but with a very small stepsize, which leads to slow convergence. On the other hand, if γ^k is too large, λ^k will oscillate and diverge. One may keep the stepsize γ^k constant at a value found by trial and error, but there are some better alternatives. One possibility that has worked well in the computational experiments of [Ber98] is to start with a fairly large γ^k (say around 1) and gradually diminish it if $h^k(n)$ changes sign frequently; for example, we may use $\gamma^k = m(\hat{k})\gamma$, where:

- (a) γ is the initial stepsize (a positive constant).
- (b) $m(\hat{k})$ is a decreasing function of \hat{k} , which is defined as the number of indexes $t \leq k$ such that $h^{t-1}(n)h^t(n) < 0$ and $|h^t(n)|$ is greater than some fixed threshold θ .

Some possibilities are

$$m(\hat{k}) = \frac{1}{\hat{k} + 1},$$

and

$$m(\hat{k}) = \beta^{\hat{k}},$$

where β is a fixed scalar from the range $(0, 1)$, so that γ^k is decreased by a factor β each time \hat{k} is incremented. Typically, in such a scheme the stepsize is reduced quickly to an appropriate level (which depends on the problem) and then stays constant for the remaining iterations.

Let us also note an improvement of the method, which guarantees that bounded iterates will be generated for any choice of stepsize. We may calculate upper and lower bounds on λ^* from iteration (4.79), and then modify iteration (4.80) to project the iterate $\lambda^k + \gamma^k h^k(n)$ on the interval of the bounds. In particular, based on the error bounds of Prop. 4.3.3 for the relative value iteration method, it can be seen that

$$\underline{\zeta}^k \leq \lambda^* \leq \bar{\zeta}^k,$$

where

$$\underline{\zeta}^k = \lambda^k + \min \left[\min_{i \neq n} [h^{k+1}(i) - h^k(i)], h^{k+1}(n) \right],$$

$$\bar{\zeta}^k = \lambda^k + \max \left[\max_{i \neq n} [h^{k+1}(i) - h^k(i)], h^{k+1}(n) \right].$$

Thus in place of the iteration $\lambda^{k+1} = \lambda^k + \gamma^k h^{k+1}(n)$ [cf. Eq. (4.80)], we may set λ^{k+1} to be the projection of $\lambda^k + \gamma^k h^{k+1}(n)$ on the interval

$$\left[\max_{m=0,\dots,k} \underline{\zeta}^m, \min_{m=0,\dots,k} \bar{\zeta}^m \right].$$

For further discussion of the implementation of the algorithm, we refer to the original source [Ber98].

The sup-norm contraction property of the mapping F can also be exploited to construct valid Gauss-Seidel and even distributed asynchronous variants of the contracting value iteration method (4.79)-(4.80), where the order of iteration of the components of h varies randomly. We refer to Bertsekas [Ber82a], [Ber83], [Ber98], and Bertsekas and Tsitsiklis [BeT89] for analysis of DP-related distributed asynchronous algorithms. On the other hand, no convergent Gauss-Seidel version of the relative value iteration method is known.

Single-Chain Value Iteration: A Summary

Several value iteration methods and variations thereof were given in this section for single-chain problems, so it is worth to summarize their properties here:

- (a) The optimal average cost vector J^* can be obtained as the limit of $(1/k)T^k h$ [cf. Eq. (4.53)]; this is true for multi-chain problems as well.
- (b) If the aperiodicity-type condition of Prop. 4.3.2 holds, then both the optimal average cost λ^* and a differential cost vector h^* satisfying $\lambda^* e + h^* = Th^*$ can be obtained in the limit via the relative value iteration (4.62).
- (c) Even when the condition in (b) fails, it can be made to hold for an equivalent problem, obtained from the original with a simple transformation that makes the associated transition probability matrices aperiodic: replacing each P_μ by $(1-\tau)I + \tau P_\mu$, where τ is some scalar from the range $(0, 1)$.
- (d) If all stationary policies are unichain, and some special state is recurrent under each stationary policy, an alternative value iteration method can be used. This method involves a weighted sup-norm contraction mapping, derived from an associated stochastic shortest path problem. As a result, it admits a Gauss-Seidel/asynchronous variant, and is not subject to the difficulties with periodic transition matrices that are inherent in relative value iteration.

4.3.2 Multi-Chain Value Iteration

We now consider the multi-chain case and the value iteration method

$$h^{k+1} = Th^k,$$

starting from an arbitrary vector h^0 . We will write more compactly T as

$$Th = \min_{\mu \in M} [g_\mu + P_\mu h],$$

with M denoting the set of all admissible stationary policies. We showed earlier in Prop. 4.3.1 that the optimal average cost vector J^* is obtained as

$$J^* = \lim_{k \rightarrow \infty} \frac{1}{k} T^k h^0.$$

However, we have yet to resolve the issue of obtaining a differential cost vector that satisfies together with J^* the coupled pair of optimality equations (4.39) and (4.40).

Unfortunately, the idea of constructing value iterates relative to some fixed state, which underlies the relative value iteration method of Section 4.3.1, does not work in the multi-chain case, because there are multiple recurrent classes, and the value iterates $h^k(i)$ will generally change with k at a rate that depends on i . We thus use a different and more sophisticated approach, which is based on the convergence of the *residual sequence*

$$r^k = h^k - k J^* = T^k h^0 - k J^*. \quad (4.82)$$

We have shown in Prop. 4.3.1(a) that $\{r^k\}$ is bounded, and we will show shortly that if $\{r^k\}$ converges, then a differential cost vector can be constructed. We will also show that under a certain aperiodicity condition, $\{r^k\}$ converges.[†]

Consider a vector \hat{h} that satisfies the modified optimality equation $J^* + \hat{h} = T\hat{h}$ (cf. Prop. 4.1.5). Then, we have [cf. Eq. (4.52)]

$$\hat{h} = T^k \hat{h} - k J^*,$$

and by combining this equation with Eq. (4.82), we obtain

$$r^k = \hat{h} + (T^k h^0 - T^k \hat{h}). \quad (4.83)$$

Thus r^k converges if and only if the difference of the optimal k -stage costs, with terminal cost vectors h^0 and \hat{h} , converges.

It turns out that the convergence of $T^k h^0 - T^k \hat{h}$ hinges on the absence of periodicities in the system, and this absence is critically important for

[†] To get a sense of the nature of r^k , consider the special case where the assumption of Section 7.4 of Vol. I holds, so the average cost problem can be converted to the associated SSP problem. Then r^k is simply the k th iterate of value iteration for the SSP problem. Thus, r^k converges to the SSP optimal cost vector, which can be viewed as a differential cost vector for the average cost problem, as discussed in Section 7.4 of Vol. I.

obtaining good estimates of a differential cost vector through value iteration. Indeed, consider the special case where there is only one stationary policy with transition probability matrix P . Then, r^k is given by

$$r^k = \hat{h} + (T^k h^0 - T^k \hat{h}) = \hat{h} + P^k(h^0 - \hat{h}),$$

and is guaranteed to converge for all initial h^0 if and only if P is aperiodic. This is reflected in the assumption of part (b) of the following proposition, which is the core of the analysis of multi-chain value iteration.

To streamline notation, we write the coupled pair of optimality equations (4.39) and (4.40) more compactly as

$$J^* = \min_{\mu \in M} P_\mu J^*, \quad J^* + h = \bar{T}h, \quad (4.84)$$

where the mapping $\bar{T} : \mathbb{R}^n \mapsto \mathbb{R}^n$ is defined by

$$\bar{T}h = \min_{\mu \in \bar{M}} [g_\mu + P_\mu h], \quad (4.85)$$

with \bar{M} being the subset of stationary policies that attain the minimum in the first optimality equation,

$$\bar{M} = \{\bar{\mu} \in M \mid J^* = P_{\bar{\mu}} J^*\}. \quad (4.86)$$

Proposition 4.3.5: Let J^* be the optimal average cost vector, and let $\{h^k\}$ be generated by the value iteration method $h^{k+1} = T^k h^k$.

(a) For all k sufficiently large, we have

$$Th^k = \bar{T}h^k,$$

and if μ is such that $T_\mu h^k = Th^k$, then $\mu \in \bar{M}$, where \bar{T} is the mapping (4.85) and \bar{M} is the set (4.86).

- (b) If μ is an optimal stationary policy that has an aperiodic transition probability matrix, then for all states i that are recurrent under μ , the i th component of the residual sequence $\{r^k\}$ converges.
- (c) If every optimal stationary policy has an aperiodic transition probability matrix, then the residual sequence $\{r^k\}$ converges. Furthermore, if μ_k is such that $T_{\mu_k} h^k = Th^k$, there exists an index \bar{k} such that μ_k is optimal for all $k \geq \bar{k}$.

Proof: (a) Let M' be the set of stationary policies μ such that $T_\mu h^k = Th^k$ for infinitely many k , and let K be such that for all $k \geq K$, we have

$T_\mu h^k = Th^k$ if and only if $\mu \in M'$. Fix $\mu \in M'$ and let \mathcal{K} be the infinite set of integers $k \geq K$ such that $Th^k = T_\mu h^k$. We have for all $k \in \mathcal{K}$,

$$h^{k+1} = g_\mu + P_\mu h^k,$$

from which

$$\frac{1}{k+1}h^{k+1} = \frac{1}{k+1}g_\mu + \frac{k}{k+1}P_\mu\left(\frac{1}{k}h^k\right).$$

Taking the limit as $k \rightarrow \infty$, $k \in \mathcal{K}$, we obtain [using the fact $(1/k)h^k \rightarrow J^*$, cf. Prop. 4.3.1(c)]

$$J^* = P_\mu J^*,$$

which implies that $\mu \in \overline{M}$. From this it also follows that $Th^k = T_\mu h^k = \overline{Th}^k$.

(b) Let (J^*, h) denote the gain-bias pair of μ , and let g and P denote the cost vector and transition probability matrix of μ . By Prop. 4.1.2,

$$J^* = PJ^*, \quad J^* + h = g + Ph. \quad (4.87)$$

Let us denote for all k

$$v^k = h^k - kJ^* - h.$$

Then, using Eq. (4.87), and the fact $h^{k+1} = Th^k \leq g + Ph^k$, we obtain

$$\begin{aligned} v^{k+1} &= h^{k+1} - (k+1)J^* - h \\ &\leq g + Ph^k - (k+1)J^* - h \\ &= g + P(h^k - kJ^* - h) + Ph - J^* - h \\ &= Pv^k. \end{aligned} \quad (4.88)$$

By Prop. 4.3.1(a), $\{v^k\}$ is bounded, so it has at least one limit point.

In the remainder of the proof, we generically denote by $z(i)$ the i th component of a vector z . Let \hat{v} be any limit point of $\{v^k\}$, let

$$\bar{v}(i) = \limsup_{k \rightarrow \infty} v^k(i), \quad i = 1, \dots, n,$$

and note that

$$\hat{v}(i) \leq \bar{v}(i), \quad i = 1, \dots, n.$$

Since P has nonnegative components, by iterating the relation $v^{k+1} \leq Pv^k$ [cf. Eq. (4.88)], we have

$$v^{k+m} \leq P^m v^k, \quad k, m = 0, 1, \dots.$$

For each i , by taking the limit as $m \rightarrow \infty$ along the subsequence that converges to $\bar{v}(i)$, we obtain

$$\bar{v}(i) \leq \sum_{j=1}^n P_{ij}^* v^k(j), \quad i = 1, \dots, n, \quad k = 0, 1, \dots,$$

where P^* is the limit of P^m , which exists since P is aperiodic by assumption. Taking also the limit as $k \rightarrow \infty$ along the subsequence that converges to \hat{v} , we have

$$\bar{v}(i) \leq \sum_{j=1}^n P_{ij}^* \hat{v}(j), \quad i = 1, \dots, n. \quad (4.89)$$

Let I be a subset of states that forms a recurrent class under μ , and let $\bar{i} \in I$ be such that $\bar{v}(\bar{i}) = \max_{i \in I} \bar{v}(i)$. From Eq. (4.89), we have

$$\bar{v}(\bar{i}) \leq \sum_{j \in I} P_{ij}^* \hat{v}(j) \leq \bar{v}(\bar{i}) \sum_{j \in I} P_{\bar{i}j}^* \leq \bar{v}(\bar{i}), \quad (4.90)$$

where the second inequality uses the fact $\hat{v}(j) \leq \bar{v}(\bar{i})$ for all $j \in I$. It follows that equality holds throughout Eq. (4.90), implying that $\hat{v}(j) = \bar{v}(\bar{i})$ for all $j \in I$, since $P_{ij}^* > 0$ for all $j \in I$ because I forms a recurrent class. Thus, $\{v^k(i)\}$ converges to $\bar{v}(\bar{i})$ for all $i \in I$, i.e., the sequence of i th components of v^k converges for all recurrent states i . Since $r^k = v^k + h$, the same is true for the sequence of i th components of the residuals r^k .

(c) Let \bar{k} be such that $Th^k = \bar{T}h^k$ for all $k \geq \bar{k}$ as in part (a). We have

$$r^{k+1} = Th^k - (k+1)J^* = \bar{T}h^k - (k+1)J^* = \min_{\mu \in \bar{M}} [g_\mu + P_\mu h^k] - (k+1)J^*$$

and finally, since $r^k = h^k - kJ^*$, and $J^* = P_\mu J^*$ for all $\mu \in \bar{M}$,

$$r^{k+1} = \min_{\mu \in \bar{M}} [g_\mu + P_\mu r^k], \quad k \geq \bar{k}, \quad (4.91)$$

where $g_\mu = g_\mu - J^*$.

Let x and y be the vectors whose i th components are

$$x(i) = \liminf_{k \rightarrow \infty} r^k(i), \quad y(i) = \limsup_{k \rightarrow \infty} r^k(i), \quad i = 1, \dots, n.$$

Note that x and y are well-defined as vectors in \mathbb{R}^n , since by Prop. 4.3.1(a), $\{r^k\}$ is bounded. Fix a state i , let $\{r^{k_m}(i)\}$ be a subsequence that converges to $y(i)$ and is such that $\{r^{k_m-1}\}$ converges to a vector $w \in \mathbb{R}^n$. For any $\epsilon > 0$, there exists $\hat{k} \geq \bar{k}$ such that

$$r^{k_m}(i) \geq y(i) - \epsilon, \quad k_m \geq \hat{k},$$

$$(P_\mu r^{k_m-1})(i) \leq (P_\mu w)(i) + \epsilon, \quad \mu \in \bar{M}, \quad k_m \geq \hat{k}.$$

Then, using also Eq. (4.91), we have

$$\begin{aligned} y(i) - \epsilon &\leq r^{k_m}(i) \\ &= \min_{\mu \in \bar{M}} [g_\mu(i) + (P_\mu r^{k_m-1})(i)] \\ &\leq \min_{\mu \in \bar{M}} [g_\mu(i) + (P_\mu w)(i)] + \epsilon \\ &\leq \min_{\mu \in \bar{M}} [g_\mu(i) + (P_\mu y)(i)] + \epsilon, \end{aligned}$$

where the last inequality holds since P_μ is nonnegative and $w \leq y$. Since this relation holds for all $i, \mu \in \overline{M}$, and $\epsilon > 0$, we obtain

$$y \leq \min_{\mu \in \overline{M}} [q_\mu + P_\mu y].$$

A similar argument can be used to show that

$$x \geq \min_{\mu \in \overline{M}} [q_\mu + P_\mu x].$$

Let $\mu \in \overline{M}$ attain the minimum in the above relation, so that we have

$$q_\mu + P_\mu x \leq x \leq y \leq q_\mu + P_\mu y. \quad (4.92)$$

From this inequality, we obtain

$$0 \leq y - x \leq P_\mu(y - x).$$

Letting $P_\mu^* = \lim_{N \rightarrow \infty} (1/N) \sum_{k=0}^{N-1} P_\mu^k$ (cf. Prop. 4.1.1), and iterating the above relation, we see that

$$0 \leq y - x \leq P_\mu^*(y - x). \quad (4.93)$$

Multiplying the left-hand side of Eq. (4.92) by P_μ^* and using the definition $q_\mu = g_\mu - J^*$, we have

$$P_\mu^*(g_\mu - J^*) + P_\mu^* P_\mu x \leq P_\mu^* x,$$

which by using the equations $J^* = P_\mu J^*$ and $P_\mu^* P_\mu = P_\mu^*$, yields

$$P_\mu^* g_\mu \leq J^*.$$

Since $P_\mu^* g_\mu = J_\mu$, we see that μ is an optimal stationary policy.

From part (b), we have that for all states i that are recurrent under μ , the limit of $r^k(i)$ exists so that $y(i) - x(i) = 0$. Using this fact, it follows from Eq. (4.93), that for all states j that are transient under μ and hence the j th column of P_μ^* is 0, we have $y(j) - x(j) = 0$. Thus, the limit of $r^k(i)$ exists for all i .

Finally, let $\tilde{k} \geq \bar{k}$ be such that for all $k \geq \tilde{k}$, the policy μ_k such that $T_{\mu_k} h^k = Th^k$ is repeated infinitely often, i.e., satisfies $T_{\mu_k} h^{k'} = Th^{k'}$ for infinitely many k' . We have for $k \geq \tilde{k}$, $\mu_k \in \overline{M}$ so that $J^* = P_{\mu_k} J^*$, and by Eq. (4.91),

$$J^* + r^{k+1} = g_{\mu_k} + P_{\mu_k} r^k.$$

Fixing μ_k , and taking the limit along the subsequence of indices k' such that $\mu_{k'} = \mu_k$, we obtain

$$J^* + r^* = g_{\mu_k} + P_{\mu_k} r^*.$$

By Prop. 4.1.9, it follows that $J_{\mu_k} = J^*$, i.e., that μ_k is optimal. **Q.E.D.**

We now explain the algorithmic significance of Prop. 4.3.5, under the aperiodicity assumption of part (c). First, it is possible to obtain the optimal average cost vector from

$$h^{k+1} - h^k \rightarrow J^*. \quad (4.94)$$

To see this, note that from the definition (4.82) of the residual sequence,

$$h^{k+1} = (k+1)J^* + r^{k+1}, \quad h^k = kJ^* + r^k,$$

so by subtraction, we obtain

$$h^{k+1} - h^k = J^* + r^{k+1} - r^k.$$

Since $\{r^k\}$ converges, we have $r^{k+1} - r^k \rightarrow 0$, so it follows that $h^{k+1} - h^k \rightarrow J^*$.

Second, it turns out that (J^*, r^*) are a solution of the coupled pair of optimality equations (4.84), where r^* is the limit of the residual sequence $\{r^k\}$. To see this, note that from Prop. 4.3.5(a), for all k sufficiently large, we have

$$h^{k+1} = Th^k = \bar{T}h^k,$$

which by using the equation $h^k = kJ^* + r^k$, can be written as

$$(k+1)J^* + r^{k+1} = \min_{\mu \in \bar{M}} [g_\mu + P_\mu(kJ^* + r^k)].$$

Since by the definition of \bar{M} , we have $J^* = P_\mu J^*$ for all $\mu \in \bar{M}$, it follows that

$$J^* + r^{k+1} = \min_{\mu \in \bar{M}} [g_\mu + P_\mu r^*],$$

from which by taking the limit as $k \rightarrow \infty$, we obtain

$$J^* + r^* = \min_{\mu \in \bar{M}} [g_\mu + P_\mu r^*].$$

We state these conclusions as a proposition.

Proposition 4.3.6: Let J^* be the optimal average cost vector, let $\{h^k\}$ be generated by the value iteration method $h^{k+1} = T^k h^0$, and assume that every optimal stationary policy has an aperiodic transition probability matrix.

(a) We have

$$J^* = \lim_{k \rightarrow \infty} (h^{k+1} - h^k).$$

(b) The residual sequence $\{h^k - kJ^*\}$ converges to some vector r^* , which together with J^* satisfies the coupled pair of optimality equations (4.84).

In the case where there is a stationary policy with a periodic transition matrix, by replacing each P_μ by $(1 - \tau)I + \tau P_\mu$ for some $\tau \in (0, 1)$, we can transform the problem to an equivalent problem where all stationary policies have aperiodic transition matrices, as discussed in connection with relative value iteration. Thus, while periodic transition matrices are an analytical concern for value iteration, from the point of view of computation they are not a significant issue.

We finally note that the error bounds of Prop. 4.3.3 can be extended to the multi-chain case, and in fact for any stationary policy μ_k such that $T_{\mu_k} h^k = Th^k$, we have

$$\underline{c}_k \leq J^*(i) \leq J_{\mu_k}(i) \leq \bar{c}_k,$$

where

$$\underline{c}_k = \min_i [h^{k+1}(i) - h^k(i)], \quad \bar{c}_k = \max_i [h^{k+1}(i) - h^k(i)]$$

(see Exercise 4.9). This is a bound on the degree of suboptimality of μ_k . Furthermore, assuming the conditions of Prop. 4.3.6, \underline{c}_k and \bar{c}_k converge to $\min_i J^*(i)$ and $\max_i J^*(i)$, respectively. Thus \underline{c}_k and \bar{c}_k converge to each other when $J^*(i)$ is the same for all i , consistently with Prop. 4.3.3. On the other hand, when $J^*(i)$ is not the same for all i , the error bound difference $\bar{c}_k - \underline{c}_k$ will not converge to 0. As a result, the bounds may be quite loose, and they may not be used as a termination criterion for value iteration.

4.4 POLICY ITERATION

The policy iteration algorithm for the average cost problem is similar to those described in Sections 1.3 and 2.4. However, the analysis is more complicated because issues relating to the structure of the associated Markov chains play an important role. The form of the algorithm is similar. Given a stationary policy, we evaluate it by solving a linear system of equations. We then obtain an improved policy by means of a minimization process until no further improvement is possible. We will first discuss the single-chain case, and then the multi-chain case.

4.4.1 Single-Chain Policy Iteration

For the single-chain case, we will assume throughout that every stationary policy encountered in the course of the algorithm is unichain. If this is not so, the multi-chain version applies.

At the k th step of the policy iteration algorithm, we have a unichain stationary policy μ^k . We then perform a *policy evaluation* step; i.e., we

obtain corresponding average and differential costs λ^k and $h^k(i)$ satisfying

$$\lambda^k + h^k(i) = g(i, \mu^k(i)) + \sum_{j=1}^n p_{ij}(\mu^k(i))h^k(j), \quad i = 1, \dots, n, \quad (4.95)$$

or equivalently

$$\lambda^k e + h^k = T_{\mu^k} h^k = g_{\mu^k} + P_{\mu^k} h^k.$$

This can be done, for example, by appending to the system (4.95) the equation

$$h^k(t) = 0,$$

where t is any state, so the solution is unique (cf. Prop. 4.2.4). Note that this solution can be obtained directly but also iteratively, using the relative value iteration method.

We subsequently perform a *policy improvement* step; i.e., we find a stationary policy μ^{k+1} , where for all i , $\mu^{k+1}(i)$ is such that

$$g(i, \mu^{k+1}(i)) + \sum_{j=1}^n p_{ij}(\mu^{k+1}(i))h^k(j) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^n p_{ij}(u)h^k(j) \right], \quad (4.96)$$

or equivalently

$$T_{\mu^{k+1}} h^k = Th^k.$$

If $\mu^{k+1} = \mu^k$, the algorithm terminates; otherwise, the process is repeated with μ^{k+1} replacing μ^k . Note that the policy μ^{k+1} does not depend on which solution h^k of the evaluation equation (4.95) we use in the policy improvement Eq. (4.96) - they are all the same up to a common scalar shift for all states (cf. Props. 4.1.9 and 4.2.4). In particular, we may use the bias of μ^k in place of h^k in Eq. (4.96).

There is an easy proof, given in Exercise 4.10, that the policy iteration algorithm terminates finitely if we assume that the Markov chain corresponding to each μ^k is irreducible (is unichain and has no transient states). To prove the result without this assumption, we impose the following restriction in the way the algorithm is operated.

Normalization Rule: If $\mu^k(i)$ attains the minimum in the policy improvement Eq. (4.96), we choose $\mu^{k+1}(i) = \mu^k(i)$, even if there are other controls attaining the minimum in addition to $\mu^k(i)$.

The following proposition establishes the validity of the policy iteration algorithm in the single-chain case.

Proposition 4.4.1: If all the generated policies are unichain, the policy iteration algorithm with the normalization rule terminates finitely with an optimal stationary policy.

It is convenient to state the main argument needed for the proof of Prop. 4.4.1 as a separate proposition:

Proposition 4.4.2: Let μ be a unichain policy with gain-bias pair (λ_μ, h_μ) , let $\bar{\mu}$ be a unichain policy obtained from μ via a policy improvement step with the normalization rule, and let $(\lambda_{\bar{\mu}}, h_{\bar{\mu}})$ be the corresponding gain-bias pair. Then if $\bar{\mu} \neq \mu$, one of the following holds:

- (1) $\lambda_{\bar{\mu}} < \lambda_\mu$.
- (2) $\lambda_{\bar{\mu}} = \lambda_\mu$ and $h_{\bar{\mu}}(i) \leq h_\mu(i)$ for all $i = 1, \dots, n$, with equality for all states i that are recurrent under $\bar{\mu}$, and strict inequality for at least one state i that is transient under $\bar{\mu}$.

Proof: To simplify notation, we denote

$$P = P_\mu, \quad \bar{P} = P_{\bar{\mu}}, \quad P^* = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k, \quad \bar{P}^* = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} \bar{P}^k,$$

$$g = g_\mu, \quad \bar{g} = g_{\bar{\mu}}, \quad \lambda = \lambda_\mu, \quad \bar{\lambda} = \lambda_{\bar{\mu}}, \quad h = h_\mu, \quad \bar{h} = h_{\bar{\mu}}.$$

Define the vector δ by

$$\delta = T_\mu h - T_{\bar{\mu}} h,$$

and note that the policy improvement step, $T_{\bar{\mu}} h = Th$, implies that

$$0 \leq \delta(i), \quad i = 1, \dots, n, \tag{4.97}$$

with strict inequality for at least one i if $\bar{\mu} \neq \mu$ (in view of the normalization rule). We have

$$T_\mu h = \lambda e + h,$$

so subtracting the equation

$$T_{\bar{\mu}} h = T_{\bar{\mu}} \bar{h} + (T_{\bar{\mu}} h - T_{\bar{\mu}} \bar{h}) = \bar{\lambda} e + \bar{h} + \bar{P}(h - \bar{h}),$$

we obtain

$$\delta = (\lambda - \bar{\lambda})e + (I - \bar{P})\Delta, \tag{4.98}$$

where

$$\Delta = h - \bar{h}.$$

By multiplying Eq. (4.98) with \bar{P}^k and by adding for $k = 0, 1, \dots, N-1$, we have

$$\begin{aligned} \sum_{k=0}^{N-1} \bar{P}^k \delta &= N(\lambda - \bar{\lambda})e + (I - \bar{P})\Delta + \dots + (\bar{P}^{N-1} - \bar{P}^N)\Delta \\ &= N(\lambda - \bar{\lambda})e + (I - \bar{P}^N)\Delta. \end{aligned} \quad (4.99)$$

Dividing by N and taking the limit as $N \rightarrow \infty$, we obtain

$$\bar{P}^* \delta = \left(\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} \bar{P}^k \right) \delta = (\lambda - \bar{\lambda})e. \quad (4.100)$$

In view of $0 \leq \delta$ [cf. Eq. (4.97)], we see that

$$\bar{\lambda} \leq \lambda.$$

If $\bar{\lambda} < \lambda$ we are done, so assume that $\lambda = \bar{\lambda}$. A state i is called \bar{P} -recurrent (\bar{P} -transient) if i belongs (does not belong, respectively) to the recurrent class of the Markov chain corresponding to \bar{P} . From Eq. (4.100), we have $\bar{P}^* \delta = 0$, and since $\delta \geq 0$ and the components of \bar{P}^* that are positive are the columns corresponding to \bar{P} -recurrent states, we obtain

$$\delta(i) = 0, \quad \text{for all } i \text{ that are } \bar{P}\text{-recurrent.} \quad (4.101)$$

In view of the normalization rule, it follows that if i is \bar{P} -recurrent, then $\bar{\mu}(i) = \mu(i)$ and the i th rows of P and \bar{P} are identical. Hence P and \bar{P} have the same recurrent states. The components of the vectors h and \bar{h} corresponding to these recurrent states are the bias vectors of the (identical) Markov chains with stage costs corresponding to μ and $\bar{\mu}$, restricted to the recurrent states. It follows that $h(i) = \bar{h}(i)$ for all i that are \bar{P} -recurrent, i.e.,

$$\Delta(i) = 0, \quad \text{for all } i \text{ that are } \bar{P}\text{-recurrent.} \quad (4.102)$$

From Eq. (4.99), since $\lambda = \bar{\lambda}$ and $\delta \geq 0$, we obtain

$$\lim_{N \rightarrow \infty} \bar{P}^N \Delta = \Delta - \lim_{N \rightarrow \infty} \sum_{k=0}^{N-1} \bar{P}^k \delta \leq \Delta - \delta.$$

The components of $\bar{P}^N \Delta$ corresponding to \bar{P} -transient states tend to zero, and in view of Eq. (4.102), the same is true for the components of $\bar{P}^N \Delta$.

corresponding to \overline{P} -recurrent states. Therefore, we have $\lim_{N \rightarrow \infty} \overline{P}^N \Delta = 0$ and it follows that

$$\delta(i) \leq \Delta(i), \quad \text{for all } i. \quad (4.103)$$

From Eqs. (4.97) and (4.101)-(4.103), we see that either $\delta = 0$, in which case $\mu = \overline{\mu}$, or else $0 \leq \Delta(i)$ for all i , with strict inequality for at least one \overline{P} -transient state i . **Q.E.D.**

From Prop. 4.4.2 it follows that no policy will be encountered more than once prior to termination. Since the number of stationary policies is finite, it follows that the policy iteration algorithm must terminate finitely. If the algorithm stops at the k th step with $\mu^{k+1} = \mu^k$, we see from Eqs. (4.95) and (4.96) that λ^k and h^k must satisfy Bellman's equation,

$$\lambda^k e + h^k = Th^k,$$

which by Prop. 4.2.1 implies that μ^k is an optimal stationary policy. This argument proves Prop. 4.4.1.

We now illustrate policy iteration with an example that we also used in the case of relative value iteration.

Example 4.4.1

Consider the problem of Example 4.3.2. Let

$$\mu^0(1) = u^1, \quad \mu^0(2) = u^2.$$

We take $t = 1$ as a reference state and obtain λ_{μ^0} , $h_{\mu^0}(1)$, and $h_{\mu^0}(2)$ from the system of equations

$$\begin{aligned} \lambda_{\mu^0} + h_{\mu^0}(1) &= g(1, u^1) + p_{11}(u^1)h_{\mu^0}(1) + p_{12}(u^1)h_{\mu^0}(2), \\ \lambda_{\mu^0} + h_{\mu^0}(2) &= g(2, u^2) + p_{21}(u^2)h_{\mu^0}(1) + p_{22}(u^2)h_{\mu^0}(2), \\ h_{\mu^0}(1) &= 0. \end{aligned}$$

Substituting the data of the problem,

$$\lambda_{\mu^0} = 2 + \frac{1}{4}h_{\mu^0}(2), \quad \lambda_{\mu^0} + h_{\mu^0}(2) = 3 + \frac{3}{4}h_{\mu^0}(2),$$

from which

$$\lambda_{\mu^0} = \frac{5}{2}, \quad h_{\mu^0}(1) = 0, \quad h_{\mu^0}(2) = 2.$$

We now find $\mu^1(1)$ and $\mu^1(2)$ by the minimization indicated in Eq. (4.96). We determine

$$\begin{aligned} \min & [g(1, u^1) + p_{11}(u^1)h_{\mu^0}(1) + p_{12}(u^1)h_{\mu^0}(2), \\ & \quad g(1, u^2) + p_{11}(u^2)h_{\mu^0}(1) + p_{12}(u^2)h_{\mu^0}(2)] \\ &= \min \left[2 + \frac{1}{4} \cdot 2, 0.5 + \frac{3}{4} \cdot 2 \right] \\ &= \min[2.5, 2] \end{aligned}$$

and

$$\begin{aligned} & \min [g(2, u^1) + p_{21}(u^1)h_{\mu^0}(1) + p_{22}(u^1)h_{\mu^0}(2), \\ & \quad g(2, u^2) + p_{21}(u^1)h_{\mu^0}(1) + p_{22}(u^2)h_{\mu^0}(2)] \\ & = \min \left[1 + \frac{1}{4} \cdot 2, 3 + \frac{3}{4} \cdot 2 \right] \\ & = \min[1.5, 4.5]. \end{aligned}$$

The minimization yields

$$\mu^1(1) = u^2, \quad \mu^1(2) = u^1.$$

We obtain λ_{μ^1} , $h_{\mu^1}(1)$, and $h_{\mu^1}(2)$ from the system of equations

$$\begin{aligned} \lambda_{\mu^1} + h_{\mu^1}(1) &= g(1, u^2) + p_{11}(u^2)h_{\mu^1}(1) + p_{12}(u^2)h_{\mu^1}(2), \\ \lambda_{\mu^1} + h_{\mu^1}(2) &= g(2, u^1) + p_{21}(u^1)h_{\mu^1}(1) + p_{22}(u^1)h_{\mu^1}(2), \\ h_{\mu^1}(1) &= 0. \end{aligned}$$

By substituting the data of the problem, we obtain

$$\lambda_{\mu^1} = \frac{3}{4}, \quad h_{\mu^1}(1) = 0, \quad h_{\mu^1}(2) = \frac{1}{3}.$$

We find $\mu^2(1)$ and $\mu^2(2)$ by determining the minimum in

$$\begin{aligned} & \min [g(1, u^1) + p_{11}(u^1)h_{\mu^1}(1) + p_{12}(u^1)h_{\mu^1}(2), \\ & \quad g(1, u^2) + p_{11}(u^2)h_{\mu^1}(1) + p_{12}(u^2)h_{\mu^1}(2)] \\ & = \min \left[2 + \frac{1}{4} \cdot \frac{1}{3}, 0.5 + \frac{3}{4} \cdot \frac{1}{3} \right] \\ & = \min[2.08, 0.75], \end{aligned}$$

and

$$\begin{aligned} & \min [g(2, u^1) + p_{21}(u^1)h_{\mu^1}(1) + p_{22}(u^1)h_{\mu^1}(2), \\ & \quad g(2, u^2) + p_{21}(u^2)h_{\mu^1}(1) + p_{22}(u^2)h_{\mu^1}(2)] \\ & = \min \left[1 + \frac{1}{4} \cdot \frac{1}{3}, 3 + \frac{3}{4} \cdot \frac{1}{3} \right] \\ & = \min[1.08, 3.25]. \end{aligned}$$

The minimization yields

$$\mu^2(1) = \mu^1(1) = u^2, \quad \mu^2(2) = \mu^1(2) = u^1,$$

and hence the preceding policy is optimal and the optimal average cost is $\lambda_{\mu^1} = 3/4$.

4.4.2 Multi-Chain Policy Iteration

The policy iteration algorithm for multi-chain problems generates a sequence of two n -dimensional vectors, gain and bias. Similar to the single-chain case, it alternates between policy evaluation and policy improvement steps. Evaluation of policy μ consists of finding a solution to the pair of equations

$$J = P_\mu J, \quad J + h = g_\mu + P_\mu h, \quad (4.104)$$

(cf. Prop. 4.1.9). Policy improvement is obtained either in the form of reduction in some component of the gain vector J (with no increase in any other component), or if this is not possible, by reduction in some component of the bias vector h (again with no increase in any other component). If no improvement is possible, the algorithm terminates with a policy that will be shown to be optimal. The improvement process is thus similar to the single-chain case (cf. Prop. 4.4.2), but it uses both optimality equations.

Let us consider first the policy evaluation step. We recall that by Prop. 4.1.9, the set of solutions (J, h) of the system (4.104) is the set of pairs of the form (J_μ, h) , where $h = h_\mu + d$ with $d = P_\mu d$. Here J_μ and h_μ are the gain and bias vectors of μ . Whereas in the single-chain case, the choice of solution did not matter, in the multi-chain case some care must be taken because the choice affects the policy obtained through the subsequent policy improvement step. In our development we will use the specific solution (J_μ, h_μ) , i.e., the algorithm computes explicitly the bias vector h_μ . Computing the bias vector using the equation $h_\mu = H_\mu g_\mu$ (cf. Prop. 4.1.2) is inefficient, because H_μ is not easily obtainable using the computationally expensive formulas of Prop. 4.1.1:

$$P_\mu^* = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} P_\mu^k, \quad H_\mu = (I - P_\mu + P_\mu^*)^{-1} - P_\mu^*.$$

Fortunately, however, one may compute h_μ by solving a linear system of $3n$ equations with $3n$ unknowns, which is given in the following proposition.

Proposition 4.4.3: (Policy Evaluation Equations) Let μ be a stationary policy with gain-bias pair (J_μ, h_μ) . The set of solutions (J, h, v) of the system of equations

$$J = P_\mu J, \quad (4.105)$$

$$J + h = g_\mu + P_\mu h, \quad (4.106)$$

$$h + v = P_\mu v, \quad (4.107)$$

is the set of triplets of the form $(J_\mu, h_\mu, -H_\mu^2 g_\mu + d)$, where d is such that $d = P_\mu d$.

Proof: To simplify notation, we drop the subscript μ , and we denote the gain and bias of μ by \bar{J} and \bar{h} , respectively, to distinguish them from the generic vectors J and h . Let (J, h, v) be a solution of the system (4.105)-(4.107). From Prop. 4.1.9, we have $J = \bar{J}$, and $h = \bar{h} + d$ for some d with $d = Pd$. We will show that $d = 0$. Indeed, the equation $d = Pd$ implies that

$$d = \frac{1}{N} \sum_{k=0}^{N-1} P^k d, \quad N = 1, 2, \dots,$$

from which by taking the limit as $N \rightarrow \infty$, we obtain

$$d = P^* d. \quad (4.108)$$

On the other hand, using also the fact $\bar{h} = Hg$, Eq. (4.107) is written as

$$Hg + d + v = Pv.$$

By multiplying with P^* and by using the facts $P^*H = 0$ and $P^*P = P^*$ [cf. Eqs. (4.7) and (4.8)], we obtain $P^*d = 0$, which together with Eq. (4.108) implies that $d = 0$ and $h = \bar{h}$.

From Eq. (4.107), we have $\bar{h} + v = Pv$ or

$$(I - P)v = -\bar{h}. \quad (4.109)$$

On the other hand, from Eq. (4.9), we have $P^* - I = (P - I)H$, from which by multiplying with \bar{h} and using also the fact $\bar{h} = Hg$, we obtain

$$P^*Hg - \bar{h} = (P - I)H^2g.$$

Since $P^*H = 0$ [cf. Eq. (4.8)], it follows that

$$-\bar{h} = -(I - P)H^2g. \quad (4.110)$$

Combining Eqs. (4.109) and (4.110), we obtain

$$(I - P)(v + H^2g) = 0,$$

i.e., the vector $v + H^2g$ is in the nullspace of $I - P$, implying that $v = -H^2g + d$ for some d satisfying $d = Pd$. Hence every solution of the system (4.105)-(4.107) has the desired form.

Conversely, a triplet $(\bar{J}, \bar{h}, -H^2g + d)$, where $d = Pd$, satisfies Eqs. (4.105) and (4.106) by Prop. 4.1.9. For such a triplet, Eq. (4.107) takes the form

$$\bar{h} - H^2g + d = -PH^2g + Pd$$

or equivalently, since $\bar{h} = Hg$ and $d = Pd$,

$$(I - H + PH)Hg = 0.$$

This equation holds because by Eqs. (4.9) and (4.8), we have $I - H + PH = P^*$ and $P^*H = 0$. **Q.E.D.**

Thus in the policy evaluation step of the policy iteration algorithm, the gain-bias pair (J_{μ^k}, h_{μ^k}) of a stationary policy μ^k can be evaluated by solving the corresponding system of Prop. 4.4.3.[†] Given (J_{μ^k}, h_{μ^k}) , one can obtain a new policy μ^{k+1} using the following policy improvement step:

Policy Improvement Step: If $\min_{\mu} P_{\mu} J_{\mu^k} \neq J_{\mu^k}$, let μ^{k+1} be such that

$$P_{\mu^{k+1}} J_{\mu^k} = \min_{\mu} P_{\mu} J_{\mu^k}.$$

Otherwise, let μ^{k+1} be such that

$$P_{\mu^{k+1}} J_{\mu^k} = \min_{\mu} P_{\mu} J_{\mu^k}, \quad T_{\mu^{k+1}} h_{\mu^k} = \min_{\mu \in \overline{M}} T_{\mu} h_{\mu^k},$$

where \overline{M} is the set of policies attaining the minimum in the equation $\min_{\mu} P_{\mu} J_{\mu^k} = J_{\mu^k}$. In both of the above minimizations, the normalization rule is observed.

Note that $\min_{\mu} P_{\mu} J_{\mu^k} \neq J_{\mu^k}$ is equivalent to having strict inequality for at least one component in the inequality

$$\min_{\mu} P_{\mu} J_{\mu^k} < J_{\mu^k},$$

which holds since μ^k satisfies $P_{\mu^k} J_{\mu^k} = J_{\mu^k}$. Note also that if the algorithm terminates with $\mu^{k+1} = \mu^k$, then μ^k satisfies the coupled pair of optimality equations and is therefore optimal by Prop. 4.1.8. Thus to prove that the policy iteration algorithm terminates finitely, it is sufficient to demonstrate that a policy cannot be repeated without reaching optimality. This is shown using the following proposition, which parallels Prop. 4.4.2, its single-chain counterpart.

Proposition 4.4.4: Let μ^k be a stationary policy with gain-bias pair (J_{μ^k}, h_{μ^k}) , let μ^{k+1} be obtained via a policy improvement step, and

[†] It is interesting to note that the term $-H_{\mu}^2 g_{\mu}$ appearing in the solution of the system (4.105)-(4.107) is the third term in the Laurent series expansion (see the footnote after Prop. 4.1.2). Indeed, the method of proof of Prop. 4.4.3 can be used to verify that all the terms in the Laurent series expansion can be progressively computed by solving linear systems of equations such as the ones of Props. 4.1.9 and 4.4.3.

let $(J_{\mu^{k+1}}, h_{\mu^{k+1}})$ be the corresponding gain-bias pair. Then if $\mu^{k+1} \neq \mu^k$, one of the following holds:

- (1) $J_{\mu^{k+1}}(i) \leq J_{\mu^k}(i)$ for all $i = 1, \dots, n$, with strict inequality for at least one state i .
- (2) $J_{\mu^{k+1}} = J_{\mu^k}$ and $h_{\mu^{k+1}}(i) \leq h_{\mu^k}(i)$ for all $i = 1, \dots, n$, with strict inequality for at least one state i that is transient under μ^{k+1} .

Proof: To simplify notation, we denote

$$P = P_{\mu^k}, \quad \bar{P} = P_{\mu^{k+1}}, \quad P^* = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k, \quad \bar{P}^* = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} \bar{P}^k,$$

$$\mu = \mu^k, \quad \bar{\mu} = \mu^{k+1}, \quad g = g_{\mu^k}, \quad \bar{g} = g_{\mu^{k+1}},$$

$$J = J_{\mu^k}, \quad \bar{J} = J_{\mu^{k+1}}, \quad h = h_{\mu^k}, \quad \bar{h} = h_{\mu^{k+1}}.$$

Assume first that $\min_{\mu} P_{\mu} J \neq J$, let

$$\gamma = J - \bar{P}J, \quad (4.111)$$

and note that $\gamma(i) \geq 0$ for all i , with strict inequality for at least one state i . By multiplying Eq. (4.111) with \bar{P}^k , $k = 1, 2, \dots$, and adding over k , we obtain

$$J = \sum_{k=0}^{N-1} \bar{P}^k \gamma + \bar{P}^N J, \quad N = 1, 2, \dots$$

and by taking the limit as $N \rightarrow \infty$,

$$\sum_{k=0}^{\infty} \bar{P}^k \gamma < \infty,$$

from which it follows that

$$\bar{P}^* \gamma = 0. \quad (4.112)$$

Let

$R = \text{set of states from which a state } j \text{ with } \gamma(j) > 0 \text{ is reachable under } \bar{P}$

and note that $i \in R$ if and only if the i th component of $\sum_{k=0}^{\infty} \bar{P}^k \gamma$ is positive. From Eq. (4.112), it follows that $[\bar{P}^*]_{ii} = 0$ if $\gamma(i) > 0$, so

i is \bar{P} -transient for all i with $\gamma(i) > 0$

and hence

$$i \text{ is } \overline{P}\text{-transient for all } i \in R.$$

States in R cannot be reached from states not in R under $\overline{\mu}$, and because of the normalization rule, the states that are not in R have the same transition probabilities under μ and $\overline{\mu}$. Thus, we have

$$J(i) = \overline{J}(i), \quad i \notin R.$$

In view of the definition $\gamma = J - \overline{P}J$, and the fact $\overline{J} = \overline{P}\overline{J}$, we have

$$J - \overline{J} = \overline{P}(J - \overline{J}) + \gamma,$$

from which it follows that

$$J - \overline{J} = d + \sum_{k=0}^{\infty} \overline{P}^k \gamma$$

where d is a vector satisfying $d = \overline{P}d$ and hence also $d = \overline{P}^*d$. Since $J(i) = \overline{J}(i)$ and $\sum_{k=0}^{\infty} (\overline{P}^k \gamma)(i) = 0$ for all $i \notin R$, we have $d(i) = 0$ for all \overline{P} -recurrent i , which together with $d = \overline{P}^*d$, implies that $d = 0$. It follows that

$$J - \overline{J} = \sum_{k=0}^{\infty} \overline{P}^k \gamma$$

from which we obtain $J(i) > \overline{J}(i)$ for all $i \in R$.

Consider next the case $\overline{P}J = J$. Proceeding as in the single-chain case, we define

$$\delta = T_{\mu}h - T_{\overline{\mu}}h, \quad \Delta = h - \overline{h},$$

and we note that $\delta(i) \geq 0$ for all i , with strict inequality for at least one i , assuming that $\overline{\mu} \neq \mu$. The proof of Prop. 4.4.2 then goes through with minor changes, the principal of which is that in deriving the analog of Eq. (4.99), we use the equality $J - \overline{J} = \overline{P}(J - \overline{J})$ (which follows from $\overline{P}J = J$ and $\overline{P}\overline{J} = \overline{J}$) in place of $(\lambda - \overline{\lambda})e = \overline{P}(\lambda - \overline{\lambda})e$. The straightforward details are left for the reader. **Q.E.D.**

Since the number of stationary policies is finite, it follows from Prop. 4.4.4 that the policy iteration method will terminate with an optimal stationary policy.

4.5 LINEAR PROGRAMMING

Let us now develop a linear programming-based solution method. For this, we will need to use some of the most well-known linear programming theory, including duality (we assume that the reader is familiar with this material;

see e.g., Dantzig [Dan63], Bertsimas and Tsitsiklis [BeT97]). We will focus on the single-chain case, and assume that the WA condition holds. There is also a linear programming method for the multi-chain case, but it is somewhat complicated and it will only be sketched; see the end-of-chapter references for more details.

Under the WA condition, the optimal average cost λ^* is independent of the initial state, and by Prop. 4.1.6, λ^* is the largest number λ that together with some vector h satisfies

$$\lambda e + h \leq g_\mu + P_\mu h$$

for all stationary policies μ . Thus, λ^* solves the following linear program, together with some vector h^* :

maximize λ

$$\text{subject to } \lambda + h(i) \leq g(i, u) + \sum_{j=1}^n p_{ij}(u)h(j), \quad i = 1, \dots, n, \quad u \in U(i). \quad (4.113)$$

Solution of this linear program will determine the optimal cost λ^* , but may not immediately yield an optimal policy; the reason is that while some optimal solution of the linear program satisfies the optimality equation

$$\lambda^* e + h^* = Th^*,$$

not all optimal solutions are guaranteed to do so (see Exercise 4.13). To address this and other issues, it is helpful to consider another linear program, which is dual to the above. In particular, the duality theory of linear programming asserts that the following (dual) linear program, whose variables are denoted $q(i, u)$,

$$\begin{aligned} & \text{minimize} \quad \sum_{i=1}^n \sum_{u \in U(i)} q(i, u)g(i, u) \\ & \text{subject to} \quad \sum_{u \in U(j)} q(j, u) = \sum_{i=1}^n \sum_{u \in U(i)} q(i, u)p_{ij}(u), \quad j = 1, \dots, n, \\ & \quad \sum_{i=1}^n \sum_{u \in U(i)} q(i, u) = 1, \\ & \quad q(i, u) \geq 0, \quad i = 1, \dots, n, \quad u \in U(i), \end{aligned} \quad (4.114)$$

has the same optimal value as the (primal) program (4.113). Duality theory asserts that the minimal value of this dual cost is λ^* , and that there exists an optimal solution of the dual program.

Suppose that we have found an optimal solution

$$\{q^*(i, u) \mid i = 1, \dots, n, u \in U(i)\}$$

of the dual program (4.114). We will show how to construct an optimal unichain policy. Consider the set of states

$$I^* = \left\{ i \mid \sum_{u \in U(i)} q^*(i, u) > 0 \right\},$$

and its complement

$$\bar{I}^* = \{i \mid i \notin I^*\},$$

and let

$$U^*(i) = \{u \in U(i) \mid q^*(i, u) > 0\}, \quad i \in I^*.$$

Note that in view of the second equality constraint of the dual problem, I^* is nonempty and $U^*(i)$ is also nonempty for all $i \in I^*$.

Consider now a deterministic stationary policy of the form

$$\mu^*(i) = \begin{cases} \text{any } u \in U^*(i) & \text{if } i \in I^*, \\ \text{any } u \in U(i) & \text{if } i \in \bar{I}^*. \end{cases} \quad (4.115)$$

We claim that the set I^* is “closed” under μ^* , in the sense that the state remains in I^* when started at a state in I^* , i.e.,

$$p_{ij}(u) = 0, \quad \text{for all } i \in I^*, u \in U^*(i), j \in \bar{I}^*. \quad (4.116)$$

Indeed, if there existed $i \in I^*$, $j \in \bar{I}^*$, and $u \in U(i)$ with $p_{ij}(u) > 0$, we would have using the first equality constraint of the dual problem,

$$0 = \sum_{u \in U^*(j)} q^*(j, u) = \sum_{i \in I^*} \sum_{u \in U^*(i)} q^*(i, u) p_{ij}(u) > 0, \quad (4.117)$$

which is a contradiction.

Using another standard result of linear programming theory, we have that the optimal solution pair (λ^*, h^*, q^*) of the primal and dual problems satisfies the following complementary slackness relation for all $i = 1, \dots, n$ and $u \in U(i)$:

$$q^*(i, u) > 0 \Rightarrow \lambda^* + h^*(i) = g(i, u) + \sum_{j=1}^n p_{ij}(u) h^*(j).$$

From this we obtain, using Eq. (4.116),

$$\lambda^* + h^*(i) = g(i, \mu^*(i)) + \sum_{j \in I^*} p_{ij}(\mu^*(i)) h^*(j), \quad i \in I^*.$$

Since the set I^* is “closed” under μ^* [cf. Eq. (4.116)], it follows from the preceding equation and Prop. 4.2.2 that the average cost of μ^* is equal to the optimal λ^* starting from all states $i \in I^*$. If μ^* is unichain (which is true in particular under the condition that all stationary policies are unichain), then I^* is its recurrent class, all states in I^* are transient, and μ^* is optimal. If μ^* is not unichain, we can identify a set of states $S \subset I^*$ that forms a recurrent class under μ^* , and use the construction given in conjunction with Prop. 4.2.6 to redefine μ^* outside the set S . This procedure works in view of the WA condition, and yields a redefined policy, which is unichain with recurrent class the set S , and with average cost for all states equal to the optimal λ^* .

Note that the preceding algorithm may be used even when it is not known whether the WA condition holds. We can first obtain an optimal solution pair (λ^*, h^*, q^*) of the primal and dual problems; it can be shown that such a pair always exists (because both the primal and the dual programs are feasible), and in fact λ^* is equal to the lowest optimal average cost among all states (see the subsequent discussion on state-action frequencies, or Exercise 4.16). Then, we can construct the nonempty set I^* and a policy μ^* of the form (4.115). As before, μ^* is optimal starting from the states in I^* , with average cost λ^* . We can then identify a set of states $S \subset I^*$ that forms a recurrent class under μ^* , and try to use the construction given in conjunction with Prop. 4.2.6 to redefine μ^* outside S . If this construction succeeds, it will yield a unichain policy with average cost equal to the optimal λ^* . If the construction fails (yielding a policy that is unichain and optimal within a strict subset of states), this means that the WA condition does not hold. In this case, to find an optimal policy for states outside I^* , a method that works with multi-chain problems should be used.

The preceding analysis has provided a method to obtain an optimal stationary policy given a dual optimal solution - an optimal primal solution is not needed. There is also an alternative approach, sketched in Exercises 4.14 and 4.15, that uses a primal optimal solution (λ^*, h^*) , assuming the WA condition holds - an optimal dual solution is not needed. The idea is to construct a “closed” set of states i for which

$$\lambda^* + h^*(i) = g(i, u) + \sum_{j=1}^n p_{ij}(u)h^*(j), \text{ for some } u \in U(i),$$

start with a recurrent class from this set, and use the construction given in conjunction with Prop. 4.2.6.

The Multi-Chain Case

Let us consider now the multi-chain version of the linear program (4.113). By Prop. 4.1.6, J^* is the “largest” vector J that together with some vector

h satisfies

$$J \leq P_\mu J, \quad J + h \leq g_\mu + P_\mu h$$

for all stationary policies μ . This leads to the linear program

$$\begin{aligned} & \text{maximize} \quad \sum_{i=1}^n \beta_i J(i) \\ & \text{subject to} \quad J(i) \leq \sum_{j=1}^n p_{ij}(u) J(j), \quad i = 1, \dots, n, \quad u \in U(i), \\ & \quad J(i) + h(i) \leq g(i, u) + \sum_{j=1}^n p_{ij}(u) h(j), \quad i = 1, \dots, n, \quad u \in U(i), \end{aligned} \tag{4.118}$$

where β_i are some positive scalars with $\sum_{i=1}^n \beta_i = 1$. The corresponding dual program is

$$\begin{aligned} & \text{minimize} \quad \sum_{i=1}^n \sum_{u \in U(i)} q(i, u) g(i, u) \\ & \text{subject to} \quad \sum_{u \in U(j)} q(j, u) = \sum_{i=1}^n \sum_{u \in U(i)} q(i, u) p_{ij}(u), \quad j = 1, \dots, n, \\ & \quad \sum_{u \in U(j)} q(j, u) + \sum_{u \in U(j)} r(j, u) = \beta_j + \sum_{i=1}^n \sum_{u \in U(i)} r(i, u) p_{ij}(u), \quad j = 1, \dots, n, \\ & \quad q(i, u), r(i, u) \geq 0, \quad i = 1, \dots, n, \quad u \in U(i). \end{aligned} \tag{4.119}$$

The dual optimization variables are $q(i, u)$ and $r(i, u)$, $i = 1, \dots, n$, $u \in U(i)$. Again by standard results of linear programming theory, it follows that the optimal values of the primal and dual programs are equal, and that the dual program has an optimal solution.

Consider an optimal solution

$$\{q^*(i, u), r^*(i, u) \mid i = 1, \dots, n, u \in U(i)\}$$

of the dual program, the sets of states

$$I^* = \left\{ i \mid \sum_{u \in U(i)} q^*(i, u) > 0 \right\}, \quad \bar{I}^* = \{i \mid i \notin I^*\}.$$

Let μ^* be any (deterministic) stationary policy of the form

$$\mu^*(i) = \begin{cases} \text{any } u \in U(i) \text{ such that } q^*(i, u) > 0 & \text{if } i \in I^*, \\ \text{any } u \in U(i) \text{ such that } r^*(i, u) > 0 & \text{if } i \in \bar{I}^*. \end{cases}$$

In general, such a policy need not be optimal starting from all initial states. However, its optimality is guaranteed if (q^*, r^*) is an *extreme point* of the dual feasible set (the simplex method can find such an optimal extreme point); see Kallenberg [Kal83], [Kal94a] (Theorem 8) and the references quoted there. Exercise 4.18 sketches the proof and provides an alternative approach based on randomized policies. In particular, it shows that the randomized policy defined by

$$P(u | i) = \begin{cases} \frac{q^*(i, u)}{\sum_{u \in U(i)} q^*(i, u)} & \text{if } i \in I^*, \\ \frac{r^*(i, u)}{\sum_{u \in U(i)} r^*(i, u)} & \text{if } i \in \bar{I}^*, \end{cases}$$

is optimal.

There are also other, more complicated methods to construct an optimal stationary policy, which do not require an extreme dual optimal solution. For these we refer to the literature; see Derman [Der70] (Chapter 6) for a method that uses the optimal primal solution.

State-Action Frequencies and Randomized Policies

The dual single-chain program (4.114) provides an interesting interpretation in connection with *randomized* policies, i.e., policies that choose at state i the control u probabilistically, by sampling the constraint set $U(i)$ according to some probabilities. Under any randomized stationary policy, we obtain a stationary Markov chain whose states are (i, u) , $i = 1, \dots, n$, $u \in U(i)$. Let R be a recurrent class of this Markov chain. Then each state-control pair $(i, u) \in R$ has a long-term frequency of occurrence $f(i, u)$ within R (also called *state-action frequency*). In particular, we have for all $(i, u) \in R$ and $(i', u') \in R$,

$$f(i, u) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N P(i_k = i, u_k = u | i_0 = i', u_0 = u'),$$

where (i_k, u_k) is the state-control pair at time k under the policy [by Prop. 4.1.1, the limit exists and is the same for all initial state-control pairs $(i', u') \in R$]. The state-action frequencies $f(i, u)$ satisfy the equations

$$f(j, v) = \nu(j, v) \sum_{(i, u) \in R} f(i, u) \tilde{p}_{ij}(u), \quad \text{for all } (j, v) \in R,$$

where $\nu(j, v)$ is the probability of applying control v at state j according to the given randomized policy [this follows from Prop. 4.1.1, cf. Eq. (4.7)]. By

adding this equation over $v \in U(j)$, we see that the scalars $q(i, u)$ defined by

$$q(i, u) = \begin{cases} f(i, u) & \text{if } (i, u) \in R, \\ 0 & \text{if } (i, u) \notin R, \end{cases}$$

satisfy the constraints of the dual program (4.114). The average cost of the given randomized policy, starting from a state-control pair in R is the dual cost

$$\sum_{i=1}^n \sum_{u \in U(i)} q(i, u) g(i, u),$$

which by the relation between primal and dual problems, is bounded below by the optimal value λ^* of the primal single-chain program (4.113).

The conclusion from this analysis is that every randomized stationary policy has average cost that is bounded below by λ^* within every one of its associated recurrent classes. Hence the average cost of the policy starting from any state-control pair is no less than λ^* (since the average cost starting from a transient state-control pair is no less than the minimal average cost starting from recurrent state-control pairs). We thus obtain a theoretically interesting fact (which was proved independently of the WA condition): *for any randomized stationary policy, the average cost starting from any initial state-control pair is no less than the optimal cost λ^* of the primal problem (4.113).* As a special case, we have that *under the WA condition, no stationary randomized policy can achieve an average cost that is less than the optimal that can be achieved with an ordinary (deterministic) stationary policy.*

4.6 INFINITE-SPACES PROBLEMS

The standing assumption in the preceding sections has been that the state and control spaces are finite. Without finiteness of one or both of these spaces, many of the results presented in the past three sections no longer hold. In particular, even if the optimal average cost is the same for all initial states, one or more of the following may happen:

- (a) Bellman's equation and/or the coupled pair of optimality equations may not have a solution.
- (b) Value iteration may fail to converge to the optimal average cost.
- (c) Policy iteration may not be valid and may not yield the optimal gain in the limit.
- (d) There may not exist an optimal stationary policy, while there is an optimal nonstationary policy.
- (e) There may not exist an optimal policy, stationary or nonstationary.

- (f) The optimal average cost may not be approachable by the costs of stationary policies.

In what follows in this section we provide examples and counterexamples illustrating the main difficulties. In subsequent sections, we discuss conditions and structure that allow the development of useful results. Still however, contrary to the case of problems with finite state and control spaces, there is no comprehensive theory for infinite space average cost problems.

Existence of Solution of Bellman's Equation

Unusual behavior can occur even in problems with a finite number of states but an infinite number of controls. In the following example, Bellman's equation does not have a solution. This example involves two states (one of them absorbing), and is obtained from the blackmail Example 2.2.1, which illustrates pathological behavior in stochastic shortest path problems. The pathological behavior reappears when the problem is viewed as an average cost problem; this is not surprising in view of the strong connection between the two types of problems.

Example 4.6.1 (The Blackmailer's Dilemma Revisited)

Suppose there are two states, 1 and t . At state 1, we can choose a control $u \in (0, 1]$, while incurring a cost $-u$; we then move to state t with probability u^2 , and stay in state 1 with probability $1 - u^2$. State t is absorbing and cost-free.

This example is derived from Example 2.2.1, a pathological stochastic shortest path problem with optimal cost $J^*(1) = -\infty$, which has no real-valued solution to Bellman's equation, has no optimal stationary policy, and yet has an optimal nonstationary policy. We note that every stationary policy leads to state t in a finite expected number of transitions, so the corresponding average cost is equal to 0. In fact it is possible to show that the optimal average cost is equal to 0 starting from either state, so all stationary policies are average cost optimal (one way to verify this is by using value iteration arguments; see the subsequent Example 4.6.3).

The coupled pair of optimality equations [cf. Eqs. (4.39) and (4.40)], has the form

$$J^*(1) = \min_{u \in (0, 1]} [(1 - u^2) J^*(1) + u^2 J^*(t)], \quad J^*(t) = J^*(t),$$

$$J^*(1) + h(1) = \min_{u \in \bar{U}(1)} [-u + (1 - u^2) h(1) + u^2 h(t)], \quad J^*(t) + h(t) = h(t),$$

where $\bar{U}(1)$ is the set of $u \in (0, 1]$ that attain the minimum in the first equation. A solution to this pair of equations must satisfy $J^*(t) = 0$ (from the last equation), and hence $J^*(1) = 0$ (from the first equation), with $\bar{U}(1) = (0, 1]$. Thus the coupled pair of optimality equations is reduced to a single equation,

which is Bellman's equation for the stochastic shortest path problem, and has a real-valued solution as mentioned earlier. What is happening here is that all policies have average cost 0, yet starting from state 1, there is an infinite advantage in total cost relative to starting from state t .

The optimal cost of the α -discounted version of the problem, starting at state 1, is the unique solution of the equation

$$J_\alpha(1) = (TJ_\alpha)(1),$$

where by using the calculations of Example 2.2.1, we have

$$(TJ)(1) = \min_{u \in (0,1]} [-u + \alpha(1-u^2)J(1)] = \begin{cases} -1 & \text{if } \alpha J(1) \geq -1/2, \\ \alpha J(1) + \frac{1}{4\alpha J(1)} & \text{if } \alpha J(1) \leq -1/2. \end{cases}$$

From this, by solving the equation $J_\alpha(1) = (TJ_\alpha)(1)$, it can be seen that

$$J_\alpha(1) = \begin{cases} -1 & \text{if } 0 \leq \alpha \leq 1/2, \\ -\frac{1}{2\sqrt{\alpha(1-\alpha)}} & \text{if } 1/2 \leq \alpha < 1, \end{cases} \quad (4.120)$$

and that the optimal α -discounted stationary policy is

$$\mu^*(1) = \begin{cases} 1 & \text{if } 0 \leq \alpha \leq 1/2, \\ \sqrt{\frac{1-\alpha}{\alpha}} & \text{if } 1/2 \leq \alpha < 1. \end{cases}$$

Thus J_α does not have a Laurent series expansion of the form of Prop. 4.1.2. This indicates that for infinite control space problems, the line of analysis based on the Laurent series expansion, Blackwell optimal policies, and the connection between average cost and discounted problems must be substantially modified in order to be useful, even if the state space is finite.

Finally, note that the difficulties just discussed also arise in the case where the control constraint is $u \in [0, 1]$. Thus compactness of the control constraint sets, and continuity with respect to u of the one-stage costs and transition probabilities are not sufficient to guarantee that the coupled pair of optimality equations or Bellman's equation has a real-valued solution, even with finite state space.

Convergence of Value Iteration

We saw in Section 4.3 that in the case of finite state and control spaces, value iteration is fairly well behaved. In particular, Prop. 4.3.1 shows that

$$\lim_{k \rightarrow \infty} \frac{1}{k} T^k h = J^*,$$

for all h , where J^* is the optimal average cost vector. Furthermore, the sequence $\{T^k h\}$ grows at the rate of k . These results fail to hold in general for infinite space problems, as the following two examples from Whittle [Whi82] demonstrate. In the first example the failure is due to another anomaly that does not occur in finite-spaces models: the optimal "upper" (or limsup) and "lower" (or liminf) costs, as defined in Section 4.1, are not equal.

Example 4.6.2

Consider a deterministic problem with countably many states and controls, where at some state, say state 0, the controls available are $1, 2, \dots$, and choice of control $u = m$ returns the state back to 0 after going through a cycle with m transitions, the first $m - 1$ of which have cost 0, and the last has cost m^2 . Thus, with choice of $u = m$ at state 0, recurrence to 0 occurs in m stages and with an average cost of m per stage. It follows that value iteration with initial function $h = 0$, yields $(T^k h)(0) = 0$ for every k , since over a finite horizon it is optimal to choose a control large enough to postpone the occurrence of a positive cost beyond the horizon. However, the optimal average cost starting from 0 [as defined by Eq. (4.1) in terms of \limsup] is clearly $J^*(0) = 1$, and is attained by the stationary policy that chooses control 1 at state 0. Thus value iteration fails in this example.

It is interesting to note here that if the cost of a policy is defined in terms of \liminf , the optimal average cost starting from 0 is equal to 0. This cost is attained with a nonstationary policy that upon return to state 0, chooses a control $u = m$ with m large enough to bring $(1/N)E\{\sum_{k=0}^{N-1} g(x_k, \mu_k(x_k))\}$ close enough to 0 for large enough N .

In the next example the value iterates $T^k h$ do not grow at the rate of k , although they do yield the optimal average cost vector in the limit.

Example 4.6.3

Consider again the blackmailer Example 4.6.1. We have

$$(Th)(1) = \min_{u \in (0,1]} [-u + (1-u^2)h(1)] = \begin{cases} -1 & \text{if } h(1) \geq -1/2, \\ h(1) + \frac{1}{4h(1)} & \text{if } h(1) \leq -1/2, \end{cases}$$

[take $\alpha = 1$ in the discounted case formula (4.120)]. From this it can be seen that starting from $h^0(1) = 0$, value iteration produces a sequence such that $(T^k h^0)(1) \rightarrow -\infty$ and that

$$(T^{k+1} h^0)(1) - (T^k h^0)(1) = \frac{1}{4(T^k h^0)(1)} \rightarrow 0.$$

Thus $|(T^k h^0)(1)|$ does not grow at the rate of k , as in the finite-spaces case (cf. Prop. 4.3.1). In fact, it can be verified by induction that there exists a constant c such that

$$-c\sqrt{k} \leq \overbrace{(T^k h^0)(1)}^{\sim} \leq 0, \quad k = 1, 2, \dots,$$

so $|(T^k h^0)(1)|$ grows at a rate of no more than \sqrt{k} .

Extension of Policy Iteration

There is a natural extension of policy iteration to average cost problems where the state space is finite, but the control constraint sets $U(i)$ are infinite. With a finite state space, the policy evaluation step remains the same, and the policy improvement step requires little modification, the main difficulty being a minimization over the infinite (rather than finite) set $U(i)$ for each i . Yet there are significant complications because the set of stationary policies is infinite, so finite termination is not expected.

In fact even for this simplest of extensions, there are examples, involving single-chain and multi-chain problems, and compact sets $U(i)$, where the gain sequence generated by policy iteration does not converge to the optimal (Dekker [Dek87]). In the single-chain case, Hordijk and Puterman [HoP87] have studied policy iteration under a certain compactness condition on the sets $U(i)$ and some other conditions, including continuity of $p_{ij}(u)$ with respect to u . Golubin [Gol03] has shown that the analysis of Hordijk and Puterman [HoP87] is flawed, and for their results to be valid, additional conditions are needed. Golubin [Gol03] and Patek [Pat04] introduced a recurrence structure, which brings to bear the connection of average cost and stochastic shortest path problems described in Section 4.3, and allows the development of convergent versions of policy iteration.

Here is another example of unusual behavior.

Example 4.6.4

Consider the blackmailer problem of Examples 4.6.1 and 4.6.3. Here all stationary policies are optimal, yet it can be verified that starting from the (optimal) stationary policy μ^0 with $\mu^0(1) = u$ where $u \in (0, 1)$, the sequence of (optimal) generated policies satisfies $\mu^k(1) = 2^{-k}u$ and does not terminate (and in fact converges to an infeasible policy).

The main reason here is that the policy improvement step cannot recognize an optimal policy, something that can also happen in the finite-spaces case. In particular, suppose that we replace $U(1)$ with the finite set

$$\{2^{-m} \mid m = 1, \dots, M\}.$$

Then, still each of the M stationary policies $\mu(1) = 2^{-m}$ is average cost optimal, but starting from $\mu^0(1) = 1/2$, the policy iteration algorithm generates all the stationary policies, and terminates with $\mu^{M-1}(1) = 2^{-M}$. However, when there are infinitely many stationary policies, as in the infinite control space case, it is possible to generate an infinite sequence of optimal policies without being able to detect optimality.

Existence of Optimal and Nearly Optimal Policies

Whereas one can restrict attention to stationary policies in finite-spaces problems, this is no longer true when the state space is infinite. The fol-

lowing example, from Ross [Ros70], shows that for a countable state space there may exist an optimal nonstationary policy, but not an optimal stationary policy.

Example 4.6.5

Let the state space be $\{1, 2, 3, \dots\}$, and let there be two controls, u^1 and u^2 . The transition probabilities and costs per stage are

$$p_{i(i+1)}(u^1) = p_{ii}(u^2) = 1,$$

$$g(i, u^1) = 1, \quad g(i, u^2) = \frac{1}{i}, \quad i = 1, 2, \dots$$

In words, at state i we may either move to state $(i + 1)$ at a cost 1 or stay at i at a cost $1/i$.

For any stationary policy μ other than the policy for which $\mu(i) = u^1$ for all i , let $n(\mu)$ be the smallest integer for which

$$\mu(n(\mu)) = u^2.$$

Then the corresponding average cost per stage satisfies

$$J_\mu(i) = \frac{1}{n(\mu)} > 0, \quad \text{for all } i \text{ with } i \leq n(\mu).$$

For the policy where $\mu(i) = u^1$ for all i , we have $J_\pi(i) = 1$ for all i . Since the optimal cost per stage cannot be less than zero, it is clear that

$$\min_{\pi} J_\pi(i) = 0, \quad i = 1, 2, \dots$$

However, the optimal cost is not attained by any stationary policy, so no stationary policy is optimal. On the other hand, consider the nonstationary policy π^* that on entering state i chooses u^2 for i consecutive times and then chooses u^1 . If the starting state is i , the sequence of costs incurred is

$$\underbrace{\frac{1}{i}, \frac{1}{i}, \dots, \frac{1}{i}}_{i \text{ times}}, \quad 1, \quad \underbrace{\frac{1}{i+1}, \frac{1}{i+1}, \dots, \frac{1}{i+1}}_{(i+1) \text{ times}}, \quad 1, \quad \frac{1}{i+2}, \frac{1}{i+2}, \dots$$

The average cost corresponding to this policy is

$$J_{\pi^*}(i) = \lim_{m \rightarrow \infty} \frac{2m}{\sum_{k=1}^m (i+k)} = 0, \quad i = 1, 2, 3, \dots$$

Hence the nonstationary policy π^* is optimal while, as shown previously, no stationary policy is optimal.

In the preceding example, it can be seen that given any $\epsilon > 0$, there exists an ϵ -optimal stationary policy, i.e., a μ_ϵ such that

$$J_{\mu_\epsilon}(i) \leq J^*(i) + \epsilon, \quad i = 1, 2, \dots$$

Ross [Ros71] gives an example where this is not so. Thus, for problems with an infinite number of states, one may not be able to approach the optimal average cost using stationary policies.

The following is an example from Ross [Ros83a], which shows that if the state space is countable, there may not exist an optimal policy.

Example 4.6.6

Let the state space be $\{1, 1', 2, 2', 3, 3', \dots\}$, and let there be two controls, u^1 and u^2 . The transition probabilities and costs per stage are

$$p_{i(i+1)}(u^1) = 1, \quad p_{ii'}(u^2) = 1, \quad i = 1, 2, \dots,$$

$$p_{i'i'}(u^1) = p_{i'i'}(u^2) = 1, \quad i = 1, 2, \dots,$$

$$g(i, u^1) = g(i, u^2) = 0, \quad i = 1, 2, \dots,$$

$$g(i', u^1) = g(i', u^2) = -1 + \frac{1}{i}, \quad i = 1, 2, \dots$$

In words, at state i we may, at a cost 0, either move to state $(i+1)$ or move to state i' , where we stay thereafter at a cost $-1 + 1/i$ per stage.

It can be seen that for every policy π and state $i = 1, 2, \dots$, we have $J_\pi(i) > -1$. However, for every state i , we can obtain an average cost per stage $-1 + 1/j$, where $j \geq i$, by moving to state j' once we get to state j . Hence, for every initial state $i = 1, 2, \dots$, an average cost per stage of -1 can be approached arbitrarily closely with a stationary policy, but cannot be attained by any policy.

Average Cost Dependence on the Initial State – Imperfect State Information Problems

While we have viewed as “typical” the case where the optimal average cost is the same for all initial states, there are important classes of problems where this viewpoint is questionable. In particular, finite-state problems with imperfect state information become (uncountably) infinite-state average cost problems when viewed in the perfect information setting of a sufficient statistic (the conditional distribution of the state). For such problems, we will demonstrate by example that the optimal average cost depends on the initial state distribution even in unexpectedly simple situations.

Consider an infinite horizon average cost version of an imperfect state information problem involving a stationary finite-state Markov chain (cf. Section 5.4.2 of Vol. I). Here, the states are denoted $1, 2, \dots, n$. When a

control u is applied, the system moves from state i to state j with probability $p_{ij}(u)$. The control u is chosen from a finite set U . Following a state transition, an observation is made by the controller. There is a finite number of possible observation outcomes, and the probability of each depends on the current state and the preceding control. The information available to the controller at stage k is the information vector

$$I_k = (z_1, \dots, z_k, u_0, \dots, u_{k-1}),$$

where for all i , z_i and u_i are the observation and control at stage i , respectively. Following the observation z_k , a control u_k is chosen by the controller, and a cost $g(x_k, u_k)$ is incurred, where x_k is the current (hidden) state.

As discussed in Section 5.4.2 of Vol. I, we can reformulate the problem into a problem of perfect state information where the objective is to control the column vector of conditional probabilities $p_k = (p_k^1, \dots, p_k^n)'$, with

$$p_k^j = P(x_k = j | I_k), \quad j = 1, \dots, n.$$

We refer to p_k as the *belief state*, and we note that it evolves according to an equation of the form

$$p_{k+1} = \Phi(p_k, u_k, z_{k+1}).$$

The function Φ represents an estimator, as discussed in Section 5.4.2 of Vol. I. The initial belief state p_0 is given.

In the case of a total cost problem with discount factor $\alpha \in (0, 1)$, the corresponding Bellman's equation has the form

$$J^*(p) = \min_{u \in U} [p' g(u) + \alpha E_z \{ J^*(\Phi(p, u, z)) | p, u \}],$$

where $g(u)$ is the column vector with components $g(1, u), \dots, g(n, u)$. The theory of Chapter 1 fully applies because the cost per stage, $p' g(u)$, is bounded. Similarly, the theory of Chapter 3 applies if there is no discounting, but the costs $g(i, u)$ are either nonnegative or nonpositive for all i and u .

On the other hand, in the case of an average cost problem, not only there are difficulties of the type discussed earlier in this section, but also the optimal average cost may depend on the initial state, even if all states communicate under all policies. The following simple example from Yu and Bertsekas [YuB06a] illustrates this fact.

Example 4.6.7

Here there are four states $\{1, 2, 3, 4\}$, two controls $\{a, b\}$, and two observations $\{c, d\}$. The choice of control cannot influence the probabilistic evolution of

the state or the observations - it can only influence the cost per stage. In particular, under any policy, the state process is a Markov chain with transition probabilities as follows:

$$p_{11} = 1/2, \quad p_{21} = 1/2,$$

$$p_{43} = 1/2, \quad p_{33} = 1/2,$$

$$p_{32} = 1, \quad p_{14} = 1,$$

(see Fig. 4.6.1). The conditional probabilities of the observations given the current state are

$$P(c | 1) = P(c | 3) = 1, \quad P(d | 2) = P(d | 4) = 1.$$

Note that the observation structure is such that states 1 and 3 are indistinguishable, and states 2 and 4 are indistinguishable. Furthermore, if we know the initial state, the state process can be inferred from the observations as if it were completely observable.

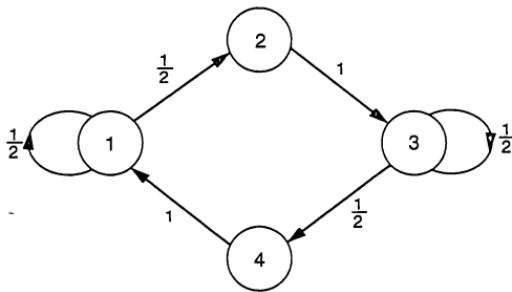


Figure 4.6.1 Transition probabilities in Example 4.6.7 (independently of the control applied).

The costs per stage are all 0 except for

$$g(1, a) = g(3, b) = 1.$$

It follows that if we start from an initial distribution p_0 with $p_0^1 = 1$ or an initial distribution with $p_0^3 = 1$, then the optimal average cost is zero, while if we start from an initial distribution p_0 with $p_0^1 = p_0^3 = 1/2$, say, then the optimal average cost is strictly greater than zero.

4.6.1 A Sufficient Condition for Optimality

Generally, the analysis of average cost problems with infinite state or control spaces poses many difficulties, and at present there is no comprehensive theory. However, there are sets of assumptions that allow a satisfactory analysis. An important tool is a straightforward extension of Prop. 4.2.1

to the case where the state and control spaces are infinite. In particular, if we can find a scalar λ and a bounded function h such that Bellman's equation (4.47) holds, then by essentially repeating the proof of Prop. 4.1.6, we can show that λ must be the optimal average cost of all initial states. We prove a somewhat more general version of this fact, and for simplicity, we restrict ourselves to the case where the state space is countably infinite. The line of proof, however, extends to more general cases.

Proposition 4.6.1: Let the state space S be countably infinite. Assume that a scalar λ and a real-valued function h solve Bellman's equation, i.e., for all states i ,

$$\lambda + h(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j \in S} p_{ij}(u)h(j) \right], \quad (4.121)$$

and furthermore h satisfies for all policies π and states i

$$\lim_{N \rightarrow \infty} \frac{1}{N} E\{h(x_N) \mid x_0 = i, \pi\} = 0. \quad (4.122)$$

Then

$$\lambda = \min_{\pi} J_{\pi}(i) = J^*(i).$$

Furthermore, if $\mu^*(i)$ attains the minimum in Eq. (4.121) for each i , the stationary policy μ^* is optimal, i.e., $J_{\mu^*}(i) = \lambda$ for all i .

Proof: Using the proof of Prop. 4.1.6, we obtain [cf. Eq. (4.36)]

$$\begin{aligned} \frac{1}{N} E\{h(x_N) \mid x_0 = i, \pi\} &+ \frac{1}{N} E \left\{ \sum_{k=0}^{N-1} g(x_k, \mu_k(x_k)) \mid x_0 = i, \pi \right\} \\ &\geq \lambda + \frac{1}{N} h(i), \end{aligned}$$

with equality if $\mu_k(i)$, $k = 0, 1, \dots$, attains the minimum in Eq. (4.121) for all i . By taking the limit as $N \rightarrow \infty$, the term $(1/N)h(i)$ vanishes while the term involving h in the left-hand side also vanishes by assumption, so that

$$J_{\pi}(i) \geq \lambda, \quad i = 1, \dots, n,$$

with equality if $\pi = \{\mu^*, \mu^*, \dots\}$ and $\mu^*(i)$ attains the minimum in Eq. (4.121) for all i . **Q.E.D.**

The assumption $E\{h(x_N) \mid x_0 = i, \pi\}/N \rightarrow 0$ [cf. Eq. (4.122)] may be viewed as a *stability condition*, requiring that $h(x_N)$ grows on the average

at a rate less than N under every policy; it is of course unnecessary if the state space is finite or more generally if it is known that h is bounded. An important analytical issue in a given problem is to delineate assumptions that imply the associated stability condition.

Some important special cases can be satisfactorily analyzed using Prop. 4.6.1. One such case, discussed in Section 4.6.2, is when the control space is infinite, while the state space is finite and some additional conditions hold. Another case, discussed in Section 4.6.3, is problems with countably infinite state space and possibly unbounded costs per stage, for which we use a line of analysis that is based on the connection of average cost and discounted problems. Countably infinite state space problems are also discussed in Section 4.6.4 for the case where there is a special state to which the system has a tendency to return under any policy, similar to the stochastic shortest path problems of Section 2.5. Still another case, involving an uncountably infinite state space, is discussed in Section 4.6.5. It is the average cost version of the linear-quadratic problem examined in Chapters 4 and 5 of Vol. I.

4.6.2 Finite State Space and Infinite Control Space

Consider the average cost problem with states $i = 1, \dots, n$ and arbitrary control constraint sets $U(i)$. We will focus on a version of the problem that involves the use of *randomized controls*, generated by randomization among (ordinary) controls from the sets $U(i)$. Our main assumption is the following strengthened version of the WA condition of Definition 4.2.2, together with a cost boundedness assumption.

Assumption 4.6.1:

- (a) **Accessibility Condition:** For every pair of states i and j , there exists a stationary policy μ and an integer k such that

$$P(x_k = j | x_0 = i, \mu) > 0.$$

- (b) **Boundedness:** The cost per stage $g(i, \cdot)$ is bounded over $U(i)$ for each i .

For each state i , we associate a control $u \in U(i)$ with the vector of transition probabilities

$$p_i(u) = (p_{i1}(u), \dots, p_{in}(u)),$$

and we denote

$$\mathcal{P}_i = \{p_i(u) \mid u \in U(i)\}.$$

We consider the convex hull of \mathcal{P}_i , denoted by $\overline{\mathcal{P}}_i$:

$$\overline{\mathcal{P}}_i = \text{conv}(\mathcal{P}_i);$$

this is the set of all convex combinations of (a finite number of) vectors in \mathcal{P}_i . Thus, \mathcal{P}_i consists of all transition probability vectors that can be generated at state i by using randomization between a finite number of controls in $U(i)$.

We denote

$$\overline{\mathcal{P}} = \overline{\mathcal{P}}_1 \times \cdots \times \overline{\mathcal{P}}_n,$$

and we associate $\overline{\mathcal{P}}$ with the set of stochastic matrices whose i th row is a transition probability vector from $\overline{\mathcal{P}}_i$. For each vector $p_i \in \overline{\mathcal{P}}_i$, we introduce an associated expected stage cost, which is a “randomized” version of our usual notion of expected cost at state i under a control. It is denoted by $\bar{g}_i(p_i)$, and it is the minimum of the expected cost of state i under randomized controls from $U(i)$ that are associated with p_i . More precisely

$$\bar{g}_i(p_i) = \min_{(\xi_1, \dots, \xi_M, u_1, \dots, u_M) \in V_i(p_i)} \sum_{m=1}^M \xi_m g(i, u_m),$$

where

$$V_i(p_i) = \left\{ (\xi_1, \dots, \xi_M, u_1, \dots, u_M) \mid u_1, \dots, u_M \in U(i), p_i = \sum_{m=1}^M \xi_m p_i(u_m), M \geq 1, \xi_i \geq 0, \sum_{m=1}^M \xi_m = 1 \right\}$$

defines the set of all “randomized” controls that give rise to the probability vector p_i . It can be shown that $\bar{g}_i(p_i)$ is a convex function, and that in the finite control case [$U(i)$ is finite], it is also polyhedral. Furthermore, $\bar{g}_i(\cdot)$ is bounded over $\overline{\mathcal{P}}_i$ under Assumption 4.6.1(b).†

† It can be seen that the epigraph of \bar{g}_i is the convex hull of the epigraphs of the extended real-valued functions $r_i(\cdot, u)$ defined by

$$r_i(p_i, u) = \begin{cases} g_i(i, u) & \text{if } p_i = p_i(u), \\ \infty & \text{otherwise,} \end{cases} \quad u \in U(i).$$

The epigraph of a function $r_i(\cdot, u)$ is the half-line

$$\{(p_i(u), \gamma) \mid g_i(i, u) \leq \gamma\} \subset \mathbb{R}^{n+1}.$$

In particular, if the set $U(i)$ is finite, the epigraph of \bar{g}_i is the convex hull of a finite number of half-lines, and is therefore a polyhedral set. Note that $\bar{g}_i(\cdot)$ is bounded over $\overline{\mathcal{P}}_i$ if the cost per stage $g_i(i, \cdot)$ is bounded over $U(i)$, which is Assumption 4.6.1(b).

For each stochastic matrix $P \in \overline{\mathcal{P}}$ of the form

$$P = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix}, \quad p_i \in \overline{\mathcal{P}}_i, i = 1, \dots, n,$$

we consider an associated n -dimensional stage cost vector

$$\bar{g}(P) = \begin{bmatrix} \bar{g}_1(p_1) \\ \vdots \\ \bar{g}_n(p_n) \end{bmatrix},$$

and average cost vector

$$J(P) = P^* \bar{g}(P),$$

where

$$P^* = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k,$$

[cf. Eq. (4.5)]. The problem is now to minimize $J(P)$ over all $P \in \overline{\mathcal{P}}$, i.e., to find

$$J^* = \min_{P \in \overline{\mathcal{P}}} J(P),$$

where the minimization is meant to be separate for each component of $J(P)$.

We recognize this as a natural extension of the average cost problem for a finite number of states and a control constraint set derived from an original (possibly infinite) control constraint set via randomization. We refer to this, somewhat loosely, in this section as the *randomized average cost problem*, to distinguish it from the original problem that involves no randomization.

We will show that there exists a scalar λ and a vector $h \in \mathbb{R}^n$ such that

$$\lambda e + h = \min_{P \in \overline{\mathcal{P}}} [\bar{g}(P) + Ph], \quad (4.123)$$

where the minimization is meant to be separate for each component. This is Bellman's equation for the randomized average cost problem. As a result, Prop. 4.6.1 can be applied to show that λ is the optimal average cost from each initial state, using randomized controls. Furthermore, an optimal stationary (randomized) policy can be obtained by minimization in the right-hand side (assuming that the minimum is attained). Such a stationary optimal policy yields a randomized control/probability vector $\zeta_i^* \in \overline{\mathcal{P}}_i$ to be applied at a state i . Each ζ_i^* is the convex combination of a finite number of probability vectors $p_i(u)$ with $u \in U(i)$ satisfying

$$\zeta_i^* = \sum_{m=1}^{M_i} \xi_{im} p_i(u_{im}), \quad \bar{g}_i(\zeta_i^*) = \sum_{m=1}^{M_i} \xi_{im} g(i, u_{im}),$$

for some $M_i \geq 1$, $u_{im} \in U(i)$, $\xi_{im} \geq 0$, $m = 1, \dots, M_i$, with $\sum_{m=1}^{M_i} \xi_{im} = 1$. Consider now the average cost problem with the control constraint set $U(i)$ replaced by the finite set

$$\bar{U}(i) = \{u_{1m}, \dots, u_{iM_i}\}.$$

This is a finite-spaces average cost problem for which the theory of Sections 4.1-4.5 applies, and from which an optimal nonrandomized policy can be extracted (cf. the methodology of Section 4.5).

We will now focus on the problem that involves randomized controls, without further reference to the (nonrandomized) original problem. Consider the mapping $T : \mathbb{R}^n \mapsto \mathbb{R}^n$ defined by

$$Th = \min_{P \in \bar{\mathcal{P}}} [\bar{g}(P) + Ph].$$

Note here that the components of Th are real-valued, since Assumption 4.6.1(b) implies that $\bar{g}_i(p_i)$, the components of $\bar{g}(P)$, are bounded. Define for all $y = (y_1, \dots, y_n) \in \mathbb{R}^n$,

$$\|y\| = \max_{i=1, \dots, n} y_i - \min_{i=1, \dots, n} y_i.$$

We refer to $\|\cdot\|$ as the *span seminorm*, and we note that while it is not a norm in the usual sense, it satisfies all the defining properties of a norm except the property that $\|y\| = 0$ implies $y = 0$. In particular, it satisfies the triangle inequality

$$\|y + z\| \leq \|y\| + \|z\|, \quad y, z \in \mathbb{R}^n.$$

Since $\|y - z\| = 0$ if and only if y and z differ by a multiple of the unit vector e , a vector h satisfies

$$\|Th - h\| = 0,$$

if and only if

$$\lambda e + h = Th$$

for some scalar λ . Hence proving existence of solution of the Bellman equation (4.123) is equivalent to proving existence of a vector h such that $\|Th - h\| = 0$.†

† We have already encountered the span seminorm in the analysis of value iteration (see the proof of Prop. 4.3.2). In fact, it can be seen that an equivalent statement of the conclusion of Prop. 4.3.2 is that the value iteration sequence $T^k h^0$ converges in the span seminorm sense, i.e., there exists h^* such that

$$\lim_{k \rightarrow \infty} \|T^k h^0 - h^*\| = 0,$$

and furthermore h^* satisfies Bellman's equation, which can be equivalently written as $\|Th^* - h^*\| = 0$.

We first derive some useful properties of the mapping T in connection with the span seminorm. For any $y \in \Re^n$, we define

$$H(y) = \max\{y_i \mid i = 1, \dots, n\}, \quad L(y) = \min\{y_i \mid i = 1, \dots, n\}.$$

We have the following lemma.

Lemma 4.6.1: For any $y, z \in \Re^n$ and $\beta, \gamma \in \Re$, we have

- (a) $H(Ty - Tz) \leq H(y - z)$.
- (b) $\|Ty - Tz\| \leq \|y - z\|$.
- (c) $\|T(\beta y) - T(\gamma z)\| \leq |\beta| \|y - z\| + |\beta - \gamma| \|z\|$.

Proof: (a) For any $\epsilon > 0$, let $P \in \overline{\mathcal{P}}$ be such that

$$\bar{g}(P) + Pz \leq Tz + \epsilon e.$$

By adding this to the relation $Ty \leq \bar{g}(P) + Py$, we obtain

$$Ty - Tz \leq P(y - z) + \epsilon e.$$

Since P is a stochastic matrix, we have $P(y - z) \leq H(y - z)$, so that

$$Ty - Tz \leq H(y - z) + \epsilon e,$$

from which

$$H(Ty - Tz) \leq H(y - z) + \epsilon e.$$

Since ϵ can be taken arbitrarily small, the desired relation follows.

(b) We have, using part (a),

$$\|Ty - Tz\| = H(Ty - Tz) + H(Tz - Ty) \leq H(y - z) + H(z - y) = \|y - z\|.$$

(c) We have, using the triangle inequality,

$$\|\beta y - \gamma z\| = \|\beta(y - z) + (\beta - \gamma)z\| \leq |\beta| \|y - z\| + |\beta - \gamma| \|z\|,$$

while from part (b),

$$\|T(\beta y) - T(\gamma z)\| \leq \|\beta y - \gamma z\|.$$

Combining these two relations, we obtain the desired result. Q.E.D.

We recall the value iteration method: $h^{k+1} = Th^k$ with $h^0 = 0$. The next lemma considers an approximate (but also more general) version of this method, where h^k is scaled by a factor $\gamma_k \in [0, 1]$ before being used in the next iteration. The lemma shows that the generated sequence $\{\|h^k\|\}$ is bounded. This fact will be used to show existence of a solution to Bellman's equation and will also be the first step for constructing a convergent relative value iteration method.

Lemma 4.6.2: Let Assumption 4.6.1 hold, and let $\{\gamma_k\}$ be a non-decreasing sequence with $\gamma_k \in [0, 1]$ for all k . Consider the sequence $\{h^k\}$ generated by

$$h^{k+1} = T(\gamma_k h^k) = \min_{P \in \overline{\mathcal{P}}} [\bar{g}(P) + \gamma_k P h^k],$$

where $h^0 = 0$. Then $\{\|h^k\|\}$ is bounded.

Proof: Without loss of generality, we assume that

$$0 \leq \bar{g}_i(p_i) \leq \beta, \quad p_i \in \overline{\mathcal{P}}_i,$$

for some scalar β [otherwise, using also the boundedness of $\bar{g}_i(p_i)$ over $p_i \in \overline{\mathcal{P}}_i$, we can add a common sufficiently large constant to every component $\bar{g}_i(p_i)$, which does not change $\|h^k\|$]. Then, it can be seen that $\{h^k\}$ is nondecreasing,

$$h^k \leq h^{k+1}, \quad k = 0, 1, \dots \quad (4.124)$$

For any pair of states i, j , let μ_{ij} be a stationary policy (for the original nonrandomized average cost problem) such that j is accessible from i under μ_{ij} , as per Assumption 6.4.1(a). Let also $P_{\mu_{ij}}$ be the corresponding transition probability matrix. Then in the Markov chain corresponding to the transition probability matrix $Q \in \overline{\mathcal{P}}$ given by

$$Q = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n P_{\mu_{ij}},$$

every state is accessible from every other state, so the states form a single recurrent class. For each pair of states i, j , we denote by τ_{ij} the expected number of transitions to reach j from i when Q is used as a stationary policy (note that τ_{ij} is finite since the states form a single recurrent class). Then, by considering the first transition, we have

$$\tau_{ij} = 1 + \sum_{l \neq j} q_{il} \tau_{lj}, \quad i, j = 1, \dots, n, \quad i \neq j, \quad (4.125)$$

where q_{il} denote the corresponding components of Q .

Let us now show by induction that

$$h_i^k \leq \beta \tau_{ij} + h_j^k, \quad i, j = 1, \dots, n, \quad i \neq j, \quad k = 0, 1, \dots \quad (4.126)$$

Indeed this relation holds for $k = 0$. Assuming it holds for a given k , using the facts $\bar{g}(P) \leq \beta e$, $\gamma_{k+1} \leq 1$, and $h^k \geq 0$, we have

$$h^{k+1} \leq \bar{g}(Q) + \gamma_{k+1} Q h^k \leq \beta e + Q h^k,$$

so for all i and j ,

$$h_i^{k+1} \leq \beta + \sum_{l=1}^n q_{il} h_l^k = \beta + \sum_{l \neq j} q_{il} h_l^k + q_{ij} h_j^k.$$

Using Eq. (4.126) and then Eq. (4.125) in the above relation, we obtain

$$\begin{aligned} h_i^{k+1} &\leq \beta + \sum_{l \neq j} q_{il} (\beta \tau_{lj} + h_j^k) + q_{ij} h_j^k \\ &= \beta \left(1 + \sum_{l \neq j} q_{il} \tau_{lj} \right) + \sum_{l \neq j} q_{il} h_j^k + q_{ij} h_j^k \\ &= \beta \tau_{ij} + h_j^k. \end{aligned}$$

Since $\{h^k\}$ is nondecreasing [cf. Eq. (4.124)], it follows that

$$h_i^{k+1} \leq \beta \tau_{ij} + h_j^{k+1},$$

thereby completing the induction.

From Eq. (4.126), we see that

$$\|h^k\| \leq \max\{\beta \tau_{ij} \mid i, j = 1, \dots, n, i \neq j\}.$$

Thus $\{\|h^k\|\}$ is bounded. **Q.E.D.**

We are now ready to show that Bellman's equation, $\lambda e + h = Th$ has a solution (or equivalently $\|Th - h\| = 0$). Consider the (approximate value iteration) sequence, generated by

$$h^{k+1} = T(\gamma_k h^k) = \min_{P \in \bar{P}} [\bar{g}(P) + \gamma_k P h^k], \quad k = 0, 1, \dots,$$

where $h^0 = 0$ and $\{\gamma_k\} \subset [0, 1]$ is a nondecreasing sequence to be determined later. Consider also its relative version

$$\tilde{h}^k = h^k - L(h^k)e, \quad k = 0, 1, \dots, \quad (4.127)$$

where $L(h^k) = \min_{i=1,\dots,n} h_i^k$. Then

$$L(\tilde{h}^k) = 0, \quad \|\tilde{h}^k\| = H(\tilde{h}^k) = \max_{i=1,\dots,n} \tilde{h}_i^k,$$

and

$$0 \leq \tilde{h}_i^k \leq \|\tilde{h}^k\| = \|h^k\|, \quad i = 1, \dots, n.$$

Since by Lemma 4.6.2, the sequence $\{\|h^k\|\}$ is bounded, it follows that the sequence $\{\tilde{h}^k\}$ is also bounded. We will prove in the following proposition that for a suitable sequence $\{\gamma_k\}$, we have $\|Th - h\| = 0$ for any limit point h of $\{\tilde{h}^k\}$, thereby showing existence of a solution of Bellman's equation.

Proposition 4.6.2: Let Assumption 4.6.1 hold. Then there exists a vector $h \in \mathbb{R}^n$ such that $\|Th - h\| = 0$, so that for some scalar λ , we have

$$\lambda e + h = Th.$$

Furthermore, if $\bar{\mu}$ satisfies $T_{\bar{\mu}}h = Th$, i.e., $\bar{\mu}(i) \in \overline{\mathcal{P}}_i$ attains the minimum in the expression

$$\min_{p_i \in \overline{\mathcal{P}}_i} [\bar{g}_i(p_i) + p_i' h], \quad i = 1, \dots, n,$$

then $\bar{\mu}$ is optimal for the randomized average cost problem.

Proof: Let h be a limit point of $\{\tilde{h}^k\}$ (as per the discussion preceding the proposition), and let $\{\tilde{h}^{k_t}\}$ be the corresponding convergent subsequence. By using Lemma 4.6.1(c), we obtain

$$\begin{aligned} \|h^{k+1} - h^k\| &= \|T(\gamma_k h^k) - T(\gamma_{k-1} h^{k-1})\| \\ &\leq \gamma_k \|h^k - h^{k-1}\| + |\gamma_k - \gamma_{k-1}| \|h^{k-1}\| \end{aligned}$$

and finally

$$\|h^{k+1} - h^k\| \leq \gamma_k \|h^k - h^{k-1}\| + B |\gamma_k - \gamma_{k-1}|, \quad (4.128)$$

where B is an upper bound to $\{\|h^k\|\}$. Let us choose

$$\gamma_0 = \frac{1}{2}, \quad \gamma_k = \underbrace{\frac{k}{k+1}}_{\sim}, \quad k = 1, 2, \dots$$

Then a straightforward induction using Eq. (4.128) and the fact $\|h^1 - h^0\| = \|h^1\| \leq B$ shows that for all k ,

$$\|h^{k+1} - h^k\| \leq \frac{B}{k+1} \left(1 + \frac{1}{2} + \dots + \frac{1}{k} \right),$$

and hence $\lim_{k \rightarrow \infty} \|h^{k+1} - h^k\| = 0$. It follows that

$$\lim_{t \rightarrow \infty} \|\tilde{h}^{k_t+1} - \tilde{h}^{k_t}\| = 0. \quad (4.129)$$

We have

$$\|Th - h\| \leq \|Th - \tilde{h}^{k_t+1}\| + \|\tilde{h}^{k_t+1} - \tilde{h}^{k_t}\| + \|\tilde{h}^{k_t} - h\|.$$

Also, by using Lemma 4.6.1(c) and the definition of \tilde{h}^k [cf. Eq. (4.127)], we obtain

$$\begin{aligned} \|Th - \tilde{h}^{k_t+1}\| &= \|Th - h^{k_t+1}\| \\ &= \left\| Th - T \left(\frac{k_t}{k_t+1} h^{k_t} \right) \right\| \\ &\leq \|h - h^{k_t}\| + \frac{1}{k_t+1} \|h^{k_t}\| \\ &= \|h - \tilde{h}^{k_t}\| + \frac{1}{k_t+1} \|\tilde{h}^{k_t}\|. \end{aligned}$$

Combining the last two equations, it follows that

$$\|Th - h\| \leq \|\tilde{h}^{k_t+1} - \tilde{h}^{k_t}\| + 2\|\tilde{h}^{k_t} - h\| + \frac{B}{k_t+1},$$

so by taking the limit as $t \rightarrow \infty$ and by using Eq. (4.129), we obtain $\|Th - h\| = 0$. The optimality of $\bar{\mu}$, when $\bar{\mu}$ satisfies $T_{\bar{\mu}}h = Th$, follows from Prop. 4.6.1. Q.E.D.

Note that a byproduct of the analysis is the validity of the relative value iteration method

$$\tilde{h}^k = h^k - L(h^k)e,$$

where $\{h^k\}$ is generated by

$$\gamma_0 = \frac{1}{2}, \quad h^0 = 0, \quad h^k = T(\gamma_{k-1} h^{k-1}), \quad \gamma_k = \frac{k}{k+1}, \quad k = 1, 2, \dots$$

The proof of Prop. 4.6.2 shows that for every limit point \tilde{h} of $\{\tilde{h}^k\}$, all the components of $Th - \tilde{h}$ are equal to a constant λ , and furthermore \tilde{h} and λ satisfy Bellman's equation. In fact it can be shown that it is not necessary to restrict oneself to the initial condition $h^0 = 0$; any initial condition can be used. The reason is that if $\{h^k\}$ and $\{\bar{h}^k\}$ are sequences generated starting from initial conditions h^0 and \bar{h}^0 , respectively, by viewing $h^0 - \bar{h}^0$ as a change in terminal cost in a k -stage optimization with discount factors $\gamma_0, \gamma_1, \dots, \gamma_{k-1}$, it can be verified that

$$\max_{i=1, \dots, n} |h^k(i) - \bar{h}^k(i)| \leq \gamma_0 \gamma_1 \cdots \gamma_{k-1} \max_{i=1, \dots, n} |h^0(i) - \bar{h}^0(i)|.$$

Since $\gamma_0 \gamma_1 \cdots \gamma_{k-1} = 1/(2k) \rightarrow 0$, it follows that the choice of initial condition is inconsequential.

Let us now show by example that the accessibility Assumption 4.6.1(a) cannot be replaced by the weaker WA condition of Section 4.2.

Example 4.6.8 (Blackmailer's Dilemma Revisited)

Consider the blackmailer Example 4.6.1, where we noted that Bellman's equation,

$$\lambda + h(1) = \min_{u \in (0,1]} [-u + (1 - u^2)h(1)], \quad \lambda + h(0) = h(0),$$

has no solution, even though the optimal average cost is equal to 0 for the two initial states. This equation can be placed in the form of this section by using the change of variables $p = 1 - u^2$:

$$\lambda + h(1) = \min_{p \in [0,1)} [-\sqrt{1-p} + ph(1)], \quad \lambda + h(0) = h(0).$$

We note that state 1 is not accessible from the termination state t , so Assumption 4.6.1(a) is violated. Yet state 1 is transient under all stationary policies, so the WA condition of Definition 4.2.2 is satisfied. Hence the accessibility Assumption 4.6.1(a) cannot be weakened and be replaced by the WA condition. This would also be true if the constraint $p \in [0, 1)$ were to be changed to $p \in [0, 1]$, a compact constraint set.

Note that the analysis of this section has not relied on convexity properties of $\bar{g}_i(p_i)$ and $\bar{\mathcal{P}}_i$. The critical assumptions have been:

- (1) The existence of a matrix $Q \in \bar{\mathcal{P}}$ that corresponds to a Markov chain with states that form a single recurrent class (cf. the proof of Lemma 4.6.2).
- (2) The boundedness of $\bar{g}(P)$ over $P \in \bar{\mathcal{P}}$.

Thus our results apply to other average cost problems with special structure where the preceding two conditions are satisfied.

4.6.3 Countable States – Vanishing Discount Approach

Consider an average cost problem where the state space is

$$S = \{0, 1, 2, \dots\}.$$

The transition probabilities are denoted $p_{ij}(u)$ for $i, j \in S$ and $u \in U(i)$, and the expected cost per stage is denoted by $g(i, u)$, $i \in S$, $u \in U(i)$. We introduce the α -discounted version of the problem and the associated Bellman equation

$$J_\alpha^*(i) = \min_{u \in U(i)} \left[g(i, u) + \alpha \sum_{j=0}^{\infty} p_{ij}(u) J_\alpha^*(j) \right].$$

Subtracting $\alpha J_\alpha^*(0)$ from both sides of this equation, and introducing the function $h_\alpha(\cdot)$ given by

$$h_\alpha(i) = J_\alpha^*(i) - J_\alpha^*(0),$$

we obtain

$$(1 - \alpha)J_\alpha^*(0) + h_\alpha(i) = \min_{u \in U(i)} \left[g(i, u) + \alpha \sum_{j=0}^{\infty} p_{ij}(u)h_\alpha(j) \right]. \quad (4.130)$$

We may view $h_\alpha(i)$ as a relative cost of state i (relative to state 0) for the α -discounted problem. Note here that there is nothing special about the reference state 0; it may be replaced by any other state. The preceding equation resembles Bellman's equation for the average cost problem. It is the starting point for a line of analysis called the *vanishing discount factor approach*.

If we take the limit of both sides as $\alpha \rightarrow 1$ in Eq. (4.130), and assume that the limits of all terms exist, we obtain Bellman's equation for the average cost problem with

$$\lambda = \lim_{\alpha \rightarrow 1} (1 - \alpha)J_\alpha^*(0), \quad h(i) = \lim_{\alpha \rightarrow 1} h_\alpha(i).$$

The following proposition shows that this is possible, provided that $h_\alpha(i)$ is uniformly bounded over i and α . While in the proposition we assume for mathematical convenience that $U(i)$ is finite, this assumption can be generalized (see the literature cited at the end of the chapter).

Proposition 4.6.3: Let the control constraint set $U(i)$ be finite for all i , and let the cost per stage $|g(i, u)|$ be bounded over i and u . Assume that $|h_\alpha(i)|$ is also bounded over i and $\alpha \in (0, 1)$. Then there exists a scalar λ and a bounded function h that solve Bellman's equation, i.e., for all i ,

$$\lambda + h(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=0}^{\infty} p_{ij}(u)h(j) \right]. \quad (4.131)$$

Furthermore,

$$\lambda = \min_{\pi} J_{\pi}(i) = J^*(i),$$

and if $\mu^*(i)$ attains the minimum in Eq. (4.131) for each i , the stationary policy μ^* is optimal.

Proof: Let $\{\alpha_k\}$ be a sequence such that $\alpha_k \rightarrow 1$. Using the boundedness of h_α , we can find a subsequence, also denoted $\{\alpha_k\}$ for simplicity, such that $\lim_{k \rightarrow \infty} h_{\alpha_k}(i) = h(i)$ for all i .[†] Since $(1 - \alpha_k)J_{\alpha_k}^*(0)$ is also bounded, in view of Eq. (4.130) and the boundedness of $|g(i, u)|$, it follows that there is a subsequence of $\{\alpha_k\}$, say $\{\alpha_{k_t}\}$, such that for some λ

$$\lim_{t \rightarrow \infty} (1 - \alpha_{k_t}) J_{\alpha_{k_t}}^*(0) = \lambda.$$

Taking the limit in Eq. (4.130) along the subsequence $\{\alpha_{k_t}\}$, and interchanging limit and minimization [using the finiteness of $U(i)$], we obtain Eq. (4.131). The last statement of the proposition follows from the sufficient condition of Prop. 4.6.1. **Q.E.D.**

As an illustration of the proposition, let us prove a result from Ross [Ros83a]. This result uses a recurrence condition, whereby a special state (by convention state 0) is reachable from all other states within bounded expected time (think of queueing/storage problems and a state 0 that corresponds to an empty system). The result has descended from papers by Derman [Der62], and Derman and Veinott [DeV67], which provided motivation for much subsequent work.

Proposition 4.6.4: Let the control constraint set $U(i)$ be finite for all i , and let the cost per stage $|g(i, u)|$ be bounded over i and u . For each initial state i , let τ_i be the first time the state reaches 0,

$$\tau_i = \min\{t \geq 1 \mid x_t = 0\},$$

and assume that for some scalar T and all stationary policies μ , that are α -discount optimal for some $\alpha \in (0, 1)$, we have

$$E\{\tau_i \mid \mu\} \leq T.$$

Then $|h_\alpha(i)|$ is bounded over i and $\alpha \in (0, 1)$, and the conclusions of Prop. 4.6.3 hold.

Proof: By adding a suitable constant to $g(i, u)$ if necessary, we may assume without loss of generality that $g(i, u)$ is nonnegative. Let B be a scalar such that for all (i, u) , we have

$$0 \leq g(i, u) \leq B.$$

[†] There is a standard procedure for doing this: sequentially, starting with $i = 0$, select for each i a subsequence $\{\alpha_k \mid k \in \mathcal{K}_i\}$, such that $\mathcal{K}_i \supset \mathcal{K}_{i+1}$ and $\{h_{\alpha_k}\}_{k \in \mathcal{K}_i}$ converges to some scalar $h(i)$. Then choose a subsequence $\{\alpha_{k_i}\}$ such that for all i , $k_i \in \mathcal{K}_i$ and $k_i < k_{i+1}$.

Fix $\alpha \in (0, 1)$ and let μ_α be optimal for the α -discounted problem. For all i , we have

$$\begin{aligned} J_\alpha^*(i) &= \lim_{N \rightarrow \infty} E \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_\alpha(x_k)) \mid x_0 = i \right\} \\ &= E \left\{ \sum_{k=0}^{\tau_i-1} \alpha^k g(x_k, \mu_\alpha(x_k)) \mid x_0 = i \right\} \\ &\quad + \lim_{N \rightarrow \infty} E \left\{ \sum_{k=0}^{N-1} \alpha^{\tau_i+k} g(x_k, \mu_\alpha(x_k)) \mid x_0 = 0 \right\}. \end{aligned} \quad (4.132)$$

From this relation, we obtain

$$J_\alpha^*(i) \leq B E \left\{ \sum_{k=0}^{\tau_i-1} \alpha^k \right\} + E \{ \alpha^{\tau_i} \} J_\alpha^*(0) \leq BT + J_\alpha^*(0). \quad (4.133)$$

Also, from Eq. (4.132) and the nonnegativity of g , we have

$$J_\alpha^*(i) \geq E \{ \alpha^{\tau_i} \} J_\alpha^*(0) \geq \alpha^{E\{\tau_i\}} J_\alpha^*(0) \geq \alpha^T J_\alpha^*(0), \quad (4.134)$$

where the second inequality is a consequence of the convexity of α^{τ_i} , viewed as a function of τ_i , and Jensen's inequality [$E\{f(Y)\} \geq f(E\{Y\})$ for any convex function f and random variable Y]. From Eq. (4.134), we have

$$J_\alpha^*(0) - J_\alpha^*(i) \leq (1 - \alpha^T) J_\alpha^*(0) \leq (1 - \alpha^T) \frac{B}{1 - \alpha} = (1 + \alpha + \dots + \alpha^{T-1}) B \leq BT. \quad (4.135)$$

From Eqs. (4.133) and (4.135), we see that $|J_\alpha^*(0) - J_\alpha^*(i)|$ is bounded over i and $\alpha \in (0, 1)$, and the result now follows from Prop. 4.6.3. Q.E.D.

The conditions of the preceding proposition have been generalized considerably; see Federgruen, Hordijk, and Tijms [FHT79], Thomas [Tho80], Whittle [Whi82] (Chapter 34), Schal [Sch93a], and the survey by Arapostathis et al. [ABF93].

4.6.4 Countable States – Contraction Approach

In this section, we discuss an alternative analysis of the countable state average cost problem of the preceding section. We follow the contraction mapping approach of Sections 1.4.3 and 2.5. While there are differences in scope and methodology with the vanishing discount factor approach of the preceding section, there is also substantial overlap, because both approaches use recurrence-type assumptions in several contexts.

We introduce a positive sequence $v = \{v_0, v_1, \dots\}$, such that

$$\inf_{i=0,1,\dots} v_i > 0,$$

and the weighted sup-norm

$$\|J\| = \max_{i=0,1,\dots} \frac{|J(i)|}{v_i}$$

in the space $B(S)$ of sequences $\{J(0), J(1), \dots\}$ such that $\|J\| < \infty$.

We assume that state 0 is special in that the system has a “tendency” to return to it under all policies. In particular, for any policy π , we denote

C_π : expected cost starting from state 0 up to the first return to 0,

N_π : expected number of stages to return to state 0 starting from 0,
and we assume the following.

Assumption 4.6.2: For every policy π , C_π and N_π are finite. Furthermore, N_π is uniformly bounded over π , i.e., for some $\bar{N} > 0$, we have $N_\pi \leq \bar{N}$ for all π .

This assumption and the ones that follow are among the simplest of a large variety of similar conditions that ensure the validity of Bellman’s equation for countable state problems. These conditions share a characteristic requirement that the system returns to a special state or a special finite subset of states infinitely often, under all or a suitable subset of policies. The corresponding line of analysis is often referred to as the *recurrence approach*.

To derive Bellman’s equation and associated results, we will follow a line of analysis similar to the one of Section 7.4 of Vol. I, which essentially converts the average cost problem to a stochastic shortest path problem. We will use instead the (countable state) stochastic shortest path results of Section 2.5. The following assumption parallels Assumption 2.5.1 in Section 2.5.

Assumption 4.6.3:

(a) The sequence $G = \{G(0), G(1), \dots\}$, where

$$G(i) = \max_{u \in U(i)} \overbrace{|g(i, u)|}, \quad i = 0, 1, \dots,$$

belongs to $B(S)$.

(b) The sequence $V = \{V(0), V(1), \dots\} \in B(S)$, where

$$V(i) = \max_{u \in U(i)} \sum_{j=0}^{\infty} p_{ij}(u) v_j, \quad i = 0, 1, \dots,$$

belongs to $B(S)$.

(c) There is a scalar $\rho \in (0, 1)$ and an integer $m \geq 1$ such that for all π and $i = 0, 1, \dots$, we have

$$\frac{\sum_{j=1}^{\infty} P(x_m = j \mid x_0 = i, \pi) v_j}{v_i} \leq \rho.$$

Note that the summation in Assumption 4.6.3(c) does not include state 0, so this assumption quantifies the tendency to return to 0 uniformly from all other states and under all policies, similar to Assumption 2.5.1(c).

For any scalar λ , we consider the mappings T_μ^λ and T^λ defined by

$$(T_\mu^\lambda J)(i) = g(i, \mu(i)) - \lambda + \sum_{j=0}^{\infty} p_{ij}(\mu(i)) J(j),$$

$$(T^\lambda J)(i) = \min_{u \in U(i)} \left[g(i, u) - \lambda + \sum_{j=0}^{\infty} p_{ij}(u) J(j) \right].$$

Similar to the analysis of Section 1.4.3, under Assumption 4.6.3, T_μ^λ and T^λ map $B(S)$ into $B(S)$ for all $J \in B(S)$ and $\lambda \in \mathbb{R}$ (cf. Prop. 1.4.2).

The following lemma prepares the ground for application of the sufficient condition of Prop. 4.6.1, which will be used to prove our main result.

Lemma 4.6.3: Let Assumption 4.6.3 hold. Then for all $h \in B(S)$, policies π , and states i

$$\lim_{N \rightarrow \infty} \frac{1}{N} E\{h(x_N) \mid x_0 = i, \pi\} = 0.$$

Proof: We will show that $E\{h(x_N) \mid x_0 = i, \pi\}$ is bounded over N . Without loss of generality, we assume that h is nonnegative (otherwise, we replace h with $|h|$). We first note that Assumption 4.6.3(c) implies that

there exists an integer $m \geq 1$ such that for every policy π and state i , we have

$$\begin{aligned} E\{v_{x_m} | x_0 = i, \pi\} &= \sum_{j=1}^{\infty} P(x_m = j | x_0 = i, \pi) v_j \\ &\quad + P(x_m = 0 | x_0 = i, \pi) v_0 \\ &\leq \rho v_i + v_0. \end{aligned} \tag{4.136}$$

Furthermore, for all $k \geq 0$,

$$\begin{aligned} E\{h(x_{km}) | x_0 = i, \pi\} &\leq \|h\| E\{v_{x_{km}} | x_0 = i, \pi\} \\ &= \|h\| E\left\{E\{v_{x_{km}} | x_{(k-1)m}, \pi\} | x_0 = i, \pi\right\} \\ &\leq \|h\| E\left\{\rho v_{x_{(k-1)m}} + v_0 | x_0 = i, \pi\right\} \\ &= \|h\| \left(\rho E\left\{v_{x_{(k-1)m}} | x_0 = i, \pi\right\} + v_0\right), \end{aligned} \tag{4.137}$$

where the second inequality follows by applying Eq. (4.136) to the inner conditional expectation term. Applying the same argument and iteratively expanding the right-hand side of inequality (4.137), we obtain

$$E\{h(x_{km}) | x_0 = i, \pi\} \leq \|h\| \left(\rho^k v_i + \frac{v_0}{1-\rho}\right). \tag{4.138}$$

Let $j \in \{0, 1, \dots, m-1\}$. By applying Eq. (4.138), we see that

$$\begin{aligned} E\{h(x_{km+j}) | x_0 = i, \pi\} &= E\left\{E\{h(x_{km+j}) | x_j, \pi\} | x_0 = i, \pi\right\} \\ &\leq \|h\| \left(\rho^k E\{v(x_j) | x_0 = i, \pi\} + \frac{v_0}{1-\rho}\right). \end{aligned} \tag{4.139}$$

By Assumption 4.6.3(b), we have

$$E\{v_{x_1} | x_0 = i, \pi\} \leq V(i) \leq \|V\| v_i, \tag{4.140}$$

and similarly, for $j = 2, \dots, m-1$,

$$\begin{aligned} E\{v_{x_j} | x_0 = i, \pi\} &= E\left\{E\{v_{x_j} | x_{j-1}, \pi\} | x_0 = i, \pi\right\} \\ &\leq \|V\| E\{v_{x_{j-1}} | x_0 = i, \pi\} \\ &\leq \|V\|^j v_i. \end{aligned} \tag{4.141}$$

Combining Eqs. (4.139)-(4.141), we obtain for all $j = 0, 1, \dots, m-1$ and $k \geq 0$,

$$E\{h(x_{km+j}) | x_0 = i, \pi\} \leq \|h\| \left(\rho^k \|V\|^j v_i + \frac{v_0}{1-\rho}\right),$$

so $E\{h(x_N) | x_0 = i, \pi\}$ is bounded over N . Q.E.D.

The following proposition provides our main result.

Proposition 4.6.5: Let Assumptions 4.6.2 and 4.6.3 hold. Then the optimal average cost, denoted λ^* , is the same for all initial states and together with some sequence $h^* = \{h^*(0), h^*(1), \dots\}$ satisfies Bellman's equation

$$\lambda^* + h^*(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=0}^{\infty} p_{ij}(u) h^*(j) \right], \quad i = 0, 1, \dots \quad (4.142)$$

Furthermore, if $\mu(i)$ attains the minimum in the above equation for all i , the stationary policy μ is optimal.

Proof: Let us denote

$$\tilde{\lambda} = \min_{\pi} \frac{C_{\pi}}{N_{\pi}},$$

where C_{π} and N_{π} have been defined earlier, and the minimum is taken over the set of all admissible policies. Consider the associated stochastic shortest path problem where the expected stage cost incurred at state i is $g(i, u) - \tilde{\lambda}$. Then Assumption 4.6.3(a),(b), and the condition $\inf_{i=0,1,\dots} v_i > 0$, can be used to show that the sequence

$$\left\{ \max_{u \in U(i)} |g(i, u) - \tilde{\lambda}| \mid i = 0, 1, \dots \right\}$$

belongs to $B(S)$, and together with Assumption 4.6.3(c) and Lemma 4.6.3, ensure that the stochastic shortest path analysis of Section 2.5 applies. It follows that the corresponding costs $h^*(0), h^*(1), \dots$ solve uniquely the corresponding Bellman's equation

$$h^*(i) = \min_{u \in U(i)} \left[g(i, u) - \tilde{\lambda} + \sum_{j=1}^{\infty} p_{ij}(u) h^*(j) \right], \quad i = 0, 1, \dots, \quad (4.143)$$

since the transition probability from i to 0 is zero in the associated stochastic shortest path problem (compare with the construction of Fig. 7.4.1 and the proof of Prop. 7.4.1 of Vol. I).

Consider now $h^*(0)$, which is the optimal cost to return to state 0 starting from state 0, when the cost per stage is $g(i, u) - \tilde{\lambda}$. Since $C_{\pi} - N_{\pi}\tilde{\lambda}$ is the corresponding cost of policy π to return to 0 starting from 0, we have

$$h^*(0) = \min_{\pi} [C_{\pi} - N_{\pi}\tilde{\lambda}].$$

Since $N_\pi \leq \bar{N}$ for all π (Assumption 4.6.2), and from the definition $\tilde{\lambda} = \min_{\pi} C_\pi/N_\pi$ we have $C_\pi - N_\pi \tilde{\lambda} \geq 0$ for all π , it follows that.

$$0 \leq h^*(0) = \min_{\pi} N_\pi \left[\frac{C_\pi}{N_\pi} - \tilde{\lambda} \right] \leq \bar{N} \min_{\pi} \left[\frac{C_\pi}{N_\pi} - \tilde{\lambda} \right] = 0.$$

Thus $h^*(0) = 0$, and Eq. (4.143) is written as

$$\tilde{\lambda} + h^*(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=0}^{\infty} p_{ij}(u) h^*(j) \right], \quad i = 0, 1, \dots$$

This relation and Lemma 4.6.3 imply the assumptions of Prop. 4.6.1, so from the conclusion of that proposition, we see that $\tilde{\lambda}$ is equal to the optimal average cost for all initial states, and also that the conclusion regarding stationary policies holds. **Q.E.D.**

4.6.5 Linear Systems with Quadratic Cost

Consider the linear-quadratic problem involving the system

$$x_{k+1} = Ax_k + Bu_k + w_k, \quad k = 0, 1, \dots,$$

and the cost function

$$J_\pi(x_0) = \lim_{N \rightarrow \infty} \frac{1}{N} \underset{\substack{w_k \\ k=0,1,\dots}}{E} \left\{ \sum_{k=0}^{N-1} (x'_k Q x_k + \mu_k(x_k)' R \mu_k(x_k)) \right\}.$$

We make the same assumptions as in Section 3.2, i.e., Q is positive semidefinite symmetric, R is positive definite symmetric, and w_k are independent, and have zero mean and finite second moments. We also assume that the pair (A, B) is controllable and that the pair (A, C) , where $Q = C'C$, is observable. Under these assumptions, it was shown in Section 4.1 of Vol. I that the Riccati equation

$$K_0 = 0,$$

$$K_{k+1} = A' (K_k - K_k B (B' K_k B + R)^{-1} B' K_k) A + Q$$

yields in the limit a matrix K ,

$$K = \lim_{k \rightarrow \infty} K_k,$$

which is the unique solution of the equation

$$K = A' (K - K B (B' K B + R)^{-1} B' K) A + Q$$

within the class of positive semidefinite symmetric matrices.

The optimal value of the N -stage costs

$$\frac{1}{N} \underset{\substack{E \\ w_k}}{\underset{k=0,1,\dots,N-1}{\sum}} \left\{ \sum_{k=0}^{N-1} (x'_k Q x_k + u'_k R u_k) \right\}$$

has been derived earlier and was seen to be equal to

$$\frac{1}{N} \left(x'_0 K_N x_0 + \sum_{k=0}^{N-1} E\{w' K_k w\} \right).$$

Since $K = \lim_{k \rightarrow \infty} K_k$ and

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} E\{w' K_k w\} = E\{w' K w\},$$

we see that the optimal N -stage costs tend to

$$\lambda = E\{w' K w\}$$

as $N \rightarrow \infty$. In addition, the N -stage optimal policy in its initial stages tends to the stationary policy

$$\mu^*(x) = -(B' K B + R)^{-1} B' K A x. \quad (4.144)$$

Furthermore, a simple calculation shows that, by the definition of λ , K , and $\mu^*(x)$, we have

$$\lambda + x' K x = \min_u E\{x' Q x + u' R u + (Ax + Bu + w)' K (Ax + Bu + w)\},$$

while the minimum in the right-hand side of this equation is attained at $u^* = \mu^*(x)$ as given by Eq. (4.144).

By repeating the proof of Prop. 4.6.1, we obtain

$$\begin{aligned} \lambda &\leq \frac{1}{N} E\{x'_N K x_N \mid x_0, \pi\} \\ &= \frac{1}{N} x'_0 K x_0 + \frac{1}{N} E \left\{ \sum_{k=0}^{N-1} (x'_k Q x_k + u'_k R u_k) \mid x_0, \pi \right\}, \end{aligned}$$

with equality if $\pi = \{\mu^*, \mu^*, \dots\}$. Hence, if π is such that $E\{x'_N K x_N \mid x_0, \pi\}$ is uniformly bounded over N , we have, by taking the limit as $N \rightarrow \infty$ in the preceding relation,

$$\lambda \leq J_\pi(x), \quad x \in \Re^n,$$

with equality if $\pi = \{\mu^*, \mu^*, \dots\}$. Thus the linear stationary policy given by Eq. (4.144) is optimal over all policies π with $E\{x'_N K x_N \mid x_0, \pi\}$ bounded uniformly over N .

4.7 NOTES, SOURCES, AND EXERCISES

Several authors have made early contributions to the average cost problem (Gillette [Gil57], Howard [How60], Brown [Bro65], Veinott [Vei66], [Vei69], Schweitzer [Sch68], Derman [Der70], Ross [Ros70]), most notably Blackwell [Bla62]. An extensive survey containing many references is given by Arapostathis et al. [ABF93]. The edited volume by Feinberg and Schwartz [FeS02] provides several surveys of average cost topics that we have not covered.

The weak accessibility condition was introduced by Platzman [Pla77a], as a modification of the accessibility condition of Section 4.6.2. The latter was introduced by Bather [Bat73], and was used to develop the analysis and results of Section 4.6.2.

The relative value iteration method of Section 4.3 is due to White [Whi63]. Its damped version that involves a stepsize parameter τ [cf. Eq. (4.72)] is due to Schweitzer [Sch71]. The error bounds of Prop. 4.3.3 are due to Odoni ([Odo69]). It is possible to analyze the relative value iteration method in terms of contraction mappings; see Federgruen, Schweitzer and Tijms [FST78], and Puterman [Put94], Section 8.5. In particular, under the assumption of Prop. 4.3.2, one can show that the mapping defining the method is an m -step contraction involving the notion of span seminorm (as formalized in Section 6.4.2). In fact, the proof of Prop. 4.3.2 is implicitly based on this line of analysis. Convergence under slightly weaker conditions than those given here is shown by Platzman [Pla77b]. The error bounds of Exercise 4.9 are due to Varaiya ([Var78]), who used them to construct a differential form of the value iteration method. Discrete-time versions of Varaiya's method are given by Poppyack, Brown, and White [PBW79]. The contracting value iteration method is due to Bertsekas [Ber98], to which we refer for a discussion of convergence and computational results. The value iteration method for multi-chain problems has been analyzed by Schweitzer [Sch71], Schweitzer and Federgruen [ScF77], [ScF78], and Federgruen, Schweitzer, and Tijms [FST78].

The policy iteration algorithm was introduced by Howard [How60], who showed finite termination under the assumption that all policies give rise to an irreducible Markov chain. Howard [How60] also considered the multi-chain case, but his proposed algorithm is flawed and may not terminate. Blackwell [Bla62] gave a convergent version, and Veinott [Vei66] gave the version presented here. For various extensions of policy iteration to infinite space problems, see Hernandez-Lerma and Lasserre [HeL97], Meyn [Mey97], Golubin [Gol03], and Patek [Pat04].

The linear programming approach was formulated by De Ghellinck [DeG60] and Manne [Man60] for single-chain problems. Denardo and Fox [DeF68], and Derman [Der70] considered the multi-chain case, and Kallenberg [Kal83] provided an extensive treatment. Derman [Der70] applied the linear programming approach to constrained average cost problems. See

Puterman [Put94] for an alternative textbook development, and Kallenberg [Kal91a], [Kal94b] for a survey. Borkar [Bor88], [Bor89], [Bor91] introduced a convex analytic approach for unconstrained and constrained average cost problems, which naturally extends the linear programming approach, and applies to problems with possibly infinite state and control spaces; see also the discussion in Arapostathis et al. [ABF93].

Since average cost measures only the asymptotic behavior of a policy, two optimal policies may have very different transient performance. For example, any unichain policy created by modifying the transition probabilities of the transient states of another unichain policy, has the same average cost as the original, yet it may have significantly different finite horizon cost when starting from the transient states. The concept of m -discount optimality is useful for differentiating between two policies with different transient performance, and is related to Blackwell optimality. In particular, given a finite-spaces problem, for an integer $m \geq -1$, a policy π^* is m -discount optimal if its discounted cost satisfies

$$\limsup_{\alpha \rightarrow 1} (1 - \alpha)^{-m} (J_{\alpha, \pi^*}(i) - J_{\alpha, \pi}(i)) \leq 0,$$

for all states i and policies π . This definition can be interpreted in the light of the Laurent series expansion, given in Section 4.1. In particular, m -discount optimality addresses the optimality of the terms of the Laurent series up to power m and no higher, and it follows that an $(m+1)$ -discount optimal policy is also k -discount optimal for all $k = -1, 0, \dots, m$. Note that a (-1) -discount optimal policy is average cost optimal (as we have used the term), while a 0-discount optimal stationary policy is average cost optimal and also minimizes the bias over all stationary policies that are average cost optimal. It can be shown that a Blackwell optimal policy is m -discount optimal for every m . There are coupled sets of $m+3$ optimality equations whose solution yields m -discount optimal policies. These resemble the coupled pair of optimality equations that we introduced in Section 4.1.3. The policy iteration algorithm can be generalized to address the computation of m -discount optimal policies, using a policy improvement step that involves $m+3$ nested minimizations. It can be shown that to obtain a Blackwell optimal policy, it is sufficient to compute an $(n-2)$ -discount optimal policy using this generalized policy iteration algorithm, where n is the number of states. We refer to Puterman [Put94], Chapter 10, for a detailed account.

Infinite-spaces models have been the focus of much research. We provide a sampling of relatively recent references. Some of these are survey papers or textbooks, and provide extensive references: Sennott [Sen86], [Sen89a], [Sen89b], [Sen91], [Sen93a], [Sen93b], [Sen98], Lasserre [Las88], Borkar [Bor88], [Bor89], [Bor91], Cavazos-Cadena [Cav89a], [Cav89b], [Cav91], Hernandez-Lerma [Her89], Fernández-Gaucherand, Arapostathis, and Marcus [FAM90], Hernandez-Lerma, Henet, and Lasserre [HHL91], Cavazos-Cadena and Sennott [CaS92], Ritt and Sennott [RiS92], Arapostathis

et al. [ABF93], Schal [Sch93a], Puterman [Put94], Hernandez-Lerma and Lasserre [HeL96], [HeL99], Meyn [Mey99], Feinberg and Schwartz [FeS02], and Guo and Rieder [GuR06].

Partially observable Markov decision problems (POMDP) can be converted to perfect state information versions in the (uncountably infinite) space of the conditional distributions of the state. Under the average cost criterion, they become challenging problems, with behavior that is interesting and not fully understood at present. In particular, the optimal average cost may depend on the initial state distribution vector p in simple and unexpected situations, as shown in Example 4.6.7. Analysis and some sufficient conditions have been given that address the question whether the optimal average cost is independent of p (Platzman [Pla77a], [Pla80], Ohnishi, Mine, and Kawai [OMK84], Fernández-Gaucherand, Arapostathis, and Marcus [FAM91], Runggaldier and Stettner [RuS94], Stettner [Ste93]); see also Arapostathis et al. [ABF93]. The paper by Yu and Bertsekas [YuB04] proposes computational algorithms based on lower bound approximation by finite-spaces (perfect observation) multi-chain average cost problems that can be solved with the algorithms of Sections 4.3-4.5. As the number of states of the approximating problem increases, the lower bound approximation converges to the optimal average cost of the original POMDP, assuming that this cost is independent of p , and that the bias is a continuous function of p . Another paper by Yu and Bertsekas [YuB06a] (see also Yu [Yu06]) considers an alternative finite-spaces approximation scheme based on the use of finite-history (fixed number of most recent observations) and finite-memory controllers, and shows that the optimal average cost can be approximated arbitrarily closely, assuming that the optimal liminf cost is independent of p .

E X E R C I S E S

4.1

Solve the average cost version ($\alpha = 1$) of the computer manufacturer's problem (Exercise 7.3, Vol. I).

4.2 [LiR71]

Consider a business providing a certain type of service to customers. The business receives at the beginning of each time period with probability p_i a proposal by a customer of type i , where $i = 1, 2, \dots, n$, who offers an amount of money M_i .

We assume that $\sum_{i=1}^n p_i \leq 1$. The business may reject the offer, in which case the customer leaves and the business remains idle during that period, or the business may accept the offer in which case the business services that customer for an amount of time k determined according to probabilities β_{ik} , where, for $k = 1, 2, \dots$,

β_{ik} = probability that the type i customer will leave after k periods, given that the customer has already been serviced for $k - 1$ periods.

The problem is to determine an acceptance-rejection policy that maximizes

$$\lim_{N \rightarrow \infty} \frac{1}{N} \{\text{Expected payment over } N \text{ periods}\}.$$

Consider two cases:

1. $\beta_{ik} = \beta_i \in (0, 1)$ for all k .
 2. For each i there exists \bar{k}_i such that $\beta_{i\bar{k}_i} = 1$.
- (a) Formulate the problem as an average cost per stage problem, and show that the optimal cost is independent of the initial state.
- (b) Show that there exists a scalar λ^* and an optimal policy that accepts the offer of a type i customer if and only if

$$\lambda^* T_i \leq M_i,$$

where T_i is the expected time of service of a type i customer given by

$$T_i = \beta_{i1} + \sum_{k=2}^{\infty} k \beta_{ik} (1 - \beta_{ik-1}) \cdots (1 - \beta_{i0}).$$

4.3

Consider an average cost problem with two states, 1 and 2, and two controls, 1 and 2. At states 1 and 2, the costs per stage are 0 and 1, respectively, regardless of the control applied. Control 1 keeps the system at the state where it is, while control 2 moves the system to the other state. Construct a nonstationary policy for which the limit as $N \rightarrow \infty$ of the average costs over N stages,

$$J_{\pi}^N(x_0) = \frac{1}{N} \sum_{k=0}^{N-1} g(x_k, \mu_k(x_k)),$$

does not exist for $x_0 = 1, 2$. Hint: Consider a nonstationary policy which switches to the other state with sufficiently decreasing frequency so that $J_{\pi}^N(x_0)$ does not converge.

4.4 (Blackwell Optimal Policies and Bellman's Equation)

Consider a deterministic system with two states 0 and 1. Upon entering state 0, the system stays there permanently at no cost. In state 1 there is a choice of staying there at no cost or moving to state 0 at cost 1.

- (a) Show that every policy is average cost optimal, but the only stationary policy that is Blackwell optimal is the one that keeps the system in the state it currently is. (Note that this policy is not unichain.)
- (b) Show that the solutions (λ, h) to Bellman's equation $\lambda e + h = Th$ are the ones for which $\lambda = 0$ and $h(1) \leq 1 + h(0)$. However, if $h(1) < 1 + h(0)$, all policies μ other than the Blackwell optimal, do not attain the minimum in the right-hand side, i.e., $T_\mu h \neq Th$.

4.5

Show that the WA condition holds if and only if there exists a randomized stationary policy that is unichain, and its transient states are transient under all stationary policies.

4.6 (Reduction to the Discounted Case)

For the finite-state average cost problem suppose there is a state t such that for some $\beta > 0$ we have $p_{it}(u) \geq \beta$ for all states i and controls u . Consider the $(1 - \beta)$ -discounted problem with the same state space, control space, and transition probabilities

$$\bar{p}_{ij}(u) = \begin{cases} (1 - \beta)^{-1} p_{ij}(u) & \text{if } j \neq t, \\ (1 - \beta)^{-1} (p_{ij}(u) - \beta) & \text{if } j = t. \end{cases}$$

Show that $\beta \bar{J}(t)$ and $\bar{J}(i)$ are optimal average and differential costs, respectively, where \bar{J} is the optimal cost function of the $(1 - \beta)$ -discounted problem.

4.7

Let h^0 be an arbitrary vector in \mathbb{R}^n , and define for all i and $k \geq 1$

$$h_i^k = T^k h^0 - (T^k h^0)(i)e,$$

$$\hat{h}^k = T^k h^0 - \frac{1}{n} \sum_{i=1}^n (T^k h^0)(i)e,$$

$$\tilde{h}^k = T^k h^0 - \min_{i=1, \dots, n} (T^k h^0)(i)e.$$

Let also $h_i^0 = \hat{h}^0 = \tilde{h}^0 = h^0$.

- (a) Show that the sequences $\{h_i^k\}$, $\{\hat{h}^k\}$, and $\{\tilde{h}^k\}$ are generated by the algorithms

$$h_i^{k+1} = Th_i^k - (Th_i^k)(i)e,$$

$$\hat{h}_i^{k+1} = T\hat{h}^k - \frac{1}{n} \sum_{i=1}^n (T\hat{h}_i^k)(i)e,$$

$$\tilde{h}_i^{k+1} = T\tilde{h}^k - \min_{i=1,\dots,n} (T\tilde{h}^k)(i)e.$$

- (b) Show that the convergence result of Prop. 4.3.2 holds for the algorithms of part (a). Hint: Proposition 4.3.2 applies to the algorithms that generate $\{h_i^k\}$. Express \hat{h}^k and \tilde{h}^k as continuous functions of $\{h_i^k\}$, $i = 1, \dots, n$.

4.8 (Variants of Relative Value Iteration)

Consider the following two variants of the relative value iteration algorithm:

$$h^{k+1}(i) = (Th^k)(i) - \lambda^k, \quad i = 1, \dots, n,$$

where

$$\lambda^k = c + \sum_{j=1}^n p_j h^k(j),$$

or

$$\lambda^k = c + \sum_{j=1}^n p_j h^{k-1}(j).$$

Here c is an arbitrary scalar and (p_1, \dots, p_n) is an arbitrary probability distribution over the states of the system. Under the assumptions of Prop. 4.3.2, show that the sequence $\{h^k\}$ converges to a vector h and the sequence $\{\lambda^k\}$ converges to a scalar λ satisfying $\lambda e + h = Th$, so that by Prop. 4.2.1, λ is equal to the optimal average cost of all initial states and h is an associated differential cost vector. Hint: Modify the problem by introducing an artificial state t' from which the system moves at a cost c to state j with probability p_j , for all u . Apply Prop. 4.3.2.

4.9 (Generalized Error Bounds)

Let h be any n -dimensional vector and let μ be such that

$$T_\mu h = Th.$$

Show that, for all i ,

$$\min_j [(Th)(j) - h(j)] \leq J^*(i) \leq J_\mu(i) \leq \max_j [(Th)(j) - h(j)],$$

regardless of whether $J^*(i)$ is independent of the initial state i . Hint: Complete the details of the following argument. Let

$$\delta(i) = (Th)(i) - h(i), \quad i = 1, \dots, n,$$

and let δ be the vector with coordinates $\delta(i)$. We have

$$T_\mu h = \delta + h, \quad T_\mu^2 h = T_\mu h + P_\mu \delta = \delta + P_\mu \delta + h$$

and, continuing in the same manner,

$$T_\mu^N h = \sum_{k=0}^{N-1} P_\mu^k \delta + h, \quad N = 1, 2, \dots$$

Hence

$$J_\mu = \lim_{N \rightarrow \infty} \frac{1}{N} T_\mu^N h = P_\mu^* \delta,$$

where

$$P_\mu^* = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} P_\mu^k,$$

proving the right-hand side of the desired relation. Also, let $\pi = \{\mu_0, \mu_1, \dots\}$ be any admissible policy. We have

$$T_{\mu_N} h \geq \delta + h$$

from which we obtain

$$T_{\mu_{N-1}} T_{\mu_N} h \geq P_{\mu_{N-1}} \delta + T_{\mu_N} h \geq P_{\mu_{N-1}} \delta + \delta + h \geq 2 \min_j \delta(j) e + h.$$

Thus, for all i ,

$$\frac{1}{N+1} (T_{\mu_0} \cdots T_{\mu_N} h)(i) \geq \min_j \delta(j) + \frac{h(i)}{N+1}$$

and, taking the limit as $N \rightarrow \infty$, we obtain

$$J_\pi(i) \geq \min_j \delta(j).$$

Since π is arbitrary, we obtain the left-hand side of the desired relation.

4.10

Use Prop. 4.1.1 to show that in the policy iteration algorithm, under the unichain assumption, we have for all k ,

$$\lambda^{k+1} e = \lambda^k e + P_{\mu^k}^* (Th^k - h^k - \lambda^k e),$$

where

$$P_{\mu^{k+1}}^* = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{m=0}^{N-1} P_{\mu^{k+1}}^m.$$

Use this fact to show that if the Markov chain corresponding to μ^{k+1} has no transient states and μ^{k+1} is not optimal, then $\lambda^{k+1} < \lambda^k$.

4.11 (Stochastic Shortest Path Solution Method)

The purpose of this exercise is to show how the average cost problem can be solved by solving a finite sequence of stochastic shortest path problems. For any scalar λ , consider the λ -SSP. Consider a one-dimensional search procedure that aims to find a zero of the function $h_\lambda(n)$ of λ by bracketing λ^* from above and below, as illustrated in Fig. 4.5.1. Show that this procedure solves the average cost problem by solving a finite number of stochastic shortest path problems.

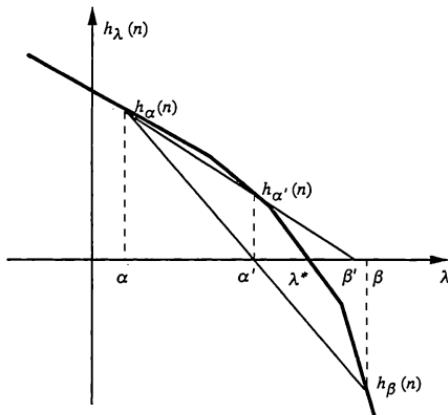


Figure 4.5.1 One dimensional iterative search procedure to find λ such that $h_\lambda(n) = 0$ (cf. Exercise 4.11). Each value $h_\lambda(n)$ is obtained by solving the associated stochastic shortest path problem with stage cost $g(i, u) - \lambda$. At the start of the typical iteration, we have scalars α and β such that $\alpha < \lambda^* < \beta$, together with the corresponding nonzero values $h_\alpha(n)$ and $h_\beta(n)$. We find α' such that

$$\frac{\alpha' - \alpha}{\alpha' - \beta} = \frac{h_\alpha(n)}{h_\beta(n)},$$

and we calculate $h_{\alpha'}(n)$. Let β' be such that

$$\frac{\beta' - \alpha'}{\beta' - \alpha} = \frac{h_{\alpha'}(n)}{h_\alpha(n)}.$$

We then replace α by α' , and if $\beta' < \beta$, we also calculate $h_{\beta'}(n)$ and we replace β by β' . We then perform another iteration. The algorithm stops if either $h_\alpha(n) = 0$ or $h_\beta(n) = 0$.

4.12 (Stochastic Shortest Path Analysis for the Unichain Case)

The purpose of this exercise is to provide an alternative line of analysis of the single-chain average cost problem based on the connection with the stochastic

shortest path problem, which is given in Section 7.4 of Vol. I. In particular, this connection is used to show that there exists a solution (λ, h) to Bellman's equation $\lambda e + h = Th$, under the assumption that every policy that is optimal within the class of stationary policies is unichain. Complete the details of the following proof:

For any stationary policy μ , let λ_μ be the average cost per stage, let $\lambda = \min_\mu \lambda_\mu$, and let $M = \{\mu \mid \lambda_\mu = \lambda\}$ be the set of optimal stationary policies. Suppose that there is a state s that is simultaneously recurrent in the Markov chains corresponding to all $\mu \in M$. Similar to Section 7.4 in Vol. I, consider an associated stochastic shortest path problem with states $1, 2, \dots, n$ and an artificial termination state t to which we move from state i with transition probability $p_{is}(u)$. The stage costs in this problem are $g(i, u) - \lambda$ for $i = 1, \dots, n$, and the transition probabilities from a state i to a state $j \neq s$ are the same as those of the original problem, while $p_{is}(u)$ is zero. Show that in this stochastic shortest path problem, every improper policy has infinite cost for some initial state, and use this fact to conclude that if $h(i)$ is the optimal cost starting at state $i = 1, \dots, n$, then λ and h satisfy $\lambda e + h = Th$. If there is no state s that is simultaneously recurrent for all $\mu \in M$, select a $\bar{\mu} \in M$ such that there is no $\mu \in M$ whose recurrent class is a strict subset of the recurrent class of $\bar{\mu}$ (it is sufficient that $\bar{\mu}$ has minimal number of recurrent states over all $\mu \in M$), change the stage cost of all states i that are not recurrent under $\bar{\mu}$ to $g(i, u) + \epsilon$, where $\epsilon > 0$, use as state s in the preceding argument any state that is recurrent under $\bar{\mu}$, and take $\epsilon \rightarrow 0$.

4.13

Construct a two-state example, with one state being cost-free and absorbing under all policies, such that an optimal solution (λ^*, h^*) of the linear program (4.113) does not satisfy the optimality equation $\lambda^* + h^* = Th^*$.

4.14 (Optimal Policy for Single-Chain Linear Programming I)

Let the WA condition hold, and let (λ^*, h^*) be an optimal solution of the single-chain linear program (4.113).

- (a) Let μ^* be an optimal stationary policy. Show that we have

$$\lambda^* + h^*(i) = g(i, \mu^*(i)) + \sum_{j=1}^n p_{ij}(\mu^*(i))h^*(j), \quad (4.145)$$

for all states i that are recurrent under μ^* . Hint: Let R be a recurrent class of the Markov chain corresponding to μ^* and let ξ_i be the long-term relative frequency of a state $i \in R$, when the initial state is in R , i.e.

$$\xi_i = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} P(i_k = i \mid i_0 = i')$$

for $i' \in R$. If Eq. (4.145) did not hold for some state in R , by multiplying with ξ_i and adding over $i \in R$, we would obtain

$$\sum_{i \in R} \xi_i (\lambda^* + h^*(i)) < \sum_{i \in R} \xi_i \left(g(i, \mu^*(i)) + \sum_{j=1}^n p_{ij}(\mu^*(i)) h^*(j) \right).$$

Show that the left-hand and right-hand sides of this relation are equal, thereby arriving at a contradiction.

(b) Let

$$I^* = \{i \mid \lambda^* + h^*(i) = (Th^*)(i)\},$$

and let $\bar{\mu}(i)$ be any stationary policy such that $(T_{\bar{\mu}}h^*)(i) = (Th^*)(i)$ for all $i \in I^*$. Show that if either $I^* = \{1, \dots, n\}$ or else all states $i \notin I^*$ are transient under $\bar{\mu}$, then $\bar{\mu}$ is optimal. In particular, if each state is recurrent under some optimal stationary policy, then $I^* = \{1, \dots, n\}$ and $\bar{\mu}$ is optimal.

4.15 (Optimal Policy for Single-Chain Linear Programming II)

Let the WA condition hold, and let (λ^*, h^*) be an optimal solution of the single-chain linear program (4.113). Denote

$$C_0 = \left\{ (i, u) \mid \lambda^* + h^*(i) = g(i, u) + \sum_{j=1}^n p_{ij}(u) h^*(j), u \in U(i) \right\},$$

$$I_0 = \{i \mid (i, u) \in C_0 \text{ for some } u \in U(i)\}.$$

Consider an algorithm that starts with (C_0, I_0) , and at the $(k+1)$ st step, given (C_k, I_k) , tries to find a pair $(i, u) \in C_k$ such that $p_{ij}(u) > 0$ for some $j \notin I_k$. If such a pair cannot be found, the algorithm stops; else the algorithm removes (i, u) from C_k to form C_{k+1} , and then defines

$$I_{k+1} = \{i \mid (i, u) \in C_{k+1} \text{ for some } u \in U(i)\}.$$

(a) Show that the algorithm stops with nonempty sets C_k and I_k . Hint: Use the complementary slackness condition to show that I_k contains the set of states

$$I^* = \left\{ i \mid \sum_{u \in U(i)} q^*(i, u) > 0 \right\},$$

where $\{q^*(i, u) \mid i = 1, \dots, n, u \in U(i)\}$ is a dual optimal solution.

(b) Consider a policy of the form

$$\mu^*(i) = \begin{cases} \text{any } u \text{ such that } (i, u) \in C_k & \text{if } i \in I_k, \\ \text{any } u \in U(i) & \text{if } i \notin I_k. \end{cases}$$

Show that μ^* is optimal starting from states in I_k , i.e., $J_\mu^*(i) = \lambda^*$ for all $i \in I_k$.

- (c) Use the construction given in conjunction with Prop. 4.2.6 to obtain an optimal policy that is unichain and coincides with μ^* on a subset of I_k that forms a recurrent class.

4.16

Consider the single-chain linear program (4.113) of Section 4.5, and let λ^* be the optimal value. Show that $\lambda^* = \min_{i=1,\dots,n} J^*(i)$, where J^* is the optimal average cost vector. Hint: Use the multi-chain linear program (4.118) to show that

$$\lambda^* \leq \inf_{\sum_{i=1}^n \beta_i = 1, \beta_i > 0, i=1,\dots,n} \sum_{i=1}^n \beta_i J^*(i).$$

Show also that $\lambda = \min_{i=1,\dots,n} J^*(i)$, together with some vector h , form a feasible solution of the single-chain linear program (4.113), so that $\lambda^* \geq \min_{i=1,\dots,n} J^*(i)$.

4.17 (Stationary Optimal Policies and Bellman's Equation)

Let the WA condition hold, and let (λ^*, h^*) satisfy Bellman's equation $\lambda^* e + h^* = Th^*$. Show that a stationary policy μ^* is optimal if and only if

$$(T_{\mu^*} h^*)(i) = (Th^*)(i)$$

for all states i that are recurrent under μ^* . Hint: Use the single-chain linear program (4.113) and the result of Exercise 4.14(a).

4.18 (Optimal Policy for Multi-Chain Linear Programming)

Consider the multi-chain linear programming case of Section 4.5, let (q^*, r^*) be an optimal dual solution, and let

$$I^* = \left\{ i \mid \sum_{u \in U(i)} q^*(i, u) > 0 \right\}, \quad \bar{I}^* = \{i \mid i \notin I^*\}.$$

- (a) Show that $\sum_{u \in U(i)} r^*(i, u) > 0$ for all $i \in \bar{I}^*$.
- (b) Consider a stationary policy μ^* such that for $i \in I^*$, $\mu^*(i)$ is any $u \in U(i)$ such that $q(i, u) > 0$, while for $i \in \bar{I}^*$, $\mu^*(i)$ is any $u \in U(i)$ such that $r(i, u) > 0$. Show that if all states in \bar{I}^* are transient under μ^* , then μ^* is optimal. Note: Kallenberg [Kal83] (Prop. 4.2.3) shows that if (q^*, r^*) is an extreme point of the dual feasible set, then the states in \bar{I}^* are transient under μ^* . Hint: Complete the details of the following argument. As in the single-chain case, under μ^* the state remains within the set I^* when

started at a state in I^* . Now note that the primal optimal and dual optimal solutions, (J^*, h^*) and (q^*, r^*) must satisfy the following complementary slackness relation for all $i = 1, \dots, n$ and $u \in U(i)$:

$$q^*(i, u) > 0 \Rightarrow J^*(i) + h^*(i) = g(i, u) + \sum_{j=1}^n p_{ij}(u)h^*(j).$$

$$r^*(i, u) > 0 \Rightarrow J^*(i) = \sum_{j=1}^n p_{ij}(u)J^*(j).$$

By applying this relation for $u = \mu^*(i)$, we obtain

$$J^*(i) + h^*(i) = g(i, \mu^*(i)) + \sum_{j=1}^n p_{ij}(\mu^*(i))h^*(j), \quad \text{for all } i \in I^*, \quad (4.146)$$

and

$$J^*(i) = \sum_{j=1}^n p_{ij}(\mu^*(i))J^*(j), \quad \text{for all } i \in \bar{I}^*.$$

Also, from the first equality constraint of the dual program, we have

$$\begin{aligned} 0 &= \sum_{j=1}^n J^*(j) \left(\sum_{u \in U(j)} q^*(j, u) - \sum_{i=1}^n \sum_{u \in U(i)} q^*(i, u)p_{ij}(u) \right) \\ &= \sum_{i=1}^n \sum_{u \in U(i)} q^*(i, u) \left(J^*(i) - \sum_{j=1}^n p_{ij}(u)J^*(j) \right). \end{aligned}$$

Since the parenthesized term in the last expression is nonpositive, it follows that $J^*(i) - \sum_{j=1}^n p_{ij}(u)J^*(j) = 0$ for all $u \in U(i)$ with $q^*(i, u) > 0$. Hence,

$$J^*(i) = \sum_{j=1}^n p_{ij}(\mu^*(i))J^*(j), \quad \text{for all } i \in I^*. \quad (4.147)$$

From the constraints of the primal problem and Eqs. (4.146), (4.147), we see that (J^*, h^*) satisfies the coupled pair of policy evaluation equations for μ^* , restricted to the set of states I^* , so it follows from Prop. 4.1.9 that $J_{\mu^*}(i) = J^*(i)$ for all $i \in I^*$, and hence μ^* is optimal for all initial states in I^* . Since the states in \bar{I}^* are transient under μ^* by assumption, we have $J_{\mu^*}(i) = J^*(i)$ for all $i = 1, \dots, n$.

(c) Consider the randomized policy defined by

$$P(u | i) = \begin{cases} \frac{q^*(i, u)}{\sum_{u \in U(i)} q^*(i, u)} & \text{if } i \in I^*, \\ \frac{r^*(i, u)}{\sum_{u \in U(i)} r^*(i, u)} & \text{if } i \in \bar{I}^*. \end{cases}$$

Show that this policy is optimal. Hint: Use the line of proof outlined for part (b). View $\sum_{u \in U(i)} q^*(i, u)$, $i = 1, \dots, n$, as steady-state probabilities, and use this to show that the states in I^* are recurrent and the states in \bar{I}^* are transient under the randomized policy.

4.19 (Policy Iteration for Linear-Quadratic Problems)

The purpose of this problem is to show that policy iteration works for linear-quadratic problems (even though neither the state space nor the control space are finite). Consider the problem of Section 4.6.5 under the usual controllability, observability, and positive (semi)definiteness assumptions. Let L_0 be an $m \times n$ matrix such that the matrix $(A + BL_0)$ is stable.

- (a) Show that the average cost per stage corresponding to the stationary policy μ^0 , where $\mu^0(x) = L_0x$, is of the form

$$J_{\mu^0} = E\{w'K_0w\},$$

where K_0 is a positive semidefinite symmetric matrix satisfying the (linear) equation

$$K_0 = (A + BL_0)'K_0(A + BL_0) + Q + L_0'RL_0.$$

- (b) Let $\mu^1(x) = L_1x = (R + B'K_0B)^{-1}B'K_0Ax$ be the control function attaining the minimum for each x in the expression

$$\min_u \{u'Ru + (Ax + Bu)'K_0(Ax + Bu)\}.$$

Show that

$$J_{\mu^1} = E\{w'K_1w\} \leq J_{\mu^0},$$

where K_1 is some positive semidefinite symmetric matrix.

- (c) Consider repeating the (policy iteration) process described in parts (a) and (b), thereby obtaining a sequence of positive semidefinite symmetric matrices $\{K_k\}$. Show that

$$K_k \rightarrow K,$$

where K is the optimal cost matrix of the problem.

4.20 (Deterministic Finite-State Systems)

Consider a deterministic finite-state system. Suppose that the system is controllable in the sense that given any two states i and j , there exists a sequence of admissible controls that drives the state of the system from i to j . Consider the problem of finding an admissible control sequence $\{u_0, u_1, \dots\}$ that minimizes

$$J_\pi(i) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} g(x_k, u_k).$$

Show that the optimal cost is independent of the initial state, and that there exist optimal control sequences, which after a certain time index are periodic.

5

Continuous-Time Problems

Contents

5.1. Uniformization	p. 288
5.2. Queueing Applications	p. 295
5.3. Semi-Markov Problems	p. 306
5.4. Notes, Sources, and Exercises	p. 317

We have considered so far problems where the cost per stage does not depend on the time required for transition from one state to the next. Such problems have a natural discrete-time representation. On the other hand, there are situations where controls are applied at discrete times but cost is continuously accumulated. Furthermore, the time between successive control choices is variable; it may be random or it may depend on the current state and the choice of control. For example, in queueing systems state transitions correspond to arrivals or departures of customers, and the corresponding times of transition are random. This chapter primarily discusses problems of this type. We restrict attention to continuous-time systems with a finite or countable number of states. Many of the practical systems of this type involve the Poisson process, so for most of the examples discussed, we assume that the reader is familiar with this process at the level of textbooks such as Ross [Ros83b], Gallager [Gal95], and Bertsekas and Tsitsiklis [BeT02].

In Section 5.1, we focus on an important class of continuous-time optimization models of the discounted type, where the times between successive transitions have an *exponential probability distribution*. We show that by using a conversion process called *uniformization*, these models can be analyzed within the discrete-time framework discussed up to now.

In Section 5.2, we discuss applications of uniformization. We concentrate on discounted-type queueing models arising in various communications and scheduling contexts.

In Section 5.3, we discuss more general continuous-time models, called *semi-Markov problems*, where the times between successive transitions need not have an exponential distribution. A simpler version of this material is also given in Chapter 7 of Vol. I.

5.1 UNIFORMIZATION

In this chapter, we restrict ourselves to continuous-time systems with a finite or a countable number of states. Here state transitions and control selections take place at discrete times, but the time from one transition to the next is random. In this section, we assume that:

1. If the system is in state i and control u is applied, the next state will be j with probability $p_{ij}(u)$.
2. The time interval τ between the transition to state i and the transition to the next state is exponentially distributed with parameter $\nu_i(u)$; i.e.,

$$P\{\text{transition time interval } > \tau \mid i, u\} = e^{-\nu_i(u)\tau},$$

or equivalently, the probability density function of τ is

$$p(\tau) = \nu_i(u)e^{-\nu_i(u)\tau}, \quad \tau \geq 0.$$

Furthermore, τ is independent of earlier transition times, states, and controls. The parameters $\nu_i(u)$ are uniformly bounded in the sense that for some ν we have

$$\nu_i(u) \leq \nu, \quad \text{for all } i, u \in U(i).$$

The parameter $\nu_i(u)$ is referred to as the *rate of transition* associated with state i and control u . It can be verified that the corresponding average transition time is

$$E\{\tau\} = \int_0^\infty \tau \nu_i(u) e^{-\nu_i(u)\tau} d\tau = \frac{1}{\nu_i(u)},$$

so $\nu_i(u)$ can be interpreted as the average number of transitions per unit time.

The state and control at any time t are denoted by $x(t)$ and $u(t)$, respectively, and stay constant between transitions. We use the following notation:

t_k : The time of occurrence of the k th transition. By convention, we denote $t_0 = 0$.

$\tau_k = t_k - t_{k-1}$: The k th transition time interval.

$x_k = x(t_k)$: We have $x(t) = x_k$ for $t_k \leq t < t_{k+1}$.

$u_k = u(t_k)$: We have $u(t) = u_k$ for $t_k \leq t < t_{k+1}$.

We consider a cost function of the form

$$\lim_{N \rightarrow \infty} E \left\{ \int_0^{t_N} e^{-\beta t} g(x(t), u(t)) dt \right\}, \quad (5.1)$$

where g is a given function and β is a given positive discount parameter. Similar to discrete-time problems, an admissible policy is a sequence $\pi = \{\mu_0, \mu_1, \dots\}$, where each μ_k is a function mapping states to controls with $\mu_k(i) \in U(i)$ for all states i . Under π , the control applied in the interval $[t_k, t_{k+1})$ is $\mu_k(x_k)$. Because states stay constant between transitions, the cost function of π is given by

$$J_\pi(x_0) = \sum_{k=0}^{\infty} E \left\{ \int_{t_k}^{t_{k+1}} e^{-\beta t} g(x_k, \mu_k(x_k)) \Big| x_0 \right\}.$$

We first consider the case where *the rate of transition is the same for all states and controls*; i.e.,

$$\nu_i(u) = \nu, \quad \text{for all } i, u.$$

A little thought shows that the problem is then essentially the same as the one where transition times are fixed, because the control cannot influence the cost of a stage by affecting the length of the next transition time interval.

Indeed, the cost (5.1) corresponding to a sequence $\{(x_k, u_k)\}$ can be expressed as

$$\sum_{k=0}^{\infty} E \left\{ \int_{t_k}^{t_{k+1}} e^{-\beta t} g(x(t), u(t)) dt \right\} = \sum_{k=0}^{\infty} E \left\{ \int_{t_k}^{t_{k+1}} e^{-\beta t} dt \right\} E\{g(x_k, u_k)\} \quad (5.2)$$

We have (using the independence of the transition time intervals)

$$\begin{aligned} E \left\{ \int_{t_k}^{t_{k+1}} e^{-\beta t} dt \right\} &= \frac{E\{e^{-\beta t_k}\}(1 - E\{e^{-\beta \tau_{k+1}}\})}{\beta} \\ &= \frac{E\{e^{-\beta(\tau_1 + \dots + \tau_k)}\}(1 - E\{e^{-\beta \tau_{k+1}}\})}{\beta} \\ &= \frac{\alpha^k(1 - \alpha)}{\beta}, \end{aligned} \quad (5.3)$$

where

$$\alpha = E\{e^{-\beta \tau}\} = \int_0^\infty e^{-\beta \tau} \nu e^{-\nu \tau} d\tau = \frac{\nu}{\beta + \nu}.$$

The above expression for α yields $(1 - \alpha)/\beta = 1/(\beta + \nu)$, so that from Eq. (5.3), we have

$$E \left\{ \int_{t_k}^{t_{k+1}} e^{-\beta t} dt \right\} = \frac{\alpha^k}{\beta + \nu}.$$

From this equation together with Eq. (5.2) it follows that the cost of the problem can be expressed as

$$\frac{1}{\beta + \nu} \sum_{k=0}^{\infty} \alpha^k E\{g(x_k, u_k)\}.$$

Thus we are faced in effect with an ordinary discrete-time problem where expected total cost is to be minimized. The effect of randomness of the transition times has been simply to appropriately scale the cost per stage.

To summarize, a continuous-time Markov chain problem with cost

$$\lim_{N \rightarrow \infty} E \left\{ \int_0^{t_N} e^{-\beta t} g(x(t), u(t)) dt \right\}$$

and rate of transition ν that is independent of state and control is equivalent to a discrete-time Markov chain problem with discount factor

$$\alpha = \frac{\nu}{\beta + \nu}, \quad (5.4)$$

and cost per stage given by

$$\bar{g}(i, u) = \frac{g(i, u)}{\beta + \nu}. \quad (5.5)$$

In particular, Bellman's equation takes the form

$$J(i) = \min_{u \in U(i)} \left[\frac{g(i, u)}{\beta + \nu} + \alpha \sum_j p_{ij}(u) J(j) \right], \quad (5.6)$$

or equivalently,

$$J(i) = \frac{1}{\beta + \nu} \min_{u \in U(i)} \left[g(i, u) + \nu \sum_j p_{ij}(u) J(j) \right].$$

In some problems, in addition to the cost (5.1), there is an extra expected stage cost $\hat{g}(i, u)$ that is incurred at the time the control u is chosen at state i , and is independent of the length of the transition interval. In that case the expected stage cost (5.5) should be changed to

$$\hat{g}(i, u) + \frac{g(i, u)}{\beta + \nu},$$

and Bellman's equation (5.6) becomes

$$J(i) = \min_{u \in U(i)} \left[\hat{g}(i, u) + \frac{g(i, u)}{\beta + \nu} + \alpha \sum_j p_{ij}(u) J(j) \right]. \quad (5.7)$$

Example 5.1.1

A manufacturer of a specialty item processes orders in batches. Orders arrive according to a Poisson process with rate ν per unit time; i.e., the successive interarrival intervals are independent and exponentially distributed with parameter ν . For each order there is a positive cost c per unit time that the order is unfilled. Costs are discounted with a discount parameter $\beta > 0$. The setup cost for processing the orders is K . Upon arrival of a new order, the manufacturer must decide whether to process the current batch or to wait for the next order.

Here the state is the number i of unfilled orders. If the decision to fill the orders at state i is made, the cost is K and the next transition will be to state 1. Otherwise, there will be an average cost $(ci)/(\beta + \nu)$ up to the transition to the next state $i + 1$ [cf. Eq. (5.5)], as shown in Fig. 5.1.1. [Note that the setup cost K is incurred immediately after a decision to process the orders is made, so K is not discounted over the time interval up to the next

transition; cf. Eq. (5.7).] We are in effect faced with a discounted discrete-time problem with positive but unbounded cost per stage. (We could also consider an alternative model where an upper bound is placed on the number of unfilled orders. We would then have a discounted discrete-time problem with bounded cost per stage.)

Since Assumption P is satisfied (cf. Section 3.1), Bellman's equation holds and takes the form

$$J(i) = \min \left[K + \alpha J(1), \frac{ci}{\beta + \nu} + \alpha J(i+1) \right], \quad i = 1, 2, \dots,$$

where $\alpha = \nu/(\beta + \nu)$ is the effective discount factor [cf. Eq. (5.4)]. Reasoning from first principles, we see that $J(i)$ is a monotonically nondecreasing function of i , so from Bellman's equation it follows that there exists a threshold i^* such that it is optimal to process the orders if and only if their number exceeds i^* .

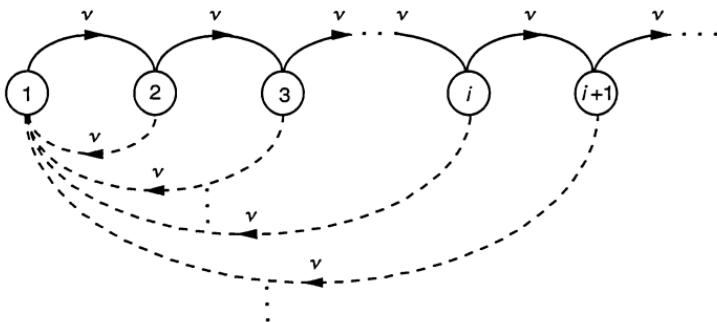


Figure 5.1.1 Transition diagram for the continuous-time Markov chain of Example 5.1.1. The transitions associated with the first control (do not fill the orders) are shown with solid lines, and the transitions associated with the second control (fill the orders) are shown with broken lines.

Nonuniform Transition Rates

We now argue that the more general case where the transition rate $\nu_i(u)$ depends on the state and the control can be converted to the previous case of uniform transition rate by using the trick of *allowing fictitious transitions from a state to itself*. Roughly, transitions that are slow on the average are speeded up with the understanding that sometimes after a transition the state may stay unchanged. To see how this works, let ν be a new uniform transition rate with $\nu_i(u) \leq \nu$ for all i and u , and define new transition

probabilities by

$$\tilde{p}_{ij}(u) = \begin{cases} \frac{\nu_i(u)}{\nu} p_{ij}(u) & \text{if } i \neq j, \\ \frac{\nu_i(u)}{\nu} p_{ii}(u) + 1 - \frac{\nu_i(u)}{\nu} & \text{if } i = j. \end{cases}$$

We refer to this process as the *uniform* version of the original (see Fig. 5.1.2). We argue now that leaving state i at a rate $\nu_i(u)$ in the original process is statistically identical to leaving state i at the faster rate ν , but returning back to i with probability $1 - \nu_i(u)/\nu$ in the new process. Equivalently, transitions are real (lead to a different state) with probability $\nu_i(u)/\nu < 1$. By statistical equivalence, we mean that, for any given policy π , initial state x_0 , time t , and state i , the probability $P\{x(t) = i | x_0, \pi\}$ is identical for the original process and its uniform version. We give a proof of this fact in Exercise 5.1 for the case of a finite number of states (see also Lippman [Lip75b], Serfozo [Ser79], and Ross [Ros83b] for further discussion).

To summarize, we can convert a continuous-time Markov chain problem with transition rates $\nu_i(u)$, transition probabilities $p_{ij}(u)$, and cost

$$\lim_{N \rightarrow \infty} E \left\{ \int_0^{t_N} e^{-\beta t} g(x(t), u(t)) dt \right\},$$

into a discrete-time Markov chain problem with discount factor

$$\alpha = \frac{\nu}{\beta + \nu},$$

where ν is a uniform transition rate chosen so that

$$\nu_i(u) \leq \nu, \quad \text{for all } i, u.$$

The transition probabilities are

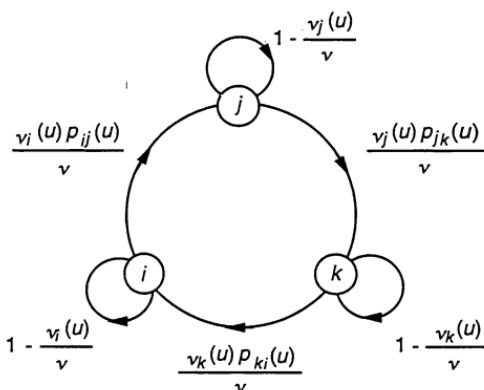
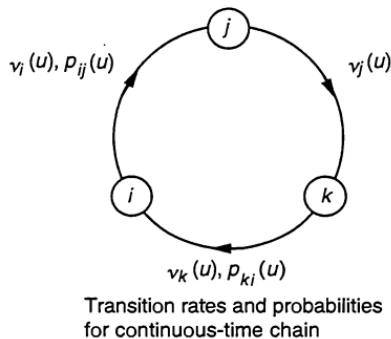
$$\tilde{p}_{ij}(u) = \begin{cases} \frac{\nu_i(u)}{\nu} p_{ij}(u) & \text{if } i \neq j, \\ \frac{\nu_i(u)}{\nu} p_{ii}(u) + 1 - \frac{\nu_i(u)}{\nu} & \text{if } i = j, \end{cases}$$

and the cost per stage is

$$\tilde{g}(i, u) = \frac{g(i, u)}{\beta + \nu}, \quad \text{for all } i, u.$$

In particular, Bellman's equation takes the form

$$J(i) = \min_{u \in U(i)} \left[\tilde{g}(i, u) + \alpha \sum_j \tilde{p}_{ij}(u) J(j) \right],$$



Transition probabilities for uniform version

Figure 5.1.2 Transforming a continuous-time Markov chain into its uniform version through the use of fictitious self-transitions. The uniform version has a uniform transition rate ν , which is an upper bound for all transition rates $\nu_i(u)$ of the original, and transition probabilities $\tilde{p}_{ij}(u) = (\nu_i(u)/\nu)p_{ij}(u)$ for $i \neq j$, and $\tilde{p}_{ii}(u) = (\nu_i(u)/\nu)p_{ii}(u) + 1 - \nu_i(u)/\nu$ for $j = i$. In the example of the figure we have $p_{ii}(u) = 0$ for all i and u .

which, after some calculation using the preceding definitions, can be written as

$$J(i) = \frac{1}{\beta + \nu} \min_{u \in U(i)} \left[g(i, u) + (\nu - \bar{\nu}_i(u)) J(i) + \nu_i(u) \sum_j p_{ij}(u) J(j) \right]. \quad (5.8)$$

In the case where there is an extra expected stage cost $\hat{g}(i, u)$ that is incurred at the time the control u is chosen at state i , Bellman's equation

becomes [cf. Eq. (5.7)]

$$J(i) = \frac{1}{\beta + \nu} \min_{u \in U(i)} \left[(\beta + \nu) \hat{g}(i, u) + g(i, u) + (\nu - \nu_i(u)) J(i) + \nu_i(u) \sum_j p_{ij}(u) J(j) \right].$$

5.2 QUEUEING APPLICATIONS

We now illustrate the theory of the preceding section through some applications involving the control of queues.

Example 5.2.1 (M/M/1 Queue with Controlled Service Rate)

Consider a single-server queueing system where customers arrive according to a Poisson process with rate λ . The service time of a customer is exponentially distributed with parameter μ (called the service rate). Service times of customers are independent and are also independent of customer interarrival times. The service rate μ can be selected from a closed subset M of an interval $[0, \bar{\mu}]$ and can be changed at the times when a customer arrives or when a customer departs from the system. There is a cost $q(\mu)$ per unit time for using rate μ and a waiting cost $c(i)$ per unit time when there are i customers in the system (waiting in queue or undergoing service). The idea is that one should be able to cut down on the customer waiting costs by choosing a faster service rate, which presumably costs more. The problem, roughly, is to select the service rate so that the service cost is optimally traded off with the customer waiting cost.

We assume the following:

- For some $\mu \in M$ we have $\mu > \lambda$. (In words, there is available a service rate that is fast enough to keep up with the arrival rate, thereby maintaining the queue length bounded.)
- The waiting cost function c is nonnegative, monotonically nondecreasing, and “convex-like” in the sense

$$c(i+2) - c(i+1) \geq c(i+1) - c(i), \quad i = 0, 1, \dots$$

- The service rate cost function q is nonnegative, and continuous on $[0, \bar{\mu}]$, with $q(0) = 0$.

The problem fits the framework of this section. The state is the number of customers in the system, and the control is the choice of service rate following a customer arrival or departure. The transition rate at state i is

$$\nu_i(\mu) = \begin{cases} \lambda & \text{if } i = 0, \\ \lambda + \mu & \text{if } i \geq 1. \end{cases}$$

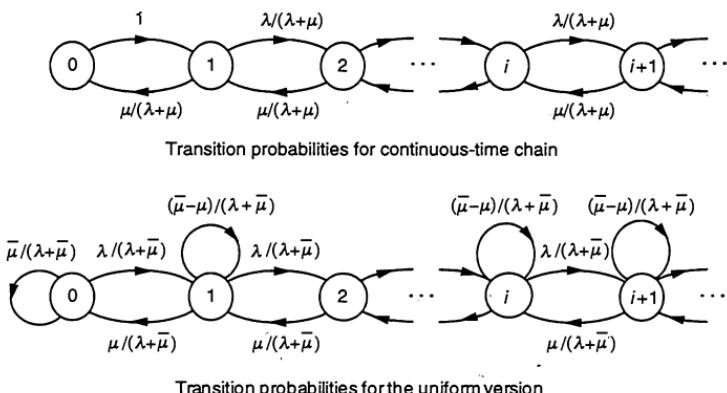


Figure 5.2.1 Continuous-time Markov chain and uniform version for Example 5.2.1 when the service rate is equal to μ . The transition rates of the original Markov chain are $\nu_i(\mu) = \lambda + \mu$ for states $i \geq 1$, and $\nu_0(\mu) = \lambda$ for state 0. The transition rate for the uniform version is $\nu = \lambda + \bar{\mu}$.

The transition probabilities of the Markov chain and its uniform version for the choice

$$\nu = \lambda + \bar{\mu}$$

are shown in Fig. 5.2.1.

The effective discount factor is

$$\alpha = \frac{\nu}{\beta + \nu}$$

and the cost per stage is

$$\frac{1}{\beta + \nu} (c(i) + q(\mu)).$$

The form of Bellman's equation is [cf. Eq. (5.8)]

$$J(0) = \frac{1}{\beta + \nu} (c(0) + (\nu - \lambda)J(0) + \lambda J(1))$$

and for $i = 1, 2, \dots$,

$$J(i) = \frac{1}{\beta + \nu} \min_{\mu \in M} [c(i) + q(\mu) + \mu J(i-1) + (\nu - \lambda - \mu)J(i) + \lambda J(i+1)]. \quad (5.9)$$

An optimal policy is to use at state i the service rate that minimizes the expression on the right. Thus it is optimal to use at state i the service rate

$$\mu^*(i) = \arg \min_{\mu \in M} \{q(\mu) - \mu \Delta(i)\}, \quad (5.10)$$

where $\Delta(i)$ is the differential of the optimal cost

$$\Delta(i) = J(i) - J(i-1), \quad i = 1, 2, \dots$$

[When the minimum in Eq. (5.10) is attained by more than one service rate μ we choose by convention the smallest.] We will demonstrate shortly that $\Delta(i)$ is monotonically nondecreasing. It will then follow from Eq. (5.10) (see Fig. 5.2.2) that the optimal service rate $\mu^*(i)$ is monotonically nondecreasing. Thus, as the queue length increases, it is optimal to use a faster service rate.

To show that $\Delta(i)$ is monotonically nondecreasing, we use the value iteration method to generate a sequence of functions J_k from the starting function

$$J_0(i) = 0, \quad i = 0, 1, \dots$$

For $k = 0, 1, \dots$, [cf. Eq. (5.9)], we have

$$J_{k+1}(0) = \frac{1}{\beta + \nu} (c(0) + (\nu - \lambda) J_k(0) + \lambda J_k(1)),$$

and for $i = 1, 2, \dots$,

$$J_{k+1}(i) = \frac{1}{\beta + \nu} \min_{\mu \in M} [c(i) + q(\mu) + \mu J_k(i-1) + (\nu - \lambda - \mu) J_k(i) + \lambda J_k(i+1)]. \quad (5.11)$$

For $k = 0, 1, \dots$ and $i = 1, 2, \dots$, let

$$\Delta_k(i) = J_k(i) - J_k(i-1).$$

For completeness of notation, define also $\Delta_k(0) = 0$. From the theory of Section 3.1 (see Prop. 3.1.7), we have $J_k(i) \rightarrow J(i)$ as $k \rightarrow \infty$. It follows that we have

$$\lim_{k \rightarrow \infty} \Delta_k(i) = \Delta(i), \quad i = 1, 2, \dots$$

Therefore, it will suffice to show that $\Delta_k(i)$ is monotonically nondecreasing for every k . For this we use induction. The assertion is trivially true for $k = 0$. Assuming that $\Delta_k(i)$ is monotonically nondecreasing, we show that the same is true for $\Delta_{k+1}(i)$. Let

$$\begin{aligned} \mu^k(0) &= 0, \\ \mu^k(i) &= \arg \min_{\mu \in M} [q(\mu) - \mu \Delta_k(i)], \quad i = 1, 2, \dots \end{aligned}$$

From Eq. (5.11) we have, for all $i = 0, 1, \dots$,

$$\begin{aligned} \Delta_{k+1}(i+1) &= J_{k+1}(i+1) - J_{k+1}(i) \\ &\geq \frac{1}{\beta + \nu} (c(i+1) + q(\mu^k(i+1)) + \mu^k(i+1) J_k(i) \\ &\quad + (\nu - \lambda - \mu^k(i+1)) J_k(i+1) \\ &\quad + \lambda J_k(i+2) - c(i) - q(\mu^k(i+1)) - \mu^k(i+1) J_k(i-1)) \quad (5.12) \\ &\quad - (\nu - \lambda - \mu^k(i+1)) J_k(i) - \lambda J_k(i+1)) \\ &= \frac{1}{\beta + \nu} (c(i+1) - c(i) + \lambda \Delta_k(i+2) + (\nu - \lambda) \Delta_k(i+1) \\ &\quad - \mu^k(i+1) (\Delta_k(i+1) - \Delta_k(i))). \end{aligned}$$

Similarly, we obtain, for $i = 1, 2, \dots$,

$$\begin{aligned}\Delta_{k+1}(i) &\leq \frac{1}{\beta + \nu} (c(i) - c(i-1) + \lambda \Delta_k(i+1) + (\nu - \lambda) \Delta_k(i)) \\ &\quad - \mu^k(i-1) (\Delta_k(i) - \Delta_k(i-1)).\end{aligned}$$

Subtracting the last two inequalities, we obtain, for $i = 1, 2, \dots$,

$$\begin{aligned}(\beta + \nu) (\Delta_{k+1}(i+1) - \Delta_{k+1}(i)) &\geq (c(i+1) - c(i)) - (c(i) - c(i-1)) \\ &\quad + \lambda (\Delta_k(i+2) - \Delta_k(i+1)) \\ &\quad + (\nu - \lambda - \mu^k(i+1)) (\Delta_k(i+1) - \Delta_k(i)) \\ &\quad + \mu^k(i-1) (\Delta_k(i) - \Delta_k(i-1)).\end{aligned}$$

Using our convexity assumption on $c(i)$, the fact $\nu - \lambda - \mu^k(i+1) = \bar{\mu} - \mu^k(i+1) \geq 0$, and the induction hypothesis, we see that every term on the right-hand side of the preceding inequality is nonnegative. Therefore, $\Delta_{k+1}(i+1) \geq \Delta_{k+1}(i)$ for $i = 1, 2, \dots$. From Eq. (5.12) we can also show that $\Delta_{k+1}(1) \geq 0 = \Delta_{k+1}(0)$, and the induction proof is complete.

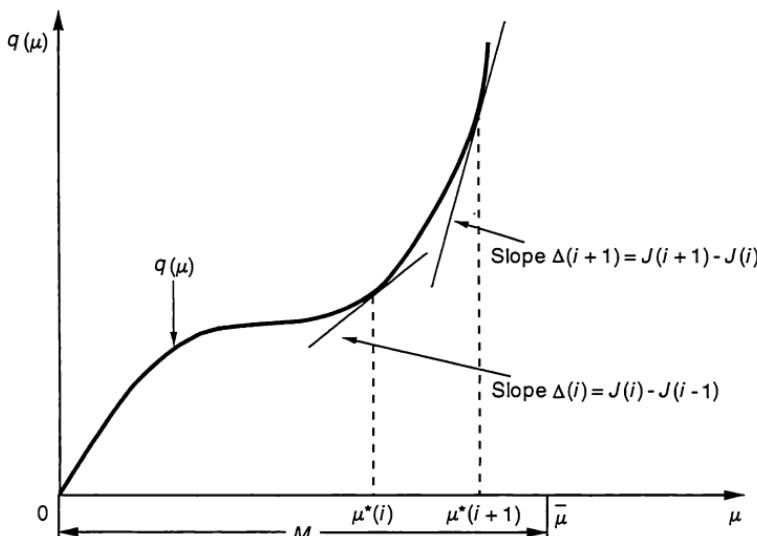


Figure 5.2.2 Determining the optimal service rate at states i and $(i+1)$ in Example 5.2.1. The optimal service rate $\underline{\mu^*(i)}$ tends to increase as the system becomes more crowded (i increases).

To summarize, the optimal service rate $\mu^*(i)$ is given by Eq. (5.10) and tends to become faster as the system becomes more crowded (i increases).

Example 5.2.2 (M/M/1 Queue with Controlled Arrival Rate)

Consider the same queueing system as in the previous example with the difference that the service rate μ is fixed, but the arrival rate λ can be controlled. We assume that λ is chosen from a closed subset Λ of an interval $[0, \bar{\lambda}]$, and there is a cost $q(\lambda)$ per unit time. All other assumptions of Example 5.2.1 are also in effect. What we have here is a problem of flow control, whereby we want to trade off optimally the cost for throttling the arrival process with the customer waiting cost.

This problem is very similar to the one of Example 5.2.1. We choose as uniform transition rate

$$\nu = \bar{\lambda} + \mu$$

and construct the uniform version of the Markov chain. Bellman's equation takes the form

$$\begin{aligned} J(0) &= \frac{1}{\beta + \nu} \min_{\lambda \in \Lambda} [c(0) + q(\lambda) + (\nu - \lambda)J(0) + \lambda J(1)], \\ J(i) &= \frac{1}{\beta + \nu} \min_{\lambda \in \Lambda} [c(i) + q(\lambda) + \mu J(i-1) + (\nu - \lambda - \mu)J(i) + \lambda J(i+1)]. \end{aligned}$$

An optimal policy is to use at state i the arrival rate

$$\lambda^*(i) = \arg \min_{\lambda \in \Lambda} [q(\lambda) + \lambda \Delta(i+1)], \quad (5.13)$$

where, as before, $\Delta(i)$ is the differential of the optimal cost

$$\Delta(i) = J(i) - J(i-1), \quad i = 1, 2, \dots$$

As in Example 5.2.1, we can show that $\Delta(i)$ is monotonically nondecreasing; so from Eq. (5.13) we see that the optimal arrival rate tends to decrease as the system becomes more crowded (i increases).

Example 5.2.3 (Priority Assignment and the μc Rule)

Consider n queues that share a single server. There is a positive cost c_i per unit time and per customer in each queue i . The service time of a customer of queue i is exponentially distributed with parameter μ_i , and all customer service times are independent. Assuming that we start with a given number of customers in each queue and no further arrivals occur, what is the optimal order for serving the customers? The cost here is

$$\lim_{N \rightarrow \infty} E \left\{ \int_0^{t_N} e^{-\beta t} \sum_{i=1}^n c_i x_i(t) dt \right\},$$

where $x_i(t)$ is the number of customers in the i th queue at time t , and β is a positive discount parameter.

We first construct the uniform version of the problem. The construction is shown in Fig. 5.2.3. The discount factor is

$$\alpha = \frac{\mu}{\beta + \mu},$$

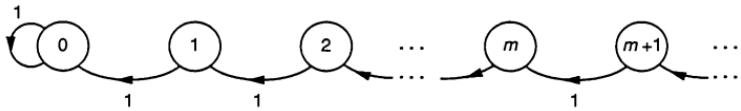
where

$$\mu = \max_i \{\mu_i\},$$

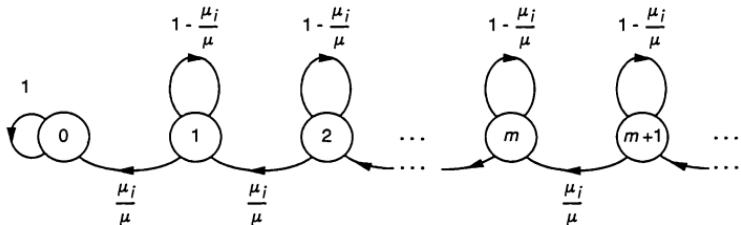
and the corresponding cost is

$$\frac{1}{\beta + \mu} \sum_{k=0}^{\infty} \alpha^k E \left\{ \sum_{i=1}^n c_i x_k^i \right\}, \quad (5.14)$$

where x_k^i is the number of customers in the i th queue after the k th transition (real or fictitious).



Transition probabilities for the i th queue when service is provided



Transition probabilities for uniform version

Figure 5.2.3 Continuous-time Markov chain and uniform version for the i th queue of Example 5.2.3 when service is provided. The transition rate for the uniform version is $\mu = \max_i \{\mu_i\}$.

We now rewrite the cost in a way that is more convenient for analysis. The idea is to transform the problem—from one of minimizing waiting costs to one of maximizing savings in waiting costs through customer service. For $k = 0, 1, \dots$, define

$$i_k = \begin{cases} i & \text{if the } k\text{th transition corresponds to a departure from queue } i, \\ 0 & \text{if the } k\text{th transition is fictitious.} \end{cases}$$

Denote also

$$c_{i_0} = 0,$$

x_0^i : the initial number of customers in queue i .

Then the cost (5.14) can also be written as

$$\begin{aligned} \frac{1}{\beta + \mu} & \left[\sum_{i=1}^n c_i x_0^i + \sum_{k=1}^{\infty} \alpha^k E \left\{ \sum_{i=1}^n c_i x_0^i - \sum_{m=0}^{k-1} c_{i_m} \right\} \right] \\ &= \frac{1}{\beta + \mu} \left[\sum_{k=0}^{\infty} \alpha^k \left(\sum_{i=1}^n c_i x_0^i \right) - E \left\{ \sum_{m=0}^{\infty} \sum_{k=m+1}^{\infty} \alpha^k c_{i_m} \right\} \right] \\ &= \frac{1}{(\beta + \mu)(1 - \alpha)} \sum_{i=1}^n c_i x_0^i - \frac{\alpha}{(\beta + \mu)(1 - \alpha)} \sum_{k=0}^{\infty} \alpha^k E\{c_{i_k}\} \\ &= \frac{1}{\beta} \sum_{i=1}^n c_i x_0^i - \frac{\alpha}{\beta} \sum_{k=0}^{\infty} \alpha^k E\{c_{i_k}\}. \end{aligned}$$

Therefore, instead of minimizing the cost (5.14), we can equivalently

$$\text{maximize } \sum_{k=0}^{\infty} \alpha^k E\{c_{i_k}\}, \quad (5.15)$$

where c_{i_k} can be viewed as the *savings in waiting cost rate* obtained from the k th transition.

We now recognize problem (5.15) as a multiarmed bandit problem. The n queues can be viewed as separate projects. At each time, a nonempty queue, say i , is selected and served. Since a customer departure occurs with probability μ_i/μ , and a fictitious transition that leaves the state unchanged occurs with probability $1 - \mu_i/\mu$, the corresponding expected reward is

$$\frac{\mu_i}{\mu} c_i. \quad (5.16)$$

It is evident that the problem falls in the deteriorating case examined at the end of Section 1.5. Therefore, after each customer departure, it is optimal to serve the queue with maximum expected reward per stage (i.e., engage the project with maximal index; cf. the end of Section 1.5). Equivalently [cf. Eq. (5.16)], it is optimal to serve the nonempty queue i for which $\mu_i c_i$ is maximum. This policy is known as the *μc rule*. It plays an important role in several other formulations of the priority assignment problem (see Baras, Dorsey, and Makowski [BDM83], and Harrison [Har75a], [Har75b]). We can view $\mu_i c_i$ as the ratio of the waiting cost rate c_i by the average time $1/\mu_i$ needed to serve a customer. Therefore, the *μc rule* amounts to serving the queue for which the savings in waiting cost rate per unit average service time are maximized.

Example 5.2.4 (Routing Policies for a Two-Station System)

Consider the system consisting of two queues shown in Fig. 5.2.4. Customers arrive according to a Poisson process with rate λ and are routed upon arrival to one of the two queues. Service times are independent and exponentially distributed with parameter μ_1 in the first queue and μ_2 in the second queue. The cost is

$$\lim_{N \rightarrow \infty} E \left\{ \int_0^{t_N} e^{-\beta t} (c_1 x_1(t) + c_2 x_2(t)) dt \right\},$$

where β , c_1 , and c_2 are given positive scalars, and $x_1(t)$ and $x_2(t)$ denote the number of customers at time t in queues 1 and 2, respectively.

As earlier, we construct the uniform version of this problem with uniform rate

$$\nu = \lambda + \mu_1 + \mu_2$$

and the transition probabilities shown in Fig. 5.2.5. We take as state space the set of pairs (i, j) of customers in queues 1 and 2. Bellman's equation takes the form

$$\begin{aligned} J(i, j) = & \frac{1}{\beta + \nu} (c_1 i + c_2 j + \mu_1 J((i - 1)^+, j) + \mu_2 J(i, (j - 1)^+)) \\ & + \frac{\lambda}{\beta + \nu} \min[J(i + 1, j), J(i, j + 1)], \end{aligned} \tag{5.17}$$

where for any x we denote

$$(x)^+ = \max(0, x).$$

From this equation we see that an optimal policy is to route an arriving customer to queue 1 if and only if the state (i, j) at the time of arrival belongs to the set

$$S_1 = \{(i, j) \mid J(i + 1, j) \leq J(i, j + 1)\}. \tag{5.18}$$

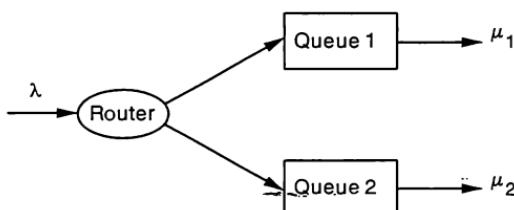
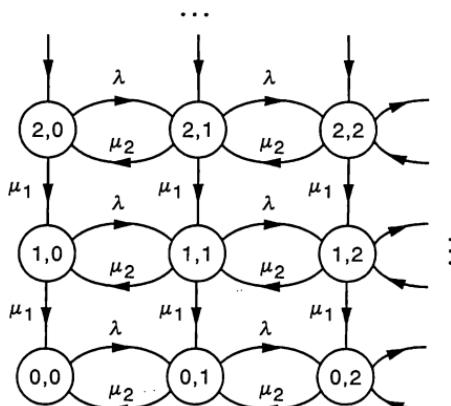
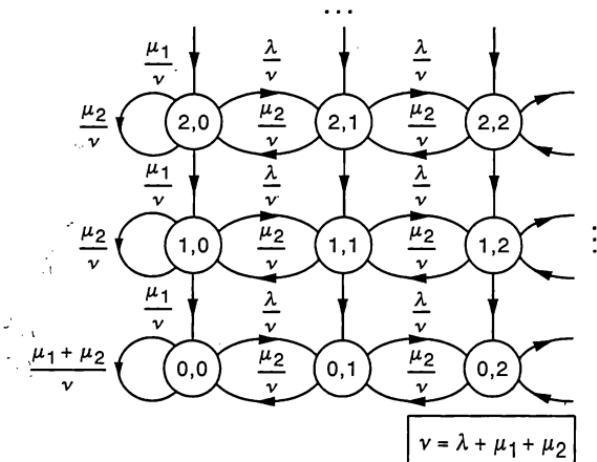


Figure 5.2.4 Queueing system of Example 5.2.4. The problem is to route each arriving customer to queue 1 or 2 so as to minimize the total average discounted waiting cost.



Components of the transition rates when customers are routed to queue 1



Transition probabilities for uniform version

Figure 5.2.5 Continuous-time Markov chain and uniform version for Example 5.2.4 when customers are routed to the first queue. The states are the pairs of customer numbers in the two queues.

This optimal policy can be characterized better by some further analysis. Intuitively, one expects that optimal routing can be achieved by sending a customer to the queue that is “less crowded” in some sense. It is therefore natural to conjecture that, if it is optimal to route to the first queue when the state is (i, j) , it must be optimal to do the same when the first queue is even less crowded; i.e., the state is $(i - m, j)$ with $m \geq 1$. This is equivalent

to saying that the set of states S_1 for which it is optimal to route to the first queue is characterized by a monotonically nondecreasing *threshold function* F by means of

$$S_1 = \{(i, u) \mid i = F(j)\} \quad (5.19)$$

(see Fig. 5.2.6). Accordingly, we call the corresponding optimal policy a *threshold policy*.

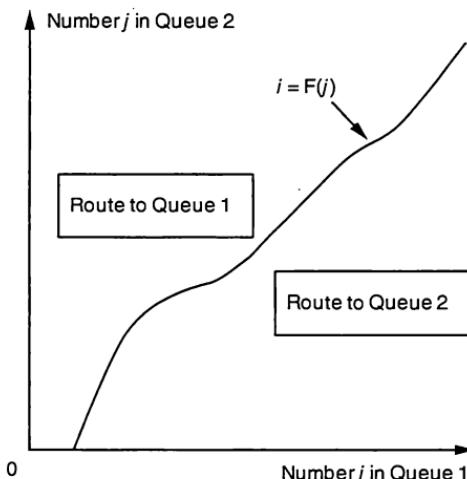


Figure 5.2.6 Typical threshold policy characterized by a threshold function F .

We will demonstrate the existence of a threshold optimal policy by showing that the functions

$$\Delta_1(i, j) = J(i + 1, j) - J(i, j + 1),$$

$$\Delta_2(i, j) = J(i, j + 1) - J(i + 1, j)$$

are monotonically nondecreasing in i for each fixed j , and in j for each fixed i , respectively. We will show this property for Δ_1 ; the proof for Δ_2 is analogous. It will be sufficient to show that for all $k = 0, 1, \dots$, the functions

$$\Delta_1^k(i, j) = J_k(i + 1, j) - J_k(i, j + 1) \quad (5.20)$$

are monotonically nondecreasing in i for each fixed j , where J_k is generated by the value iteration method starting from the zero function; i.e., $J_{k+1}(i, j) = (TJ_k)(i, j)$, where T is the DP mapping defining Bellman's equation (5.17) and $J_0 = 0$. This is true because $J_k(i, j) \rightarrow J(i, j)$ for all i, j as $k \rightarrow \infty$ (Prop. 3.1.6). To prove that $\Delta_1^k(i, j)$ has the desired property, it is useful to first verify that $J_k(i, j)$ is monotonically nondecreasing in i (or j) for fixed j

(or i). This is simple to show by induction or by arguing from first principles using the fact that $J_k(i, j)$ has a k -stage optimal cost interpretation. Next we use Eqs. (5.17) and (5.20) to write

$$\begin{aligned} (\beta + \nu) \Delta_1^{k+1}(i, j) &= c_1 - c_2 \\ &\quad + \mu_1 (J_k(i, j) - J_k((i-1)^+, j+1)) \\ &\quad + \mu_2 (J_k(i+1, (j-1)^+) - J_k(i, j)) \\ &\quad + \lambda (\min [J_k(i+2, j), J_k(i+1, j+1)] \\ &\quad - \min [J_k(i+1, j+1), J_k(i, j+2)]). \end{aligned} \quad (5.21)$$

We now argue by induction. We have $\Delta_1^0(i, j) = 0$ for all (i, j) . We assume that $\Delta_1^k(i, j)$ is monotonically nondecreasing in i for fixed j , and show that the same is true for $\Delta_1^{k+1}(i, j)$. This can be verified by showing that each of the terms in the right-hand side of Eq. (5.21) is monotonically nondecreasing in i for fixed j . Indeed, the first term is constant, and the second and third terms are seen to be monotonically nondecreasing in i using the induction hypothesis for the case where $i, j > 0$ and the earlier shown fact that $J_k(i, j)$ is monotonically nondecreasing in i for the case where $i = 0$ or $j = 0$. The last term on the right-hand side of Eq. (5.21) can be written as

$$\begin{aligned} &\lambda (J_k(i+1, j+1) + \min [J_k(i+2, j) - J_k(i+1, j+1), 0] \\ &\quad - J_k(i+1, j+1) - \min [0, J_k(i, j+2) - J_k(i+1, j+1)]) \\ &= \lambda (\min [0, J_k(i+1, j) - J_k(i+1, j+1)] \\ &\quad + \max [0, J_k(i+1, j+1) - J_k(i, j+2)]) \\ &= \lambda (\min [0, \Delta_1^k(i+1, j)] + \max [0, \Delta_1^k(i, j+1)]). \end{aligned}$$

Since $\Delta_1^k(i+1, j)$ and $\Delta_1^k(i, j+1)$ are monotonically nondecreasing in i by the induction hypothesis, the same is true for the preceding expression. Therefore, each of the terms on the right-hand side of Eq. (5.21) is monotonically nondecreasing in i , and the induction proof is complete. Thus the existence of an optimal threshold policy is established.

There are a number of generalizations of the routing problem of this example that admit a similar analysis and for which there exist optimal policies of the threshold type. For example, suppose that there are additional Poisson arrival processes with rates λ_1 and λ_2 at queues 1 and 2, respectively. The existence of an optimal threshold policy can be shown by a nearly verbatim repetition of our analysis. A more substantive extension is obtained when there is additional service capacity μ that can be switched at the times of transition due to an arrival or service completion to serve a customer in queue 1 or 2. Then we can similarly prove that it is optimal to route to queue 1 if and only if $(i, j) \in S_1$ and to switch the additional service capacity to queue 2 if and only if $(i+1, j+1) \in S_1$, where S_1 is given by Eq. (5.18) and is characterized by a threshold function as in Eq. (5.19). For a proof of this and further extensions, we refer to Hajek [Haj84], which generalizes and unifies several earlier results on the subject.

5.3 SEMI-MARKOV PROBLEMS

We now consider a more general version of the continuous-time problem of Section 5.1. We still have a finite or a countable number of states, but we replace transition probabilities with *transition distributions* $Q_{ij}(\tau, u)$ that, for a given pair (i, u) , specify the joint distribution of the transition interval and the next state:

$$Q_{ij}(\tau, u) = P\{t_{k+1} - t_k \leq \tau, x_{k+1} = j \mid x_k = i, u_k = u\}.$$

We assume that for all states i and j , and controls $u \in U(i)$, $Q_{ij}(\tau, u)$ is known and that the average transition time is finite:

$$\int_0^\infty \tau Q_{ij}(\tau, u) d\tau < \infty.$$

Note that the transition distributions specify the ordinary transition probabilities via

$$p_{ij}(u) = P\{x_{k+1} = j \mid x_k = i, u_k = u\} = \lim_{\tau \rightarrow \infty} Q_{ij}(\tau, u).$$

The difference from the model of Section 5.1 is that $Q_{ij}(\tau, u)$ need not be an exponential distribution. Note that a more elementary version of the subsequent analysis is given in Section 7.5 of Vol. I.

Discounted Problems

Let us first consider a cost function of the form

$$\lim_{N \rightarrow \infty} E \left\{ \int_0^{t_N} e^{-\beta t} g(x(t), u(t)) dt \right\}, \quad (5.22)$$

where t_N is the completion time of the N th transition, and the function g and the positive discount parameter β are given. The cost function of an admissible N -stage policy $\pi = \{\mu_0, \mu_1, \dots, \mu_{N-1}\}$ is given by

$$J_\pi^N(i) = \sum_{k=0}^{N-1} E \left\{ \underbrace{\int_{t_k}^{t_{k+1}} e^{-\beta t} g(x_k, \mu_k(x_k)) dt}_{-} \mid x_0 = i \right\}.$$

We see that for all states i we have

$$J_\pi^N(i) = G(i, \mu_0(i)) + \sum_j \int_0^\infty e^{-\beta \tau} Q_{ij}(d\tau, \mu(i)) J_{\pi_1}^{N-1}(j), \quad (5.23)$$

where $J_{\pi_1}^{N-1}(j)$ is the $(N-1)$ -stage cost of the policy $\pi_1 = \{\mu_1, \mu_2, \dots, \mu_{N-1}\}$ that is used after the first stage, and $G(i, u)$ is the expected single stage cost corresponding to (i, u) . This latter cost is given by

$$G(i, u) = g(i, u) \sum_j \int_0^\infty \left(\int_0^\tau e^{-\beta t} dt \right) Q_{ij}(d\tau, u),$$

or equivalently, since $\int_0^\tau e^{-\beta t} dt = (1 - e^{-\beta \tau})/\beta$,

$$G(i, u) = g(i, u) \sum_j \int_0^\infty \frac{1 - e^{-\beta \tau}}{\beta} Q_{ij}(d\tau, u). \quad (5.24)$$

If we denote

$$m_{ij}(u) = \int_0^\infty e^{-\beta \tau} Q_{ij}(d\tau, u), \quad (5.25)$$

we see that Eq. (5.23) can be written in the form

$$J_\pi^N(i) = G(i, \mu_0(i)) + \sum_j m_{ij}(\mu_0(i)) J_{\pi_1}^{N-1}(j), \quad (5.26)$$

which is similar to the corresponding equation for discounted discrete-time problems [we have $m_{ij}(u)$ in place of $\alpha p_{ij}(u)$].

The expression (5.26) motivates the use of mappings T and T_μ that are similar to those used in Chapter 1 for discounted problems. Let us define for a function J and a stationary policy μ ,

$$(T_\mu J)(i) = G(i, \mu(i)) + \sum_j m_{ij}(\mu(i)) J(j), \quad (5.27)$$

$$(TJ)(i) = \min_{u \in U(i)} \left[G(i, u) + \sum_j m_{ij}(u) J(j) \right]. \quad (5.28)$$

Then by using Eq. (5.26), it can be seen that the cost function J_π of an infinite horizon policy $\pi = \{\mu_0, \mu_1, \dots\}$ can be expressed as

$$J_\pi(i) = \lim_{N \rightarrow \infty} J_\pi^N(i) = \lim_{N \rightarrow \infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} J_0)(i),$$

where J_0 is the zero function [$J_0(i) = 0$ for all i]. The cost of a stationary policy μ can be expressed as

$$J_\mu(i) = \lim_{N \rightarrow \infty} (T_\mu^N J_0)(i).$$

The discounted cost analysis of Section 1.2 carries through in its entirety, provided we assume that:

- (a) $g(i, u)$ [and hence also $G(i, u)$] is a bounded function of i and u .
 (b) The maximum over (i, u) of the sum $\sum_j m_{ij}(u)$ is less than one; i.e.,

$$\rho = \max_{i, u \in U(i)} \sum_j m_{ij}(u) < 1. \quad (5.29)$$

Under these circumstances, the mappings T and T_μ can be shown to be contraction mappings with modulus of contraction ρ (compare also with Prop. 1.2.4). Using this fact, analogs of Props. 1.2.1-1.2.3 can be readily shown. In particular, the optimal cost function J^* is the unique bounded solution of Bellman's equation $J = TJ$ or

$$J(i) = \min_{u \in U(i)} \left[G(i, u) + \sum_j m_{ij}(u) J(j) \right].$$

In addition, there are analogs of several of the computational methods of Section 1.3, including policy iteration and linear programming.

What is happening here is that essentially we have the equivalent of a discrete-time discounted problem where the discount factor depends on i and u . In fact, in Exercise 1.12 of Chapter 1, a data transformation is given, which converts such a problem to an ordinary discrete-time discounted problem where the discount factor is the same for all i and u . With a little thought it can be seen that this data transformation is very similar to the uniformization process we discussed in Section 5.1.

We note that for the contraction property $\rho < 1$ [cf. Eq. (5.29)] to hold, it is sufficient that there exist $\bar{\tau} > 0$ and $\epsilon > 0$ such that the transition time is greater than $\bar{\tau}$ with probability greater than $\epsilon > 0$; i.e., we have for all i and $u \in U(i)$,

$$1 - \sum_j Q_{ij}(\bar{\tau}, u) = \sum_j P\{\tau \geq \bar{\tau} \mid i, u, j\} \geq \epsilon. \quad (5.30)$$

In the case where the state space is countably infinite and the function $g(i, u)$ is not bounded, the mappings T and T_μ are not contraction mappings, and a discounted cost analysis that parallels the one of Section 1.2 is not possible. Even in this case, however, analogs of the results of Section 3.1 can often be shown under appropriate conditions that parallel Assumptions P and N of that section.

We finally note that in some problems, in addition to the cost (5.22), there is an extra expected stage-cost $\hat{g}(i, u)$ that is incurred at the time the control u is chosen at state i , and is independent of the length of the transition interval. In that case the mappings T and T_μ should be changed to

$$(T_\mu J)(i) = \hat{g}(i, \mu(i)) + G(i, \mu(i)) + \sum_j m_{ij}(\mu(i)) J(j),$$

$$(TJ)(i) = \min_{u \in U(i)} \left[\hat{g}(i, u) + G(i, u) + \sum_j m_{ij}(u) J(j) \right]. \quad (5.31)$$

Another problem variation arises when the cost per unit time g depends on the next state j . In this problem formulation, once the system goes into state i , a control $u \in U(i)$ is selected, the next state is determined to be j with probability $p_{ij}(u)$, and the cost of the next transition is $g(i, u, j)\tau_{ij}(u)$ where $\tau_{ij}(u)$ is random with distribution $Q_{ij}(\tau, u)/p_{ij}(u)$. In this case, $G(i, u)$ should be defined by

$$G(i, u) = \sum_j \int_0^\infty g(i, u, j) \frac{1 - e^{-\beta\tau}}{\beta} Q_{ij}(d\tau, u),$$

[cf. Eq. (5.24)] and the preceding development goes through without modification.

Example 5.3.1 (Control of an M/D/1 Queue)

Consider a single server queue where customers arrive according to a Poisson process with rate λ . The service time of a customer is deterministic and is equal to $1/\mu$ where μ is the service rate provided. The arrival and service rates λ and μ can be selected from given subsets Λ and M , and can be changed only when a customer departs from the system. There are costs $q(\lambda)$ and $r(\mu)$ per unit time for using rates λ and μ , respectively, and there is a waiting cost $c(i)$ per unit time when there are i customers in the system (waiting in queue or undergoing service). We wish to find a rate-setting policy that minimizes the total cost when there is a positive discount parameter β .

This problem bears similarity with Examples 2.1 and 2.2 of Section 5.2. Note, however, that while in those examples the rates can be changed both when a customer arrives and when a customer departs, here the rates can be changed only when a customer departs. Because the service time distribution is not exponential, it is necessary to make this restriction in order to be able to use as state the number of customers in the system; if we allowed the arrival rate to also change when a customer arrives, the time already spent in service by the customer found in service by the arriving customer would have to be part of the state.

The transition distributions are given by

$$Q_{0j}(\tau, \lambda, \mu) = \begin{cases} 1 - e^{-\lambda\tau} & \text{if } j = 1, \\ 0 & \text{otherwise,} \end{cases}$$

$$Q_{ij}(\tau, \lambda, \mu) = \begin{cases} p_{ij}(\lambda, \mu) & \text{if } 1/\mu \leq \tau, \\ 0 & \text{otherwise,} \end{cases} \quad i \geq 1,$$

where $p_{ij}(\lambda, \mu)$ are the state transition probabilities. It can be seen that for $i \geq 1$ and $j \geq i-1$, $p_{ij}(\lambda, \mu)$ can be calculated as the probability that $j-i+1$ arrivals will occur in an interval of length $[0, 1/\mu]$. In particular, we have

$$p_{ij}(\lambda, \mu) = \begin{cases} \frac{e^{-\lambda/\mu} (\lambda/\mu)^{(j-i+1)}}{(j-i+1)!} & \text{if } j \geq i-1, \\ 0 & \text{otherwise,} \end{cases} \quad i \geq 1.$$

Using the above formulas and Eqs. (5.24)-(5.25) and (5.27)-(5.28), one can write Bellman's equation and solve the problem as if it were essentially a discrete-time discounted problem.

Average Cost Problems

A natural cost function for the continuous-time average cost problem would be

$$\lim_{T \rightarrow \infty} \frac{1}{T} E \left\{ \int_0^T g(x(t), u(t)) dt \right\}. \quad (5.32)$$

However, we will use instead the cost function

$$\lim_{N \rightarrow \infty} \frac{1}{E\{t_N\}} E \left\{ \int_0^{t_N} g(x(t), u(t)) dt \right\}, \quad (5.33)$$

where t_N is the completion time of the N th transition. This cost function is also reasonable and turns out to be analytically convenient. We note, however, that the cost functions (5.31) and (5.33) are equivalent under the conditions of the subsequent analysis, although a rigorous justification of this is beyond our scope (see [Ros70], p. 52 and p. 160 for related analysis).

We assume that there are n states, denoted $1, \dots, n$, and that the control constraint set $U(i)$ is finite for each state i . For each pair (i, u) , we denote by $G(i, u)$ the single stage expected cost corresponding to state i and control u . We have

$$G(i, u) = g(i, u) \bar{\tau}_i(u), \quad (5.34)$$

where $\bar{\tau}_i(u)$ is the expected value of the transition time corresponding to (i, u) :

$$\bar{\tau}_i(u) = \sum_{j=1}^n \int_0^\infty \tau Q_{ij}(d\tau, u). \quad (5.35)$$

[If the cost per unit time g depends on the next state j , the expected transition cost $G(i, u)$ should be defined by

$$G(i, u) = \sum_{j=1}^n \int_0^\infty g(i, u, j) \tau Q_{ij}(d\tau, u),$$

and the following analysis and results go through without modification.] We assume throughout the remainder of this section that

$$0 < \bar{\tau}_i(u) < \infty, \quad i = 1, \dots, n, \quad u \in U(i). \quad (5.36)$$

The cost function of an admissible policy $\pi = \{\mu_0, \mu_1, \dots\}$ is given by

$$J_\pi(i) = \lim_{N \rightarrow \infty} \frac{1}{E\{t_N \mid x_0 = i, \pi\}} E \left\{ \sum_{k=0}^{N-1} G(x_k, \mu_k(x_k)) \mid x_0 = i \right\}.$$

Our earlier analysis of the discrete-time average cost problem in Chapter 4 suggests that under assumptions similar to those of Section 4.2, the cost $J_\mu(i)$ of a stationary policy μ , as well as the optimal average cost per stage $J^*(i)$, are independent of the initial state i . Indeed, we will see that the character of the solution of the problem is determined by the structure of the *embedded Markov chain*, which is the controlled discrete-time Markov chain whose transition probabilities are

$$p_{ij}(u) = \lim_{\tau \rightarrow \infty} Q_{ij}(\tau, u).$$

In particular, we will show that $J_\mu(i)$ and $J^*(i)$ are independent of i if and only if the same is true for the embedded Markov chain problem. For example, we will show that $J_\mu(i)$ and $J^*(i)$, are independent of i if all stationary policies μ are unichain; i.e., the Markov chain with transition probabilities $p_{ij}(\mu(i))$ has a single recurrent class.

We will also show that Bellman's equation for average cost semi-Markov problems resembles the corresponding discrete-time equation, and takes the form

$$h(i) = \min_{u \in U(i)} \left[G(i, u) - \lambda \bar{\tau}_i(u) + \sum_{j=1}^n p_{ij}(u) h(j) \right]. \quad (5.37)$$

As a special case, when $\bar{\tau}_i(u) = 1$ for all (i, u) , we obtain the corresponding discrete-time equation of Chapter 4. We illustrate Bellman's equation (5.37) for the case of a single unichain policy with the stochastic shortest path argument that we used to prove Prop. 4.2.5.

Consider a unichain policy μ and without loss of generality assume that state n is a recurrent state in the Markov chain corresponding to μ . For each state $i \neq n$ let C_i and T_i be the expected cost and the expected time, respectively, up to reaching state n for the first time starting from i . Let also C_n and T_n be the expected cost and the expected time, respectively, up to returning to n for the first time starting from n . We can view C_i as the costs corresponding to μ in a stochastic shortest path problem where n is a termination state and the costs are $G(i, \mu(i))$. Since μ is a proper policy for this problem, from Prop. 2.1.1, we have that the scalars C_i solve uniquely the system of equations

$$C_i = G(i, \mu(i)) + \sum_{j=1, j \neq n}^n p_{ij}(\mu(i)) C_j, \quad i = 1, \dots, n. \quad (5.38)$$

Similarly, we can view T_i as the costs corresponding to μ in a stochastic shortest path problem where n is a termination state and the costs are $\bar{\tau}_i(\mu(i))$, so that the T_i solve uniquely the system of equations

$$T_i = \bar{\tau}_i(\mu(i)) + \sum_{j=1, j \neq n}^n p_{ij}(\mu(i)) T_j, \quad i = 1, \dots, n. \quad (5.39)$$

Denote

$$\lambda_\mu = \frac{C_n}{T_n}. \quad (5.40)$$

Multiplying Eq. (5.39) by λ_μ and subtracting it from Eq. (5.38), we obtain for all $i = 1, \dots, n$,

$$C_i - \lambda_\mu T_i = G(i, \mu(i)) - \lambda_\mu \bar{\tau}_i(\mu(i)) + \sum_{j=1, j \neq n}^n p_{ij}(\mu(i))(C_j - \lambda_\mu T_j).$$

By defining

$$h_\mu(i) = C_i - \lambda_\mu T_i, \quad i = 1, \dots, n, \quad (5.41)$$

and by noting that from Eq. (5.40) we have

$$h_\mu(n) = 0,$$

we obtain for all $i = 1, \dots, n$,

$$h_\mu(i) = G(i, \mu(i)) - \lambda_\mu \bar{\tau}_i(\mu(i)) + \sum_{j=1}^n p_{ij}(\mu(i)) h_\mu(j), \quad (5.42)$$

which is Bellman's equation (5.37) for the case of a single stationary policy μ .

We have not yet proved that the scalar λ_μ of Eq. (5.40) is the average cost per stage corresponding to μ . This fact will follow from the following proposition, which parallels Prop. 4.2.1 and shows that if Bellman's equation (5.37) has a solution (λ, h) , then the optimal average cost is equal to λ and is independent of the initial state.

Proposition 5.3.1: If a scalar λ and an n -dimensional vector h satisfy

$$h(i) = \min_{u \in U(i)} \left[G(i, u) - \lambda \bar{\tau}_i(u) + \sum_{j=1}^n p_{ij}(u) h(j) \right], \quad i = 1, \dots, n, \quad (5.43)$$

then λ is the optimal average cost per stage $J^*(i)$ for all i ,

$$\lambda = \min_{\pi} J_{\pi}(i) = J^*(i), \quad i = 1, \dots, n. \quad (5.44)$$

Furthermore, if $\mu^*(i)$ attains the minimum in Eq. (5.43) for each i , the stationary policy μ^* is optimal; i.e., $J_{\mu^*}(i) = \lambda$ for all i .

Proof: For any μ consider the mapping $T_{\mu} : \mathbb{R}^n \mapsto \mathbb{R}^n$ given by

$$(T_{\mu} h)(i) = G(i, \mu(i)) + \sum_{j=1}^n p_{ij}(\mu(i)) h(j), \quad i = 1, \dots, n,$$

and the vector $\bar{\tau}(\mu)$ and matrix P_{μ} given by

$$\bar{\tau}(\mu) = \begin{bmatrix} \bar{\tau}_1(\mu(1)) \\ \vdots \\ \bar{\tau}_n(\mu(n)) \end{bmatrix}, \quad P_{\mu} = \begin{bmatrix} p_{11}(\mu(1)) & \dots & p_{1n}(\mu(1)) \\ \dots & \dots & \dots \\ p_{n1}(\mu(n)) & \dots & p_{nn}(\mu(n)) \end{bmatrix}.$$

Let $\pi = \{\mu_0, \mu_1, \dots\}$ be any admissible policy and N be a positive integer. We have from Eq. (5.43),

$$T_{\mu_{N-1}} h \geq \lambda \bar{\tau}(\mu_{N-1}) + h.$$

By applying $T_{\mu_{N-2}}$ to both sides of this relation, and by using the monotonicity of $T_{\mu_{N-2}}$ and Eq. (5.43), we see that

$$\begin{aligned} T_{\mu_{N-2}} T_{\mu_{N-1}} h &\geq T_{\mu_{N-2}} (\lambda \bar{\tau}(\mu_{N-1}) + h) \\ &= \lambda P_{\mu_{N-2}} \bar{\tau}(\mu_{N-1}) + T_{\mu_{N-2}} h \\ &\geq \lambda P_{\mu_{N-2}} \bar{\tau}(\mu_{N-1}) + \lambda \bar{\tau}(\mu_{N-2}) + h. \end{aligned}$$

Continuing in the same manner, we finally obtain

$$T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} h \geq \lambda \bar{t}_N(\pi) + h, \quad (5.45)$$

where $\bar{t}_N(\pi)$ is given by

$$\begin{aligned} \bar{t}_N(\pi) &= P_{\mu_0} \cdots P_{\mu_{N-2}} \bar{\tau}(\mu_{N-1}) \\ &\quad + P_{\mu_0} \cdots P_{\mu_{N-3}} \bar{\tau}(\mu_{N-2}) + \cdots + \bar{\tau}(\mu_0). \end{aligned}$$

Note that the i th component of the vector $\bar{t}_N(\pi)$ is $E\{t_N \mid x_0 = i, \pi\}$, the expected value of the completion time of the N th transition when the initial state is i and π is used. Note also that equality holds in Eq. (5.45)

if $\mu_k(i)$ attains the minimum in Eq. (5.43) for all k and i . It can be seen that

$$(T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} h)(i) = E \left\{ h(x_N) + \int_0^{t_N} g(x(t), u(t)) dt \mid x_0 = i, \pi \right\}.$$

Using this relation in Eq. (5.45) and dividing by $E\{t_N \mid x_0 = i, \pi\}$, we obtain for all i

$$\begin{aligned} \frac{E\{h(x_N) \mid x_0 = i, \pi\}}{E\{t_N \mid x_0 = i, \pi\}} + \frac{E\left\{\int_0^{t_N} g(x(t), u(t)) dt \mid x_0 = i, \pi\right\}}{E\{t_N \mid x_0 = i, \pi\}} \\ \geq \lambda + \frac{h(i)}{E\{t_N \mid x_0 = i, \pi\}}. \end{aligned}$$

Taking the limit as $N \rightarrow \infty$ and using the fact $\lim_{N \rightarrow \infty} E\{t_N \mid x_0 = i, \pi\} = \infty$ [cf. Eq. (5.36)], we see that

$$\lim_{N \rightarrow \infty} \frac{E\left\{\int_0^{t_N} g(x(t), u(t)) dt \mid x_0 = i, \pi\right\}}{E\{t_N \mid x_0 = i, \pi\}} = J_\pi(i) \geq \lambda, \quad i = 1, \dots, n,$$

with equality if $\mu_k(i)$ attains the minimum in Eq. (5.43) for all k and i . Q.E.D.

By combining Prop. 5.3.1 with Eq. (5.42), we obtain the following:

Proposition 5.3.2: Let μ be a unichain policy. Then:

- (a) There exists a scalar λ_μ and a vector h_μ such that

$$J_\mu(i) = \lambda_\mu, \quad i = 1, \dots, n, \quad (5.46)$$

and

$$h_\mu(i) = G(i, \mu(i)) - \lambda_\mu \bar{\tau}_i(\mu(i)) + \sum_{j=1}^n p_{ij}(\mu(i)) h_\mu(j), \quad i = 1, \dots, n. \quad (5.47)$$

- (b) Let t be a fixed state. The system of the $n+1$ linear equations

$$h(i) = G(i, \mu(i)) - \lambda_\mu \bar{\tau}_i(\mu(i)) + \sum_{j=1}^n p_{ij}(\mu(i)) h(j), \quad i = 1, \dots, n, \quad (5.48)$$

$$h(t) = 0, \quad (5.49)$$

in the $n+1$ unknowns $\lambda, h(1), \dots, h(n)$ has a unique solution.

Proof: Part (a) follows from Prop. 5.3.1 and Eq. (5.42). The proof of part (b) is identical to the proof of Prop. 4.2.5(b). **Q.E.D.**

To establish conditions under which there exists a solution (λ, h) to Bellman's equation (5.43), we formulate a corresponding discrete-time average cost problem. Let γ be any scalar such that

$$0 < \gamma < \frac{\bar{\tau}_i(u)}{1 - p_{ii}(u)}$$

for all i and $u \in U(i)$ with $p_{ii}(u) < 1$. Define also for all i and $u \in U(i)$,

$$\tilde{p}_{ij}(u) = \begin{cases} \frac{\gamma p_{ij}(u)}{\bar{\tau}_i(u)} & \text{if } j \neq i, \\ 1 - \frac{\gamma(1-p_{ii}(u))}{\bar{\tau}_i(u)} & \text{if } j = i. \end{cases} \quad (5.50)$$

It can be seen that we have for all i and j

$$0 \leq \tilde{p}_{ij}(u), \quad \sum_{j=1}^n \tilde{p}_{ij}(u) = 1,$$

$$\tilde{p}_{ij}(u) = 0 \quad \text{if and only if} \quad p_{ij}(u) = 0. \quad (5.51)$$

We view $\tilde{p}_{ij}(u)$ as the transition probabilities of the discrete-time average cost problem whose expected stage cost corresponding to (i, u) is

$$\tilde{G}(i, u) = \frac{G(i, u)}{\bar{\tau}_i(u)}. \quad (5.52)$$

We call this the *auxiliary discrete-time average cost problem*. The following proposition shows that this problem is essentially equivalent with our original semi-Markov average cost problem.

Proposition 5.3.3 If the scalar λ and the vector \tilde{h} satisfy

$$\tilde{h}(i) = \min_{u \in U(i)} \left[\tilde{G}(i, u) - \lambda + \sum_{j=1}^n \tilde{p}_{ij}(u) \tilde{h}(j) \right], \quad i = 1, \dots, n, \quad (5.53)$$

then λ and the vector h with components

$$h(i) = \gamma \tilde{h}(i), \quad i = 1, \dots, n, \quad (5.54)$$

satisfy Bellman's equation

$$h(i) = \min_{u \in U(i)} \left[G(i, u) - \lambda \bar{\tau}_i(u) + \sum_{j=1}^n p_{ij}(u) h(j) \right], \quad i = 1, \dots, n, \quad (5.55)$$

for the semi-Markov average cost problem.

Proof: By substituting Eqs. (5.50), (5.52), and (5.54) in Eq. (5.53), we obtain after a straightforward calculation

$$0 = \min_{u \in U(i)} \frac{1}{\bar{\tau}_i(u)} \left[G(i, u) - \lambda \bar{\tau}_i(u) + \sum_{j=1}^n p_{ij}(u) h(j) - h(i) \right], \quad i = 1, \dots, n.$$

This implies that the minimum of the expression within brackets in the right-hand side above is zero, which is equivalent to Bellman's equation (5.55). **Q.E.D.**

Note that in view of Eq. (5.51), the auxiliary discrete-time average cost problem and the semi-Markov average cost problem have the same probabilistic structure. In particular, if all stationary policies are unichain or the WA condition holds for one problem, the same is true for the other. Thus, the results and algorithms of Sections 4.2 and 4.3, when applied to the auxiliary discrete-time problem, yield results and algorithms for the semi-Markov problem. For example, value iteration, policy iteration, and linear programming can be applied to the auxiliary problem in order to solve the semi-Markov problem. In particular, we have the following sufficient condition for the optimal average cost to be independent of the initial state, based on the WA condition of Section 4.2.

Proposition 5.3.4: Consider the semi-Markov average cost problem, and assume either one of the following two conditions:

- (1) The set of states can be partitioned into two sets S_t and S_c such that all states in S_t are transient under every stationary policy, while for every two states $i, j \in S_c$, there exists a stationary policy μ (depending on i and j) such that, for some k ,

$$P(x_k = j \mid x_0 = i, \mu) > 0.$$

- (2) Every policy that is optimal within the class of stationary policies is unichain. Then the optimal average cost per stage has the same

value λ for all initial states i . Furthermore, λ together with a vector h satisfies Bellman's equation (5.55) for the semi-Markov average cost problem.

Proof: By Prop. 4.2.3 and 4.2.5, under either condition stated, Bellman's equation (5.53) for the auxiliary discrete-time average cost problem has a solution (λ, \tilde{h}) , from which a solution to Bellman's equation (5.55) can be extracted according to Prop. 5.3.3. **Q.E.D.**

5.4 NOTES, SOURCES, AND EXERCISES

The idea of using uniformization to convert continuous-time stochastic control problems involving Markov chains into discrete-time problems gained wide attention following the paper by Lippman [Lip75b]. For a more recent discussion, see Beutler and Ross [BeR87].

Control of queueing systems has been researched extensively. For additional material on the problem of control of arrival rate or service rate (cf. Examples 5.2.1 and 5.2.2), see Blanc et. al. [BWN92], Courcoubetis and Reiman [CoR87], Courcoubetis and Varaiya [CoV84], Rosberg, Varaiya, and Walrand [RVW82], Sobel [Sob82], Stidham and Prabhu [StP74], and Stidham [Sti85]. For more on priority assignment and routing (cf. Examples 5.2.3 and 5.2.4), see Baras, Dorsey, and Makowski [BDM83], Baras and Dorsey [BaD81], Bertsekas and Tsitsiklis [BeT89], Bhattacharya and Ephremides [BhE91], Courcoubetis and Varaiya [CoV84], Harrison [Har75a], [Har75b], Pattipati and Kleinman [PaK81], Suk and Cassandras [SuC91], Ayoun and Rosberg [AyR91], Cruz and Chuah [CrC91], Ephremides, Varaiya, and Walrand [EVW80], Ephremides and Verd'u [EpV89], Hajek [Haj84], Lin and Kumar [LiK84], Towsley, Sparaggis, and Cassandras [TSC92], Viniotis and Ephremides [ViE88], respectively.

Semi-Markov decision models were introduced by Jewell [Jew63] and are also discussed by Ross [Ros70].

E X E R C I S E S

5.1 (Proof of Validity of Uniformization)

Complete the details of the following argument, showing the validity of the uniformization procedure for the case of a finite number of states $i = 1, \dots, n$. We

fix a policy, and for notational simplicity we do not show the dependence of transition rates on the control. Let $p(t)$ be the row vector with coordinates

$$p_i(t) = P\{x(t) = i \mid x_0\}, \quad i = 1, \dots, n.$$

We have

$$dp(t)/dt = p(t)A,$$

where $p(0)$ is the row vector with i th coordinate equal to one if $x_0 = i$ and zero otherwise, and the matrix A has elements

$$a_{ij} = \begin{cases} \nu_i p_{ij} & \text{if } i \neq j, \\ -\nu_i & \text{if } i = j. \end{cases}$$

From this we obtain

$$p(t) = p(0)e^{At},$$

where

$$e^{At} = \sum_{k=0}^{\infty} \frac{(At)^k}{k!}.$$

Consider the transition probability matrix B of the uniform version

$$B = I + \frac{A}{\nu},$$

where $\nu \geq \nu_i$, $i = 1, \dots, n$. Consider also the following equation:

$$e^{At} = e^{-\nu t} e^{B\nu t} = e^{-\nu t} \sum_{k=0}^{\infty} \frac{(B\nu t)^k}{k!}.$$

Use these relations to write

$$p(t) = p(0) \sum_{k=0}^{\infty} \Gamma(k, t) B^k,$$

where

$$\Gamma(k, t) = \frac{(\nu t)^k}{k!} e^{-\nu t} = \text{Prob}\{k \text{ transitions occur between 0 and } t \text{ in the uniform Markov chain}\}.$$

Verify that for $i = 1, \dots, n$ we have $\overbrace{\hspace{1cm}}^{\sim} -$

$$p_i(t) = \text{Prob}\{x(t) = i \text{ in the uniform Markov chain}\}.$$

5.2

Consider the $M/M/1$ queueing problem with variable service rate (Example 5.2.1). Assume that no arrivals are allowed ($\lambda = 0$), and one can either serve a customer at rate μ or refuse service ($M = \{0, \mu\}$). Let the cost rates for customer waiting and service be $c(i) = ci$ and $q(\mu)$, respectively, with $q(0) = 0$.

- (a) Show that an optimal policy is to always serve an available customer if

$$\frac{q(\mu)}{\mu} \leq \frac{c}{\beta},$$

and to always refuse service otherwise.

- (b) Analyze the problem when the cost rate for waiting is instead $c(i) = ci^2$.

5.3

A person has an asset to sell for which she receives offers that can take one of n values. The times between successive offers are random, independent, and identically distributed with given distribution. Find the offer acceptance policy that maximizes $E\{\alpha^T s\}$, where T is the time of sale, s is the sale price, and $\alpha \in (0, 1)$ is a discount factor.

5.4

Analyze the priority assignment problem of Example 5.2.3 within the semi-Markov context of Section 5.3, assuming that the customer service times are independent but not exponentially distributed. Consider both the discounted and the average cost cases.

Approximate Dynamic Programming

Contents

6.1. Cost Approximation	p. 325
6.2. Approximate Policy Iteration – Direct Policy Evaluation	p. 329
6.2.1. Gradient Methods for Direct Policy Evaluation	p. 332
6.2.2. TD(λ)	p. 336
6.2.3. Optimistic Policy Iteration	p. 337
6.2.4. Approximate Policy Iteration Based on Q -Factors	p. 338
6.3. Indirect Methods for Policy Evaluation	p. 340
6.3.1. Policy Evaluation by Projected Value Iteration	p. 341
6.3.2. Least Squares Policy Evaluation (LSPE)	p. 346
6.3.3. PVI(λ) and LSPE(λ)	p. 348
6.3.4. The LSTD(λ) Algorithm	p. 355
6.3.5. The TD(λ) Algorithm	p. 357
6.3.6. Summary and Examples	p. 359
6.4. Q -Learning	p. 363
6.4.1. Q -Factor Approximations	p. 364
6.4.2. Q -Learning for Optimal Stopping Problems	p. 366
6.5. Stochastic Shortest Path Problems	p. 369
6.6. Average Cost Problems	p. 375
6.6.1. Approximate Policy Evaluation	p. 375
6.6.2. Approximate Policy Iteration	p. 386
6.6.3. Q -Learning for Average Cost Problems	p. 389
6.7. Approximation in Policy Space	p. 392
6.8. Notes, Sources, and Exercises	p. 399

In this chapter we consider approximation methods for challenging, computationally intensive DP problems. We discussed a number of such methods in Chapter 6 of Vol. I and Chapter I of the present volume, such as for example rollout and other one-step lookahead approaches. Here our focus will be on algorithms that are patterned after two principal methods of infinite horizon DP: policy and value iteration. These algorithms form the core of a methodology known as *approximate dynamic programming*, or *neuro-dynamic programming*, or *reinforcement learning*.

The policy and value iteration methods of the preceding chapters apply when there is a mathematical model of the transition probabilities and the cost structure of the system. In many problems, however, such a model may be hard to construct, but instead, the system and cost structure may be simulated (think, for example, of a queueing network with complicated but well-defined service disciplines at the queues). The assumption here is that there is a computer program that simulates, for a given control u , the probabilistic transitions from any given state i to a successor state j according to the transition probabilities $p_{ij}(u)$, and also generates a corresponding transition cost $g(i, u, j)$. It may then be possible to use repeated simulation to calculate (at least approximately) the transition probabilities of the system and the expected stage costs by averaging, and then to apply the methods discussed earlier.

The methods of this chapter, however, are geared towards an alternative possibility, which is much more attractive when one is faced with a large and complex system, and one contemplates approximations. Rather than estimate explicitly the transition probabilities and costs, we aim to approximate the cost function of a given policy or even the optimal cost-to-go function by generating one or more simulated system trajectories and associated costs, and by using some form of “least squares fit.” In another type of method, which we will discuss only briefly, we use a gradient method and simulation data to approximate directly an optimal policy with a policy of a given parametric form.

Our main focus will be on two types of methods: *policy evaluation algorithms*, which deal with approximation of the cost of a single policy, and *Q-learning algorithms*, which deal with approximation of the optimal cost. Let us summarize each type of method, focusing for concreteness on the finite-state discounted case.

Policy Evaluation Algorithms

With this class of methods, we aim to approximate the cost function $J_\mu(i)$ of a policy μ with a parametric architecture of the form $\tilde{J}(i, r)$, where r is a parameter vector (cf. Section 6.3.5 of Vol. I). This approximation may be carried out repeatedly, for a sequence of policies, in the context of a policy

iteration scheme. We discuss two types of methods:[†]

- (a) **Direct methods**, where we use simulation to collect samples of costs for various initial states, and fit the architecture \tilde{J} to the samples through some least squares problem. This problem may be solved by several possible algorithms, but we will focus on gradient methods. A certain type of gradient method, called *incremental*, has been used extensively, and will be described in some detail in Section 6.2.1. A particular implementation of the incremental gradient method uses the notion of *temporal difference* (TD for short), which provides a convenient way to simplify various formulas. The resulting method, called TD(1), motivates a broader class of methods, called TD(λ), that are parameterized by a scalar $\lambda \in [0, 1]$.
- (b) **Indirect methods**, where we obtain r by solving an approximate version of Bellman's equation

$$\tilde{J} = \tilde{T}\tilde{J},$$

where \tilde{T} is an approximation to the DP mapping T_μ , designed so that the above equation has a solution. We will focus exclusively on the case of a *linear architecture*, where \tilde{J} is of the form Φr , and Φ is a matrix whose columns can be viewed as basis functions (cf. Section 6.3.5 of Vol. I). In this case, we obtain the parameter vector r by solving the equation

$$\Phi r = \Pi T(\Phi r), \quad (6.1)$$

where Π denotes projection with respect to a suitable norm on the subspace of vectors of the form Φr , and T is either the mapping T_μ or a related mapping, which also has J_μ as its unique fixed point. We can view Eq. (6.1) as a form of *projected Bellman equation*. For a special choice of the norm of the projection, ΠT is a contraction mapping, so the projected Bellman equation has a unique solution Φr^* . We will discuss several iterative methods for finding r^* in Section 6.3. All these methods use simulation and can be shown to converge under reasonable assumptions to r^* , so they produce the same approximate cost function. However, they differ in their speed of convergence and in their suitability for various problem contexts. They all depend on a parameter $\lambda \in [0, 1]$, which enters in the definition of the mapping

[†] In a third type of method, often called the *Bellman equation error* approach, which we will not discuss here, the parameter vector r is determined by minimizing a measure of error in satisfying Bellman's equation; for example, by minimizing over r

$$\|\tilde{J} - T\tilde{J}\|,$$

where $\|\cdot\|$ is some norm. If $\|\cdot\|$ is a Euclidean norm, and $\tilde{J}(i, r)$ is linear in r , this minimization is a linear least squares problem.

\tilde{T} . Here are the methods that we will focus on in Section 6.3 for discounted problems, and also in Sections 6.5 and 6.6 for other types of problems.

- (1) $TD(\lambda)$ or *temporal differences method*. This algorithm descends from $TD(1)$, which as noted earlier, is a gradient method for solving least squares problems in the context of the direct approach. While $TD(\lambda)$ resembles gradient methods, when $\lambda < 1$ it is more properly viewed as a stochastic iterative method for solving the projected Bellman equation (6.1). $TD(\lambda)$ embodies important ideas and has played an important role in the development of the subject, but in practical terms, it is inferior to the next two methods, so it will be discussed in less detail.
- (2) $LSPE(\lambda)$ or *least squares policy evaluation method*. This algorithm is based on the idea of executing value iteration within the lower dimensional space spanned by the basis functions. It has the form

$$\Phi r_{k+1} = \Pi T(\Phi r_k) + \text{simulation noise}, \quad (6.2)$$

i.e., the current value iterate $T(\Phi r_k)$ is projected on S and is suitably approximated by simulation. The simulation noise tends to 0 asymptotically, so the method converges to the solution of the projected Bellman equation (6.1).

- (3) $LSTD(\lambda)$ or *least squares temporal differences method*. This algorithm computes and solves a progressively more refined simulation-based approximation to the projected Bellman equation (6.1). In fact the $LSPE(\lambda)$ iteration (6.2) can be viewed as a single-iteration attempt to solve the approximation to the projected Bellman equation solved by $LSTD(\lambda)$. The rate of convergence of $LSTD(\lambda)$ is comparable to that of $LSPE(\lambda)$, and its theory is relatively simple, so it will be discussed in less detail.

Q-Learning Algorithms

With this class of methods, we aim to compute, without any approximation, the optimal cost function (not just the cost function of a single policy).† The letter “*Q*” stands for nothing special - it just refers to notation used in the original proposal of this method!

Q-learning maintains and updates for each state-control pair (i, u) an estimate of the expression that is minimized in the right-hand side of

† Another class of methods, which aim to approximate the optimal cost function, is based on linear programming. They were discussed in Section 1.3.4, and will not be considered further here (see also the references cited in Chapter 1).

Bellman's equation. This is called the *Q-factor* of the pair (i, u) , and is denoted by $Q^*(i, u)$. The *Q*-factors are updated with what may be viewed as a simulation-based form of value iteration, as will be explained in Section 6.4. An important advantage of using *Q*-factors is that when they are available, they can be used to obtain an optimal control at any state i simply by minimizing $Q^*(i, u)$ over $u \in U(i)$, so the transition probabilities of the problem are not needed. In addition, for mathematical reasons to be explained later, there is no counterpart of *Q*-learning that can be used to calculate directly the (exact) optimal costs $J^*(i)$.

On the other hand, for problems involving a large number of state-control pairs, *Q*-learning is often impractical because there may be simply too many *Q*-factors to update. As a result, the algorithm is primarily suitable for systems with a small number of states (or for aggregated/few-state versions of more complex systems). There are also algorithms that use parametric approximations for the *Q*-factors (see Section 6.4.1), but they are either tailored to special classes of problems, or else they lack a firm theoretical basis. Still, however, these methods are used widely, and often with success.

Chapter Organization

Throughout this chapter, we will focus on perfect state information problems, involving a Markov chain with a finite number of states i , transition probabilities $p_{ij}(u)$, and single stage costs $g(i, u, j)$. We will consider first, in Sections 6.1-6.4, the discounted problem using the notation of Section 1.3. Section 6.1 provides a broad overview of cost approximation architectures and their uses. Section 6.2 focuses on approximate policy iteration, and direct methods for policy evaluation. Section 6.3 is a long section on the major indirect methods for policy evaluation. Section 6.4 discusses *Q*-learning and its variations. Extensions to stochastic shortest path and average cost problems are given in Sections 6.5 and 6.6, respectively. Methods involving the parametric approximation of policies are discussed in Section 6.7.

We note that the books by Bertsekas and Tsitsiklis [BeT96], and by Sutton and Barto [SuB98] provide a more detailed account of the subject, and consider more general problems. However, the present volume provides a different perspective and includes new research, performed after the appearance of these books.

6.1 COST APPROXIMATION

The major use of cost approximation is for obtaining a one-step lookahead suboptimal policy (cf. Section 6.3 of Vol. I). In particular, suppose that we use $\tilde{J}(j, r)$ as an approximation to the optimal cost of the finite-state

discounted problem of Section 1.3. Here \tilde{J} is a function of some chosen form (the approximation architecture) and r is a parameter/weight vector. Once r is determined, it yields a suboptimal control at any state i via the minimization

$$\tilde{\mu}(i) = \arg \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u)(g(i, u, j) + \alpha \tilde{J}(j, r)). \quad (6.3)$$

An alternative possibility is to obtain a parametric approximation $\tilde{Q}(i, u, r)$ of the Q -factor of the pair (i, u) , defined in terms of the optimal cost function J^* as

$$Q^*(i, u) = \sum_{j=1}^n p_{ij}(u)(g(i, u, j) + \alpha J^*(j)).$$

Since $Q^*(i, u)$ is the expression minimized in Bellman's equation, given the approximation $\tilde{Q}(i, u, r)$, we can generate a suboptimal control at any state i via

$$\tilde{\mu}(i) = \arg \min_{u \in U(i)} \tilde{Q}(i, u, r).$$

The advantage of using Q -factors is that the transition probabilities $p_{ij}(u)$ are not needed in the above minimization; this is in contrast with the minimization (6.3).

Note that we may similarly use approximations to the cost functions J_μ and Q -factors $Q_\mu(i, u)$ of specific policies μ . A major use of such approximations is in the context of an approximate policy iteration scheme; see Section 6.2.

The choice of architecture is very significant for the success of the approximation approach. One possibility is to use the *linear* form

$$\tilde{J}(i, r) = \sum_{k=1}^s r_k \phi_k(i), \quad (6.4)$$

where $r = (r_1, \dots, r_s)$ is the parameter vector, and $\phi_k(i)$ are some known scalars that depend on the state i . This amounts to cost function approximation by a vector in the subspace

$$S = \{\Phi r \mid r \in \mathbb{R}^s\},$$

where Φ is the $n \times s$ matrix whose columns are the vectors

$$(\phi_k(1), \dots, \phi_k(n))', \quad k = 1, \dots, s.$$

We can view these columns as basis functions, and we can view Φr as a linear combination of basis functions.

For each state i , the approximate cost $\tilde{J}(i, r)$ is the inner product $\phi(i)'r$ of r and

$$\phi(i) = (\phi_1(i), \dots, \phi_s(i))'.$$

The components of $\phi(i)$ are called *features*, and $\phi(i)$ is called the *feature vector* of i . For example, in computer chess (Section 6.3.5 of Vol. I) where the state is the current board position, appropriate features are material balance, piece mobility, king safety, and other positional factors. Features, when well-crafted, can capture the dominant nonlinearities of the cost function, and their linear combination may work very well as an approximation architecture (see Fig. 6.1.1).

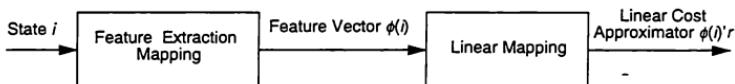


Figure 6.1.1. A linear feature-based architecture. It combines a mapping that extracts the feature vector $\phi(i) = (\phi_1(i), \dots, \phi_s(i))'$ associated with state i , and a parameter vector r to form a linear cost approximator.

Example 6.1.1 (Polynomial Approximations)

An important example of linear cost approximation is based on polynomial basis functions. Suppose that the state consists of q integer components x_1, \dots, x_q , each taking values within some limited range of integers. For example, in a queueing system, x_k may represent the number of customers in the k th queue, where $k = 1, \dots, q$. Suppose that we want to use an approximating function that is quadratic in the components x_k . Then we can define a total of $1 + q + q^2$ basis functions that depend on the state $x = (x_1, \dots, x_q)$ via

$$\phi_0(x) = 1, \quad \phi_k(x) = x_k, \quad \phi_{km}(x) = x_k x_m, \quad k, m = 1, \dots, q.$$

A linear approximation architecture that uses these functions is given by

$$\tilde{J}(x, r) = r_0 + \sum_{k=1}^q r_k x_k + \sum_{k=1}^q \sum_{m=k}^q r_{km} x_k x_m,$$

where the parameter vector r has components r_0 , r_k , and r_{km} , with $k = 1, \dots, q$, $m = k, \dots, q$. In fact, any kind of approximating function that is polynomial in the components x_1, \dots, x_q can be constructed similarly.

There are also interesting nonlinear approximation architectures, including those defined by neural networks, perhaps in combination with

feature extraction mappings (see Bertsekas and Tsitsiklis [BeT96], or Sutton and Barto [SuB98] for further discussion). In this chapter, we will not address the choice of the structure of $\tilde{J}(i, r)$ or associated basis functions, but rather focus on various iterative algorithms for obtaining a suitable parameter vector r . However, we will mostly focus on the case of linear architectures, because many of the policy evaluation algorithms of this chapter are guaranteed to converge only for that case.

Direct and Indirect Approximation Approaches

Let us consider the problem of approximating the cost function of a fixed stationary policy μ within the subspace $S = \{\Phi r \mid r \in \mathbb{R}^s\}$. The most straightforward approach to approximate J_μ , referred to as *direct*, is to find an approximation $\tilde{J} \in S$ that matches best J_μ in some normed error sense, i.e.,

$$\min_{\tilde{J} \in S} \|J_\mu - \tilde{J}\|,$$

or equivalently,

$$\min_{r \in \mathbb{R}^s} \|J_\mu - \Phi r\|$$

(see the left-hand side of Fig. 6.1.2). Here, $\|\cdot\|$ is usually some (possibly weighted) Euclidean norm, in which case the approximation problem is a linear least squares problem, whose solution, denoted r^* , can in principle be obtained in closed form by solving the associated quadratic minimization problem. If the matrix Φ has linearly independent columns, the solution is unique and can also be represented as

$$\Phi r^* = \Pi \tilde{J},$$

where Π denotes projection with respect to $\|\cdot\|$ on the subspace S .† A major difficulty is that specific cost function values $J_\mu(i)$ can only be estimated through their simulation-generated cost samples, hence the need for stochastic iterative methods to carry out the above minimization. We will discuss some methods of this type in the next section.

An alternative approach, referred to as *indirect*, is to try to approximate the solution of Bellman's equation $J = T_\mu J$ on the subspace S (see the right-hand side of Fig. 6.1.2). This leads to the problem of finding a vector r^* such that

$$\Phi r^* = \Pi T_\mu(\Phi r^*). \quad (6.5)$$

We can view this equation as a *projected form of Bellman's equation*.

† In what follows in this chapter, we will not distinguish between the linear operation of projection and the corresponding matrix representation, denoting them both by Π . The meaning should be clear from the context.

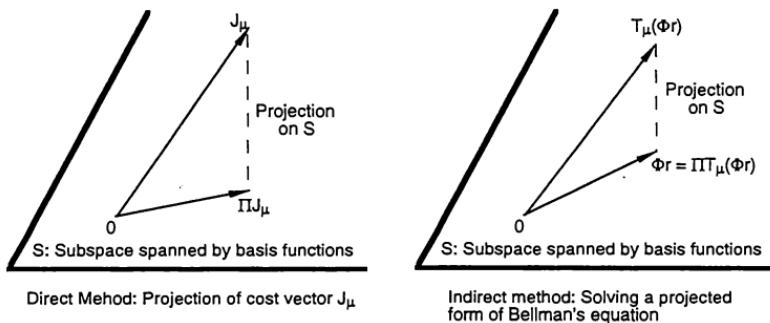


Figure 6.1.2. Two methods for approximating the cost function J_μ as a linear combination of basis functions (subspace S). In the direct method (figure on the left), J_μ is projected on S . In the indirect method (figure on the right), the approximation is found by solving $\Phi r = \Pi T_\mu(\Phi r)$, a projected form of Bellman's equation.

An important fact here is that ΠT_μ is a contraction, provided we use a special weighted Euclidean norm for projection, as will be proved in Section 6.3 for discounted problems (Prop. 6.3.1). In this case, Eq. (6.5) has a unique solution, leading to algorithms such as LSPE(λ) and LSTD(λ), which are discussed in Sections 6.3.1 and 6.3.4, respectively. Unfortunately, the contraction property of ΠT_μ does not extend to the case where T_μ is replaced by T , the DP mapping corresponding to multiple/all policies, although there are some interesting exceptions.

6.2 APPROXIMATE POLICY ITERATION – DIRECT POLICY EVALUATION

In this section, we consider a form of approximate policy iteration, where we compute simulation-based approximations $\tilde{J}(\cdot, r)$ to the cost functions J_μ of stationary policies μ , and we use them to compute new policies based on (approximate) policy improvement. We impose no constraints on the approximation architecture, so $\tilde{J}(i, r)$ may be linear or nonlinear in r .

The method is illustrated in Fig. 6.2.1. Its theoretical basis was discussed in Section 1.3 (cf. Prop. 1.3.6), where it was shown that if the policy evaluation is accurate to within ϵ (in the sup-norm sense), then for an α -discounted problem, the method will yield in the limit (after infinitely many policy evaluations) a stationary policy that is optimal to within

$$\frac{\epsilon}{(1 - \alpha)^2},$$

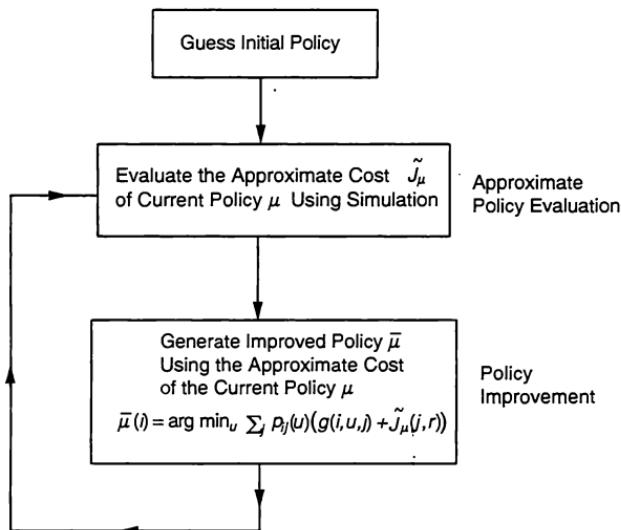


Figure 6.2.1 Block diagram of approximate policy iteration.

where α is the discount factor. Experimental evidence indicates that this bound is usually conservative. Furthermore, often just a few policy evaluations are needed before the bound is attained.

The policy evaluation portion of approximate policy iteration can be implemented with either the direct or the indirect approach (cf. the discussion in the preceding section and Fig. 6.1.2). However, in this section we will focus exclusively on the direct approach. In particular, suppose that the current policy is μ , and for a given r , $\tilde{J}(i, r)$ is an approximation of $J_\mu(i)$. We generate an “improved” policy $\bar{\mu}$ using the formula

$$\bar{\mu}(i) = \arg \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u)(g(i, u, j) + \alpha \tilde{J}(j, r)), \quad \text{for all } i. \quad (6.6)$$

To evaluate approximately $J_{\bar{\mu}}$, we select a subset of “representative” states \tilde{S} (perhaps chosen in the course of a simulation), and for each $i \in \tilde{S}$, we obtain $M(i)$ samples of the cost $J_{\bar{\mu}}(i)$. The m th such sample is denoted by $c(i, m)$, and mathematically, it can be viewed as being $J_{\bar{\mu}}(i)$ plus some simulation error/noise. Then we obtain the corresponding parameter vector \bar{r} by solving the following least squares problem

$$\min_{\bar{r}} \sum_{i \in \tilde{S}} \sum_{m=1}^{M(i)} (\tilde{J}(i, \bar{r}) - c(i, m))^2, \quad (6.7)$$

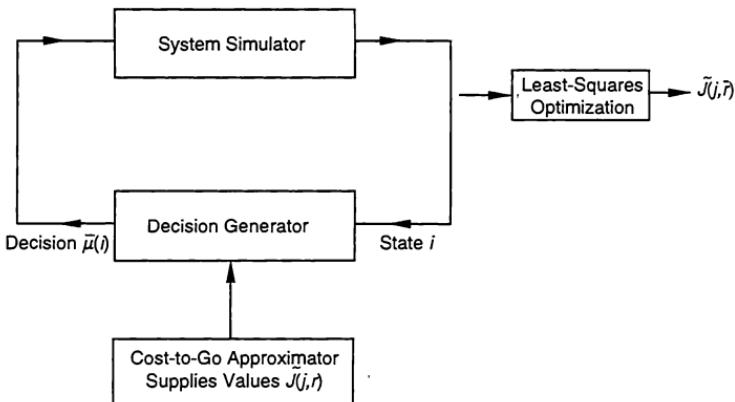


Figure 6.2.2 Structure of approximate policy iteration algorithm with direct policy evaluation. It illustrates how given the approximation $\tilde{J}(i, r)$, we generate cost samples of the “improved” policy $\bar{\mu}$ by simulation (the “decision generator” module). We use these samples to generate by least squares the approximator $\tilde{J}(i, \bar{r})$ of $\bar{\mu}$.

and we repeat the process with $\bar{\mu}$ and \bar{r} replacing μ and r , respectively (see Fig. 6.2.1).

A possible implementation of the algorithm is illustrated in Fig. 6.2.2. It consists of four modules:

- The *simulator*, which given a state-control pair (i, u) , generates the next state j according to the system’s transition probabilities.
- The *decision generator*, which generates the control $\bar{\mu}(i)$ of the improved policy at the current state i [cf. Eq. (6.6)] for use in the simulator.
- The *cost-to-go approximator*, which is the function $\tilde{J}(j, r)$ that is used by the decision generator in the minimization of Eq. (6.6).
- The *least squares optimizer*, which accepts as input the cost samples $c(i, m)$ produced by the simulator and solves the problem (6.7) to obtain the approximation $\tilde{J}(\cdot, \bar{r})$ of the cost of $\bar{\mu}$.

Note that there are two policies μ and $\bar{\mu}$, and parameter vectors r and \bar{r} , which are simultaneously involved in this algorithm. In particular, r corresponds to the current policy μ , and the approximation $\tilde{J}(\cdot, r)$ is used in the policy improvement Eq. (6.6) to generate the new policy $\bar{\mu}$. At the same time, $\bar{\mu}$ drives the simulation to generate the cost samples $c(i, m)$ for the least squares minimization (6.7), which determines the parameter \bar{r} that corresponds to $\bar{\mu}$ and will be used in the next policy iteration.

The Issue of Exploration

We will discuss shortly methods for solving the least squares problem (6.7), but before doing so, we note a generic difficulty with simulation-based policy iteration: to evaluate a policy μ , we need to generate cost samples using that policy, but this biases the simulation by underrepresenting states that are unlikely to occur under μ . As a result, the cost-to-go estimates of these underrepresented states may be highly inaccurate, causing potentially serious errors in the calculation of the improved control policy $\bar{\mu}$ via the policy improvement Eq. (6.6).

The difficulty just described is known as *inadequate exploration* of the system's dynamics because of the use of a fixed policy. It is a particularly acute difficulty when the system is deterministic, or when the randomness embodied in the transition probabilities is "relatively small." One possibility for guaranteeing adequate exploration of the state space is to frequently restart the simulation and to ensure that the initial states employed form a rich and representative subset. Another possibility is to artificially introduce some extra randomization in the simulation, by occasionally generating transitions that use a randomly selected control rather than the one dictated by the policy μ .

6.2.1 Gradient Methods for Direct Policy Evaluation

We will now consider specific algorithms for finding an approximation $\tilde{J}(\cdot, \bar{r})$ of the cost function $J_{\bar{\mu}}$ of the "improved" policy $\bar{\mu}$ defined by Eq. (6.6). In particular, we will discuss gradient-like methods for formulating and solving the least squares problem (6.7).

Batch Gradient Methods for Policy Evaluation

Let us focus on an N -transition portion (i_0, \dots, i_N) of a simulated trajectory, also called a *batch*. We view the numbers

$$\sum_{t=k}^{N-1} \alpha^{t-k} g(i_t, \bar{\mu}(i_t), i_{t+1}), \quad k = 0, \dots, N-1,$$

as cost samples, one per initial state i_0, \dots, i_{N-1} , which can be used for least squares approximation of the parametric architecture $\tilde{J}(i, \bar{r})$ [cf. Eq. (6.7)]:

$$\min_{\bar{r}} \sum_{k=0}^{N-1} \frac{1}{2} \left(\tilde{J}(i_k, \bar{r}) - \sum_{t=k}^{N-1} \alpha^{t-k} g(i_t, \bar{\mu}(i_t), i_{t+1}) \right)^2. \quad (6.8)$$

One way to solve this least squares problem is to use a gradient method, whereby the parameter \bar{r} associated with $\bar{\mu}$ is updated at time N by

$$\bar{r} := \bar{r} - \gamma \sum_{k=0}^{N-1} \nabla \tilde{J}(i_k, \bar{r}) \left(\tilde{J}(i_k, \bar{r}) - \sum_{t=k}^{N-1} \alpha^{t-k} g(i_t, \bar{\mu}(i_t), i_{t+1}) \right). \quad (6.9)$$

Here, $\nabla \tilde{J}$ denotes gradient with respect to \bar{r} and γ is a positive stepsize, which is usually diminishing over time (we leave its precise choice open for the moment). Each of the N terms in the summation in the right-hand side above is the gradient of a corresponding term in the least squares summation of problem (6.8). Note that the update of \bar{r} is done after processing the entire batch, and that the gradients $\nabla \tilde{J}(i_k, \bar{r})$ are evaluated at the preexisting value of \bar{r} , i.e., the one before the update.

In a traditional gradient method, the gradient iteration (6.9) is repeated, until convergence to the solution of the least squares problem (6.8), i.e., a single N -transition batch is used. However, there is an important tradeoff relating to the size N of the batch: in order to reduce simulation error and generate multiple cost samples for a representatively large subset of states, it is necessary to use a large N , yet to keep the work per gradient iteration small it is necessary to use a small N .

To address the issue of size of N , an expanded view of the gradient method is preferable in practice, whereby batches may be changed after one or more iterations. Thus, in this more general method, the N -transition batch used in a given gradient iteration comes from a potentially longer simulated trajectory, or from one of many simulated trajectories. A sequence of gradient iterations is performed, with each iteration using cost samples formed from batches collected in a variety of different ways and whose length N may vary. Batches may also overlap to a substantial degree.

We leave the method for generating simulated trajectories and forming batches open for the moment, but we note that it influences strongly the result of the corresponding least squares optimization (6.7), providing better approximations for the states that arise most frequently in the batches used. This is related to the issue of ensuring that the state space is adequately “explored,” with an adequately broad selection of states being represented in the least squares optimization, cf. our earlier discussion on the exploration issue.

The gradient method (6.9) is simple, widely known, and easily understood. There are extensive convergence analyses of this method and its variations, for which we refer to the literature cited at the end of the chapter. These analyses often involve considerable mathematical sophistication, particularly when multiple batches are involved, because of the stochastic nature of the simulation and the complex correlations between the cost samples. However, qualitatively, the conclusions of these analyses are consistent among themselves as well as with practical experience, and indicate that:

- (1) Under some reasonable technical assumptions, convergence to a limiting value of \bar{r} that is a local minimum of the associated optimization problem is expected.
- (2) For convergence, it is essential to gradually reduce the stepsize to 0, the most popular choice being to use a stepsize proportional to $1/m$,

while processing the m th batch. In practice, considerable trial and error may be needed to settle on an effective stepsize choice method. Sometimes it is possible to improve performance by using a different stepsize (or scaling factor) for each component of the gradient.

- (3) The rate of convergence is often very slow, and depends among other things on the initial choice of \bar{r} , the number of states and the dynamics of the associated Markov chain, the level of simulation error, and the method for stepsize choice. In fact, the rate of convergence is sometimes so slow, that practical convergence is infeasible, even if theoretical convergence is guaranteed.

Incremental Gradient Methods for Policy Evaluation

We will now consider a variant of the gradient method called *incremental*. This method can also be described through the use of N -transition batches, but we will see that (contrary to the batch version discussed earlier) the method is suitable for use with very long batches, including the possibility of a single very long simulated trajectory, viewed as a single batch.

For a given N -transition batch (i_0, \dots, i_N) , the batch gradient method processes the N transitions all at once, and updates \bar{r} using Eq. (6.9). The incremental method updates \bar{r} a total of N times, once after each transition. Each time it adds to \bar{r} the corresponding portion of the gradient in the right-hand side of Eq. (6.9) that can be calculated using the newly available simulation data. Thus, after each transition (i_k, i_{k+1}) :

- (1) We evaluate the gradient $\nabla \tilde{J}(i_k, \bar{r})$ at the current value of \bar{r} .
- (2) We sum all the terms in the right-hand side of Eq. (6.9) that involve the transition (i_k, i_{k+1}) , and we update \bar{r} by making a correction along their sum:

$$\bar{r} := \bar{r} - \gamma \left(\nabla \tilde{J}(i_k, \bar{r}) \tilde{J}(i_k, \bar{r}) - \left(\sum_{t=0}^k \alpha^{k-t} \nabla \tilde{J}(i_t, \bar{r}) \right) g(i_k, \bar{\mu}(i_k), i_{k+1}) \right). \quad (6.10)$$

By adding the parenthesized “incremental” correction terms in the above iteration, we see that after N transitions, all the terms of the batch iteration (6.9) will have been accumulated, but there is a difference: in the incremental version, \bar{r} is changed during the processing of the batch, and the gradient $\nabla \tilde{J}(i_t, \bar{r})$ is evaluated at the most recent value of \bar{r} [after the transition (i_t, i_{t+1})]. By contrast, in the batch version these gradients are evaluated at the value of \bar{r} prevailing at the beginning of the batch. Note that the gradient sum in the right-hand side of Eq. (6.10) can be conveniently updated following each transition, thereby resulting in an efficient implementation.

It can now be seen that because \bar{r} is updated at intermediate transitions within a batch (rather than at the end of the batch), the location of the end of the batch becomes less relevant. It is thus possible to have very long batches, and indeed the algorithm can be operated with a single very long simulated trajectory and a single batch. In this case, for each state i , we will have one cost sample for every time when state i is encountered in the simulation. Accordingly state i will be weighted in the least squares optimization in proportion to the frequency of its occurrence within the simulated trajectory.

Generally, within the least squares/policy evaluation context of this section, the incremental versions of the gradient methods can be implemented more flexibly and tend to converge faster than their batch counterparts, so they will be adopted as the default in our discussion. The book by Bertsekas and Tsitsiklis [BeT96] contains an extensive analysis of the theoretical convergence properties of incremental gradient methods (they are fairly similar to those of batch methods), and provides some insight into the reasons for their superior performance relative to the batch versions; see also the author’s nonlinear programming book [Ber99] (Section 1.5.2), and the paper by Bertsekas and Tsitsiklis [BeT00]. Still, however, the rate of convergence can be very slow.

Implementation Using Temporal Differences – TD(1)

We now introduce an alternative, mathematically equivalent, implementation of the batch and incremental gradient iterations (6.9) and (6.10), which is described with cleaner formulas. It uses the notion of *temporal difference* (TD for short) given by

$$d_k = g(i_k, \bar{\mu}(i_k), i_{k+1}) + \alpha \tilde{J}(i_{k+1}, \bar{r}) - \tilde{J}(i_k, \bar{r}), \quad k = 0, \dots, N-2, \quad (6.11)$$

$$d_{N-1} = g(i_{N-1}, \bar{\mu}(i_{N-1}), i_N) - \tilde{J}(i_{N-1}, \bar{r}). \quad (6.12)$$

In particular, by noting that the parenthesized term multiplying $\nabla \tilde{J}(i_k, \bar{r})$ in Eq. (6.9) is equal to

$$-(d_k + \alpha d_{k+1} + \dots + \alpha^{N-1-k} d_{N-1}),$$

we can verify by adding the equations below that iteration (6.9) can also be implemented as follows:

After the state transition (i_0, i_1) , set

$$\bar{r} := \bar{r} + \gamma d_0 \nabla \tilde{J}(i_0, \bar{r}).$$

After the state transition (i_1, i_2) , set

$$\bar{r} := \bar{r} + \gamma d_1 (\alpha \nabla \tilde{J}(i_0, \bar{r}) + \nabla \tilde{J}(i_1, \bar{r})).$$

Proceeding similarly, after the state transition (i_{N-1}, t) , set

$$\bar{r} := \bar{r} + \gamma d_{N-1} (\alpha^{N-1} \nabla \tilde{J}(i_0, \bar{r}) + \alpha^{N-2} \nabla \tilde{J}(i_1, \bar{r}) + \cdots + \nabla \tilde{J}(i_{N-1}, \bar{r})).$$

The batch version (6.9) is obtained if the gradients $\nabla \tilde{J}(i_k, \bar{r})$ are all evaluated at the value of \bar{r} that prevails at the beginning of the batch. The incremental version (6.10) is obtained if each gradient $\nabla \tilde{J}(i_k, \bar{r})$ is evaluated at the value of \bar{r} that prevails when the transition (i_k, i_{k+1}) is processed.

In particular, for the incremental version, we start with some vector r_0 , and following the transition (i_k, i_{k+1}) , $k = 0, \dots, N-1$, we set

$$r_{k+1} = r_k + \gamma_k d_k \sum_{t=0}^k \alpha^{k-t} \nabla \tilde{J}(i_t, r_t). \quad (6.13)$$

This algorithm is known as TD(1). In the important case of a linear approximation architecture of the form

$$\tilde{J}(i, r) = \phi(i)' r, \quad i = 1, \dots, n,$$

where $\phi(i) \in \mathbb{R}^s$ are some fixed vectors, TD(1) takes the form

$$r_{k+1} = r_k + \gamma_k d_k \sum_{t=0}^k \alpha^{k-t} \phi(i_t).$$

Note that in the incremental version it is possible to vary the stepsize γ_k from one transition to the next.

6.2.2 TD(λ)

A variant of TD(1) is TD(λ), which uses a parameter $\lambda \in [0, 1]$ in the formula (6.13). For $k = 0, \dots, N-1$, following the state transition (i_k, i_{k+1}) , it sets

$$r_{k+1} = r_k + \gamma_k d_k \sum_{t=0}^k (\alpha \lambda)^{k-t} \nabla \tilde{J}(i_t, r_t), \quad (6.14)$$

where d_k is the TD given by Eqs. (6.11), (6.12).

At this point it may be best to view TD(λ) simply as an approximation to TD(1), a gradient-type method that has solid convergence properties. It is natural to expect that for $\lambda \approx 1$, convergence would be guaranteed, and that the limit, while dependent on λ , should be close to the limit of TD(1).

The main advantage of using $\lambda < 1$ is that the sum in Eq. (6.14) exhibits less dependence on earlier states, and generally has smaller variance than in the case where $\lambda = 1$. This in turn enhances the practical

convergence properties of the method (even though it still tends to be slow for any value of λ). On the other hand, there are disadvantages to using $\lambda < 1$. In particular, it turns out that as λ is reduced towards 0, the quality of approximation provided by the method tends to deteriorate. We will discuss this fact and we will appropriately reinterpret $\text{TD}(\lambda)$ in the next section.

Consider now the incremental version of $\text{TD}(\lambda)$ in conjunction with a single infinitely long simulated trajectory (i_0, i_1, \dots) , and a linear approximation architecture of the form

$$\tilde{J}(i, r) = \phi(i)'r, \quad i = 1, \dots, n,$$

where $\phi(i)$ is the feature vector of state i . We note that this is the only case of $\text{TD}(\lambda)$ for which there is convergence analysis when $\lambda < 1$; see the discussion in Section 6.3.5 and the references at the end of the chapter.

Since $\nabla \tilde{J}(i_t, r_t) = \phi(i_t)$, the algorithm becomes

$$r_{k+1} = r_k + \gamma_k d_k \sum_{t=0}^k (\alpha\lambda)^{k-t} \phi(i_t), \quad (6.15)$$

where γ_k is the stepsize used following the transition (i_k, i_{k+1}) , and d_k is the TD given by [cf. Eq. (6.11)]

$$d_k = g(i_k, \bar{\mu}(i_k), i_{k+1}) + \alpha\phi(i_{k+1})'r_k - \phi(i_k)'r_k, \quad k = 0, 1, \dots$$

A more convenient formula is obtained if we introduce the vector

$$z_k = \sum_{t=0}^k (\alpha\lambda)^{k-t} \phi(i_t),$$

which is updated according to

$$z_{k+1} = \alpha\lambda z_k + \phi(i_{k+1}). \quad (6.16)$$

Then $\text{TD}(\lambda)$ takes the simple form

$$r_{k+1} = r_k + \gamma_k d_k z_k. \quad (6.17)$$

6.2.3 Optimistic Policy Iteration

In the approximate policy iteration approach discussed so far, the least squares optimization used for evaluation of the cost of the improved policy $\bar{\mu}$ must be solved completely for the vector \bar{r} . An alternative, known as *optimistic policy iteration*, is to solve this problem approximately and

replace the policy μ with the policy $\bar{\mu}$ after only a few simulation samples have been processed. An extreme possibility is to replace μ with $\bar{\mu}$ at the end of each state transition of a single infinitely long simulated trajectory (i_0, i_1, \dots) , as in the following optimistic version of $\text{TD}(\lambda)$, which is the same as the one of Eq. (6.14), except that the policy is updated with each transition.

After the state transition (i_k, i_{k+1}) , set

$$r_{k+1} = r_k + \gamma_k d_k \sum_{t=0}^k (\alpha\lambda)^{k-t} \nabla \tilde{J}(i_t, r_t), \quad (6.18)$$

and generate the next transition (i_{k+1}, i_{k+2}) by simulation using the control

$$\bar{\mu}(i_{k+1}) = \arg \min_{u \in U(i_{k+1})} \sum_{j=1}^n p_{i_{k+1}j}(u) (g(i_{k+1}, u, j) + \alpha \tilde{J}(j, r_{k+1})). \quad (6.19)$$

For $\lambda = 0$, we obtain optimistic $\text{TD}(0)$, which has the simple form

$$r_{k+1} = r_k + \gamma_k d_k \nabla \tilde{J}(i_k, r_k) \quad (6.20)$$

[cf. Eq. (6.18)].

Optimistic policy iteration is discussed extensively in the book by Bertsekas and Tsitsiklis [BeT96], together with several variants. It has been successfully used, among others, in an impressive backgammon application (Tesauro [Tes92]). However, the associated theoretical convergence properties are not fully understood. As shown in Section 6.4.2 of [BeT96], optimistic policy iteration can exhibit fascinating and counterintuitive behavior, including a natural tendency for a phenomenon called *chattering*, whereby the generated parameter sequence $\{r_k\}$ converges, while the generated policy sequence oscillates because the limit of $\{r_k\}$ corresponds to multiple policies.

We note that optimistic policy iteration tends to deal better with the problem of exploration discussed earlier, because with rapid changes of policy, there is less tendency to bias the simulation towards particular states that are favored by a particular policy. Nonetheless, some enforced exploration in the form of occasional use of randomly chosen control actions may be helpful.

6.2.4 Approximate Policy Iteration Based on Q -Factors

The approximate policy iteration method discussed so far relies on the calculation of the approximation $\tilde{J}(\cdot, r)$ to the cost function J_μ of the current policy, which is then used for policy improvement using the minimization

$$\bar{\mu}(i) = \arg \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u) (g(i, u, j) + \alpha \tilde{J}(j, r)).$$

Carrying out this minimization requires knowledge of the transition probabilities $p_{ij}(u)$ and calculation of the associated expected values for all controls $u \in U(i)$ (otherwise a time-consuming simulation of these expected values is needed). An interesting alternative is to compute *approximate Q-factors*

$$\tilde{Q}(i, u, r) \approx \sum_{j=1}^n p_{ij}(u)(g(i, u, j) + \alpha J_\mu(j)), \quad (6.21)$$

and use the minimization

$$\bar{\mu}(i) = \arg \min_{u \in U(i)} \tilde{Q}(i, u, r) \quad (6.22)$$

for policy improvement. Here, r is an adjustable parameter vector and $\tilde{Q}(i, u, r)$ is a parametric architecture, possibly of the linear form

$$\tilde{Q}(i, u, r) = \sum_{k=1}^m r_k \phi_k(i, u),$$

where $\phi_k(i, u)$ are basis functions that depend on both state and control [cf. Eq. (6.4)].

The important point here is that given the current policy μ , we can construct *Q-factor approximations* $\tilde{Q}(i, u, r)$ using any method for constructing cost approximations $\tilde{J}(i, r)$, including the versions of TD(λ) discussed earlier. The way to do this is to apply the latter method to the Markov chain whose states are the pairs (i, u) , and the probability of transition from (i, u) to (j, v) is

$$p_{ij}(u) \quad \text{if } v = \mu(j),$$

and is 0 otherwise. This is the probabilistic mechanism by which state-control pairs evolve under the stationary policy μ .

A major concern with this approach is that the state-control pairs (i, u) with $u \neq \mu(i)$ are never generated in this Markov chain, so they are not represented in the cost samples used to construct the approximation $\tilde{Q}(i, u, r)$. This creates an acute difficulty due to diminished exploration, as discussed earlier. To address this difficulty, it is common to use a form of randomization to modify the preceding transition probabilities by using a small positive parameter ϵ , so that a more representative collection of state-control pairs are generated during the simulation. In this modified Markov chain, at the typical state i , the control $\mu(i)$ is applied with probability $1 - \epsilon$ and a randomly chosen control u from the feasible set is applied with probability ϵ .

We will return to the use of *Q-factors* in Section 6.4, where we will discuss optimistic versions of TD(0), as well as exact and approximate implementations of the *Q*-learning algorithm.

6.3 INDIRECT METHODS FOR POLICY EVALUATION

The policy evaluation method discussed in the preceding section is based on the direct approach of cost function approximation by using a least squares fit of a parametric architecture to simulation-generated sample costs (cf. the left-hand side of Fig. 6.1.2). In this section, we consider the alternative indirect approach, whereby the policy evaluation is based on solving a projected form of Bellman's equation (cf. the right-hand side of Fig. 6.1.2). We will derive some methods that also use simulation, a least squares fit, and temporal differences.

We will be dealing with a single stationary policy μ , so we suppress in our notation the dependence on control of the transition probabilities and the cost per stage. We thus consider a stationary finite-state Markov chain, and we denote the states by $i = 1, \dots, n$, the transition probabilities by p_{ij} , $i, j = 1, \dots, n$, and the stage costs by $g(i, j)$. We want to evaluate the expected cost of μ corresponding to each initial state i , given by

$$J_\mu(i) = \lim_{N \rightarrow \infty} E \left\{ \sum_{k=0}^{N-1} \alpha^k g(i_k, i_{k+1}) \mid i_0 = i \right\}, \quad i = 1, \dots, n,$$

where i_k denotes the state at time k , and $\alpha \in (0, 1)$ is the discount factor.

We approximate $J_\mu(i)$ with a linear architecture of the form

$$\tilde{J}(i, r) = \phi(i)'r, \quad i = 1, \dots, n, \quad (6.23)$$

where r is a parameter vector and $\phi(i)$ is an s -dimensional feature vector associated with the state i . (Throughout this section, vectors are viewed as column vectors, and a prime denotes transposition.) As earlier, we also write the vector

$$(\tilde{J}(1, r), \dots, \tilde{J}(n, r))'$$

in the compact form Φr , where Φ is the $n \times s$ matrix that has as rows the feature vectors $\phi(i)$, $i = 1, \dots, n$. Thus, we want to approximate J_μ within

$$S = \{\Phi r \mid r \in \mathbb{R}^s\},$$

the subspace spanned by s basis functions, the columns of Φ . Our standing assumptions in this section are the following:

Assumption 6.3.1: The Markov chain has steady-state probabilities ξ_1, \dots, ξ_n , which are positive, i.e., for all $i = 1, \dots, n$,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N P(i_k = j \mid i_0 = i) = \xi_j > 0, \quad j = 1, \dots, n.$$

Assumption 6.3.2: The matrix Φ has rank s .

Assumption 6.3.1 is equivalent to assuming that the Markov chain has a single recurrent class and no transient states. Assumption 6.3.2 is equivalent to the basis functions (the columns of Φ) being linearly independent, and is analytically convenient because it implies that each vector J in the subspace S is represented in the form Φr with a unique vector r .

6.3.1 Policy Evaluation by Projected Value Iteration

In this section, we discuss a method that emulates the value iteration algorithm for solving a projected form of Bellman's equation. Consider a weighted Euclidean norm on \mathbb{R}^n of the form

$$\|J\|_v = \sqrt{\sum_{i=1}^n v_i (J(i))^2},$$

where v is a vector of positive weights v_1, \dots, v_n , and let Π denote the projection operation onto S with respect to this norm. Thus for any $J \in \mathbb{R}^n$, ΠJ is the unique vector in S that minimizes $\|J - \hat{J}\|$ over all $\hat{J} \in S$. It can also be written as

$$\Pi J = \Phi r_J,$$

where

$$r_J = \arg \min_{r \in \mathbb{R}^s} \|J - \Phi r\|_v, \quad J \in \mathbb{R}^n. \quad (6.24)$$

This is because Φ has rank s by Assumption 6.3.2, so a vector in S is uniquely written in the form Φr .

Note that Π and r_J can be written explicitly in closed form.[†] The corresponding projection formulas are well-known from linear algebra, and can be obtained by carrying out explicitly the minimization of the quadratic function $\|J - \Phi r\|_v^2$ over $r \in \mathbb{R}^s$ [cf. Eq. (6.24)]. However, they will not be needed in our analysis, so they will not be derived.

Consider the mapping T given by

$$(TJ)(i) = \sum_{i=1}^n p_{ij} (g(i, j) + \alpha J(j)), \quad i = 1, \dots, n,$$

[†] We have

$$\Pi = \Phi(\Phi' V \Phi)^{-1} \Phi' V, \quad r_J = (\Phi' V \Phi)^{-1} \Phi' V J,$$

where V is the diagonal matrix with v_i , $i = 1, \dots, n$, along the diagonal.

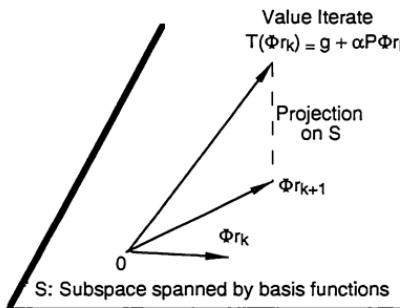


Figure 6.3.1. Illustration of the projected value iteration (PVI) method

$$\Phi r_{k+1} = \Pi T(\Phi r_k).$$

At the typical iteration k , the current iterate Φr_k is operated on with T , and the generated vector $T(\Phi r_k)$ is projected onto S , to yield the new iterate Φr_{k+1} .

or in more compact notation,

$$TJ = g + \alpha P J, \quad (6.25)$$

where g is the vector with components $\sum_{i=1}^n p_{ij} g(i, j)$, $i = 1, \dots, n$, and P is the matrix with components p_{ij} . An approximate version of the value iteration $J_{k+1} = TJ_k$ is restricted within the subspace S , and involves projection of the value iterates onto S . It is given by

$$\Phi r_{k+1} = \Pi T(\Phi r_k), \quad k = 0, 1, \dots \quad (6.26)$$

Thus at iteration k , the current iterate Φr_k is operated on with T , and the generated vector $T(\Phi r_k)$ (which does not necessarily lie in S) is projected onto S , to yield the new iterate Φr_{k+1} (see Fig. 6.3.1).

We refer to iteration (6.26) as *projected value iteration* (PVI for short), and we note that if it converges to some vector r^* , then Φr^* must be a fixed point of the mapping ΠT (the composition of Π with T):

$$\Phi r^* = \Pi T(\Phi r^*). \quad (6.27)$$

We may view this as a projected/approximate form of Bellman's equation. Some natural questions arise:

- (a) Under what conditions on the norm $\|\cdot\|_v$ does PVI converge? A related question is under what conditions is the mapping ΠT a contraction?
- (b) Assuming ΠT has a unique fixed point Φr^* , how close is Φr^* to J_μ ? Is there a bound on the error $\|J_\mu - \Phi r^*\|_v$?

- (c) Is there a way to implement PVI using simulation, and without the need for calculating explicitly the n -dimensional value iterates $T(\Phi r_k)$? (This is critical if n is very large.)

These questions have interesting and largely satisfactory answers, as we will explain in what follows.

We first note an important property of the projection Π on S with respect to the weighted Euclidean norm $\|\cdot\|_v$, namely that the following form of the *Pythagorean Theorem* holds:

$$\|J - \bar{J}\|_v^2 = \|J - \Pi J\|_v^2 + \|\Pi J - \bar{J}\|_v^2, \quad \text{for all } J \in \mathbb{R}^n, \bar{J} \in S. \quad (6.28)$$

Geometrically, the vectors $(J - \Pi J)$ and $(\Pi J - \bar{J})$ are orthogonal in the scaled geometry of the norm $\|\cdot\|_v$, where two vectors $x, y \in \mathbb{R}^n$ are said to be orthogonal if $\sum_{i=1}^n v_i x_i y_i = 0$. This orthogonality property, characteristic of Euclidean projections, implies the Pythagorean Theorem, which in turn implies another property of projections: they are *nonexpansive*, in the sense

$$\|\Pi J - \Pi \bar{J}\|_v \leq \|J - \bar{J}\|_v, \quad \text{for all } J, \bar{J} \in \mathbb{R}^n.$$

To see this, note that

$$\|\Pi(J - \bar{J})\|_v^2 \leq \|\Pi(J - \bar{J})\|_v^2 + \|(I - \Pi)(J - \bar{J})\|_v^2 = \|J - \bar{J}\|_v^2,$$

where the equality follows from the Pythagorean Theorem. Thus, for ΠT to be a contraction with respect to $\|\cdot\|_v$, it is sufficient that T be a contraction with respect to $\|\cdot\|_v$, as can be seen from the relation

$$\|\Pi T J - \Pi T \bar{J}\|_v \leq \|T J - T \bar{J}\|_v \leq \beta \|J - \bar{J}\|_v,$$

where β is the modulus of contraction of T with respect to $\|\cdot\|_v$ (see Fig. 6.3.2).

Let us now discuss the contraction properties of ΠT . We know from Section 1.4 that T is a contraction with respect to the sup-norm, but unfortunately this does not necessarily imply that T is a contraction with respect to the norm $\|\cdot\|_v$. We will next show an important fact: if v is chosen to be the steady-state probability vector ξ , then T is a contraction with respect to $\|\cdot\|_\xi$, with modulus α . The critical part of the proof is addressed in the following lemma.

Lemma 6.3.1: For any $n \times n$ stochastic matrix P that has a steady-state probability vector $\xi = (\xi_1, \dots, \xi_n)$ with positive components, we have

$$\|Pz\|_\xi \leq \|z\|_\xi, \quad z \in \mathbb{R}^n.$$

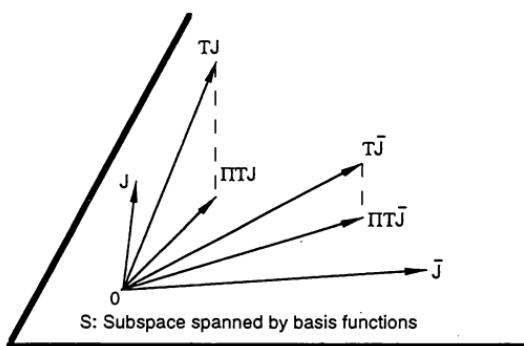


Figure 6.3.2. Illustration of the contraction property of ΠT due to the nonexpansiveness of Π . If T is a contraction with respect to $\|\cdot\|_v$, the Euclidean norm used in the projection, then ΠT is also a contraction with respect to that norm, since Π is nonexpansive and we have

$$\|\Pi T J - \Pi T \bar{J}\|_v \leq \|T J - T \bar{J}\|_v \leq \beta \|J - \bar{J}\|_v,$$

where β is the modulus of contraction of T with respect to $\|\cdot\|_v$.

Proof: Let p_{ij} be the components of P . For all $z \in \Re^n$, we have

$$\begin{aligned} \|Pz\|_\xi^2 &= \sum_{i=1}^n \xi_i \left(\sum_{j=1}^n p_{ij} z_j \right)^2 \\ &\leq \sum_{i=1}^n \xi_i \sum_{j=1}^n p_{ij} z_j^2 \\ &= \sum_{j=1}^n \sum_{i=1}^n \xi_i p_{ij} z_j^2 \\ &= \sum_{j=1}^n \xi_j z_j^2 \\ &= \|z\|_\xi^2, \end{aligned}$$

where the inequality follows from the convexity of the quadratic function, and the next to last equality follows from the defining property $\sum_{i=1}^n \xi_i p_{ij} = \xi_j$ of the steady-state probabilities. **Q.E.D.**

Proposition 6.3.1: The mappings T and ΠT are contractions of modulus α with respect to the weighted Euclidean norm $\|\cdot\|_\xi$, where ξ is the steady-state probability vector of the Markov chain.

Proof: Using the definition $TJ = g + \alpha PJ$ [cf. Eq. (6.25)], we have for all $J, \bar{J} \in \mathbb{R}^n$,

$$TJ - T\bar{J} = \alpha P(J - \bar{J}).$$

We thus obtain

$$\|TJ - T\bar{J}\|_\xi = \alpha \|P(J - \bar{J})\|_\xi \leq \alpha \|J - \bar{J}\|_\xi,$$

where the inequality follows from Lemma 6.3.1. Hence T is a contraction of modulus α . The contraction property of ΠT follows from the contraction property of T and the nonexpansiveness property of Π noted earlier. **Q.E.D.**

The next proposition gives an estimate of the difference between J_μ and the fixed point of ΠT .

Proposition 6.3.2: Let Φr^* be the fixed point of ΠT . We have

$$\|J_\mu - \Phi r^*\|_\xi \leq \frac{1}{\sqrt{1-\alpha^2}} \|J_\mu - \Pi J_\mu\|_\xi.$$

Proof: We have

$$\begin{aligned} \|J_\mu - \Phi r^*\|_\xi^2 &= \|J_\mu - \Pi J_\mu\|_\xi^2 + \|\Pi J_\mu - \Phi r^*\|_\xi^2 \\ &= \|J_\mu - \Pi J_\mu\|_\xi^2 + \|\Pi T J_\mu - \Pi T(\Phi r^*)\|_\xi^2 \\ &\leq \|J_\mu - \Pi J_\mu\|_\xi^2 + \alpha^2 \|J_\mu - \Phi r^*\|_\xi^2, \end{aligned}$$

where the first equality uses the Pythagorean Theorem [cf. Eq. (6.28)], the second equality holds because J_μ is the fixed point of T and Φr^* is the fixed point of ΠT , and the inequality uses the contraction property of ΠT . From this relation, the result follows. **Q.E.D.**

Since ΠT is a contraction, it follows that the sequence $\{\Phi r_k\}$ generated by the PVI algorithm

$$\Phi r_{k+1} = \Pi T(\Phi r_k)$$

converges to the unique fixed point Φr^* of ΠT , and that the approximation error to J_μ can be estimated as in Prop. 6.3.2. The critical assumption here is that the projection is carried out with respect to the weighted Euclidean norm $\|\cdot\|_\xi$ corresponding to the steady-state probability vector ξ . Indeed, Props. 6.3.1 and 6.3.2 hold if T is any (possibly nonlinear) contraction with respect to the Euclidean norm of the projection (cf. Fig. 6.3.2).

Unfortunately, however, while PVI is an important conceptual starting point for our methodology, it is not a practical algorithm for our purposes. The reason is that the vector $T(\Phi r_k)$ is n -dimensional and its calculation is prohibitive for the large problems that we aim to address. Furthermore, even if $T(\Phi r_k)$ were calculated, its projection on S requires knowledge of the steady-state probabilities ξ_1, \dots, ξ_n , which are generally unknown. Fortunately, both of these difficulties can be dealt with through the use of simulation, as we discuss in the next section.

6.3.2 Least Squares Policy Evaluation (LSPE)

We will develop a simulation-based implementation of the PVI iteration

$$\Phi r_{k+1} = \Pi T(\Phi r_k).$$

By expressing the projection as a least squares minimization, we see that r_{k+1} is given by

$$r_{k+1} = \arg \min_{r \in \mathbb{R}^s} \| \Phi r - T(\Phi r_k) \|_\xi^2,$$

or equivalently

$$r_{k+1} = \arg \min_{r \in \mathbb{R}^s} \sum_{i=1}^n \xi_i \left(\phi(i)' r - \sum_{j=1}^n p_{ij} (g(i, j) + \alpha \phi(j)' r_k) \right)^2. \quad (6.29)$$

We approximate this optimization by generating an infinitely long trajectory (i_0, i_1, \dots) and by updating r_k after each transition (i_k, i_{k+1}) according to

$$r_{k+1} = \arg \min_{r \in \mathbb{R}^s} \sum_{t=0}^k (\phi(i_t)' r - g(i_t, i_{t+1}) - \alpha \phi(i_{t+1})' r_k)^2. \quad (6.30)$$

We call this iteration *least squares policy evaluation* (LSPE for short).

The similarity of PVI [Eq. (6.29)] and LSPE [Eq. (6.30)] can be seen by explicitly calculating the solutions of the associated least squares problems. For PVI, by setting the gradient of the cost function in Eq. (6.29) to 0 and using a straightforward calculation, we have

$$r_{k+1} = \left(\sum_{i=1}^n \xi_i \phi(i) \phi(i)' \right)^{-1} \left(\underbrace{\sum_{i=1}^n \xi_i \phi(i) \sum_{j=1}^n p_{ij} (g(i, j) + \alpha \phi(j)' r_k)}_{=} \right). \quad (6.31)$$

For LSPE, we similarly have from Eq. (6.30)

$$r_{k+1} = \left(\sum_{t=0}^k \phi(i_t) \phi(i_t)' \right)^{-1} \left(\sum_{t=0}^k \phi(i_t) (g(i_t, i_{t+1}) + \alpha \phi(i_{t+1})' r_k) \right),$$

which can equivalently be written as

$$r_{k+1} = \left(\sum_{i=1}^n \hat{\xi}_{i,k} \phi(i) \phi(i)' \right)^{-1} \left(\sum_{i=1}^n \hat{\xi}_{i,k} \phi(i) \sum_{j=1}^n \hat{p}_{ij,k} (g(i,j) + \alpha \phi(j)' r_k) \right), \quad (6.32)$$

where $\hat{\xi}_{i,k}$ and $\hat{p}_{ij,k}$ are empirical frequencies of state i and transition (i,j) , defined by

$$\hat{\xi}_{i,k} = \frac{\sum_{t=0}^k \delta(i_t = i)}{k+1}, \quad \hat{p}_{ij,k} = \frac{\sum_{t=0}^k \delta(i_t = i, i_{t+1} = j)}{\sum_{t=0}^k \delta(i_t = i)}. \quad (6.33)$$

(We will discuss later the question of existence of the matrix inverses in the preceding equations.) Here, $\delta(\cdot)$ denotes the indicator function [$\delta(E) = 1$ if the event E has occurred and $\delta(E) = 0$ otherwise], so for example, $\hat{\xi}_{i,k}$ is the fraction of time that state i has occurred within (i_0, \dots, i_k) , the initial $(k+1)$ -state portion of the simulated trajectory. By comparing Eqs. (6.31) and (6.32), we see that they asymptotically coincide, since the empirical frequencies $\hat{\xi}_{i,k}$ and $\hat{p}_{ij,k}$ asymptotically converge (with probability 1) to the probabilities ξ_i and p_{ij} , respectively.

Thus, LSPE may be viewed as PVI with simulation error added in the right-hand side (see Fig. 6.3.3). Since the empirical frequencies $\hat{\xi}_{i,k}$ and $\hat{p}_{ij,k}$ converge to the probabilities ξ_i and p_{ij} , the error asymptotically diminishes to 0 (assuming the iterates produced by LSPE are bounded). Because of this diminishing nature of the error and the contraction property of ΠT , it is intuitively clear and can be rigorously shown that LSPE converges to the same limit as PVI (for a proof and a detailed analysis, see Bertsekas et al. [BBN04], and Yu and Bertsekas [YuB06]). The limit is the unique r^* satisfying the equation

$$\Phi r^* = \Pi T(\Phi r^*)$$

[cf. Eq. (6.27)], and the error estimate of Prop. 6.3.2 applies.

Optimistic LSPE

In the LSPE method discussed so far, the underlying assumption is that the policy evaluated remains constant throughout the simulation. There is also an optimistic version (cf. Section 6.2.3), where the policy μ is replaced by an “improved” policy $\bar{\mu}$ after only a few simulation samples have been processed. In the extreme case of a single sample between updates, the algorithm takes the following form:

Following the state transition (i_k, i_{k+1}) , set

$$r_{k+1} = \arg \min_r \sum_{t=0}^k (\phi(i_t)' r - g(i_t, u_t, i_{t+1}) - \alpha \phi(i_{t+1})' r_k)^2, \quad (6.34)$$

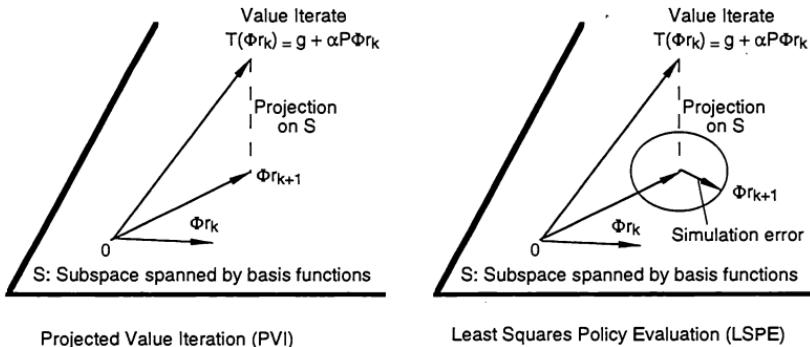


Figure 6.3.3. Illustration of LSPE. At the typical iteration k , the current value iterate Φr_k is operated on with the mapping T . The generated vector $T(\Phi r_k)$ is projected onto S with some simulation error, to yield the new iterate Φr_{k+1} . The error asymptotically tends to 0 as $k \rightarrow \infty$. Thus LSPE can be viewed as PVI, with added simulation error, which asymptotically diminishes to 0. This is the basis for showing that LSPE and PVI converge to the same limit, the fixed point of ΠT .

and generate the next transition (i_{k+1}, i_{k+2}) by simulation using the control

$$u_{k+1} = \arg \min_{u \in U(i_{k+1})} \sum_{j=1}^n p_{i_{k+1} j}(u) (g(i_{k+1}, u, j) + \alpha \phi(j)' r_{k+1}). \quad (6.35)$$

Similar to other versions of optimistic policy iteration, to enhance exploration, one may occasionally replace the control u_{k+1} by a control selected at random from the constraint set $U(i_{k+1})$. As noted earlier, optimistic variants of policy evaluation methods with function approximation are popular in practice, but the associated convergence behavior is complex and little understood at present (see the discussion in Section 6.4 of Bertsekas and Tsitsiklis [BeT96]).

6.3.3 PVI(λ) and LSPE(λ)

We now introduce a family of projected value iteration methods, parameterized by the scalar λ , which takes values in the interval $[0, 1]$. The method that corresponds to λ is called PVI(λ). The PVI method of Section 6.3.1 corresponds to $\lambda = 0$. We will see that PVI(λ) is really PVI applied to a multistep version of the mapping T and the associated Bellman equation.

For $\lambda \in [0, 1]$, define the mapping

$$T^{(\lambda)} = (1 - \lambda) \sum_{t=0}^{\infty} \lambda^t T^{t+1}. \quad (6.36)$$

Note that, for $t \geq 1$, T^t is a contraction mapping with modulus α^t , and has J_μ as its unique fixed point. Furthermore, $T^{(\lambda)}$ can be viewed as a “geometric average” of T^{t+1} [cf. Eq. (6.36)], and it can be shown to be a contraction mapping with modulus $\alpha(1 - \lambda)/(1 - \alpha\lambda)$ (see the subsequent Prop. 6.3.3). Thus, its unique fixed point is again J_μ . In fact, we may interpret

$$J = T^t J$$

as a t -stage version of Bellman’s equation, and we may view

$$J = T^{(\lambda)} J$$

as a version of Bellman’s equation with number of stages that is geometrically distributed with parameter λ .†

To obtain an alternative formula for $T^{(\lambda)}$, let us view $T^{t+1}J$ as the vector of costs over a horizon of $(t+1)$ stages with the terminal cost function being J , and write

$$T^{t+1}J = \alpha^{t+1}P^{t+1}J + \sum_{k=0}^t \alpha^k P^k g. \quad (6.37)$$

Combining Eqs. (6.36) and (6.37), we obtain

$$\begin{aligned} T^{(\lambda)}J &= (1 - \lambda) \sum_{t=0}^{\infty} \alpha^{t+1} \lambda^t P^{t+1}J + (1 - \lambda) \sum_{t=0}^{\infty} \lambda^t \sum_{k=0}^t \alpha^k P^k g \\ &= (1 - \lambda) \sum_{t=0}^{\infty} \alpha^{t+1} \lambda^t P^{t+1}J + \sum_{t=0}^{\infty} \alpha^t \lambda^t P^t g, \end{aligned}$$

where the second equality follows by straightforward calculation. More compactly, we have

$$T^{(\lambda)}J = P^{(\lambda)}J + (I - \alpha\lambda P)^{-1}g, \quad (6.38)$$

where

$$P^{(\lambda)} = (1 - \lambda) \sum_{t=0}^{\infty} \alpha^{t+1} \lambda^t P^{t+1}. \quad (6.39)$$

We will now prove that $\Pi T^{(\lambda)}$ is a contraction with respect to $\|\cdot\|_\xi$, and provide an estimate between its fixed point and J_μ . This is in analogy with Props. 6.3.1 and 6.3.2, which correspond to the case $\lambda = 0$.

† One may consider alternative methods to the ones of this section, which use in place of $T^{(\lambda)}$, the t -step mapping T^t for some $t \geq 1$. These methods behave similarly to the one of the present section, with the parameter t playing the role of λ (increasing values of t correspond to increasing values of λ); see [BBN04].

Proposition 6.3.3: The mappings $T^{(\lambda)}$ and $\Pi T^{(\lambda)}$ are contractions of modulus

$$\alpha_\lambda = \frac{\alpha(1-\lambda)}{1-\alpha\lambda}$$

with respect to the weighted Euclidean norm $\|\cdot\|_\xi$, where ξ is the steady-state probability vector of the Markov chain. Furthermore

$$\|J_\mu - \Phi r_\lambda^*\|_\xi \leq \frac{1}{\sqrt{1-\alpha_\lambda^2}} \|J_\mu - \Pi J_\mu\|_\xi, \quad (6.40)$$

where Φr_λ^* is the fixed point of $\Pi T^{(\lambda)}$.

Proof: Using Lemma 6.3.1, we have

$$\begin{aligned} \|P^{(\lambda)} z\|_\xi &\leq (1-\lambda) \sum_{t=0}^{\infty} \alpha^{t+1} \lambda^t \|P^{t+1} z\|_\xi \\ &\leq (1-\lambda) \alpha \sum_{t=0}^{\infty} \alpha^t \lambda^t \|z\|_\xi \\ &= \frac{\alpha(1-\lambda)}{1-\alpha\lambda} \|z\|_\xi. \end{aligned}$$

Since $T^{(\lambda)}$ is linear with associated matrix $P^{(\lambda)}$ [cf. Eq. (6.38)], it follows that $T^{(\lambda)}$ is a contraction with modulus $\alpha(1-\lambda)/(1-\alpha\lambda)$. The estimate (6.40) follows similar to the proof of Prop. 6.3.2. **Q.E.D.**

Since T and $T^{(\lambda)}$ have the same unique fixed point, the cost vector J_μ , we may consider a method like PVI, but with $T^{(\lambda)}$ replacing T , i.e., the method

$$\Phi r_{k+1} = \Pi T^{(\lambda)}(\Phi r_k), \quad k = 0, 1, \dots,$$

where Π denotes projection with respect to the distribution norm $\|\cdot\|_\xi$ [cf. Eq. (6.26)]. We call this method PVI(λ), and we will next derive a simulation-based approximation of it, LSPE(λ), through a process similar to the one we used to derive LSPE from PVI [in fact LSPE coincides with LSPE(0)].

Note that the convergence rate of PVI(λ) is governed by the contraction modulus

$$\alpha_\lambda = \frac{\alpha(1-\lambda)}{1-\alpha\lambda}$$

(cf. Prop. 6.3.3). It can be verified that α_λ decreases as λ increases, and $\alpha_\lambda \rightarrow 0$ as $\lambda \rightarrow 1$. Furthermore, the error bound (6.40) also becomes better as λ increases. This argues for large values of λ . On the other hand, we

will later argue that when $\text{PVI}(\lambda)$ is approximated by using simulation, the effects of simulation noise become more pronounced as λ increases. Furthermore, it should be recalled that in the context of approximate policy iteration, the objective is not just to approximate well the cost of the current policy, but rather to use the approximate cost to obtain the next “improved” policy. We are ultimately interested in a “good” next policy, and there is no consistent experimental or theoretical evidence that this is achieved solely by good cost approximation of the current policy. Thus, in practice, some trial and error with the value of λ may be useful.

Another interesting fact, which follows from the property $\alpha_\lambda \rightarrow 0$ as $\lambda \rightarrow 1$, is that given any norm, the mapping $T^{(\lambda)}$ is a contraction (with arbitrarily small modulus) with respect to that norm for λ sufficiently close to 1. This is a consequence of the norm equivalence property in \Re^n (any norm is bounded by a constant multiple of any other norm). As a result, for any weighted Euclidean norm of projection, the mapping $\Pi T^{(\lambda)}$ is a contraction provided λ is sufficiently close to 1.

From $\text{PVI}(\lambda)$ to $\text{LSPE}(\lambda)$

Let us rewrite the mapping $T^{(\lambda)}$ in a way that involves temporal differences. From Eq. (6.37), we have

$$(T^{t+1}J)(i) = E \left\{ \alpha^{t+1} J(i_{t+1}) + \sum_{k=0}^t \alpha^k g(i_k, i_{k+1}) \mid i_0 = i \right\}.$$

As a result the mapping $T^{(\lambda)}$ appearing in Eq. (6.36) can be expressed as

$$(T^{(\lambda)}J)(i) = \sum_{t=0}^{\infty} (1-\lambda) \lambda^t E \left\{ \alpha^{t+1} J(i_{t+1}) + \sum_{k=0}^t \alpha^k g(i_k, i_{k+1}) \mid i_0 = i \right\},$$

which can be written as

$$\begin{aligned} (T^{(\lambda)}J)(i) &= J(i) + (1-\lambda) \\ &\quad \cdot \sum_{t=0}^{\infty} \sum_{k=0}^t \lambda^t \alpha^k E \{ g(i_k, i_{k+1}) + \alpha J_t(i_{k+1}) - J_t(i_k) \mid i_0 = i \} \\ &= J(i) + (1-\lambda) \\ &\quad \cdot \sum_{k=0}^{\infty} \left(\sum_{t=k}^{\infty} \lambda^t \right) \alpha^k E \{ g(i_k, i_{k+1}) + \alpha J(i_{k+1}) - J(i_k) \mid i_0 = i \} \end{aligned}$$

and finally,

$$(T^{(\lambda)}J)(i) = J(i) + \sum_{t=0}^{\infty} (\alpha\lambda)^t E \{ g(i_t, i_{t+1}) + \alpha J(i_{t+1}) - J(i_t) \mid i_0 = i \}.$$

Using this equation, we can write the $\text{PVI}(\lambda)$ iteration

$$\Phi r_{k+1} = \Pi T^{(\lambda)}(\Phi r_k)$$

as

$$r_{k+1} = \arg \min_{r \in \mathbb{R}^s} \sum_{i=1}^n \xi_i \left(\phi(i)'r - \phi(i)'r_k - \sum_{t=0}^{\infty} (\alpha\lambda)^t E\{g(i_t, i_{t+1}) + \alpha\phi(i_{t+1})'r_k - \phi(i_t)'r_k \mid i_0 = i\} \right)^2$$

and by introducing the temporal differences

$$d_k(i_t, i_{t+1}) = g(i_t, i_{t+1}) + \alpha\phi(i_{t+1})'r_k - \phi(i_t)'r_k,$$

we finally obtain $\text{PVI}(\lambda)$ in the form

$$r_{k+1} = \arg \min_{r \in \mathbb{R}^s} \sum_{i=1}^n \xi_i \left(\phi(i)'r - \phi(i)'r_k - \sum_{t=0}^{\infty} (\alpha\lambda)^t E\{d_k(i_t, i_{t+1}) \mid i_0 = i\} \right)^2. \quad (6.41)$$

The LSPE(λ) method is a simulation-based approximation to the above $\text{PVI}(\lambda)$ iteration. It has the form

$$r_{k+1} = \arg \min_{r \in \mathbb{R}^s} \sum_{t=0}^k \left(\phi(i_t)'r - \phi(i_t)'r_k - \sum_{m=t}^k (\alpha\lambda)^{m-t} d_k(i_m, i_{m+1}) \right)^2, \quad (6.42)$$

where (i_0, i_1, \dots) is an infinitely long trajectory generated by simulation. The justification is that the solution of the least squares problem in the $\text{PVI}(\lambda)$ iteration (6.41) is approximately equal to the solution of the least squares problem in the LSPE(λ) iteration (6.42). Similar to the case $\lambda = 0$ [cf. Eqs. (6.29) and (6.30)], the approximation is due to:

- (a) The substitution of the steady-state probabilities ξ_i and transition probabilities p_{ij} with the empirical frequencies $\hat{\xi}_{i,k}$ and $\hat{p}_{ij,k}$ defined by Eq. (6.33).
- (b) The approximation of the infinite discounted sum of temporal differences in Eq. (6.41) with the finite discounted sum in Eq. (6.42), which also uses an approximation of the conditional probabilities of the transitions (i_t, i_{t+1}) with corresponding empirical frequencies.

Since as $k \rightarrow \infty$, the empirical frequencies converge to the true probabilities and the finite discounted sums converge to the infinite discounted sums, it follows that $\text{PVI}(\lambda)$ and LSPE(λ) asymptotically coincide. From this it can be shown that the sequence $\{\Phi r_k\}$ generated by LSPE(λ) converges (with probability 1) to Φr_λ^* , the fixed point of $\Pi T^{(\lambda)}$ and limit of $\text{PVI}(\lambda)$ (see Bertsekas et al. [BBN04]).

The Matrix Form of LSPE(λ)

Let us now streamline the calculations of LSPE(λ) and express it in terms of convenient vector-matrix formulas. It is straightforward to verify that the least squares solution of Eq. (6.42) can be written as

$$r_{k+1} = r_k + \bar{B}_k^{-1}(\bar{A}_k r_k + \bar{b}_k), \quad (6.43)$$

where

$$\bar{A}_k = \frac{A_k}{k+1}, \quad \bar{B}_k = \frac{B_k}{k+1}, \quad \bar{b}_k = \frac{b_k}{t+1}, \quad (6.44)$$

and the matrices A_k , B_k , and vector b_k are defined by†

$$A_k = \sum_{t=0}^k z_t (\phi(i_{t+1}) - \phi(i_t))', \quad B_k = \sum_{t=0}^k \phi(i_t) \phi(i_t)', \quad (6.45)$$

$$b_k = \sum_{t=0}^k z_t g(i_t, i_{t+1}), \quad z_t = \sum_{m=0}^t (\alpha\lambda)^{t-m} \phi(i_m). \quad (6.46)$$

Furthermore, because of the rank assumption on Φ (Assumption 6.3.2), the inverse of \bar{B}_k exists for k sufficiently large (as a practical matter, it is common to add a small positive multiple of the identity to \bar{B}_k , to ensure that

† By setting to 0 the gradient of the function minimized in Eq. (6.42), we see that r_{k+1} satisfies

$$\sum_{t=0}^k \phi(i_t) \left(\phi(i_t)' r_{k+1} - \phi(i_t)' r_k - \sum_{m=t}^k (\alpha\lambda)^{m-t} d_k(i_m, i_{m+1}) \right) = 0,$$

from which we obtain

$$\begin{aligned} r_{k+1} &= \left(\sum_{t=0}^k \phi(i_t) \phi(i_t)' \right)^{-1} \sum_{t=0}^k \left(\phi(i_t) \phi(i_t)' r_k \right. \\ &\quad \left. + \phi(i_t) \sum_{m=t}^k (\alpha\lambda)^{m-t} (g(i_m, i_{m+1}) + \alpha \phi(i_{m+1})' r_k - \phi(i_m)' r_k) \right) \\ &= r_k + \left(\sum_{t=0}^k \phi(i_t) \phi(i_t)' \right)^{-1} \\ &\quad \cdot \sum_{t=0}^k \phi(i_t) \sum_{m=t}^k (\alpha\lambda)^{m-t} (g(i_m, i_{m+1}) + \alpha \phi(i_{m+1})' r_k - \phi(i_m)' r_k). \end{aligned}$$

From this equation, it is straightforward to obtain the vector-matrix formulas (6.43) and (6.44) by appropriately collecting terms.

its inverse exists - we will ignore the issue of potential noninvertibility of \bar{B}_k as it can be easily addressed without essentially affecting our algorithms or analysis). Note that A_k , B_k , b_k , and z_k can be conveniently updated by means of recursive formulas. We have

$$A_k = A_{k-1} + z_k (\phi(i_{k+1}) - \phi(i_k))', \quad B_k = B_{k-1} + \phi(i_k) \phi(i_k)', \quad (6.47)$$

$$b_k = b_{k-1} + z_k g(i_k, i_{k+1}), \quad z_k = \alpha \lambda z_{k-1} + \phi(i_k). \quad (6.48)$$

• It can be shown that the matrices \bar{A}_k , \bar{B}_k , and vector \bar{b}_k converge to limits with probability 1. In particular,

$$\bar{A}_k \rightarrow A, \quad \bar{B}_k \rightarrow B, \quad \bar{b}_k \rightarrow b,$$

where

$$A = \Phi' \Xi (P^{(\lambda)} - I) \Phi, \quad B = \Phi' \Xi \Phi, \quad b = \Phi' \Xi (I - \alpha \lambda P)^{-1} g, \quad (6.49)$$

the matrix $P^{(\lambda)}$ is defined by Eq. (6.39), and Ξ is the diagonal matrix with diagonal entries ξ_1, \dots, ξ_n :

$$\Xi = \text{diag}(\xi_1, \dots, \xi_n).$$

This involves straightforward but tedious law-of-large-numbers type of arguments, using the fact that the empirical frequencies $\hat{\xi}_{i,k}$ and $\hat{p}_{ij,k}$ defined by Eq. (6.33) converge with probability 1 to the steady-state probabilities ξ_i and transition probabilities p_{ij} . We refer to the book by Bertsekas and Tsitsiklis [BeT96], and the papers by Tsitsiklis and Van Roy [TsV97], and Nedić and Bertsekas [NeB03] for more details.

From the iteration formula (6.43), it also follows that the limit r_λ^* of LSPE(λ) must satisfy $Ar_\lambda^* + b = 0$, so it is given by

$$r_\lambda^* = -A^{-1}b. \quad (6.50)$$

It can be seen that the matrix A is invertible: if it were not, then $\Pi T^{(\lambda)}$ would have multiple fixed points, which contradicts the contraction property of $\Pi T^{(\lambda)}$ (cf. Prop. 6.3.3).

LSPE(1)

In the limiting case where $\lambda = 1$, the LSPE algorithm takes the form

$$r_{k+1} = \arg \min_{r \in \mathbb{R}^s} \sum_{t=0}^k \left(\phi(i_t)' r - \phi(i_t)' r_k - \sum_{m=t}^k \alpha^{m-t} d_k(i_m, i_{m+1}) \right)^2,$$

where (i_0, i_1, \dots) is an infinitely long trajectory generated by simulation [cf. Eq. (6.42)]. We refer to this as LSPE(1) and we will now explain that it is a sound algorithm.

Using the definition of temporal differences

$$d_k(i_m, i_{m+1}) = g(i_m, i_{m+1}) + \alpha \phi(i_{m+1})' r_k - \phi(i_m)' r_k,$$

we can write LSPE(1) as

$$r_{k+1} = \arg \min_{r \in \mathbb{R}^s} \sum_{t=0}^k \left(\phi(i_t)' r - \alpha^{k-t+1} \phi(i_{k+1})' r_k - \sum_{m=t}^k \alpha^{m-t} g(i_m, i_{m+1}) \right)^2$$

We recognize the term

$$\alpha^{k-t+1} \phi(i_{k+1})' r_k + \sum_{m=t}^k \alpha^{m-t} g(i_m, i_{m+1}) \quad (6.51)$$

in the preceding equation as a simulation-based cost sample associated with initial state i_t . Using this fact, we can view LSPE(1) as a least square fit of the linear architecture Φr to cost samples of the form (6.51), which are extracted from a single infinitely long simulated trajectory. Thus, LSPE(1) is a special case of the (direct) policy evaluation method of Section 6.2.1 [cf. Eq. (6.7)].

6.3.4 The LSTD(λ) Algorithm

The closed form expression $r_\lambda^* = -A^{-1}b$ defining the fixed point Φr_λ^* of $\Pi T^{(\lambda)}$ [cf. Eq. (6.50)] motivates a different least squares TD algorithm. Since

$$\bar{A}_k \rightarrow A, \quad \bar{b}_k \rightarrow b$$

where \bar{A}_k and \bar{b}_k are given by Eq. (6.44)-(6.46), it is natural to try to obtain r_λ^* as the limit of

$$\hat{r}_k = -\bar{A}_k^{-1} \bar{b}_k.$$

This method is called *least squares temporal differences method*, or LSTD(λ) for short.

The inverse \bar{A}_k^{-1} exists for k sufficiently large, since \bar{A}_k converges to A , which is invertible. Thus, the LSTD(λ) estimate is well-defined for large enough k , and converges with probability 1 to $r_\lambda^* = -A^{-1}b$, the same limit as LSPE(λ) [and also the fixed point of the projected value iteration mapping $\Pi T^{(\lambda)}$].

The method generates a single infinitely long simulated trajectory $(i_0, i_1, \dots, i_k, \dots)$ and with each transition (i_k, i_{k+1}) , it updates the matrix \bar{A}_k and vector \bar{b}_k using Eqs. (6.47) and (6.48), and calculates $\hat{r}_k = -\bar{A}_k^{-1} \bar{b}_k$

[it is actually possible to directly update \bar{A}_k^{-1} using a formula that is similar to Eq. (6.47)].

The preceding development suggests that the rationales behind LSPE and LSTD are quite different: while the former approximates the PVI(λ) iteration

$$\Phi r_{k+1} = \Pi T^{(\lambda)}(\Phi r_k)$$

by introducing asymptotically diminishing simulation noise in its right-hand side, the latter solves at each iteration an increasingly accurate simulation-based approximation to the equation

$$\Phi r = \Pi T^{(\lambda)}(\Phi r).$$

Yet the two methods produce similar iterates, and in fact it can be shown that the difference $r_{k+1} - \hat{r}_k$ converges to 0 faster than r_{k+1} and \hat{r}_k converge to their limit r_λ^* . Thus, for large k , the iterates of LSPE(λ) and LSTD(λ) essentially coincide (although the early iterates may differ considerably).

For an intuitive explanation, consider the LSPE(λ) iteration

$$r_{k+1} = r_k + \bar{B}_k^{-1}(\bar{A}_k r_k + \bar{b}_k) \quad (6.52)$$

[cf. Eq. (6.43)]. Since \bar{B}_k , \bar{A}_k , and \bar{b}_k are simulation-based estimates of their eventual limits \bar{B} , \bar{A} , and \bar{b} , they converge at a slower time scale than the geometric convergence rate that the iteration (6.52) would have if \bar{B}_k , \bar{A}_k , and \bar{b}_k were constant (since the eigenvalues of $I + \bar{B}^{-1}\bar{A}$ lie strictly within the unit circle, the same is true for the eigenvalues of $I + \bar{B}_k^{-1}\bar{A}_k$ for large enough k). This means that the LSPE(λ) iteration (6.52) “sees \bar{B}_k , \bar{A}_k , and \bar{b}_k as essentially constant,” so that for large k , r_{k+1} is essentially equal to the corresponding limit of iteration (6.52) with \bar{B}_k , \bar{A}_k , and \bar{b}_k held fixed. This limit is $-\bar{A}_k^{-1}\bar{b}_k$, i.e., \hat{r}_k . It follows that \hat{r}_k converges to r_{k+1} faster than \hat{r}_k converges to r_λ^* . A rigorous proof of this property is given in the paper by Yu and Bertsekas [YuB06b], which also provides results of supporting computational experiments.

Some further insight into the connection of LSPE and LSTD can be obtained by verifying that the LSTD estimate \hat{r}_{k+1} is also the unique vector \hat{r} satisfying

$$\hat{r} = \arg \min_{r \in \mathbb{R}^s} \sum_{t=0}^k \left(\phi(i_t)'r - \phi(i_t)'\hat{r} - \sum_{m=t}^k (\alpha\lambda)^{m-t} \hat{d}(t) \right)^2, \quad (6.53)$$

where

$$\hat{d}(t) = g(i_t, i_{t+1}) + \overbrace{\alpha \phi(i_{t+1})' \hat{r}}^{\sim} - \phi(i_t)' \hat{r}.$$

Note that solving Eq. (6.53) is not a least squares problem, because the sum in the right-hand side involves \hat{r} . Yet, the similarity with the least squares problem solved by LSPE(λ) [cf. Eq. (6.42)] is evident. Indeed, LSTD(λ)

produces \hat{r}_{k+1} , the unique fixed point of the “arg min” mapping in the right-hand-side of Eq. (6.53), while LSPE(λ) approximates this fixed point with a single iteration starting from r_k [cf. Eq. (6.42)]. In fact if at time k we were to stop obtaining new simulation samples but continued the iterations of LSPE(λ) [i.e., for a fixed k , we used the iteration $r_{t+1} = r_t + \bar{B}_k^{-1}(\bar{A}_k r_t + \bar{b}_k)$, $t \geq k$, in place of Eq. (6.52)], we would obtain in the limit the LSTD(λ) iterate \hat{r}_{k+1} .

The preceding discussion has shown that LSPE(λ) and LSTD(λ) behave similarly in an asymptotic sense for the case of a single policy evaluation. However, they may behave quite differently in the short term and also in the context of optimistic policy iteration schemes, where the number of samples collected between policy updates is relatively small. Some experimentation is needed to clarify this issue.

A further difference between the LSPE and LSTD approaches manifests itself in the more general case where T is a *nonlinear* mapping (corresponding for example to the minimum over several policies), and we seek to find a fixed point of ΠT . A simulation-based LSPE approach can then be readily applied with two qualifications:

- (a) The existence of a fixed point may not be guaranteed because ΠT may not be a contraction with respect to the projection norm.
- (b) While the least squares problem, to be solved with each simulation transition, is still linear, its solution may not readily benefit from an efficient recursive implementation that uses the formulas (6.45) and (6.46).

On the other hand, the LSTD approach becomes very cumbersome because the associated system of equations that needs to be solved at each transition is nonlinear. See also the discussion in Section 6.4 on Q -learning with function approximation.

6.3.5 The TD(λ) Algorithm

Let us now discuss TD(λ), as defined in Section 6.2.2, and its relation to LSPE(λ). We assume a linear architecture [$\tilde{J}(i, r) = \phi(i)'r$ for all i], in conjunction with simulation of an infinitely long trajectory. The algorithm has the form

$$r_{k+1} = r_k + \gamma_k d_k z_k, \quad (6.54)$$

where γ_k is a positive stepsize,

$$d_k = g(i_k, i_{k+1}) + \alpha \phi(i_{k+1})' r_k - \phi(i_k)' r_k, \quad k = 0, 1, \dots,$$

and

$$z_k = \sum_{t=0}^k (\alpha \lambda)^{k-t} \phi(i_t)$$

[cf. Eq. (6.17)]. We will now briefly discuss the convergence mechanism of TD(λ), while referring to the paper by Tsitsiklis and Van Roy [TsV97] (also the book [BeT96]) for a detailed analysis.

We may view TD(λ) as a stochastic iterative algorithm for solving the projected Bellman equation $\Phi r = \Pi \Phi r$, or equivalently $Ar + b = 0$, where A and b are given by Eq. (6.49). The essence of the analysis of Tsitsiklis and Van Roy [TsV97] is to write the algorithm as

$$r_{k+1} = r_k + \gamma_k(Ar_k + b) + \gamma_k(\Xi_k r_k + \xi_k), \quad t = 0, 1, \dots, \quad (6.55)$$

where Ξ_k and ξ_k are some sequences of random matrices and vectors, respectively, that depend only on the simulated trajectory (so they are independent of r_k), and asymptotically have zero mean. [This requires a straightforward but tedious calculation to rewrite Eq. (6.54) into the form of Eq. (6.55), using Eq. (6.49).] A key to the convergence proof is that the matrix A can be proved to be negative definite (in the sense $z'Az < 0$ for all $z \neq 0$), which implies that it has eigenvalues with negative real parts. This in turn shows that the matrix $I + \gamma_k A$ has eigenvalues strictly within the unit circle for sufficiently small γ_k . In TD(λ), it is also essential that the stepsize γ_k be diminishing to 0, both because a small γ_k is needed to keep the eigenvalues of $I + \gamma_k A$ within the unit circle, and also because Ξ_k and ξ_k do not converge to 0.

Under Assumptions 6.3.1 and 6.3.2, and some additional technical conditions, Tsitsiklis and Van Roy [TsV97] use the preceding line of analysis to show convergence (with probability 1) of the sequence $\{r_k\}$ generated by TD(λ). Note from Eq. (6.55) that when TD(λ) converges, its limit must satisfy the equation $Ar + b = 0$. Hence, the limit of TD(λ) is $-A^{-1}b$, the unique solution of the projected Bellman equation [also the limit of LSPE(λ) and LSTD(λ)].

Note also that since $\bar{A}_k \rightarrow A$ and $\bar{b}_k \rightarrow b$, where \bar{A}_k , \bar{b}_k are given by Eqs. (6.47) and (6.48), we can write the TD(λ) iteration (6.55) in the form

$$r_{k+1} = r_k + \gamma_k(\bar{A}_k r_k + \bar{b}_k) + \gamma_k(\tilde{\Xi}_k r_k + \tilde{\xi}_k),$$

where $\tilde{\Xi}_k$ and $\tilde{\xi}_k$ asymptotically have zero mean. This iteration should be compared to the LSPE(λ) iteration

$$r_{k+1} = r_k + \bar{B}_k^{-1}(\bar{A}_k r_k + \bar{b}_k),$$

[cf. Eq. (6.43)]. A key difference of the two iterations is that TD(λ) uses the γ_k -multiple of the identity in place of \bar{B}_k^{-1} as a “scaling” matrix. This is reminiscent of a connection between the gradient/steepest descent method and the Gauss-Newton method for solving least squares problems (see nonlinear programming books, e.g., [Ber99], Section 1.5).

For another view of the connection between TD and LSPE, we note that the gradient of the least squares sum of $\text{LSPE}(\lambda)$ in Eq. (6.42) is

$$-2 \sum_{t=0}^k \phi(i_t) \sum_{m=t}^k (\alpha\lambda)^{m-t} d_k(i_m, i_{m+1}).$$

This gradient, after some calculation, can be written as

$$-2(z_0 d_k(i_0, i_1) + \dots + z_k d_k(i_k, i_{k+1})), \quad (6.56)$$

where

$$z_t = \sum_{m=0}^t (\alpha\lambda)^{t-m} \phi(i_m), \quad t = 0, \dots, k.$$

We now note that asymptotically, for large k and t , the expected values of the terms

$$z_t d_k(i_t, i_{t+1}), z_{t+1} d_k(i_{t+1}, i_{t+2}), \dots,$$

in the gradient sum (6.56) are nearly equal and proportional to the expected value of the term $z_k d_k(i_k, i_{k+1})$ in the $\text{TD}(\lambda)$ iteration (6.54). From this it can be seen that $\text{TD}(\lambda)$ updates r_k along a direction, which is proportional to the gradient of the least squares sum of $\text{LSPE}(\lambda)$, plus stochastic noise that asymptotically has zero mean.

We finally note that $\text{TD}(\lambda)$ has significant drawbacks relative to $\text{LSPE}(\lambda)$ and $\text{LSTD}(\lambda)$: it tends to converge much slower (as experiments confirm and intuition suggests, based on the generic slow convergence of gradient-like methods), and it is sensitive to the method of stepsize choice [while $\text{LSPE}(\lambda)$ and $\text{LSTD}(\lambda)$ do not require a stepsize choice].

6.3.6 Summary and Examples

Several algorithms for cost evaluation have been given so far for finite-state discounted problems, under a variety of assumptions, and we will now summarize the analysis (either given here, or discussed here and referred to other sources). We will also explain what can go wrong when the assumptions of this analysis are violated.

The algorithms considered so far for approximate evaluation of the cost vector J_μ of a single stationary policy μ are of two types:

- (1) Direct methods, such as the batch and incremental gradient methods of Section 6.2, including $\text{TD}(1)$. These methods allow for a nonlinear approximation architecture, and for a lot of flexibility in the collection of the cost samples that are used in the least squares optimization. The drawbacks of these methods are that they can be very slow, a generic characteristic of gradient-like methods, and that they are not well-suited for problems with large variance of simulation “noise.”

The latter difficulty is due to the lack of the parameter λ , which is used in other methods to reduce the variance in the parameter update formulas. Indeed, the noise in a simulation sample of a t -stages cost $T^t J$ tends to be larger as t increases, and from the formula $T^{(\lambda)} = (1 - \lambda) \sum_{t=0}^{\infty} \lambda^t T^{t+1}$ [cf. Eq. (6.36)], it can be seen that the simulation samples of $T^{(\lambda)}(\Phi r_k)$, used by LSPE(λ) and LSTD(λ), tend to contain more noise as λ increases.

- (2) Indirect methods that are based on solution of a projected version of Bellman's equation. These methods include TD(λ), LSPE(λ), and LSTD(λ), and have been discussed in Sections 6.3.1-6.3.5.

The salient characteristics of indirect methods are the following:

- (a) All these algorithms aim to converge to r_λ^* , the unique solution of the projected Bellman equation $\Phi r = \Pi T^{(\lambda)}(\Phi r)$. A key fact for convergence is that $T^{(\lambda)}$ is a contraction with respect to the projection norm $\|\cdot\|_\xi$, which implies that $\Pi T^{(\lambda)}$ is also a contraction with respect to the same norm.
- (b) The limit r_λ^* depends on λ . The estimate of Prop. 6.3.3 indicates that the approximation error $\|J_\mu - \Phi r_\lambda^*\|_\xi$ increases as the distance $\|J_\mu - \Pi J\|_\xi$ from the subspace S becomes larger, and also increases as λ becomes smaller. Indeed, the error degradation may be very significant for small values of λ , as shown by an example in Bertsekas [Ber95b] (also reproduced in Bertsekas and Tsitsiklis [BeT96], Example 6.5). Note, however, that in the context of approximate policy iteration, the correlation between approximation error in the cost of the current policy and the performance of the next policy is somewhat unclear in practice.
- (c) Experience has shown that the algorithms tend to be faster and more reliable in practice when λ takes smaller values (or at least when λ is not too close to 1), a plausible reason being that the influence of simulation noise is reduced with smaller values of λ . There is no rule of thumb for selecting λ , which is usually chosen with some trial and error.
- (d) TD(λ) is much slower than LSPE(λ) and LSTD(λ), which in turn have comparable convergence rates. From a practical point of view, it is hard to see circumstances where one would prefer to use TD(λ). Still TD(λ) is simpler and embodies important ideas, which we did not cover sufficiently in our presentation. We refer to the convergence analysis by Tsitsiklis and Van Roy [TsV97], and the subsequent papers [TsV99a] and [TsV02], for extensive discussions of TD(λ).
- (e) For all methods, the assumptions under which convergence can be shown include:

- (i) The existence of a steady-state distribution vector ξ , with positive components.
- (ii) The use of a linear approximation architecture Φr , with Φ satisfying the rank Assumption 6.3.2.
- (iii) The use of data from a single infinitely long simulated trajectory of the associated Markov chain.
- (iv) The explicit or implicit use of projection with respect to the norm $\|\cdot\|_\xi$.
- (v) The use of a diminishing stepsize for $\text{TD}(\lambda)$; the other methods do not require a stepsize choice.
- (vi) The use of a single policy, unchanged during the simulation; convergence does not extend to the case where T involves a minimization over multiple policies, or optimistic variants, where the policy used to generate the simulation data is changed after one or more transitions.

Let us now discuss the above assumptions (i)-(vi). Regarding (i), if a steady-state distribution exists but has some components that are 0, the corresponding states are transient, so they will not appear in the simulation after a finite number of transitions. Once this happens, the algorithms will operate as if the Markov chain consists of just the recurrent states, and convergence will not be affected. However, the transient states would be underrepresented in the cost approximation. If on the other hand there is no steady-state distribution, there must be multiple recurrent classes, so the results of the algorithms would depend on the initial state of the simulated trajectory (more precisely on the recurrent class of this initial state). In particular, states from other recurrent classes, and transient states would be underrepresented in the cost approximation obtained. This may be remedied by using multiple trajectories, with initial states from all the recurrent classes, so that all these classes are represented in the simulation.

Regarding (ii), even if Φ does not have rank s , the mapping $\Pi T^{(\lambda)}$ will be a contraction with respect to $\|\cdot\|_\xi$, so $\text{PVI}(\lambda)$ will converge to the unique fixed point of $\Pi T^{(\lambda)}$. However, this fixed point would not correspond to a unique parameter vector, so the convergence of $\text{TD}(\lambda)$, $\text{LSPE}(\lambda)$, and $\text{LSTD}(\lambda)$ is unclear. What is needed for convergence is to discard a maximal subset of columns of Φ that are linearly dependent on the remaining columns, thereby reducing Φ to a rank s matrix. Also there are no convergence guarantees for methods that use nonlinear architectures. In particular, an example by Tsitsiklis and Van Roy [TsV97] (also replicated in Bertsekas and Tsitsiklis [BeT96], Example 6.6) shows that $\text{TD}(\lambda)$ may diverge if a nonlinear architecture is used.

The key issue regarding (iii) is whether the empirical frequencies at which sample costs of states are collected are consistent with the corre-

sponding steady-state probabilities. If there is a discrepancy, $\text{LSPE}(\lambda)$ will approximate a projected value iteration method where the projection is with respect to a norm different from $\|\cdot\|_\xi$, in which case $\Pi T^{(\lambda)}$ may not be a contraction, with divergence potentially resulting. An example of divergence in a related algorithmic context is given in Bertsekas and Tsitsiklis [BeT96] (Example 6.7).

Regarding (iv), unless the projection is with respect to $\|\cdot\|_\xi$, there is no guarantee that $\Pi T^{(\lambda)}$ is a contraction, and hence no guarantee of convergence of any of the algorithms. Exercise 6.3 gives an example where Π is projection with respect to the standard Euclidean norm and ΠT is not a contraction while $\text{PVI}(0)$ diverges. Tsitsiklis and Van Roy [TsV96] give an example of divergence, which involves a projection with respect to a non-Euclidean norm. On the other hand, as noted earlier, $\Pi T^{(\lambda)}$ is a contraction for any Euclidean projection norm, provided λ is sufficiently close to 1.

Regarding (v), the method for stepsize choice is critical for $\text{TD}(\lambda)$; both for convergence and for performance. On the other hand, $\text{LSPE}(\lambda)$ and $\text{LSTD}(\lambda)$ work without a stepsize. It is possible to introduce a stepsize $\gamma \in (0, 1)$ in $\text{LSPE}(\lambda)$. For example, the following “damped” version of $\text{PVI}(\lambda)$,

$$\Phi r_{k+1} = (1 - \gamma)\Phi r_k + \gamma \Pi T^{(\lambda)}(\Phi r_k),$$

is convergent for $\gamma \in (0, 1)$, and the same is true for the corresponding $\text{LSPE}(\lambda)$ algorithm. In Section 6.6, we will see that for average cost problems there is an exceptional case where a stepsize less than 1 is essential for the convergence of $\text{LSPE}(0)$.

Regarding (vi), once multiple policies are introduced, or optimistic variants are used, the behavior of the methods becomes quite peculiar and unpredictable. For instance, there are examples where $\Pi T^{(\lambda)}$ has no fixed point, and examples where it has multiple fixed points; see Bertsekas and Tsitsiklis [BeT96] (Example 6.9), and de Farias and Van Roy [DFV00]. Section 6.4.2 of Bertsekas and Tsitsiklis [BeT96] discusses the interesting *chattering* phenomenon for optimistic $\text{TD}(\lambda)$, whereby there is simultaneously oscillation in policy space and convergence in parameter space. It seems that chattering occurs generically in optimistic methods, including optimistic variants of $\text{LSPE}(\lambda)$ and $\text{LSTD}(\lambda)$, but an argument given in Section 6.4.2 of [BeT96] indicates that for many problems, its effects may not be serious. Generally, the issues associated with the asymptotic behavior of optimistic methods, or even (nonoptimistic) approximate policy iteration, are not well understood at present. On the other hand, there are limited classes of problems, involving multiple policies, where the mapping $\Pi T^{(\lambda)}$ is a contraction, and for such problems, versions of $\text{TD}(\lambda)$ and $\text{LSPE}(\lambda)$ that work with multiple policies are valid. An example, optimal stopping problems, is discussed in Section 6.4.2.

6.4 Q-LEARNING

We now introduce another method for discounted problems, which is suitable for cases where there is no explicit model of the system and the cost structure. This method is analogous to value iteration and has the additional advantage that it can be used directly in the case of multiple policies. Instead of approximating the cost function of a particular policy, it updates the Q -factors associated with an *optimal* policy, thereby avoiding the multiple policy evaluation steps of the policy iteration method.

In the discounted problem, the Q -factors are defined, for all pairs (i, u) , by

$$Q^*(i, u) = \sum_{j=1}^n p_{ij}(u) (g(i, u, j) + \alpha J^*(j)).$$

Using Bellman's equation, we see that the Q -factors satisfy for all (i, u) ,

$$Q^*(i, u) = \sum_{j=1}^n p_{ij}(u) \left(g(i, u, j) + \alpha \min_{u' \in U(j)} Q^*(j, u') \right), \quad (6.57)$$

and can be shown to be the unique solution of this set of equations. The proof is essentially the same as the proof of existence and uniqueness of solution of Bellman's equation. In fact, by introducing a system whose states are the original states $1, \dots, n$, together with all the pairs (i, u) , the above set of equations can be seen to be a special case of Bellman's equation (see Exercise 6.2). Because of this interpretation, the Q -factors can be obtained by the value iteration

$$Q(i, u) := \sum_{j=1}^n p_{ij}(u) \left(g(i, u, j) + \alpha \min_{u' \in U(j)} Q(j, u') \right), \quad \text{for all } (i, u).$$

A “damped” version of this iteration is

$$Q(i, u) := (1 - \gamma)Q(i, u) + \gamma \sum_{j=1}^n p_{ij}(u) \left(g(i, u, j) + \alpha \min_{u' \in U(j)} Q(j, u') \right),$$

where γ is a stepsize parameter with $\gamma \in (0, 1]$, that may change from one iteration to the next. The *Q -learning method* is an approximate version of this iteration, whereby the expected value is replaced by a single sample:

$$Q(i, u) := Q(i, u) + \gamma \left(g(i, u, j) + \alpha \min_{u' \in U(j)} Q(j, u') - Q(i, u) \right).$$

Here j and $g(i, u, j)$ are generated from the pair (i, u) by simulation, i.e., according to the transition probabilities $p_{ij}(u)$. Thus Q -learning can be viewed as a combination of value iteration and simulation.

An important point here is that the above Q -learning algorithm works in conjunction with simulation because in the Q -factor version of Bellman's equation [Eq. (6.57)], the order of expectation and minimization is reversed relatively to the ordinary cost function version of Bellman's equation:

$$J^*(i) = \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u) (g(i, u, j) + \alpha J^*(j)).$$

This, together with the monotonicity of the Q -learning mapping, defined by Eq. (6.57), are the key mathematical reasons why Q -learning works, in the sense that it has some guaranteed convergence properties.

To guarantee the convergence of Q -learning to the optimal Q -factors, some conditions must be satisfied. Chief among these conditions are that all state-control pairs (i, u) must be visited infinitely often, and the stepsize γ should be chosen in some special way. In particular, if the k th iteration corresponds to the m th visit of the pair (i, u) , one may use in the Q -learning iteration the stepsize

$$\gamma_k = \frac{c}{m},$$

where c is a positive constant. We refer to Tsitsiklis [Tsi94b] for a proof of convergence of Q -learning under quite general conditions.

6.4.1 Q -Factor Approximations

We will now discuss methods that aim to compute approximate Q -factors. The methods are similar to the optimistic approximate policy iteration methods based on TD(0), LSPE(0), and LSTD(0), which we discussed earlier. In particular, we introduce a parametric architecture $\tilde{Q}(i, u, r)$, possibly of the linear form

$$\tilde{Q}(i, u, r) = \phi(i, u)'r, \quad (6.58)$$

where $\phi(i, u)$ is a feature vector that depends on both state and control.

Let us consider an algorithm that is very similar to the optimistic version of TD(0) [cf. Eqs. (6.19), (6.20)], the difference being that it uses approximate Q -factors rather than approximate costs, and generates transitions of state-control pairs rather than transitions of states (of course this is entirely appropriate as discussed in Section 6.2.4). In particular, we use a single infinitely long simulated trajectory (i_0, i_1, \dots) and we update the parameter vector by the following algorithm.

At the start of iteration k , we have the current parameter vector r_k , we are at some state i_k , and we have chosen a control u_k . Then:

- (1) We simulate the next transition (i_k, i_{k+1}) using the transition probabilities $p_{i_k j}(u_k)$.

(2) We generate the control u_{k+1} from the minimization

$$u_{k+1} = \arg \min_{u \in U(i_{k+1})} \tilde{Q}(i_{k+1}, u, r_k). \quad (6.59)$$

(3) We calculate the TD

$$d_k = g(i_k, u_k, i_{k+1}) + \alpha \tilde{Q}(i_{k+1}, u_{k+1}, r_k) - \tilde{Q}(i_k, u_k, r_k).$$

(4) We update the parameter vector via

$$r_{k+1} = r_k + \gamma_k d_k \nabla \tilde{Q}(i_k, u_k, r_k),$$

where γ_k is a positive stepsize.

The process is now repeated with r_{k+1} , i_{k+1} , and u_{k+1} replacing r_k , i_k , and u_k , respectively.

Note that steps (1) and (2) represent the simulated transition from state-control pair (i_k, u_k) to state-control pair (i_{k+1}, u_{k+1}) , while steps (3) and (4) update r_k via an optimistic TD(0) iteration for Q -factors [cf. Eq. (6.20)].

There is also an optimistic version of LSPE(0) with linear function approximation [cf. Eq. (6.58)]. It is given by

$$r_{k+1} = \arg \min_{r \in \mathbb{R}^s} \sum_{t=0}^k (\phi(i_t, u_t)' r - g(i_t, u_t, i_{t+1}) - \alpha \phi(i_{t+1}, u_{t+1})' r_k)^2, \quad (6.60)$$

or equivalently,

$$\begin{aligned} r_{k+1} &= \left(\sum_{t=0}^k \phi(i_t, u_t) \phi(i_t, u_t)' \right)^{-1} \\ &\quad \sum_{t=0}^k \phi(i_t, u_t) \left(g(i_t, u_t, i_{t+1}) + \alpha \phi(i_{t+1}, u_{t+1})' r_k \right). \end{aligned} \quad (6.61)$$

where the controls u_0, \dots, u_{k+1} are generated using the minimization (6.59) [cf. Eqs. (6.34), (6.35)].

From the above optimistic version of LSPE(0), we can infer a corresponding optimistic version of LSTD(0). Similar to our discussion in Section 6.3.4, LSTD(0) finds at time k the unique point \hat{r} such that

$$\hat{r} = \arg \min_{r \in \mathbb{R}^s} \sum_{t=0}^k (\phi(i_t, u_t)' r - g(i_t, u_t, i_{t+1}) - \alpha \phi(i_{t+1}, u_{t+1})' \hat{r})^2,$$

cf. Eq. (6.53). The LSPE(0) iteration (6.60) can be viewed as a single fixed point iteration (starting from r_k) for the fixed point equation that is exactly solved by LSTD(0).

In more effective implementations of the preceding algorithms, which address the issue of exploration, the control u_{k+1} is replaced by a randomized control, which uses a small parameter $\epsilon > 0$: it applies u_{k+1} as given by Eq. (6.59) with probability $1 - \epsilon$, and a random control from $U(i_{k+1})$ with probability ϵ . As in all forms of optimistic policy iteration, the behavior of these algorithms is very complex, and there is no guarantee of success. However, the approach is often tried because of its simplicity and its model-free character [it does not require knowledge of the transition probabilities $p_{ij}(u)$].

6.4.2 Q -Learning for Optimal Stopping Problems

The policy evaluation algorithms of Section 6.3, such as TD(λ), LSPE(λ), and LSTD(λ), apply when there is a single policy to be evaluated in the context of approximate policy iteration. One can in principle extend the PVI method to the case of multiple policies. It takes the form

$$\Phi r_{k+1} = \Pi T(\Phi r_k) \quad (6.62)$$

[cf. Eq. (6.26)], where T is a DP mapping that may involve minimization over multiple controls. One may then obtain a simulation-based approximation, similar to LSPE. However, there are two difficulties:

- (a) The mapping ΠT may not in general be a contraction with respect to any norm, so the iteration (6.62) may diverge, and by extension, the associated simulation-based approximation may also diverge.
- (b) The least squares problem to be solved in the implementation of LSPE may not admit an efficient recursive implementation because $T(\Phi r_k)$ may be a nonlinear function of r_k . This will be so if T involves minimization over multiple controls.

In this section we discuss the extension of projected value iteration ideas for an optimal stopping problem where the two difficulties noted above can be largely overcome.

Optimal stopping problems are a special case of DP problems where we can only choose whether to terminate at the current state or not. Examples are problems of search, sequential hypothesis testing, and pricing of derivative financial instruments (see Section 4.4 of Vol. I, and Section 3.4 of the present volume).

We are given a Markov chain with state space $\{1, \dots, n\}$, described by transition probabilities p_{ij} . We assume that the states form a single recurrent class, so that the steady-state distribution vector $\xi = (\xi_1, \dots, \xi_n)$ satisfies $\xi_i > 0$ for all i , as in Section 6.3. Given the current state i , we

assume that we have two options: to stop and incur a cost $c(i)$, or to continue and incur a cost $g(i, j)$, where j is the next state (there is no control to affect the corresponding transition probabilities). The problem is to minimize the associated α -discounted infinite horizon cost.

We associate a Q -factor with each of the two possible decisions. The Q -factor for the decision to stop is equal to $c(i)$. The Q -factor for the decision to continue is denoted by $Q(i)$, and satisfies Bellman's equation

$$Q(i) = \sum_{j=1}^n p_{ij} \left(g(i, j) + \alpha \min \{c(j), Q(j)\} \right). \quad (6.63)$$

The Q -learning algorithm is

$$Q(i) := Q(i) + \gamma \left(g(i, j) + \alpha \min \{c(j), Q(j)\} - Q(i) \right),$$

where i is the state at which we update the Q -factor and j is a successor state, generated randomly according to the transition probabilities p_{ij} . The convergence of this algorithm is addressed by the general theory of Q -learning discussed earlier. Once the Q -factors are calculated, an optimal policy can be implemented by stopping at state i if and only if $c(i) \leq Q(i)$.

Let us now consider the approximate evaluation of $Q(i)$. We introduce the mapping $F : \mathbb{R}^n \mapsto \mathbb{R}^n$ given by

$$(FQ)(i) = \sum_{j=1}^n p_{ij} \left(g(i, j) + \alpha \min \{c(j), Q(j)\} \right),$$

and we note that the (exact) Q -factor for the choice to continue is the unique fixed point of F [cf. Eq. (6.63)]. Let $\|\cdot\|_\xi$ be the weighted Euclidean norm associated with the steady-state probability vector ξ . We claim that F is a contraction with respect to this norm. Indeed, for any two vectors Q and \overline{Q} , we have

$$\begin{aligned} |(FQ)(i) - (F\overline{Q})(i)| &\leq \alpha \sum_{j=1}^n p_{ij} |\min\{c(j), Q(j)\} - \min\{c(j), \overline{Q}(j)\}| \\ &\leq \alpha \sum_{j=1}^n p_{ij} |Q(j) - \overline{Q}(j)|, \end{aligned}$$

or

$$|FQ - F\overline{Q}| \leq \alpha P |Q - \overline{Q}|,$$

where we use the notation $|x|$ to denote a vector whose components are the absolute values of the components of x . Hence,

$$\|FQ - F\overline{Q}\|_\xi \leq \alpha \|P|Q - \overline{Q}|\|_\xi \leq \alpha \|Q - \overline{Q}\|_\xi,$$

where the last step follows from the inequality $\|PJ\|_\xi \leq \|J\|_\xi$, which holds for every vector J (cf. Lemma 6.3.1). We conclude that F is a contraction with respect to $\|\cdot\|_\xi$, with modulus α .

We will now consider Q -factor approximations, using a linear approximation architecture

$$\tilde{Q}(i, r) = \phi(i)'r,$$

where $\phi(i)$ is an s -dimensional feature vector associated with state i . We also write the vector

$$(\tilde{Q}(1, r), \dots, \tilde{Q}(n, r))'$$

in the compact form Φr , where as in Section 6.3, Φ is the $n \times s$ matrix whose rows are $\phi(i)'$, $i = 1, \dots, n$. We assume that Φ has rank s , and we adopt some of the notation of Sections 6.3.1 and 6.3.2. In particular, we denote by Π the projection mapping with respect to $\|\cdot\|_\xi$ on the subspace $S = \{\Phi r \mid r \in \mathbb{R}^s\}$.

Because F is a contraction with respect to $\|\cdot\|_\xi$ with modulus α , and Π is nonexpansive, the mapping ΠF is a contraction with respect to $\|\cdot\|_\xi$ with modulus α . Therefore, the algorithm

$$\Phi r_{k+1} = \Pi F(\Phi r_k)$$

converges to the unique fixed point of ΠF . We will now consider a simulation-based approximation that is similar to LSPE(0).

We generate a single infinitely long simulated trajectory (i_0, i_1, \dots) corresponding to an unstopped system, i.e., using the transition probabilities p_{ij} . Following the transition (i_k, i_{k+1}) , we find

$$r_{k+1} = \arg \min_{r \in \mathbb{R}^s} \sum_{t=0}^k \left(\phi(i_t)'r - g(i_t, i_{t+1}) - \alpha \min\{c(i_{t+1}), \phi(i_{t+1})'r_k\} \right)^2, \quad (6.64)$$

or equivalently,

$$r_{k+1} = \left(\sum_{t=0}^k \phi(i_t)\phi(i_t)' \right)^{-1} \sum_{t=0}^k \phi(i_t) \left(g(i_t, i_{t+1}) + \alpha \min\{c(i_{t+1}), \phi(i_{t+1})'r_k\} \right). \quad (6.65)$$

Similar to the calculations involving the relation between PVI and LSPE [cf. Eqs. (6.29)-(6.33)], it can be shown that r_{k+1} as given by this iteration is equal to the iterate produced by the iteration $\Phi r_{k+1} = \Pi F(\Phi r_k)$ plus a simulation-induced error that asymptotically converges to 0 with probability 1 (see the paper by Yu and Bertsekas [YuB06c], to which we refer for further analysis). As a result, the generated sequence $\{\Phi r_k\}$ asymptotically converges to the unique fixed point of ΠF .

In comparing the Q -learning iteration (6.64) with the alternative optimistic LSPE version (6.60), we note that it has considerably higher computation overhead. In the process of updating r_{k+1} via Eq. (6.65), we can compute the matrix $\sum_{t=0}^k \phi(i_t)\phi(i_t)'$ and the vector $\sum_{t=0}^k \phi(i_t)g(i_t, i_{t+1})$ iteratively as in the LSPE algorithms of Section 6.3. However, the terms $\min\{c(i_{t+1}), \phi(i_{t+1})'r_k\}$ need to be recomputed for all the samples i_{t+1} , $t \leq k$. Intuitively, this computation corresponds to repartitioning the states into those at which to stop and those at which to continue, based on the current approximate Q -factors Φr_k . By contrast, in the corresponding optimistic LSPE version (6.61), there is no repartitioning, and these terms are replaced by $\tilde{q}(i_{t+1}, r_k)$, given by

$$\tilde{q}(i_{t+1}, r_k) = \begin{cases} c(i_{t+1}) & \text{if } t \in T, \\ \phi(i_{t+1})'r_k & \text{if } t \notin T, \end{cases}$$

where $T = \{t \mid c(i_{t+1}) \leq \phi(i_{t+1})'r_t\}$ is the set of states to stop based on the approximate Q -factors Φr_t . In particular, the term

$$\sum_{t=0}^k \phi(i_t) \min\{c(i_{t+1}), \phi(i_{t+1})'r_k\}$$

in Eq. (6.65) is replaced by

$$\sum_{t=0}^k \phi(i_t) \tilde{q}(i_{t+1}, r_k) = \sum_{t=0}^k \phi(i_t) \left(\sum_{t \leq k, t \in T} c(i_{t+1}) + \left(\sum_{t \leq k, t \notin T} \phi(i_{t+1})' \right) r_k \right), \quad (6.66)$$

which can be efficiently updated at each time k . It can be seen that the optimistic algorithm that uses the expression (6.66) (no repartitioning) can only converge to the same limit as the nonoptimistic version (6.65).

Another variant of the algorithm is obtained by simply replacing the term $\phi(i_{t+1})'r_k$ in Eq. (6.65) by $\phi(i_{t+1})'r_t$, thereby eliminating the need for repartitioning. We refer to the paper by Yu and Bertsekas [YuB06c] for further discussion of the associated convergence issues.

6.5 STOCHASTIC SHORTEST PATH PROBLEMS

In this section we consider policy evaluation for finite-state stochastic shortest path (SSP) problems (cf. Chapter 2). We assume that there is no discounting ($\alpha = 1$), and that the states are $0, 1, \dots, n$, where state 0 is a special cost-free termination state. We focus on a fixed proper policy μ , under which all the states $1, \dots, n$ are transient.

We will consider a natural extension of the PVI(λ) and LSPE(λ) algorithms. We introduce a linear approximation architecture of the form

$$\tilde{J}(i, r) = \phi(i)'r, \quad i = 0, 1, \dots, n,$$

and the subspace

$$S = \{\Phi r \mid r \in \mathbb{R}^s\},$$

where, as in Section 6.3, Φ is the $n \times s$ matrix whose rows are $\phi(i)'$, $i = 1, \dots, n$. We assume that Φ has rank s . Also, for notational convenience in the subsequent formulas, we define $\phi(0) = 0$.

We assume that a sequence of simulated trajectories is generated, each of the form (i_0, i_1, \dots, i_N) , where $i_N = 0$, and $i_t \neq 0$ for $t < N$. Once a trajectory is completed, an initial state i_0 for the next trajectory is chosen according to a fixed probability distribution $q_0 = (q_0(1), \dots, q_0(n))$, where

$$q_0(i) = P(i_0 = i) > 0, \quad i = 1, \dots, n, \quad (6.67)$$

and the process is repeated.

The LSPE(λ) algorithm updates the parameter vector after each trajectory is fully generated. Let $(i_{0,l}, i_{1,l}, \dots, i_{N_l,l})$ be the l th trajectory (with $i_{N_l,l} = 0$), and let r_k be the parameter vector after k trajectories. We set

$$r_{k+1} = \arg \min_r \sum_{l=1}^{k+1} \sum_{t=0}^{N_l-1} \left(\phi(i_{t,l})'r - \phi(i_{t,l})'r_k - \sum_{m=t}^{N_l-1} \lambda^{m-t} d_k(i_{m,l}, i_{m+1,l}) \right)^2, \quad (6.68)$$

where for all k , m , and l , the temporal differences are given by

$$d_k(i_{m,l}, i_{m+1,l}) = g(i_{m,l}, i_{m+1,l}) + \phi(i_{m+1,l})'r_k - \phi(i_{m,l})'r_k$$

[cf. Eq. (6.42)]. An efficient recursive implementation of this algorithm is possible along the lines of the discounted case [cf. Eqs. (6.45)-(6.48)].

It is also possible to use a variant of the algorithm where the parameter vector is updated at every transition rather than at the end of each trajectory. In this case, the sum of squared terms corresponding to the last trajectory in Eq. (6.68) should be adjusted to include just the terms up to the current transition. Such a variant may be preferable in cases where the trajectories tend to be very long.

The convergence analysis of the LSPE(λ) algorithm (6.68) follows similar lines to the one for discounted problems in Section 6.3, and will only be sketched. For a trajectory i_0, i_1, \dots , of the SSP problem consider the probabilities

$$q_t(i) = P(i_t = i), \quad i = 1, \dots, n, \quad t = 0, 1, \dots$$

Note that $q_t(i)$ diminishes to 0 as $t \rightarrow \infty$ at the rate of a geometric progression (cf. Section 2.1), so the limits

$$q(i) = \sum_{t=0}^{\infty} q_t(i), \quad i = 1, \dots, n,$$

are finite. Let q be the vector with components $q(1), \dots, q(n)$. We introduce the norm

$$\|J\|_q = \sqrt{\sum_{i=1}^n q(i)(J(i))^2},$$

and we denote by Π the projection onto the subspace S with respect to this norm. In the context of the SSP problem, the projection norm $\|\cdot\|_q$ plays a role similar to the one played by the steady-state distribution norm $\|\cdot\|_\xi$ for discounted problems (cf. Section 6.3).

Let P be the $n \times n$ matrix with components p_{ij} , $i, j = 1, \dots, n$. Consider also the mapping $T : \mathbb{R}^n \mapsto \mathbb{R}^n$ given by

$$TJ = g + PJ,$$

where g is the vector with components $\sum_{j=0}^n p_{ij}g(j)$, $i = 1, \dots, n$. For $\lambda \in [0, 1]$, define the mapping

$$T^{(\lambda)} = (1 - \lambda) \sum_{t=0}^{\infty} \lambda^t T^{t+1}$$

[cf. Eq. (6.36)]. Similar to Section 6.3.3, we have

$$T^{(\lambda)}J = P^{(\lambda)}J + (I - \lambda P)^{-1}g,$$

where

$$P^{(\lambda)} = (1 - \lambda) \sum_{t=0}^{\infty} \lambda^t P^{t+1} \tag{6.69}$$

[cf. Eq. (6.39)].

We consider PVI(λ):

$$\Phi r_{k+1} = \Pi T^{(\lambda)}(\Phi r_k), \tag{6.70}$$

which as in Section 6.3.3 [cf. Eq. (6.41)], can be written as

$$\begin{aligned} r_{k+1} &= \arg \min_{r \in \mathbb{R}^n} \sum_{i=1}^n q(i) \left(\phi(i)'r - \phi(i)'r_k \right. \\ &\quad \left. - \sum_{t=0}^{\infty} \lambda^t E\{d_k(i_t, i_{t+1}) \mid i_0 = i\} \right)^2, \end{aligned} \tag{6.71}$$

where $d_k(i_t, i_{t+1})$ are the temporal differences

$$d_k(i_t, i_{t+1}) = g(i_t, i_{t+1}) + \phi(i_{t+1})' r_k - \phi(i_t)' r_k.$$

Similar to Section 6.3.3, it can be shown that the LSPE(λ) iteration (6.68) is a simulation-based approximation of PVI(λ) [cf. Eq. (6.71), or equivalently Eq. (6.70)]. For a more detailed view, note that for any $b \geq 1$, the b -stages PVI formula $\Phi r_{k+1} = \Pi T^b(\Phi r_k)$, can be written as

$$r_{k+1} = \arg \min_r \sum_{i=1}^n q(i) \left(\phi(i)' r - E \{ \text{cost of } b \text{ stages with terminal cost function } \Phi r_k \mid i_0 = i \} \right)^2,$$

or

$$r_{k+1} = \left(\sum_{i=1}^n q(i) \phi(i) \phi(i)' \right)^{-1} \sum_{i=1}^n q(i) \phi(i) \cdot E \{ \text{cost of } b \text{ stages with terminal cost function } \Phi r_k \mid i_0 = i \}. \quad (6.72)$$

Its simulation-based approximation (b -stages LSPE) is

$$r_{k+1} = \arg \min_r \sum_{l=1}^{k+1} \sum_{t=0}^{N_l-1} \left(\phi(i_{t,l})' r - \left(\sum_{m=t}^{t+b-1} g(i_{m,l}, i_{m+1,l}) + \phi(i_{t+b})' r_k \right) \right)^2$$

or

$$r_{k+1} = \left(\sum_{l=1}^{k+1} \sum_{t=0}^{N_l-1} \phi(i_{t,l}) \phi(i_{t,l})' \right)^{-1} \cdot \sum_{l=1}^{k+1} \sum_{t=0}^{N_l-1} \phi(i_{t,l}) \left(\sum_{m=t}^{t+b-1} g(i_{m,l}, i_{m+1,l}) + \phi(i_{t+b})' r_k \right). \quad (6.73)$$

It can be seen that iteration (6.73) is a valid approximation to the iteration (6.72), the approximation being due to the use of empirical frequencies in place of the probabilities $q(i)$ and p_{ij} .

We now note that PVI(λ) [cf. Eq. (6.71)] is the same as Eq. (6.72), but with

$$E \{ \text{cost of } b \text{ stages with terminal cost function } \Phi r_k \mid i_0 = i \}$$

replaced by

$$(1 - \lambda)$$

$$\cdot \sum_{b=0}^{\infty} \lambda^b E \{ \text{cost of } b+1 \text{ stages with terminal cost function } \Phi r_k \mid i_0 = i \}.$$

Similarly LSPE(λ) [cf. Eq. (6.68)] is the same as Eq. (6.73), but with

$$\sum_{m=t}^{t+b-1} g(i_{m,l}, i_{m+1,l}) + \phi(i_{t+b})' r_k$$

replaced by

$$(1 - \lambda) \sum_{b=0}^{\infty} \lambda^b \sum_{m=t}^{t+b} g(i_{m,l}, i_{m+1,l}) + \phi(i_{t+b})' r_k.$$

After some straightforward calculation, the resulting formula becomes Eq. (6.68). The conclusion is that LSPE(λ) is a valid approximation of PVI(λ), where empirical frequencies are used in place of the probabilities $q(i)$ and p_{ij} .

We will now show that $\Pi T^{(\lambda)}$ is a contraction, so that PVI(λ), and hence also LSPE(λ), converges to the unique fixed point of $\Pi T^{(\lambda)}$.

Proposition 6.5.1: For all $\lambda \in [0, 1]$, $\Pi T^{(\lambda)}$ is a contraction with respect to some norm.

Proof: Let $\lambda > 0$. We will show that $T^{(\lambda)}$ is a contraction with respect to the projection norm $\|\cdot\|_q$, so the same is true for $\Pi T^{(\lambda)}$, since Π is nonexpansive. Indeed, let us first note that with an argument like the one in the proof of Lemma 6.3.1, we have

$$\|PJ\|_q \leq \|J\|_q, \quad J \in \Re^n,$$

from which it follows that

$$\|P^t J\|_q \leq \|J\|_q, \quad J \in \Re^n, t = 0, 1, \dots$$

Thus, by using the definition (6.69) of $P^{(\lambda)}$, we also have

$$\|P^{(\lambda)} J\|_q \leq \|J\|_q, \quad J \in \Re^n.$$

Since $\lim_{t \rightarrow \infty} P^t J = 0$ for any $J \in \Re^n$, it follows that $\|P^t J\|_q < \|J\|_q$ for all $J \neq 0$ and t sufficiently large. Therefore,

$$\|P^{(\lambda)} J\|_q < \|J\|_q, \quad \text{for all } J \neq 0. \quad (6.74)$$

We now define

$$\beta = \max\{\|P^{(\lambda)} J\|_q \mid \|J\|_q = 1\}$$

and note that since the maximum in the definition of β is attained by the Weierstrass Theorem (a continuous function attains a maximum over a compact set), we have $\beta < 1$ in view of Eq. (6.74). Since

$$\|P^{(\lambda)}J\|_q \leq \beta \|J\|_q, \quad J \in \mathbb{R}^n,$$

it follows that $P^{(\lambda)}$ is a contraction of modulus β with respect to $\|\cdot\|_q$.

Let $\lambda = 0$. We use a different argument because T is not necessarily a contraction with respect to $\|\cdot\|_q$. We show that ΠT is a contraction with respect to a different norm by showing that the eigenvalues of ΠP lie strictly within the unit circle.[†] Indeed, with an argument like the one used to prove Lemma 6.3.1, we have $\|PJ\|_q \leq \|J\|_q$ for all J , which implies that $\|\Pi PJ\|_q \leq \|J\|_q$, so the eigenvalues of ΠP cannot be outside the unit circle. Assume to arrive at a contradiction that ν is an eigenvalue of ΠP with $|\nu| = 1$, and let ζ be a corresponding eigenvector. We claim that $P\zeta$ must have both real and imaginary components in the subspace S . If this were not so, we would have $P\zeta = \Pi P\zeta$, so that

$$\|P\zeta\|_q > \|\Pi P\zeta\|_q = \|\nu\zeta\|_q = |\nu| \|\zeta\|_q = \|\zeta\|_q,$$

which contradicts the fact $\|PJ\|_q \leq \|J\|_q$ for all J . Thus, the real and imaginary components of $P\zeta$ are in S , which implies that $P\zeta = \Pi P\zeta = \nu\zeta$, so that ν is an eigenvalue of P . This is a contradiction because $|\nu| = 1$ while the eigenvalues of P are strictly within the unit circle, since the policy being evaluated is proper. Q.E.D.

The preceding proof has shown that $\Pi T^{(\lambda)}$ is a contraction with respect to $\|\cdot\|_q$ when $\lambda > 0$. As a result, similar to Prop. 6.3.3, we can obtain the error bound

$$\|J_\mu - \Phi r_\lambda^*\|_q \leq \frac{1}{\sqrt{1 - \alpha_\lambda^2}} \|J_\mu - \Pi J_\mu\|_q, \quad \lambda > 0,$$

where Φr_λ^* and α_λ are the fixed point and contraction modulus of $\Pi T^{(\lambda)}$, respectively. When $\lambda = 0$, we have

$$\begin{aligned} \|J_\mu - \Phi r_0^*\| &\leq \|J_\mu - \Pi J_\mu\| + \|\Pi J_\mu - \Phi r_0^*\| \\ &= \|J_\mu - \Pi J_\mu\| + \|\Pi T J_\mu - \Pi T(\Phi r_0^*)\| \\ &= \|J_\mu - \Pi J_\mu\| + \alpha_0 \|J_\mu - \Phi r_0^*\|, \end{aligned}$$

[†] We use here the fact that if a square matrix has eigenvalues strictly within the unit circle, then there exists a norm with respect to which the linear mapping defined by the matrix is a contraction. Also in the following argument, the projection Πz of a complex vector z is obtained by separately projecting the real and the imaginary components of z on S . The projection norm for a complex vector $x + iy$ is defined by

$$\|x + iy\|_q = \sqrt{\|x\|_q^2 + \|y\|_q^2}.$$

where $\|\cdot\|$ is the norm with respect to which ΠT is a contraction (cf. Prop. 6.5.1), and Φr_0^* and α_0 are the fixed point and contraction modulus of ΠT . We thus have the error bound

$$\|J_\mu - \Phi r_0^*\| \leq \frac{1}{1 - \alpha_0} \|J_\mu - \Pi J_\mu\|.$$

We finally note that we can implement LSPE(λ) efficiently and recursively, as in the discounted and average cost cases. Similarly, we can construct a version of the LSTD(λ) method for SSP problems. Following the transition (i_k, i_{k+1}) , this method computes \hat{r}_{k+1} , which is the unique vector \hat{r} satisfying

$$\hat{r} = \arg \min_r \sum_{l=1}^{k+1} \sum_{t=0}^{N_l-1} \left(\phi(i_{t,l})' r - \phi(i_{t,l})' \hat{r} - \sum_{m=t}^{N_l-1} \lambda^{m-t} \hat{d}(i_{m,i}) \right)^2, \quad (6.75)$$

where

$$\hat{d}(t) = g(i_t, i_{t+1}) + \phi(i_{t+1})' \hat{r} - \phi(i_t)' \hat{r}$$

[cf. Eq. (6.53)]. Note that LSTD(λ) produces the unique fixed point of the “arg min” mapping in the right-hand-side of Eq. (6.75), while LSPE(λ) approximates this fixed point with a single iteration starting from r_k [cf. Eq. (6.68)]. The two methods have similar convergence rate, and in fact their iterates can be shown to converge to each other faster than they converge to their limit, similar to the discounted problem case.

6.6 AVERAGE COST PROBLEMS

In this section we consider average cost problems and related approximations: policy evaluation algorithms such as LSPE(λ) and LSTD(λ), approximate policy iteration, and Q -learning. We assume throughout the finite state model of Section 4.1, with the optimal average cost being the same for all initial states (cf. Section 4.2).

6.6.1 Approximate Policy Evaluation

Let us consider the problem of approximate evaluation of a stationary policy μ . As in the discounted case (Section 6.3), we consider a stationary finite-state Markov chain with states $i = 1, \dots, n$, transition probabilities p_{ij} , $i, j = 1, \dots, n$, and stage costs $g(i, j)$. We assume that the states form a single recurrent class. An equivalent way to express this assumption is the following.

Assumption 6.6.1: The Markov chain has a steady-state probability vector $\xi = (\xi_1, \dots, \xi_n)$ with positive components, i.e., for all $i = 1, \dots, n$,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N P(i_k = j \mid i_0 = i) = \xi_j > 0, \quad j = 1, \dots, n.$$

From Section 4.2, we know that under Assumption 6.6.1, the average cost, denoted by η , is independent of the initial state

$$\eta = \lim_{N \rightarrow \infty} \frac{1}{N} E \left\{ \sum_{k=0}^{N-1} g(x_k, x_{k+1}) \mid x_0 = i \right\}, \quad i = 1, \dots, n, \quad (6.76)$$

and satisfies

$$\eta = \xi' g,$$

where g is the vector whose i th component is the expected stage cost $\sum_{j=1}^n p_{ij} g(i, j)$. (In Chapter 4 we denoted the average cost by λ , but in the present chapter, with apologies to the readers, we reserve λ for use in the TD, LSPE, and LSTD algorithms, hence the change in notation.) Together with a differential cost vector $h = (h(1), \dots, h(n))'$, the average cost η satisfies Bellman's equation

$$h(i) = \sum_{j=1}^n p_{ij} (g(i, j) - \eta + h(j)), \quad i = 1, \dots, n.$$

The solution is unique up to a constant shift for the components of h , and can be made unique by eliminating one degree of freedom, such as fixing the differential cost of a single state to 0 (cf. Prop. 4.2.4).

We consider a linear architecture for the differential costs of the form

$$\tilde{h}(i, r) = \phi(i)' r, \quad i = 1, \dots, n.$$

where $r \in \mathbb{R}^s$ is a parameter vector and $\phi(i)$ is a feature vector associated with state i . These feature vectors define the subspace

$$S = \{\Phi \tilde{r} \mid \tilde{r} \in \mathbb{R}^s\},$$

where as in Section 6.3, Φ is the $n \times s$ matrix whose rows are $\phi(i)'$, $i = 1, \dots, n$. We will thus aim to approximate h by a vector in S , similar to Section 6.3, which dealt with cost approximation in the discounted case.

We introduce the mapping $F : \mathbb{R}^n \mapsto \mathbb{R}^n$ defined by

$$FJ = g - \eta e + PJ,$$

where P is the transition probability matrix and e is the unit vector. Note that the definition of F uses the *exact* average cost η , as given by Eq. (6.76). With this notation, Bellman's equation becomes

$$h = Fh,$$

so if we know η , we can try to find or approximate a fixed point of F .

Similar to Section 6.3, the projected value iteration (PVI) algorithm for average cost is

$$\Phi r_{k+1} = \Pi F(\Phi r_k),$$

where Π is projection on the subspace S with respect to the norm $\|\cdot\|_\xi$. The LSPE algorithm for average cost is a simulation-based approximation of the PVI, where Π and F are approximated by using a least squares minimization and simulation. An important issue is whether ΠF is a contraction. For this it is necessary to make the following assumption.

Assumption 6.6.2: The columns of the matrix Φ together with the unit vector $e = (1, \dots, 1)'$ form a linearly independent set of vectors.

Note the difference with the corresponding Assumption 6.3.2 for the discounted case in Section 6.3. Here, in addition to Φ having rank s , we require that e does not belong to the subspace S . To get a sense why this is needed, observe that if $e \in S$, then ΠF cannot be a contraction, since any scalar multiple of e when added to a fixed point of ΠF would also be a fixed point.

PVI(λ) and LSPE(λ)

We will consider the more general versions of PVI and LSPE, parameterized by $\lambda \in [0, 1]$, as in Section 6.3.3. In particular, PVI(λ) is given by

$$\Phi r_{k+1} = \Pi F^{(\lambda)}(\Phi r_k), \quad (6.77)$$

where $F^{(\lambda)}$ uses multiple-step versions of F and combines them with geometrically decreasing weights that depend on λ [cf. Eq. (6.36)]:

$$F^{(\lambda)} = (1 - \lambda) \sum_{t=0}^{\infty} \lambda^t F^{t+1}.$$

In matrix notation, the mapping $F^{(\lambda)}$ can be written as

$$F^{(\lambda)}J = (1 - \lambda) \sum_{t=0}^{\infty} \lambda^t P^{t+1}J + \sum_{t=0}^{\infty} \lambda^t P^t(g - \eta e),$$

or more compactly as

$$F^{(\lambda)}J = P^{(\lambda)}J + (I - \lambda P)^{-1}(g - \eta e), \quad (6.78)$$

where the matrix $P^{(\lambda)}$ is defined by

$$P^{(\lambda)} = (1 - \lambda) \sum_{t=0}^{\infty} \lambda^t P^{t+1} \quad (6.79)$$

[cf. Eq. (6.39)]. Note that for $\lambda = 0$, we have $F^{(0)} = F$ and $P^{(0)} = P$.

Similar to Section 6.3.3, we approximate PVI(λ) by generating an infinitely long simulated trajectory (i_0, i_1, \dots) . Following each transition (i_k, i_{k+1}) , we estimate η by

$$\eta_k = \frac{1}{k+1} \sum_{t=0}^k g(i_t, i_{t+1}), \quad (6.80)$$

and we update r_k by

$$r_{k+1} = \arg \min_{r \in \mathcal{R}^s} \sum_{t=0}^k \left(\phi(i_t)'r - \phi(i_t)'r_k - \sum_{m=t}^k \lambda^{m-t} d_k(m) \right)^2, \quad (6.81)$$

where $d_k(m)$ are the temporal differences†

$$d_k(m) = g(i_m, i_{m+1}) - \eta_m + \phi(i_{m+1})'r_k - \phi(i_m)'r_k. \quad (6.82)$$

This is the LSPE(λ) algorithm for average cost, and it should be compared with the discounted case version [cf. Eq. (6.42)]. The two algorithms are very similar, the differences being that in the average cost case:

- (1) The temporal differences include the estimate η_m [cf. Eq. (6.82)].
- (2) There is no discount factor α .

Since η_k converges to the average cost value η with probability 1, for large k it can be viewed as a constant, which is essentially inconsequential as

† It is generally preferable to use an estimate η_k as close as possible to the true average cost η , so if a better estimate than the one of Eq. (6.80) is known, it should be used in the definition of $d_k(m)$. A related idea is to backtrack and make retroactive corrections on earlier iterates r_t , $t < k$, that account for more accurate estimates of η , which are available at time k . These corrections can be organized efficiently, but we will not go into the details.

far as the convergence analysis is concerned. On the other hand the absence of a discount factor is significant because it impacts on the contraction properties of the mappings $F^{(\lambda)}$ and $\Pi F^{(\lambda)}$, which were the key to the analysis of the discounted versions of PVI(λ) and LSPE(λ). Hence, we will now focus on these contraction properties.

Let us note that the estimate Φr_{k+1} produced by the LSPE(λ) iteration (6.81)-(6.82) is equal to the PVI(λ) iterate

$$\Phi r_{k+1} = \Pi F^{(\lambda)}(\Phi r_k),$$

plus stochastic simulation noise, which asymptotically diminishes to 0 with probability 1 [assuming the iterates produced by LSPE(λ) are bounded]. This is in complete analogy with the discounted case (Section 6.3.3) and rests on the fact that the LSPE(λ) iteration involves empirical frequencies, which asymptotically converge to the steady-state probabilities ξ_i and the transition probabilities p_{ij} that the PVI(λ) iteration involves. As a result, to prove convergence of LSPE(λ), it is sufficient to prove convergence of PVI(λ). The most natural way to do this is to show that the mapping $\Pi F^{(\lambda)}$ is a contraction. In particular, we may show that $F^{(\lambda)}$ is a contraction with respect to the norm $\|\cdot\|_\xi$, and then use the nonexpansiveness of Π with respect to that norm.[†] The following proposition relates to the composition of general linear mappings with Euclidean projections, and captures the essence of the argument.

Proposition 6.6.1: Let S be a subspace of \mathbb{R}^n and let $H : \mathbb{R}^n \mapsto \mathbb{R}^n$ be a linear mapping,

$$H(x) = Cx + d,$$

where C is an $n \times n$ matrix and d is a vector in \mathbb{R}^n . Let $\|\cdot\|$ be a weighted Euclidean norm with respect to which H is nonexpansive, and let Π denote projection onto S with respect to that norm.

- (a) ΠH has a unique fixed point if and only if either 1 is not an eigenvalue of C , or else the eigenvectors corresponding to the eigenvalue 1 do not belong to S .
- (b) If ΠH has a unique fixed point, then for all $\gamma \in (0, 1)$, the mapping

[†] A more general approach is to show that the eigenvalues of the matrix $\Pi P^{(\lambda)}$ associated with the linear mapping $\Pi F^{(\lambda)}$ lie strictly within the unit circle, thereby showing that $\Pi F^{(\lambda)}$ is a contraction with respect to some norm (not necessarily $\|\cdot\|_\xi$). This approach has been pursued in the paper by Yu and Bertsekas [YuB06b], which obtains sharper convergence results than the ones given here.

$$G_\gamma = (1 - \gamma)I + \gamma \Pi H$$

is a contraction, i.e., for some scalar $\rho_\gamma \in (0, 1)$, we have

$$\|G_\gamma x - G_\gamma y\| \leq \rho_\gamma \|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

Proof: (a) Assume that ΠH has a unique fixed point, or equivalently (in view of the linearity of H) that 0 is the unique fixed point of ΠC . If 1 is an eigenvalue of C with a corresponding eigenvector z that belongs to S , then $Cz = z$ and $\Pi Cz = \Pi z = z$. Thus, z is a fixed point of ΠC with $z \neq 0$, a contradiction. Hence, either 1 is not an eigenvalue of C , or else the eigenvectors corresponding to the eigenvalue 1 do not belong to S .

Conversely, assume that either 1 is not an eigenvalue of C , or else the eigenvectors corresponding to the eigenvalue 1 do not belong to S . We will show that the mapping $\Pi(I - C)$ is one-to-one from S to S , and hence the fixed point of ΠH is the unique vector $x^* \in S$ satisfying $\Pi(I - C)x^* = \Pi d$. Indeed, assume the contrary, i.e., that $\Pi(I - C)$ has a nontrivial nullspace in S , so that some $z \in S$ with $z \neq 0$ is a fixed point of ΠC . Then, either $Cz = z$ (which is impossible since then 1 is an eigenvalue of C , and z is a corresponding eigenvector that belongs to S), or $Cz \neq z$, in which case Cz differs from its projection ΠCz and

$$\|z\| = \|\Pi Cz\| < \|Cz\| \leq \|C\| \|z\|,$$

so that $1 < \|C\|$ (which is impossible since H is nonexpansive, and therefore $\|C\| \leq 1$), thereby arriving at a contradiction.

(b) Since ΠH has a unique fixed point, we have $z \neq \Pi Cz$ for all $z \neq 0$. Hence, for all $z \in \mathbb{R}^n$, we have

$$\|(1 - \gamma)z + \gamma \Pi Cz\| < (1 - \gamma)\|z\| + \gamma\|\Pi Cz\| \leq (1 - \gamma)\|z\| + \gamma\|z\| = \|z\|, \quad (6.83)$$

where the strict inequality follows from the strict convexity of the norm, and the weak inequality follows from the nonexpansiveness of ΠC . If we define

$$\rho_\gamma = \sup\{\|(1 - \gamma)z + \gamma \Pi Cz\| \mid \|z\| \leq 1\},$$

and note that the supremum above is attained by the Weierstrass Theorem (a continuous function attains a minimum over a compact set), we see that Eq. (6.83) yields $\rho_\gamma < 1$ and

$$\|(1 - \gamma)z + \gamma \Pi Cz\| \leq \rho_\gamma \|z\|, \quad z \in \mathbb{R}^n.$$

By letting $z = x - y$, with $x, y \in \mathbb{R}^n$, and by using the definition of G_γ , we have

$$G_\gamma x - G_\gamma y = G_\gamma(x - y) = (1 - \gamma)(x - y) + \gamma \Pi C(x - y) = (1 - \gamma)z + \gamma \Pi Cz,$$

so by combining the preceding two relations, we obtain

$$\|G_\gamma x - G_\gamma y\| \leq \rho_\gamma \|x - y\|, \quad x, y \in \mathbb{R}^n.$$

Q.E.D.

We can now derive the conditions under which the mapping underlying the LSPE iteration is a contraction with respect to $\|\cdot\|_\xi$.

Proposition 6.6.2: The mapping

$$F_{\gamma, \lambda} = (1 - \gamma)I + \gamma \Pi F^{(\lambda)}$$

is a contraction with respect to $\|\cdot\|_\xi$ if one of the following is true:

- (i) $\lambda \in (0, 1)$ and $\gamma \in (0, 1]$,
- (ii) $\lambda = 0$ and $\gamma \in (0, 1)$.

Proof: Consider first the case, $\gamma = 1$ and $\lambda \in (0, 1)$. Then $F^{(\lambda)}$ is a linear mapping involving the matrix $P^{(\lambda)}$. Since $0 < \lambda$ and all states form a single recurrent class, all entries of $P^{(\lambda)}$ are positive. Thus $P^{(\lambda)}$ can be expressed as a convex combination

$$P^{(\lambda)} = (1 - \beta)I + \beta \bar{P}$$

for some $\beta \in (0, 1)$, where \bar{P} is a stochastic matrix with positive entries. We make the following observations:

- (i) \bar{P} corresponds to a nonexpansive mapping with respect to the norm $\|\cdot\|_\xi$. The reason is that the steady-state distribution of \bar{P} is ξ [as can be seen by multiplying the relation $P^{(\lambda)} = (1 - \beta)I + \beta \bar{P}$ with ξ , and by using the relation $\xi' = \xi' P^{(\lambda)}$ to verify that $\xi' = \xi' \bar{P}$]. Thus, we have $\|\bar{P}z\|_\xi \leq \|z\|_\xi$ for all $z \in \mathbb{R}^n$ (cf. Lemma 6.3.1), implying that \bar{P} has the nonexpansiveness property mentioned.
- (ii) Since \bar{P} has positive entries, the states of the Markov chain corresponding to \bar{P} form a single recurrent class. If z is an eigenvector of \bar{P} corresponding to the eigenvalue 1, we have $z = \bar{P}^k z$ for all $k \geq 0$, so $z = \bar{P}^* z$, where $\bar{P}^* = \lim_{N \rightarrow \infty} (1/N) \sum_{k=0}^{N-1} \bar{P}^k$ (cf. Prop. 4.1.2). The rows of \bar{P}^* are all equal to ξ' since the steady-state distribution of \bar{P} is ξ , so the equation $z = \bar{P}^* z$ implies that z is a nonzero multiple of e . Using Assumption 6.6.2, it follows that z does not belong to the subspace S , and from Prop. 6.6.1 (with \bar{P} in place of C , and β in place of γ), we see that $\Pi P^{(\lambda)}$ is a contraction with respect to the norm $\|\cdot\|_\xi$. This implies that $\Pi F^{(\lambda)}$ is also a contraction.

Consider next the case, $\gamma \in (0, 1)$ and $\lambda \in (0, 1)$. Since $\Pi F^{(\lambda)}$ is a contraction with respect to $\|\cdot\|_\xi$, as just shown, we have for any $J, \bar{J} \in \Re^n$,

$$\begin{aligned}\|F_{\gamma,\lambda}J - F_{\gamma,\lambda}\bar{J}\|_\xi &\leq (1-\gamma)\|J - \bar{J}\|_\xi + \gamma\|\Pi F^{(\lambda)}J - \Pi F^{(\lambda)}\bar{J}\|_\xi \\ &\leq (1-\gamma + \gamma\beta)\|J - \bar{J}\|_\xi,\end{aligned}$$

where β is the contraction modulus of $F^{(\lambda)}$. Hence, $F_{\gamma,\lambda}$ is a contraction.

Finally, consider the case $\gamma \in (0, 1)$ and $\lambda = 0$. We will show that the mapping ΠF has a unique fixed point, by showing that either 1 is not an eigenvalue of P , or else the eigenvectors corresponding to the eigenvalue 1 do not belong to S [cf. Prop. 6.6.1(a)]. Assume the contrary, i.e., that some $z \in S$ with $z \neq 0$ is an eigenvector corresponding to 1. We then have $z = Pz$. From this it follows that $z = P^k z$ for all $k \geq 0$, so $z = P^* z$, where $P^* = \lim_{N \rightarrow \infty} (1/N) \sum_{k=0}^{N-1} P^k$ (cf. Prop. 4.1.2). The rows of P^* are all equal to ξ' , so the equation $z = P^* z$ implies that z is a nonzero multiple of e . Hence, by Assumption 6.6.2, z cannot belong to S - a contradiction. Thus ΠF has a unique fixed point, and the contraction property of $F_{\gamma,\lambda}$ for $\gamma \in (0, 1)$ and $\lambda = 0$ follows from Prop. 6.6.1(b). **Q.E.D.**

Proposition 6.6.2 implies that for $\lambda > 0$, the mapping $\Pi F^{(\lambda)}$ has a unique fixed point, denoted Φr_λ^* . The LSPE(λ) algorithm can be shown to converge to r_λ^* with probability 1, by using the contraction property of $\Pi F^{(\lambda)}$. When $\lambda = 0$, the mapping $\Pi F^{(0)}$, which is ΠF , need not be a contraction, but because, according to Prop. 6.6.2, the mapping

$$F_{\gamma,0} = (1-\gamma)I + \gamma\Pi F$$

is a contraction for all $\gamma \in (0, 1)$, there is a vector r_0^* such that Φr_0^* is the (common) unique fixed point of all the mappings $F_{\gamma,0}$, $\gamma \in (0, 1)$. The *damped* version of PVI(0), which is the algorithm

$$r_{k+1} = (1-\gamma)r_k + \gamma\Pi F r_k,$$

converges to Φr_0^* for any $\gamma \in (0, 1)$. The corresponding damped version of LSPE(0) is the algorithm

$$r_{k+1} = (1-\gamma)r_k + \gamma\tilde{r}_k,$$

where \tilde{r}_k is what is obtained with an LSPE(0) iteration starting from r_k . It similarly converges to Φr_0^* for all values of $\gamma \in (0, 1)$.

As a final note, it can be shown that the convergence of LSPE(0) is guaranteed even in the case where a stepsize $\gamma = 1$ is used, provided P is aperiodic. The reason is that the mapping $F_{1,0}$ is a contraction in this case, but with respect to a different norm, not the norm $\|\cdot\|_\xi$ (see Yu and Bertsekas [YuB06b]). As a result the LSPE(0) algorithm converges without the need to reduce the stepsize from the value of 1. However, an example given in [YuB06b] shows that LSPE(0) with a unit stepsize may not converge if P is periodic, so $\gamma < 1$ is required for convergence of LSPE(0) when P is periodic.

Error Estimate

We have shown that for each $\lambda \in [0, 1)$, there is a vector Φr_λ^* , the unique fixed point of $\Pi F_{\gamma, \lambda}$, $\gamma \in (0, 1)$, which is the limit of $\text{LSPE}(\lambda)$ (cf. Prop. 6.6.2). Let h be any differential cost vector, and let $\beta_{\gamma, \lambda}$ be the modulus of contraction of $\Pi F_{\gamma, \lambda}$, with respect to $\|\cdot\|_\xi$. Similar to the proof of Prop. 6.3.2 for the discounted case, we have

$$\begin{aligned}\|h - \Phi r_\lambda^*\|_\xi^2 &= \|h - \Pi h\|_\xi^2 + \|\Pi h - \Phi r_\lambda^*\|_\xi^2 \\ &= \|h - \Pi h\|_\xi^2 + \|\Pi F_{\gamma, \lambda} h - \Pi F_{\gamma, \lambda}(\Phi r_\lambda^*)\|_\xi^2 \\ &\leq \|h - \Pi h\|_\xi^2 + \beta_{\gamma, \lambda} \|h - \Phi r_\lambda^*\|_\xi^2.\end{aligned}$$

It follows that

$$\|h - \Phi r_\lambda^*\|_\xi \leq \frac{1}{\sqrt{1 - \beta_{\gamma, \lambda}^2}} \|h - \Pi h\|_\xi, \quad \lambda \in [0, 1), \quad \gamma \in (0, 1), \quad (6.84)$$

for all differential cost vector vectors h .

This estimate is a little peculiar because the differential cost vector is not unique. The set of differential cost vectors is

$$D = \{h^* + \gamma e \mid \gamma \in \mathbb{R}\},$$

where h^* is the bias of the policy evaluated (cf. Section 4.1, and Props. 4.1.1 and 4.1.2). In particular, h^* is the unique $h \in D$ that satisfies $\xi' h = 0$ or equivalently $P^* h = 0$, where

$$P^* = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k.$$

Usually, in average cost policy evaluation, we are interested in obtaining a small error $(h - \Phi r_\lambda^*)$ with the choice of h being immaterial (see the discussion of the next section on approximate policy iteration). It follows that since the estimate (6.84) holds for all $h \in D$, a better error bound can be obtained by using an optimal choice of h in the left-hand side and an optimal choice of γ in the right-hand side. Indeed, Tsitsiklis and Van Roy [TsV99a] have obtained such an optimized error estimate. It has the form

$$\min_{h \in D} \|h - \Phi r_\lambda^*\|_\xi = \|h^* - (I - P^*) \Phi r_\lambda^*\|_\xi \leq \frac{1}{\sqrt{1 - \alpha_\lambda^2}} \|\Pi^* h^* - h^*\|_\xi, \quad (6.85)$$

where h^* is the bias vector, Π^* denotes projection with respect to $\|\cdot\|_\xi$ onto the subspace

$$S^* = \{(I - P^*)y \mid y \in S\},$$

and α_λ is the minimum over $\gamma \in (0, 1)$ of the contraction modulus of the mapping $\Pi^* F_{\gamma, \lambda}$:

$$\alpha_\lambda = \min_{\gamma \in (0, 1)} \max_{\|y\|_\xi=1} \|\Pi^* P_{\gamma, \lambda} y\|_\xi,$$

where $P_{\gamma, \lambda} = (1 - \gamma)I + \gamma \Pi^* P^{(\lambda)}$. Note that this error bound has similar form with the one for discounted problems (cf. Prop. 6.3.3), but S has been replaced by S^* and Π has been replaced by Π^* . It can be shown that the scalar α_λ decreases as λ increases, and approaches 0 as $\lambda \uparrow 1$. This is consistent with the corresponding error bound for discounted problems (cf. Prop. 6.3.3), and is also consistent with empirical observations, which suggest that smaller values of λ lead to larger approximation errors.

Figure 6.6.1 illustrates and explains the projection operation Π^* , the distance of the bias h^* from its projection $\Pi^* h^*$, and the other terms in the error bound (6.85).

LSTD(λ)

The LSTD(λ) algorithm for average cost is a straightforward extension of the discounted version, and will only be summarized. One can verify that the LSPE(λ) iteration can be written [similar to the discounted case, cf. Eq. (6.43)] as

$$r_{k+1} = r_k + \bar{B}_k^{-1} (\bar{A}_k r_k + \bar{b}_k), \quad (6.86)$$

where

$$\bar{A}_k = \frac{A_k}{k+1}, \quad \bar{B}_k = \frac{B_k}{k+1}, \quad \bar{b}_k = \frac{b_k}{t+1},$$

and the matrices A_k , B_k , and vector b_k are defined by

$$A_k = \sum_{t=0}^k z_t (\phi(i_{t+1})' - \phi(i_t)'), \quad B_k = \sum_{t=0}^k \phi(i_t) \phi(i_t)',$$

$$b_k = \sum_{t=0}^k z_t (g(i_t, i_{t+1}) - \eta_t), \quad z_t = \sum_{m=0}^t (\lambda)^{t-m} \phi(i_m)'.$$

The matrices \bar{A}_k , \bar{B}_k , and vector \bar{b}_k can be shown to converge to limits:

$$\bar{A}_k \rightarrow A, \quad \bar{B}_k \rightarrow B, \quad \bar{b}_k \rightarrow b,$$

where

$$A = \Phi' \Xi (P^{(\lambda)} - I) \Phi, \quad B = \Phi' \Xi \Phi, \quad b = \Phi' \Xi (I - \lambda P)^{-1} (g - \eta e), \quad (6.87)$$

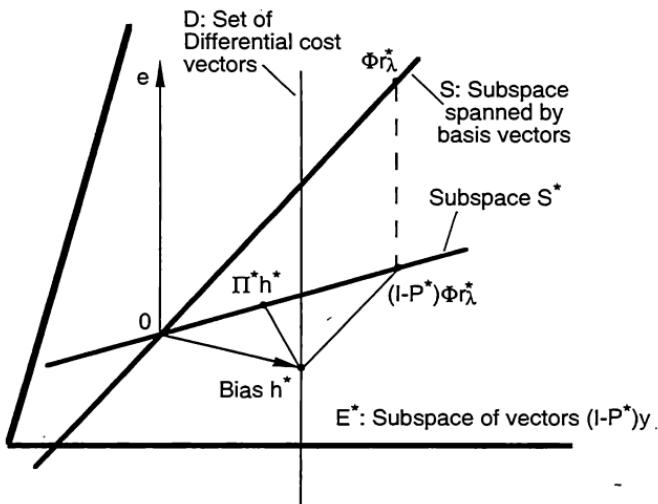


Figure 6.6.1: Illustration of the estimate (6.85). Consider the subspace

$$E^* = \{(I - P^*)y \mid y \in \mathbb{R}^n\}.$$

Let Ξ be the diagonal matrix with ξ_1, \dots, ξ_n on the diagonal. Note that:

- (a) E^* is the subspace that is orthogonal to the unit vector e in the scaled geometry of the norm $\|\cdot\|_\xi$, in the sense that $e'\Xi z = 0$ for all $z \in E^*$. Indeed we have

$$e' \Xi (I - P^*)y = 0, \quad \text{for all } y \in \mathbb{R}^n,$$

because $e'\Xi = \xi'$ and $\xi'(I - P^*) = 0$ as can be easily verified from the fact that the rows of P^* are all equal to ξ' .

- (b) Projection onto E^* with respect to the norm $\|\cdot\|_\xi$ is simply multiplication with $(I - P^*)$ (since $P^*y = \xi'y$, so P^*y is orthogonal to E^* in the scaled geometry of the norm $\|\cdot\|_\xi$). Thus, S^* is the projection of S onto E^* .
- (c) We have $h^* \in E^*$ since $(I - P^*)h^* = h^*$ in view of $P^*h^* = 0$.
- (d) The equation

$$\min_{h \in D} \|h - \Phi r_\lambda^*\|_\xi = \|h^* - (I - P^*)\Phi r_\lambda^*\|_\xi$$

is geometrically evident from the figure. Also, the term $\|\Pi^*h^* - h^*\|_\xi$ of the error bound is the minimum possible error given that h^* is approximated with an element of S^* .

- (e) The estimate (6.85), is the analog of the discounted estimate of Prop. 6.3.3, with E^* playing the role of the entire space, and with the “geometry of the problem” projected onto E^* . Thus, S^* plays the role of S , h^* plays the role of J_μ , $(I - P^*)\Phi r_\lambda^*$ plays the role of Φr_λ^* , and Π^* plays the role of Π . Finally, α_λ is the best possible contraction modulus of $\Pi^*F_{\gamma, \lambda}$ over $\gamma \in (0, 1)$ and within E^* (see the paper [TsV99a] for a detailed analysis).

the matrix $P^{(\lambda)}$ is defined by Eq. (6.79), and Ξ is the diagonal matrix with diagonal entries ξ_1, \dots, ξ_n :

$$\Xi = \text{diag}(\xi_1, \dots, \xi_n).$$

From the iteration formula (6.86), it also follows that the limit r_λ^* of LSPE(λ) is given by

$$r_\lambda^* = -A^{-1}b, \quad (6.88)$$

and using the formula of Eq. (6.87), one may show that A is invertible. The LSTD(λ) algorithm is given by

$$\hat{r}_k = -\bar{A}_k^{-1}\bar{b}_k.$$

Much of our discounted case discussion regarding the performance comparison of LSTD(λ) and LSPE(λ) applies. In particular, it can be shown that \hat{r}_k converges to r_{k+1} faster than \hat{r}_k converges to r_λ^* (see Yu and Bertsekas [YuB06b]). Thus, the two methods perform very similarly, in an asymptotic sense, at least in the context of single policy evaluation.

6.6.2 Approximate Policy Iteration

Let us consider an approximate policy iteration method that involves approximate policy evaluation and approximate policy improvement. We assume that *all stationary policies are unichain, and a special state s is recurrent in the Markov chain corresponding to each stationary policy*. As in Section 4.3.1, we consider the stochastic shortest path problem obtained by leaving unchanged all transition probabilities $p_{ij}(u)$ for $j \neq s$, by setting all transition probabilities $p_{is}(u)$ to 0, and by introducing an artificial termination state t to which we move from each state i with probability $p_{it}(u)$. The one-stage cost is equal to $g(i, u) - \eta$, where η is a scalar parameter. We refer to this stochastic shortest path problem as the η -SSP (cf. Fig 4.3.3).

The method generates a sequence of stationary policies μ_k , a corresponding sequence of gains η_{μ_k} , and a sequence of cost vectors h_k . We assume that for some $\epsilon > 0$, we have

$$\max_{i=1,\dots,n} |h_k(i) - h_{\mu_k, \eta_k}(i)| \leq \epsilon, \quad k = 0, 1, \dots,$$

where

$$\eta_k = \min_{m=0,1,\dots,k} \eta_{\mu_m},$$

$h_{\mu_k, \eta_k}(i)$ is the cost-to-go from state i to the reference state s for the η_k -SSP under policy μ_k , and ϵ is a positive scalar quantifying the accuracy of evaluation of the cost-to-go function of the η_k -SSP. Note that we assume

exact calculation of the gains η_{μ_k} . Note also that we may calculate approximate differential costs $\tilde{h}_k(i, r)$ that depend on a parameter vector r without regard to the reference state s . These differential costs may then be replaced by

$$h_k(i) = \tilde{h}_k(i, r) - \tilde{h}(s, r), \quad i = 1, \dots, n.$$

We assume that policy improvement is carried out by approximate minimization in the DP mapping. In particular, we assume that there exists a tolerance $\delta > 0$ such that for all i and k , $\mu_{k+1}(i)$ attains the minimum in the expression

$$\min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u) (g(i, u, j) + h_k(j)),$$

within a tolerance $\delta > 0$.

We now note that since η_k is monotonically nonincreasing and is bounded below by the optimal gain η^* , it must converge to some scalar $\bar{\eta}$. Since η_k can take only one of the finite number of values η_μ corresponding to the finite number of stationary policies μ , we see that η_k must converge finitely to $\bar{\eta}$; that is, for some \bar{k} , we have

$$\eta_k = \bar{\eta}, \quad k \geq \bar{k}.$$

Let $h_{\bar{\eta}}(s)$ denote the optimal cost-to-go from state s in the $\bar{\eta}$ -SSP. Then, by using Prop. 2.4.1, we have

$$\limsup_{k \rightarrow \infty} (h_{\mu_k, \bar{\eta}}(s) - h_{\bar{\eta}}(s)) \leq \frac{n(1 - \rho + n)(\delta + 2\epsilon)}{(1 - \rho)^2}, \quad (6.89)$$

where

$$\rho = \max_{i=1, \dots, n, \mu} P(i_k \neq s, k = 1, \dots, n \mid i_0 = i, \mu),$$

and i_k denotes the state of the system after k stages. On the other hand, as can also be seen from Fig. 6.6.2, the relation

$$\bar{\eta} \leq \eta_{\mu_k}$$

implies that

$$h_{\mu_k, \bar{\eta}}(s) \geq h_{\mu_k, \eta_{\mu_k}}(s) = 0.$$

It follows, using also Fig. 6.6.2, that

$$h_{\mu_k, \bar{\eta}}(s) - h_{\bar{\eta}}(s) \geq -h_{\bar{\eta}}(s) \geq -h_{\mu^*, \bar{\eta}}(s) = (\bar{\eta} - \eta^*) N_{\mu^*}, \quad (6.90)$$

where μ^* is an optimal policy for the η^* -SSP (and hence also for the original average cost per stage problem) and N_{μ^*} is the expected number of stages

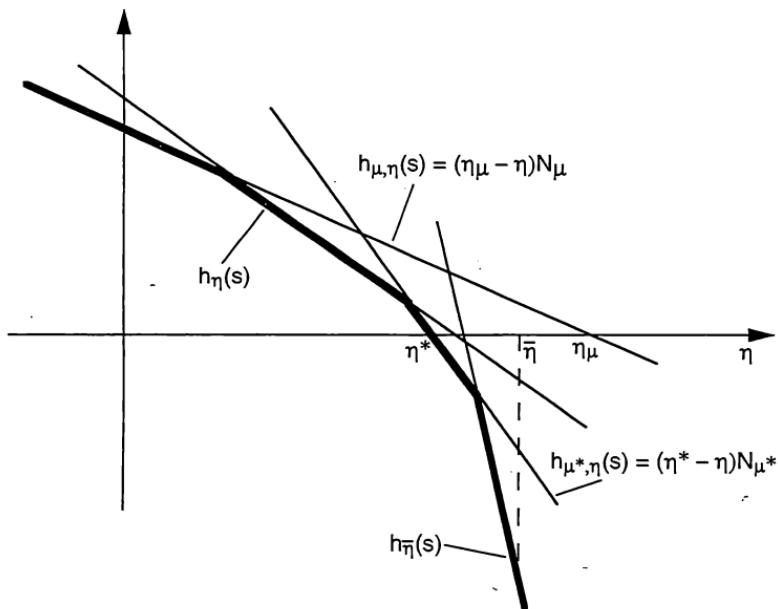


Figure 6.6.2: Relation of the costs of stationary policies for the η -SSP in the approximate policy iteration method (cf. Fig. 4.3.3). Here, N_μ is the expected number of stages to return to state s , starting from s and using μ . Since $\eta_{\mu_k} \geq \bar{\eta}$, we have

$$h_{\mu_k, \bar{\eta}}(s) \geq h_{\mu_k, \eta_{\mu_k}}(s) = 0.$$

Furthermore, if μ^* is an optimal policy for the η^* -SSP, we have

$$h_{\bar{\eta}}(s) \leq h_{\mu^*, \bar{\eta}}(s) = (\eta^* - \bar{\eta})N_{\mu^*}.$$

to return to state s , starting from s and using μ^* . Thus, from Eqs. (6.89) and (6.90), we have

$$\bar{\eta} - \eta^* \leq \frac{n(1 - \rho + n)(\delta + 2\epsilon)}{N_{\mu^*}(1 - \rho)^2}. \quad (6.91)$$

This relation provides an estimate on the steady-state error of the approximate policy iteration method.

We finally note that optimistic versions of the preceding approximate policy iteration method are harder to implement than their discounted cost counterparts. The reason is our assumption that the gain η_μ of every generated policy μ is exactly calculated; in an optimistic method the current policy μ may not remain constant for sufficiently long time to estimate accurately η_μ . One may consider schemes where an optimistic version of policy iteration is used to solve the η -SSP for a fixed η . The value of η

may occasionally be adjusted downward by calculating “exactly” through simulation the gain η_μ of some of the (presumably most promising) generated policies μ , and by then updating η according to $\eta := \min\{\eta, \eta_\mu\}$. An alternative is to approximate the average cost problem with a discounted problem, for which an optimistic version of approximate policy iteration can be readily implemented.

6.6.3 Q-Learning for Average Cost Problems

To derive the appropriate form of the Q -learning algorithm, we form an auxiliary average cost problem by augmenting the original system with one additional state for each possible pair (i, u) with $u \in U(i)$. Thus, the states of the auxiliary problem are those of the original problem, $i = 1, \dots, n$, together with the additional states (i, u) , $i = 1, \dots, n$, $u \in U(i)$. The probabilistic transition mechanism from an original state i is the same as for the original problem [probability $p_{ij}(u)$ of moving to state j], while the probabilistic transition mechanism from a state (i, u) is that we move only to states j of the original problem with corresponding probabilities $p_{ij}(u)$ and costs $g(i, u, j)$.

It can be seen that the auxiliary problem has the same optimal average cost per stage η as the original, and that the corresponding Bellman’s equation is

$$\eta + h(i) = \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u)(g(i, u, j) + h(j)), \quad i = 1, \dots, n, \quad (6.92)$$

$$\eta + Q(i, u) = \sum_{j=1}^n p_{ij}(u)(g(i, u, j) + h(j)), \quad i = 1, \dots, n, \quad u \in U(i), \quad (6.93)$$

where $Q(i, u)$ is the differential cost corresponding to (i, u) . Taking the minimum over u in Eq. (6.93) and comparing with Eq. (6.92), we obtain

$$h(i) = \min_{u \in U(i)} Q(i, u), \quad i = 1, \dots, n.$$

Substituting the above form of $h(i)$ in Eq. (6.93), we obtain Bellman’s equation in a form that exclusively involves the Q -factors:

$$\eta + Q(i, u) = \sum_{j=1}^n p_{ij}(u) \left(g(i, u, j) + \min_{u' \in U(j)} Q(j, u') \right), \quad i = 1, \dots, n, \quad u \in U(i).$$

Let us now apply to the auxiliary problem the following variant of the relative value iteration

$$h^{k+1} = Th^k - h^k(s)e,$$

where s is a special state. We then obtain the iteration [cf. Eqs. (6.92) and (6.93)]

$$\begin{aligned} h^{k+1}(i) &= \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u) (g(i, u, j) + h^k(j)) - h^k(s), \quad i = 1, \dots, n, \\ Q^{k+1}(i, u) &= \sum_{j=1}^n p_{ij}(u) (g(i, u, j) + h^k(j)) - h^k(s), \quad i = 1, \dots, n, u \in U(i). \end{aligned} \quad (6.94)$$

From these equations, we have that

$$h^k(i) = \min_{u \in U(i)} Q^k(i, u), \quad i = 1, \dots, n,$$

and by substituting the above form of h^k in Eq. (6.94), we obtain the following relative value iteration for the Q -factors

$$Q^{k+1}(i, u) = \sum_{j=1}^n p_{ij}(u) \left(g(i, u, j) + \min_{u' \in U(j)} Q^k(j, u') \right) - \min_{u' \in U(s)} Q^k(s, u').$$

The sequence of values $\min_{u \in U(s)} Q^k(s, u)$ is expected to converge to the optimal average cost per stage and the sequences of values $\min_{u \in U(i)} Q(i, u)$ are expected to converge to differential costs $h(i)$.

An incremental version of the preceding iteration that involves a positive stepsize γ is given by

$$\begin{aligned} Q(i, u) := (1 - \gamma)Q(i, u) + \gamma \left(\sum_{j=1}^n p_{ij}(u) \left(g(i, u, j) + \min_{u' \in U(j)} Q(j, u') \right) \right. \\ \left. - \min_{u' \in U(s)} Q(s, u') \right). \end{aligned}$$

The natural form of the Q -learning method for the average cost problem is an approximate version of this iteration, whereby the expected value is replaced by a single sample, i.e.,

$$\begin{aligned} Q(i, u) := Q(i, u) + \gamma \left(g(i, u, j) + \min_{u' \in U(j)} Q(j, u') - \min_{u' \in U(s)} Q(s, u') \right. \\ \left. - Q(i, u) \right), \end{aligned}$$

where j and $g(i, u, j)$ are generated from the pair (i, u) by simulation. A convergence analysis of this method can be found in the paper by Abounadi, Bertsekas, and Borkar [ABB01].

Q-Learning Based on the Contracting Value Iteration

We now consider an alternative Q -learning method, which is based on the contracting value iteration method of Section 4.3. If we apply this method to the auxiliary problem used above, we obtain the following algorithm

$$h^{k+1}(i) = \min_{u \in U(i)} \left[\sum_{j=1}^n p_{ij}(u)g(i, u, j) + \sum_{\substack{j=1 \\ j \neq s}}^n p_{ij}(u)h^k(j) \right] - \eta^k, \quad (6.95)$$

$$Q^{k+1}(i, u) = \sum_{j=1}^n p_{ij}(u)g(i, u, j) + \sum_{\substack{j=1 \\ j \neq s}}^n p_{ij}(u)h^k(j) - \eta^k, \quad (6.96)$$

$$\eta^{k+1} = \eta^k + \alpha^k h^{k+1}(s).$$

From these equations, we have that

$$h^k(i) = \min_{u \in U(i)} Q^k(i, u),$$

and by substituting the above form of h^k in Eq. (6.96), we obtain

$$Q^{k+1}(i, u) = \sum_{j=1}^n p_{ij}(u)g(i, u, j) + \sum_{\substack{j=1 \\ j \neq s}}^n p_{ij}(u) \min_{v \in U(j)} Q^k(j, v) - \eta^k,$$

$$\eta^{k+1} = \eta^k + \alpha^k \min_{v \in U(s)} Q^{k+1}(s, v).$$

A small-stepsize version of this iteration is given by

$$Q(i, u) := (1 - \gamma)Q(i, u) + \gamma \left(\sum_{j=1}^n p_{ij}(u)g(i, u, j) + \sum_{\substack{j=1 \\ j \neq s}}^n p_{ij}(u) \min_{v \in U(j)} Q(j, v) - \eta \right),$$

$$\eta := \eta + \alpha \min_{v \in U(s)} Q(s, v),$$

where γ and α are positive stepsizes. A natural form of Q -learning based on this iteration is obtained by replacing the expected values by a single sample, i.e.,

$$Q(i, u) := (1 - \gamma)Q(i, u) + \gamma \left(g(i, u, j) + \min_{v \in U(j)} \hat{Q}(j, v) - \eta \right), \quad (6.97)$$

$$\eta := \eta + \alpha \min_{v \in U(s)} Q(s, v), \quad (6.98)$$

where

$$\hat{Q}(j, v) = \begin{cases} Q(j, v), & \text{if } j \neq s, \\ 0, & \text{otherwise,} \end{cases}$$

and j and $g(i, u, j)$ are generated from the pair (i, u) by simulation. Here the stepsizes γ and α should be diminishing, but α should diminish “faster” than γ ; i.e., the ratio of the stepsizes α/γ should converge to zero. For example, we may use $\gamma = C/k$ and $\alpha = c/k \log k$, where C and c are positive constants and k is the number of iterations performed on the corresponding pair (i, u) or η , respectively.

The algorithm has two components: the iteration (6.97), which is essentially a Q -learning method that aims to solve the η -SSP for the current value of η , and the iteration (6.98), which updates η towards its correct value η^* . However, η is updated at a slower rate than Q , since the stepsize ratio α/γ converges to zero. The effect is that the Q -learning iteration (6.97) is fast enough to keep pace with the slower changing η -SSP. A convergence analysis of this method can also be found in the paper [ABB01].

6.7 APPROXIMATION IN POLICY SPACE

Our approach so far in this chapter has been to use an approximation architecture for some cost function, differential cost, or Q -factor. Sometimes this is called *approximation in value space*, to indicate that a cost or value function is being approximated. In an important alternative, called *approximation in policy space*, we parameterize the set of policies by a vector $r = (r_1, \dots, r_s)$ and we optimize the cost over this vector. In particular, we consider randomized stationary policies of a given parametric form $\tilde{\mu}_u(i, r)$, where $\tilde{\mu}_u(i, r)$ denotes the probability that control u is applied when the state is i . Each value of r defines a randomized stationary policy, which in turn defines the cost of interest as a function of r . We then choose r to minimize this cost.

In an important special case of this approach, the parameterization of the policies is indirect, through an approximate cost function. In particular, a cost approximation architecture parameterized by r , defines a policy dependent on r via the minimization in Bellman’s equation. For example, Q -factor approximations $\tilde{Q}(i, u, r)$, define a parameterization of policies by letting $\tilde{\mu}_u(i, r) = 1$ for some u that minimizes $\tilde{Q}(i, u, r)$ over $u \in U(i)$, and $\tilde{\mu}_u(i, r) = 0$ for all other u . This parameterization is discontinuous in r , but in practice it is smoothed by replacing the minimization operation with a smooth exponential-based approximation; we refer to the literature for the details. Also in a more abstract and general view of approximation in policy space, rather than parameterizing policies or Q -factors, we can simply parameterize by r the problem data (stage costs and transition

probabilities), and optimize the corresponding cost function over r . Thus, in this more general formulation, we may aim to select some parameters of a given system to optimize performance.

We will focus on the finite spaces average cost problem of the preceding section (a similar development is possible for discounted and SSP problems). Let the cost per stage vector and transition probability matrix be given as functions of r : $G(r)$ and $P(r)$, respectively. Assume that the states form a single recurrent class under each $P(r)$, and let $\xi(r)$ be the corresponding steady-state probability vector. We denote by $G_i(r)$, $P_{ij}(r)$, and $\xi_i(r)$ the components of $G(r)$, $P(r)$, and $\xi(r)$, respectively. Each value of r defines an average cost $\eta(r)$, and the problem is to find

$$\min_{r \in \mathbb{R}^s} \eta(r).$$

Assuming that $\eta(r)$ is differentiable with respect to r (something that must be independently verified), one may use a gradient method for this minimization:

$$r_{k+1} = r_k - \gamma_k \nabla \eta(r_k),$$

where γ_k is a positive stepsize. This is known as a *policy gradient method*.

The Gradient Formula

We will now show that a convenient formula for the gradients $\nabla \eta(r)$ can be obtained by differentiating Bellman's equation

$$\eta(r) + h_i(r) = G_i(r) + \sum_{j=1}^n P_{ij}(r)h_j(r), \quad i = 1, \dots, n, \quad (6.99)$$

with respect to the components of r , where $h_i(r)$ are the differential costs. Taking the partial derivative with respect to r_m , we obtain for all i and m ,

$$\frac{\partial \eta}{\partial r_m} + \frac{\partial h_i}{\partial r_m} = \frac{\partial G_i}{\partial r_m} + \sum_{j=1}^n \frac{\partial P_{ij}}{\partial r_m} h_j + \sum_{j=1}^n P_{ij} \frac{\partial h_j}{\partial r_m}.$$

(In what follows we assume that the partial derivatives with respect to components of r appearing in various equations exist. The argument at which they are evaluated, is often suppressed to simplify notation.) By multiplying this equation with $\xi_i(r)$, adding over i , and using the fact $\sum_{i=1}^n \xi_i(r) = 1$, we obtain

$$\frac{\partial \eta}{\partial r_m} + \sum_{i=1}^n \xi_i \frac{\partial h_i}{\partial r_m} = \sum_{i=1}^n \xi_i \frac{\partial G_i}{\partial r_m} + \sum_{i=1}^n \xi_i \sum_{j=1}^n \frac{\partial P_{ij}}{\partial r_m} h_j + \sum_{i=1}^n \xi_i \sum_{j=1}^n P_{ij} \frac{\partial h_j}{\partial r_m}.$$

The last summation on the right-hand side cancels the last summation on the left-hand side, because from the defining property of the steady-state probabilities, we have

$$\sum_{i=1}^n \xi_i \sum_{j=1}^n P_{ij} \frac{\partial h_j}{\partial r_m} = \sum_{j=1}^n \left(\sum_{i=1}^n \xi_i P_{ij} \right) \frac{\partial h_j}{\partial r_m} = \sum_{j=1}^n \xi_j \frac{\partial h_j}{\partial r_m}.$$

We thus obtain

$$\frac{\partial \eta(r)}{\partial r_m} = \sum_{i=1}^n \xi_i(r) \left(\frac{\partial G_i(r)}{\partial r_m} + \sum_{j=1}^n \frac{\partial P_{ij}(r)}{\partial r_m} h_j(r) \right), \quad m = 1, \dots, s, \quad (6.100)$$

or in more compact form

$$\nabla \eta(r) = \sum_{i=1}^n \xi_i(r) \left(\nabla G_i(r) + \sum_{j=1}^n \nabla P_{ij}(r) h_j(r) \right), \quad (6.101)$$

where all the gradients are column vectors of dimension s .

Computing the Gradient by Simulation

Despite its relative simplicity, the gradient formula (6.101) involves formidable computations to obtain $\nabla \eta(r)$ at just a single value of r . The reason is that neither the steady-state probability vector $\xi(r)$ nor the bias vector $h(r)$ are readily available, so they must be computed or approximated in some way. Furthermore, $h(r)$ is a vector of dimension n , so for large n , it can only be approximated either through its simulation samples or by using a parametric architecture and an algorithm such as LSPE or LSTD (see the references cited at the end of the chapter).

The possibility to approximate h using a parametric architecture ushers a connection between approximation in policy space and approximation in value space. It also raises the question whether approximations introduced in the gradient calculation may affect the convergence guarantees of the policy gradient method. Fortunately, however, gradient algorithms tend to be robust and maintain their convergence properties, even in the presence of significant error in the calculation of the gradient.

In the literature, algorithms where both μ and h are parameterized are sometimes called *actor-critic* methods. Algorithms where just μ is parameterized and h is not parameterized but rather estimated explicitly or implicitly by simulation, are called *actor-only* methods, while algorithms where just h is parameterized and μ is obtained by one-step lookahead minimization, are called *critic-only* methods.

We will now discuss some possibilities of using simulation to approximate $\nabla\eta(r)$. Let us introduce for all i and j such that $P_{ij}(r) > 0$, the function

$$L_{ij}(r) = \frac{\nabla P_{ij}(r)}{P_{ij}(r)}.$$

Then, suppressing the dependence on r , we write the partial derivative formula (6.101) in the form

$$\nabla\eta = \sum_{i=1}^n \xi_i \left(\nabla G_i + \sum_{j=1}^n P_{ij} L_{ij} h_j \right). \quad (6.102)$$

We assume that for all states i and possible transitions (i, j) , we can calculate ∇G_i and L_{ij} . Suppose now that we generate a single infinitely long simulated trajectory (i_0, i_1, \dots) . We can then estimate the average cost η as

$$\tilde{\eta} = \frac{1}{k} \sum_{t=0}^{k-1} G_{i_t},$$

where k is large. Then, given an estimate $\tilde{\eta}$, we can estimate the bias components h_j by using simulation-based approximations to the formula

$$h_{i_0} = \lim_{N \rightarrow \infty} E \left\{ \sum_{t=0}^N (G_{i_t} - \eta) \right\},$$

[which holds from general properties of the bias vector when $P(r)$ is aperiodic – see the discussion following Prop. 4.1.2]. Alternatively, we can estimate h_j by using the LSPE or LSTD algorithms of Section 6.6.1 [note here that if the feature subspace contains the bias vector, the LSPE and LSTD algorithms will find exact values of h_j in the limit, so with a sufficiently rich set of features, an asymptotically exact calculation of h_j , and hence also $\nabla\eta(r)$, is possible]. Finally, given estimates $\tilde{\eta}$ and \tilde{h}_j , we can estimate the gradient $\nabla\eta$ with a vector δ_η given by

$$\delta_\eta = \frac{1}{k} \sum_{t=0}^{k-1} (\nabla G_{i_t} + L_{i_t i_{t+1}} \tilde{h}_{i_{t+1}}). \quad (6.103)$$

This can be seen by a comparison of Eqs. (6.102) and (6.103): if we replace the expected values of ∇G_i and L_{ij} by empirical averages, and we replace h_j by \tilde{h}_j , we obtain the estimate δ_η .

The estimation-by-simulation procedure outlined above provides a conceptual starting point for more practical gradient estimation methods. For example, in such methods, the estimation of η and h_j may be done simultaneously with the estimation of the gradient via Eq. (6.103), and with a variety of different algorithms. We refer to the literature cited at the end of the chapter.

Essential Features of Critics

We will now develop an alternative (but mathematically equivalent) expression for the gradient $\nabla\eta(r)$ that involves Q -factors instead of differential costs. Let us consider randomized policies where $\tilde{\mu}_u(i, r)$ denotes the probability that control u is applied at state i . We assume that $\tilde{\mu}_u(i, r)$ is differentiable with respect to r for each i and u . Then the corresponding stage costs and transition probabilities are given by

$$G_i(r) = \sum_{u \in U(i)} \tilde{\mu}_u(i, r) \sum_{j=1}^n p_{ij}(u) g(i, u, j), \quad i = 1, \dots, n,$$

$$P_{ij}(r) = \sum_{u \in U(i)} \tilde{\mu}_u(i, r) p_{ij}(u), \quad i, j = 1, \dots, n.$$

Differentiating these equations with respect to r , we obtain

$$\nabla G_i(r) = \sum_{u \in U(i)} \nabla \tilde{\mu}_u(i, r) \sum_{j=1}^n p_{ij}(u) g(i, u, j), \quad (6.104)$$

$$\nabla P_{ij}(r) = \sum_{u \in U(i)} \nabla \tilde{\mu}_u(i, r) p_{ij}(u), \quad i, j = 1, \dots, n. \quad (6.105)$$

Since $\sum_{u \in U(i)} \tilde{\mu}_u(i, r) = 1$ for all r , we have $\sum_{u \in U(i)} \nabla \tilde{\mu}_u(i, r) = 0$, so Eq. (6.104) yields

$$\nabla G_i(r) = \sum_{u \in U(i)} \nabla \tilde{\mu}_u(i, r) \left(\sum_{j=1}^n p_{ij}(u) g(i, u, j) - \eta(r) \right).$$

Also, by multiplying with $h_j(r)$ and adding over j , Eq. (6.105) yields

$$\sum_{j=1}^n \nabla P_{ij}(r) h_j(r) = \sum_{j=1}^n \sum_{u \in U(i)} \nabla \tilde{\mu}_u(i, r) p_{ij}(u) h_j(r).$$

By using the preceding two equations to rewrite the gradient formula (6.101), we obtain

$$\begin{aligned} \nabla \eta(r) &= \sum_{i=1}^n \xi_i(r) \left(\nabla G_i(r) + \sum_{j=1}^n \nabla P_{ij}(r) h_j(r) \right) \\ &= \sum_{i=1}^n \xi_i(r) \sum_{u \in U(i)} \nabla \tilde{\mu}_u(i, r) \sum_{j=1}^n p_{ij}(u) (g(i, u, j) - \eta(r) + h_j(r)), \end{aligned}$$

and finally

$$\nabla \eta(r) = \sum_{i=1}^n \sum_{u \in U(i)} \xi_i(r) \tilde{Q}(i, u, r) \nabla \tilde{\mu}_u(i, r), \quad (6.106)$$

where $\tilde{Q}(i, u, r)$ are the approximate Q -factors corresponding to r :

$$\tilde{Q}(i, u, r) = \sum_{j=1}^n p_{ij}(u) (g(i, u, j) - \eta(r) + h_j(r)).$$

Let us now express the formula (6.106) in a way that is amenable to proper interpretation. In particular, by writing

$$\nabla \eta(r) = \sum_{i=1}^n \sum_{\{u \in U(i) | \tilde{\mu}_u(i, r) > 0\}} \xi_i(r) \tilde{\mu}_u(i, r) \tilde{Q}(i, u, r) \frac{\nabla \tilde{\mu}_u(i, r)}{\tilde{\mu}_u(i, r)},$$

and by introducing the function

$$\psi_r(i, u) = \frac{\nabla \tilde{\mu}_u(i, r)}{\tilde{\mu}_u(i, r)},$$

we obtain

$$\nabla \eta(r) = \sum_{i=1}^n \sum_{\{u \in U(i) | \tilde{\mu}_u(i, r) > 0\}} \zeta_r(i, u) \tilde{Q}(i, u, r) \psi_r(i, u), \quad (6.107)$$

where $\zeta_r(i, u)$ are the steady-state probabilities of the pairs (i, u) under r :

$$\zeta_r(i, u) = \xi_i(r) \tilde{\mu}_u(i, r).$$

Note that for each (i, u) , $\psi_r(i, u)$ is a vector of dimension s , the dimension of the parameter vector r . We denote by $\psi_r^m(i, u)$, $m = 1, \dots, s$, the components of this vector.

Equation (6.107) can form the basis for policy gradient methods that estimate $\tilde{Q}(i, u, r)$ by simulation, thereby leading to actor-only algorithms. An alternative suggested by Konda and Tsitsiklis [KoT99], [KoT03], is to interpret the formula as an inner product, thereby leading to a different set of algorithms. In particular, for a given r , we define the inner product of two real-valued functions Q_1, Q_2 of (i, u) , by

$$\langle Q_1, Q_2 \rangle_r = \sum_{i=1}^n \sum_{\{u \in U(i) | \tilde{\mu}_u(i, r) > 0\}} \zeta_r(i, u) Q_1(i, u) Q_2(i, u).$$

With this notation, we can rewrite Eq. (6.107) as

$$\frac{\partial \eta(r)}{\partial r_m} = \langle \tilde{Q}(\cdot, \cdot, r), \psi_r^m(\cdot, \cdot) \rangle_r, \quad m = 1, \dots, s.$$

An important observation is that although $\nabla\eta(r)$ depends on $\tilde{Q}(i, u, r)$, which has a number of components equal to the number of state-control pairs (i, u) , the dependence is only through its inner products with the s functions $\psi_r^m(\cdot, \cdot)$, $m = 1, \dots, s$.

Now let $\|\cdot\|_r$ be the norm induced by this inner product, i.e.,

$$\|Q\|_r^2 = \langle Q, Q \rangle_r.$$

Let also S_r be the subspace that is spanned by the functions $\psi_r^m(\cdot, \cdot)$, $m = 1, \dots, s$, and let Π_r denote projection with respect to this norm onto S_r . Since

$$\langle \tilde{Q}(\cdot, \cdot, r), \psi_r^m(\cdot, \cdot) \rangle_r = \langle \Pi_r \tilde{Q}(\cdot, \cdot, r), \psi_r^m(\cdot, \cdot) \rangle_r, \quad m = 1, \dots, s,$$

it is sufficient to know the projection of $\tilde{Q}(\cdot, \cdot, r)$ onto S_r in order to compute $\nabla\eta(r)$. Thus S_r defines a subspace of *essential features*, i.e., features the knowledge of which is essential for the calculation of the gradient $\nabla\eta(r)$. As discussed in Section 6.2, the projection of $\tilde{Q}(\cdot, \cdot, r)$ onto S_r can be done in an approximate sense with TD(λ), LSPE(λ), or LSTD(λ) for $\lambda \approx 1$. We refer to the papers by Konda and Tsitsiklis [KoT99], [KoT03], and Sutton, McAllester, Singh, and Mansour [SMS99] for further discussion.

Approximation in Policy Space vs. Approximation in Value Space

Let us now provide a comparative assessment of approximation in policy and value space. We first note that in comparing approaches, one must bear in mind that specific problems may admit natural parametrizations that favor one type of approximation over the other. For example, in inventory control problems, it is natural to consider policy parametrizations that resemble the (s, S) policies that are optimal for special cases, but also make intuitive sense in a broader context.

Policy gradient methods for approximation in policy space are supported by interesting theory and aim directly at finding an optimal policy within the given parametric class (as opposed to aiming for policy evaluation in the context of an approximate policy iteration scheme). However, they suffer from a drawback that is well-known to practitioners of nonlinear optimization: slow convergence, which unless improved through the use of effective scaling of the gradient (with an appropriate diagonal or nondiagonal matrix), all too often leads to jamming (no visible progress) and complete breakdown. Unfortunately, there has been no proposal of a demonstrably effective scheme to scale the gradient in policy gradient methods (see, however, Kakade [Kak01] for an interesting attempt to address this issue, based on the work of Amari [Ama98]). Furthermore, the performance and reliability of policy gradient methods are susceptible to

degradation by large variance of simulation noise. Thus, while policy gradient methods are supported by convergence guarantees in theory, attaining convergence in practice is often challenging. In addition, there is the generic difficulty that gradient methods have with local minima, which is not well-understood at present in the context of approximation in policy space.

A major difficulty for approximation in value space is that a good choice of basis functions/features is often far from evident. Furthermore, even when good features are available, the indirect approach of $\text{TD}(\lambda)$, $\text{LSPE}(\lambda)$, and $\text{LSTD}(\lambda)$ may neither yield the best possible approximation of the cost function or the Q -factors of a policy within the feature subspace, nor yield the best possible performance of the associated one-step-lookahead policy. In the case of a fixed policy, $\text{LSPE}(\lambda)$ and $\text{LSTD}(\lambda)$ are quite reliable algorithms, in the sense that they ordinarily achieve their theoretical guarantees in approximating the associated cost function or Q -factors: they involve solution of systems of linear equations, simulation (with convergence governed by the law of large numbers), and contraction iterations (with favorable contraction modulus when λ is not too close to 0). However, within the multiple policy context of an approximate policy iteration scheme, approximation in value space has additional difficulties (regardless of the algorithm used for policy evaluation): the need for adequate exploration, the chattering phenomenon and the associated issue of policy oscillation, and the lack of convergence guarantees for both optimistic and nonoptimistic schemes.

6.8 NOTES, SOURCES, AND EXERCISES

There has been intensive interest in simulation-based methods for approximate DP since the early 90s, in view of their promise to address the dual curses of DP: the curse of dimensionality (the explosion of the computation needed to solve the problem as the number of states increases), and the curse of modeling (the need for an exact model of the system's dynamics). We have used the name *approximate dynamic programming* to collectively refer to these methods. Two other popular names are *reinforcement learning* and *neuro-dynamic programming*. The latter name, adopted by Bertsekas and Tsitsiklis [BeT96], comes from the strong connections with DP as well as with methods traditionally developed in the field of neural networks, such as the training of approximation architectures using empirical or simulation data.

Two books have been written so far on the subject, one by Sutton and Barto [SuB98], which reflects an artificial intelligence viewpoint, and another by Bertsekas and Tsitsiklis [BeT96], which is more mathematical and reflects an optimal control/operations research viewpoint. We refer to the latter book for a broader discussion of some of the topics of this chapter,

for related material on approximation architectures, batch and incremental gradient methods, and neural network training, as well as for an extensive overview of the history and bibliography of the subject.

Several survey papers in the volume by Si, Barto, Powell, and Wunsch [SBP04] describe recent work and approximation methodology that we have not covered in this chapter: linear programming-based approaches (De Farias [DeF04]), large-scale resource allocation methods (Powell and Van Roy [PoV04]), and deterministic optimal control approaches (Ferrari and Stengel [FeS04], and Si, Yang, and Liu [SYL04]).

Temporal differences originated in reinforcement learning, where they are viewed as a means to encode the error in predicting future costs, which is associated with an approximation architecture. An interpretation commonly adopted within this context is that $\text{TD}(\lambda)$ aims to keep this error small, and involves “looking back in time and correcting previous predictions.” The influential paper by Sutton [Sut88] proposed $\text{TD}(\lambda)$, albeit without a convergence analysis. The convergence of $\text{TD}(\lambda)$ and related methods was considered for discounted problems by several authors, including Dayan [Day92], Gurvits, Lin, and Hanson [GLH94], Jaakkola, Jordan, and Singh [JJS94], Pineda [Pin97], Tsitsiklis and Van Roy [TsV97], and Van Roy [Van98]. The analysis of Tsitsiklis and Van Roy [TsV97] was based on the contraction property of ΠT (cf. Lemma 6.3.1 and Prop. 6.3.1), which is the starting point of our analysis of Section 6.3.

The LSPE(λ) algorithm, was first proposed for stochastic shortest path problems by Bertsekas and Ioffe [BeI96], and was described in the book by Bertsekas and Tsitsiklis [BeT96], without a convergence analysis. The convergence of the method for discounted problems was given by Nedić and Bertsekas [NeB03] (for a diminishing stepsize), and by Bertsekas, Borkar, and Nedić [BBN04] (for a unit stepsize). In the paper [BeI96] and the book [BeT96], the LSPE method was related to a version of the policy iteration method, called λ -policy iteration (see also Exercise 1.19), and it was applied to the game of tetris.

The LSTD(λ) algorithm was first proposed by Bradtko and Barto [BrB96] for $\lambda = 0$, and later extended by Boyan [Boy02] for $\lambda > 0$. It has been discussed within the context of approximate policy iteration by Lagoudakis and Parr [LaP03]. A proof of convergence of LSTD for discounted problems was given by Nedić and Bertsekas [NeB03]. The rate of convergence of LSTD was analyzed by Konda [Kon02], who showed that LSTD has optimal rate of convergence within a broad class of temporal difference methods. Bertsekas, Borkar, and Nedić [BBN04] compared informally LSPE and LSTD for discounted problems, and suggested that they asymptotically coincide in the sense described in Section 6.3.4. Yu and Bertsekas [YuB06b] provided a mathematical proof of this for both discounted and average cost problems.

Q -learning was proposed by Watkins [Wat89], who explained insightfully the essence of the method, but did not provide a rigorous convergence

analysis; see also Watkins and Dayan [WaD92]. A convergence proof was given by Tsitsiklis [Tsi94b]. Some of the assumptions of Tsitsiklis for SSP problems with improper policies were relaxed by Abounadi, Bertsekas, and Borkar [ABB02], using an alternative line of proof. For a survey of related methods, which also includes many historical and other references, see Barto, Bradtke, and Singh [BBS95]. A variant of Q -learning is the method of advantage updating, developed by Baird [Bai93], [Bai94], [Bai95], and Harmon, Baird, and Klopff [HBK94] (see the book [BeT96]).

Approximation methods for the optimal stopping problem (Section 6.4.2) were first investigated by Tsitsiklis and Van Roy [TsV99b], [Van98], who noted that Q -learning with linear function approximation could be applied because the associated mapping F is a contraction with respect to the norm $\|\cdot\|_\xi$. They proved the convergence of a corresponding Q -learning method, and they applied it to a problem of pricing financial derivatives. The LSPE algorithm given in Section 6.4.2 for this problem is due to Yu and Bertsekas [YuB06c], to which we refer for additional analysis. An alternative algorithm with some similarity to LSPE is given by Choi and Van Roy [ChV06], and is also applied to the optimal stopping problem. We note that approximate dynamic programming and simulation methods for stopping problems have become popular in the finance area, within the context of pricing options; see e.g., Longstaff and Schwartz [LoS01].

The LSPE algorithm for SSP problems in Section 6.5 is the one originally proposed by Bertsekas and Ioffe [BeI96], and applied to a challenging problem: learning an optimal strategy to play the game of tetris (see also Bertsekas and Tsitsiklis [BeT96], Section 8.3). The associated contraction mapping analysis (Prop. 6.5.1) is based on the convergence analysis for $TD(\lambda)$ given in Bertsekas and Tsitsiklis [BeT96], Section 6.3.4.

The $TD(\lambda)$ algorithm was extended to the average cost problem, and its convergence was proved by Tsitsiklis and Van Roy [TsV99a] (see also [TsV02]). The average cost analysis of LSPE in Section 6.6.1 is due to Yu and Bertsekas [YuB06b], and made substantial use of the framework established by Tsitsiklis and Van Roy [TsV99a]. An alternative to the LSPE and LSTD algorithms of Section 6.6.1 is based on the relation between average cost and SSP problems, and the associated contracting value iteration method discussed in Section 4.4.1. The idea is to convert the average cost problem into a time-varying SSP, which however converges to the correct one as the gain of the policy is estimated correctly by simulation. The SSP algorithms of Section 6.5 can then be used with the estimated gain of the policy η_k [cf. Eq. (6.80)] replacing the true gain η . The convergence analysis of Section 6.5 fully applies to these algorithms.

While the convergence analysis of the policy evaluation methods of Sections 6.3, 6.5, and 6.6 is based on contraction mapping arguments, a different type of convergence analysis is necessary for Q -learning algorithms for average cost problems (as well as for SSP problems where there may exist some improper policies). The reason is that there may not be an

underlying contraction, so the nonexpansive property of the DP mapping must be used instead. As a result, the analysis is more complicated, and a different method of proof has been employed, based on the so-called ODE approach; see Abounadi, Bertsekas, and Borkar [ABB01], [ABB02]. In particular, the Q -learning algorithms of Section 6.6.3 were proposed and analyzed in these references. They are also discussed in the book [BeT96] (Section 7.1.5). Alternative algorithms of the Q -learning type were given without convergence proof by Schwartz [Sch93b], Singh [Sin94], and Mahadevan [Mah96]; see also Gosavi [Gos04].

There is a large literature on policy gradient methods for average cost problems. The formula for the gradient of the average cost has been given in different forms and within a variety of different contexts: see Cao and Chen [CaC97], Cao and Wan [CaW98], Cao [Cao99], [Cao05], Fu and Hu [FuH94], Glynn [Gly87], Jaakkola, Singh, and Jordan [JSJ95], L'Ecuyer [L'Ec91], and Williams [Wil92]. We follow the derivations of Marbach and Tsitsiklis [MaT01]. The inner product expression of $\partial\eta(r)/\partial r_m$ was used to delineate essential features for gradient calculation by Konda and Tsitsiklis [KoT99], [KoT03], and Sutton, McAllester, Singh, and Mansour [SMS99].

Several implementations of policy gradient methods, some of which use cost approximations, have been proposed: see Cao [Cao04], Grudic and Ungar [GrU04], He [He02], He, Fu, and Marcus [HFM05], Kakade [Kak01], Konda [Kon02], Konda and Borkar [KoB99], Konda and Tsitsiklis [KoT99], [KoT03], Marbach and Tsitsiklis [MaT01], [MaT03], Sutton, McAllester, Singh, and Mansour [SMS99], and Williams [Wil92].

Approximate DP methods for partially observed Markov decision problems are not as well-developed as their perfect observation counterparts. Approximations obtained by solving finite-spaces discounted or average cost problems have been proposed by Yu and Bertsekas [YuB04], [YuB06a]. Policy gradient methods of the actor-only type have been given by Baxter and Bartlett [BaB01], and Aberdeen and Baxter [AbB00]. An alternative method, which is of the actor-critic type, has been proposed by Yu [Yu05], [Yu06]. See also Singh, Jaakkola, and Jordan [SJ94].

Many problems have special structure, which can be exploited in approximate DP. For some representative work, see Guestrin et al. [GKP03], and Koller and Parr [KoP00].

E X E R C I S E S

6.1 (Multiple State Visits in Monte Carlo Simulation)

Argue that the Monte-Carlo simulation formula

$$J_\mu(i) = \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M c(i, m)$$

is valid even if a state may be revisited within the same sample trajectory. Hint: Suppose the M cost samples are generated from N trajectories, and that the k th trajectory involves n_k visits to state i and generates n_k corresponding cost samples. Denote $m_k = n_1 + \dots + n_k$. Write

$$\begin{aligned} \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M c(i, m) &= \lim_{N \rightarrow \infty} \frac{\frac{1}{N} \sum_{k=1}^N \sum_{m=m_{k-1}+1}^{m_k} c(i, m)}{\frac{1}{N} (n_1 + \dots + n_N)} \\ &= \frac{E \left\{ \sum_{m=m_{k-1}+1}^{m_k} c(i, m) \right\}}{E \{n_k\}}, \end{aligned}$$

and argue that

$$E \left\{ \sum_{m=m_{k-1}+1}^{m_k} c(i, m) \right\} = E \{n_k\} J_\mu(i),$$

(or see Ross [Ros83b], Cor. 7.2.3 for a closely related result).

6.2 (Viewing Q -Factors as Optimal Costs)

Consider the stochastic shortest path problem under Assumptions 2.1.1 and 2.1.2. Show that the Q -factors $Q(i, u)$ can be viewed as state costs associated with a modified stochastic shortest path problem. Use this fact to show that the Q -factors $Q(i, u)$ are the unique solution of the system of equations

$$Q(i, u) = \sum_j p_{ij}(u) \left(g(i, u, j) + \min_{u' \in U(j)} Q(j, u') \right).$$

Hint: Introduce a new state for each pair (i, u) , with transition probabilities $p_{ij}(u)$ to the states $j = 1, \dots, n, t$.

6.3

This exercise provides a counterexample to the convergence of PVI for discounted problems when the projection is with respect to a norm other than $\|\cdot\|_\xi$. Consider the mapping $TJ = g + \alpha PJ$ and the algorithm $\Phi r_{k+1} = \Pi T(\Phi r_k)$, where P and Φ satisfy Assumptions 6.3.1 and 6.3.2. Here Π denotes projection on the subspace spanned by Φ with respect to the weighted Euclidean norm $\|J\|_v = \sqrt{J'VJ}$, where V is a diagonal matrix with positive components. Use the formula $\Pi = \Phi(\Phi'V\Phi)^{-1}\Phi'V$ to show that in the single basis function case (Φ is an $n \times 1$ vector) the algorithm is written as

$$r_{k+1} = \frac{\Phi'Vg}{\Phi'V\Phi} + \frac{\alpha\Phi'VP\Phi}{\Phi'V\Phi} r_k.$$

Show that the algorithm diverges if and only if either $\|P\Phi\|_v > \alpha$ or $\|P\Phi\|_v = \alpha$ and $\Phi'Vg \neq \Phi'V\Phi$. Construct choices of α , g , P , Φ , and V for which the algorithm diverges.

6.4 (LSPE(0) for Average Cost Problems [YuB06b])

Show the convergence of LSPE(0) for average cost problems with unit stepsize, assuming that P is aperiodic, by showing that the eigenvalues of the matrix ΠF lie strictly within the unit circle.

6.5 (Relation of Discounted and Average Cost Approximations [TsV02])

Consider the finite-state α -discounted and average cost frameworks of Sections 6.3 and 6.6 for a fixed stationary policy with cost per stage g and transition probability matrix P . Assume that the states form a single recurrent class, let J_α be the α -discounted cost vector, let (η^*, h^*) be the gain-bias pair, let ξ be the steady-state probability vector, let Ξ be the diagonal matrix with diagonal elements the components of ξ , and let

$$P^* = \lim_{N \rightarrow \infty} \sum_{k=0}^{N-1} P^k.$$

Show that:

- (a) $\eta^* = (1 - \alpha)\xi' J_\alpha$ and $P^* J_\alpha = (1 - \alpha)^{-1} \eta^* e$.
- (b) $h^* = \lim_{\alpha \rightarrow 1} (I - P^*) J_\alpha$. Hint: Use the Laurent series expansion of J_α (cf. Prop. 4.1.2).
- (c) Consider the subspace

$$E^* = \{(I - P^*)y \mid y \in \mathbb{R}^n\},$$

which is orthogonal to the unit vector e in the scaled geometry where x and y are orthogonal if $x'\Xi y = 0$ (cf. Fig. 6.6.1). Verify that J_α can be decomposed into the sum of two vectors that are orthogonal (in the scaled geometry): $P^* J_\alpha$, which is the projection of J_α onto the line defined by e , and $(I - P^*) J_\alpha$, which is the projection of J_α onto E^* and converges to h^* as $\alpha \rightarrow 1$.

- (d) Use part (c) to show that the limit $r_{\lambda, \alpha}^*$ of PVI(λ) for the α -discounted problem converges to the limit r_λ^* of PVI(λ) for the average cost problem as $\alpha \rightarrow 1$.

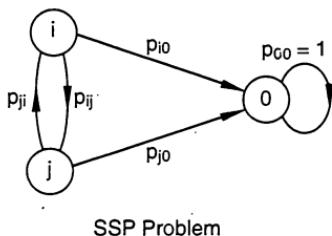
6.6 (Conversion of SSP to Average Cost Policy Evaluation)

We have often used the transformation of an average cost problem to an SSP problem (cf. Section 4.3.1, and Chapter 7 of Vol. I). The purpose of this exercise (unpublished collaboration of J. Yu and the author) is to show that a reverse transformation is possible, from SSP to average cost, at least in the case where all policies are proper. As a result, analysis, insights, and algorithms for average cost policy evaluation can be applied to policy evaluation of a SSP problem.

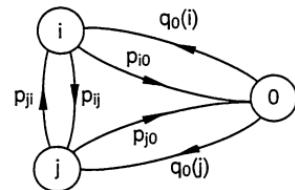
Consider the SSP problem, a single proper stationary policy μ , and the probability distribution $q_0 = (q_0(1), \dots, q_0(n))$ used for restarting simulated trajectories [cf. Eq. (6.67)]. Let us modify the Markov chain by eliminating the self-transition from state 0 to itself, and substituting instead transitions from 0 to i with probabilities $q_0(i)$,

$$\tilde{p}_{0i} = q_0(i),$$

each with a fixed transition cost β , where β is a scalar parameter. All other transitions and costs remain the same (cf. Fig. 6.8.1). We call the corresponding average cost problem β -AC. Denote by J_μ the SSP cost vector of μ , and by η_β and $h_\beta(i)$ the average and differential costs of β -AC, respectively.



SSP Problem



Average Cost Problem

Figure 6.8.1. Transformation of a SSP problem to an average cost problem. The transitions from 0 to each $i = 1, \dots, n$, have cost β .

- (a) Show that η_β can be expressed as the average cost per stage of the cycle that starts at state 0 and returns to 0, i.e.,

$$\eta_\beta = \frac{\beta + \sum_{i=1}^n q_0(i) J_\mu(i)}{T},$$

where T is the expected time to return to 0 starting from 0.

- (b) Show that for the special value

$$\beta^* = - \sum_{i=1}^n q_0(i) J_\mu(i),$$

we have $\eta_{\beta^*} = 0$, and

$$J_\mu(i) = h_{\beta^*}(i) - h_{\beta^*}(0), \quad i = 1, \dots, n.$$

Hint: Since the states of β -AC form a single recurrent class, we have from Bellman's equation

$$\eta_\beta + h_\beta(i) = \sum_{j=0}^n p_{ij} (g(i, j) + h_\beta(j)), \quad i = 1, \dots, n, \quad (6.108)$$

$$\eta_\beta + h_\beta(0) = \beta + \sum_{i=1}^n q_0(i) h_\beta(i). \quad (6.109)$$

From Eq. (6.108) it follows that if $\beta = \beta^*$, we have $\eta_{\beta^*} = 0$, and

$$\delta(i) = \sum_{j=0}^n p_{ij} g(i, j) + \sum_{j=1}^n p_{ij} \delta(j), \quad i = 1, \dots, n, \quad (6.110)$$

where

$$\delta(i) = h_{\beta^*}(i) - h_{\beta^*}(0), \quad i = 1, \dots, n.$$

Since Eq. (6.110) is Bellman's equation for the SSP problem, we see that $\delta(i) = J_\mu(i)$ for all i .

- (c) Derive a transformation to convert an average cost policy evaluation problem into another average cost policy evaluation problem where the transition probabilities out of a single state are modified in any way such that the states of the resulting Markov chain form a single recurrent class. The two average cost problems should have the same differential cost vectors, except for a constant shift. *Note:* This conversion may be useful if the transformed problem has more favorable properties.

6.7 (Policy Gradient Formulas for SSP)

Consider the SSP context, and let the cost per stage and transition probability matrix be given as functions of a parameter vector r . Denote by $g_i(r)$, $i = 1, \dots, n$, the expected cost starting at state i , and by $p_{ij}(r)$ the transition probabilities. Each value of r defines a stationary policy, which is assumed proper. For each r , the expected costs starting at states i are denoted by $J_i(r)$. We wish to calculate the gradient of a weighted sum of the costs $J_i(r)$, i.e.,

$$\bar{J}(r) = \sum_{i=1}^n q(i) J_i(r),$$

where $q = (q(1), \dots, q(n))$ is some probability distribution over the states. Consider a single scalar component r_m of r , and differentiate Bellman's equation to show that

$$\frac{\partial J_i}{\partial r_m} = \frac{\partial g_i}{\partial r_m} + \sum_{j=1}^n \frac{\partial p_{ij}}{\partial r_m} J_j + \sum_{j=1}^n p_{ij} \frac{\partial J_j}{\partial r_m}, \quad i = 1, \dots, n,$$

where the argument r at which the partial derivatives are computed is suppressed. Interpret the above equation as Bellman's equation for a SSP problem.

APPENDIX A:

Measure Theoretic Issues in Dynamic Programming

A general theory of stochastic dynamic programming must deal with the formidable mathematical questions that arise from the presence of uncountable probability spaces. The purpose of this appendix is to orient the mathematically advanced reader on these questions.[†]

The appendix is based on the research monograph by Bertsekas and Shreve [BeS78] (freely available from the internet), to which we refer for a detailed analysis, for references to earlier research, and for the development of mathematical background and terminology on Borel spaces and related subjects. We will explore here the main questions by means of a simple two-stage example described in Section A.1. In Section A.2, we develop a framework, based on universally measurable policies, for the rigorous mathematical development of the standard DP results for this example and for more general finite horizon models.

A.1 A TWO-STAGE EXAMPLE

Suppose that the initial state x_0 is a point on the real line \mathbb{R} . Knowing x_0 , we must choose a control $u_0 \in \mathcal{R}$. Then the new state x_1 is generated

[†] The style and terminology of this appendix assume a mathematically sophisticated reader, who has knowledge of the basic notions of measure theory and is also familiar with finite horizon DP. In particular, we freely use basic notions of measurability and integration. We also use “inf” notation rather than “min” in various optimization equations, when the infimum is not known to be attained.

according to a transition probability measure $p(dx_1 | x_0, u_0)$ on the Borel σ -algebra of \mathfrak{R} (the one generated by the open sets of \mathfrak{R}). Then, knowing x_1 , we must choose a control $u_1 \in \mathfrak{R}$ and incur a cost $g(x_1, u_1)$, where g is a real-valued function that is bounded either above or below. Thus a cost is incurred only at the second stage.

A policy $\pi = \{\mu_0, \mu_1\}$ is a pair of functions from state to control, i.e., if π is employed and x_0 is the initial state, then $u_0 = \mu_0(x_0)$, and if x_1 is the subsequent state, then $u_1 = \mu_1(x_1)$. The expected value of the cost corresponding to π when x_0 is the initial state is given by

$$J_\pi(x_0) = \int g(x_1, \mu_1(x_1)) p(dx_1 | x_0, \mu_0(x_0)). \quad (\text{A.1})$$

We wish to find π to minimize $J_\pi(x_0)$.

To formulate the problem properly, we must insure that the integral in Eq. (A.1) is defined. Various sufficient conditions can be used for this; for example it is sufficient that g , μ_0 , and μ_1 be Borel measurable, and that $p(B | x_0, u_0)$ is a Borel measurable function of (x_0, u_0) for every Borel set B (see [BeS78]). However, our aim in this example is to discuss the necessary measure theoretic framework not only for the cost $J_\pi(x_0)$ to be defined, but also for the major DP-related results to hold. We thus leave unspecified for the moment the assumptions on the problem data and the measurability restrictions on the policy π .

The optimal cost is

$$J^*(x_0) = \inf_{\pi} J_\pi(x_0),$$

where the infimum is over all policies $\pi = \{\mu_0, \mu_1\}$ such that μ_0 and μ_1 are measurable functions from \mathfrak{R} to \mathfrak{R} with respect to σ -algebras to be specified later. Given $\epsilon > 0$, a policy π is ϵ -optimal if

$$J_\pi(x_0) \leq J^*(x_0) + \epsilon, \quad \forall x_0 \in \mathfrak{R}.$$

A policy π is optimal if

$$J_\pi(x_0) = J^*(x_0), \quad \forall x_0 \in \mathfrak{R}.$$

The DP Algorithm

The DP algorithm for the preceding two-stage problem takes the form

$$J_1(x_1) = \inf_{u_1 \in \mathfrak{R}} g(x_1, u_1), \quad \forall x_1 \in \mathfrak{R}, \quad (\text{A.2})$$

$$J_0(x_0) = \inf_{u_0 \in \mathfrak{R}} \int J_1(x_1) p(dx_1 | x_0, u_0), \quad \forall x_0 \in \mathfrak{R}, \quad (\text{A.3})$$

and assuming that

$$J_0(x_0) > -\infty, \quad \forall x_0 \in \mathfrak{R}, \quad J_1(x_1) > -\infty, \quad \forall x_1 \in \mathfrak{R},$$

the results we expect to be able to prove are:

R.1: There holds

$$J^*(x_0) = J_0(x_0), \quad \forall x_0 \in \mathfrak{R}.$$

R.2: Given any $\epsilon > 0$, there is an ϵ -optimal policy.

R.3: If $\mu_1^*(x_1)$ and $\mu_0^*(x_0)$ attain the infimum in the DP algorithm (A.2), (A.3) for all $x_1 \in \mathfrak{R}$ and $x_0 \in \mathfrak{R}$, respectively, then $\pi^* = \{\mu_0^*, \mu_1^*\}$ is optimal.

We will see that to establish these results, we will need to address two main issues:

- (1) The cost function J_π of a policy π , and the functions J_0 and J_1 produced by DP should be well-defined, with a mathematical framework, which ensures that the integrals in Eqs. (A.1)-(A.3) make sense.
- (2) Since $J_0(x_0)$ is easily seen to be a lower bound to $J_\pi(x_0)$ for all x_0 and $\pi = \{\mu_0, \mu_1\}$, the equality of J_0 and J^* will be ensured if the class of policies has an ϵ -selection property, which guarantees that the minima in Eqs. (A.2) and (A.3) can be nearly attained by $\mu_1(x_1)$ and $\mu_0(x_0)$ for all x_1 and x_0 , respectively.

To get a better sense of these issues, consider the following informal derivation of R.1:

$$\begin{aligned} J^*(x_0) &= \inf_{\pi} J_{\pi}(x_0) \\ &= \inf_{\mu_0} \inf_{\mu_1} \int g(x_1, \mu_1(x_1)) p(dx_1 | x_0, \mu_0(x_0)) \end{aligned} \quad (\text{A.4a})$$

$$= \inf_{\mu_0} \int \left\{ \inf_{\mu_1} g(x_1, \mu_1(x_1)) \right\} p(dx_1 | x_0, \mu_0(x_0)) \quad (\text{A.4b})$$

$$= \inf_{\mu_0} \int \left\{ \inf_{u_1} g(x_1, u_1) \right\} p(dx_1 | x_0, \mu_0(x_0)) \quad (\text{A.4c})$$

$$= \inf_{\mu_0} \int J_1(x_1) p(dx_1 | x_0, \mu_0(x_0)) \quad (\text{A.4d})$$

$$= \inf_{\mu_0} \int J_1(x_1) p(dx_1 | x_0, u_0) \\ = J_0(x_0).$$

In order to make this derivation meaningful and mathematically rigorous, the following points need to be justified:

- (a) g and μ_1 must be such that $g(x_1, \mu_1(x_1))$ can be integrated in a well-defined manner in Eq. (A.4a).

- (b) The interchange of infimization and integration in Eq. (A.4b) must be legitimate.
- (c) g must be such that the function

$$J_1(x_1) = \inf_{u_1} g(x_1, u_1)$$

can be integrated in a well-defined manner in Eq. (A.4c).

We first discuss these points in the easier context where the state space is essentially countable.

Countable Space Problems

We observe that if for each (x_0, u_0) , the measure $p(dx_1 | x_0, u_0)$ has *countable support*, i.e., is concentrated on a countable number of points, then for a fixed policy π and initial state x_0 , the integral defining the cost $J_\pi(x_0)$ of Eq. (A.1) is defined in terms of (possibly infinite) summation. Similarly, the DP algorithm (A.2), (A.3) is defined in terms of summation, and the same is true for the integrals in Eqs. (A.4a)-(A.4d). Thus, there is no need to impose measurability restrictions of any kind for the integrals to make sense, and for the summations/integrations to be well-defined, it is sufficient that g is bounded either above or below.

It can also be shown that the interchange of infimization and summation in Eq. (A.4b) is justified in view of the assumption

$$\inf_{u_1} g(x_1, u_1) > -\infty, \quad \forall x_1 \in \mathfrak{R}.$$

To see this, for any $\epsilon > 0$, select $\bar{\mu}_1 : \mathfrak{R} \mapsto \mathfrak{R}$ such that

$$g(x_1, \bar{\mu}_1(x_1)) \leq \inf_{u_1} g(x_1, u_1) + \epsilon, \quad \forall x_1 \in \mathfrak{R}. \quad (\text{A.5})$$

Then

$$\begin{aligned} \inf_{\mu_1} \int g(x_1, \mu_1(x_1)) p(dx_1 | x_0, \mu_0(x_0)) \\ \leq \int g(x_1, \bar{\mu}_1(x_1)) p(dx_1 | x_0, \mu_0(x_0)) \\ \leq \int \underbrace{\inf_{u_1} g(x_1, u_1)}_{-} p(dx_1 | x_0, \mu_0(x_0)) + \epsilon. \end{aligned}$$

Since $\epsilon > 0$ is arbitrary, it follows that

$$\inf_{\mu_1} \int g(x_1, \mu_1(x_1)) p(dx_1 | x_0, \mu_0(x_0)) \leq \int \inf_{u_1} g(x_1, u_1) p(dx_1 | x_0, \mu_0(x_0)).$$

The reverse inequality also holds, since for all μ_1 , we can write

$$\int \inf_{u_1} g(x_1, u_1) p(dx_1 | x_0, \mu_0(x_0)) \leq \int g(x_1, \mu_1(x_1)) p(dx_1 | x_0, \mu_0(x_0)),$$

and then we can take the infimum over μ_1 . It follows that the interchange of infimization and summation in Eq. (A.4b) is justified, with the ϵ -optimal selection property of Eq. (A.5) being the key step in the proof.

We have thus shown that when the measure $p(dx_1 | x_0, u_0)$ has countable support, g is bounded either above or below, and $J_0(x_0) > -\infty$ for all x_0 and $J_1(x_1) > -\infty$ for all x_1 , the derivation of Eq. (A.4) is valid and proves that the DP algorithm produces the optimal cost function J^* (cf. property R.1).† A similar argument proves the existence of an ϵ -optimal policy (cf. R.2); it uses the ϵ -optimal selection (A.5) for the second stage and a similar ϵ -optimal selection for the first stage, i.e., the existence of a $\bar{\mu}_0 : \mathcal{R} \mapsto \mathcal{R}$ such that

$$\int J_1(x_1) p(dx_1 | x_0, \bar{\mu}_0(x_0)) \leq \inf_{u_0} \int J_1(x_1) p(dx_1 | x_0, u_0) + \epsilon. \quad (\text{A.6})$$

Also R.3 follows easily using the fact that there are no measurability restrictions on μ_0 and μ_1 .

Approaches for Uncountable Space Problems

To address the case where $p(dx_1 | x_0, u_0)$ does not have countable support, two approaches have been used. The first is to *expand the notion of integration*, and the second is to place *appropriate measurability restrictions on g , p , and $\{\mu_0, \mu_1\}$* . Expanding the notion of integration is possible by interpreting the integrals appearing in the preceding equations as outer integrals. Since the outer integral can be defined for any function, measurable or not, there is no need to impose any measurability assumptions, and the arguments given above go through just as in the countable disturbance case. We do not discuss this approach further except to mention that the Bertsekas and Shreve book [BeS78] shows that the basic results for finite and infinite horizon problems of perfect state information carry through within an outer integration framework. However, there are inherent limitations in this approach centering around the pathologies of outer integration, as discussed in [BeS78].

The second approach is to impose a suitable measurability structure that allows the key proof steps of the validity of the DP algorithm. These are:

† The condition that g is bounded either above or below may be replaced by any condition that guarantees that the infinite sum/integral of J_1 in Eq. (A.3) is well-defined. Note also that if g is bounded below, then the assumption that $J_0(x_0) > -\infty$ for all x_0 and $J_1(x_1) > -\infty$ for all x_1 is automatically satisfied.

- (a) Properly interpreting the integrals in the definition (A.2)-(A.3) of the DP algorithm and the derivation (A.4).
- (b) The ϵ -optimal selection property (A.5), which in turn justifies the interchange of infimization and integration in Eq. (A.4b).

To enable (a), the required properties of the problem structure must include the preservation of measurability under partial minimization. In particular, it is necessary that when g is measurable in some sense, the partial minimum function

$$J_1(x_1) = \inf_{u_1} g(x_1, u_1)$$

is also measurable in the same sense, so that the integration in Eq. (A.3) is well-defined. It turns out that this is a major difficulty with Borel measurability, which may appear to be a natural framework for formulating the problem: *J₁ need not be Borel measurable even when g is Borel measurable.* For this reason it is necessary to pass to a larger class of measurable functions, which is closed under the key operation of partial minimization (and also under some other common operations, such as addition and functional composition).†

One such class is *lower semianalytic functions* and the related class of *universally measurable functions*, which will be the focus of the next section. They are the basis for a problem formulation that enables a DP theory as powerful as the one for problems where measurability is of no concern (e.g., those where the state and control spaces are countable).

A.2

RESOLUTION OF THE MEASURABILITY ISSUES

The example of the preceding section indicates that if measurability restrictions are necessary for the problem data and policies, then measurable selection and preservation of measurability under partial minimization, become crucial parts of the analysis. We will discuss measurability frameworks that are favorable in this regard, and to this end, we will use the theory of Borel spaces.

† It is also possible to use a smaller class of functions that is closed under the same operations. This has led to the so-called *semicontinuous models*, where the state and control spaces are Borel spaces, and g and p have certain semicontinuity and other properties. These models are also analyzed in detail in the Bertsekas and Shreve book [BeS78] (Section 8.3). However, they are not as useful and widely applicable as the universally measurable models we will focus on, because they involve assumptions that may be restrictive and/or hard to verify. By contrast, the universally measurable models are simple and very general. They allow a problem formulation that brings to bear the power of DP analysis under minimal assumptions. This analysis can in turn be used to prove more specific results based on special characteristics of the model.

Borel Spaces and Analytic Sets

Given a topological space Y , we denote by \mathcal{B}_Y the σ -algebra generated by the open subsets of Y , and refer to the members of \mathcal{B}_Y as the *Borel subsets* of Y . A topological space Y is a *Borel space* if it is homeomorphic to a Borel subset of a complete separable metric space. The concept of Borel space is quite broad, containing any “reasonable” subset of n -dimensional Euclidean space. Any Borel subset of a Borel space is again a Borel space, as is any homeomorphic image of a Borel space and any finite or countable Cartesian product of Borel spaces. Let Y and Z be Borel spaces, and consider a function $h : Y \mapsto Z$. We say that h is *Borel measurable* if $h^{-1}(B) \in \mathcal{B}_Y$ for every $B \in \mathcal{B}_Z$.

Borel spaces have a deficiency in the context of optimization: even in the unit square, there exist Borel sets whose projections onto an axis are not Borel subsets of that axis. In fact, this is the source of the difficulty we mentioned earlier regarding Borel measurability in the DP context: if $g(x_1, u_1)$ is Borel measurable, the partial minimum function

$$J_1(x_1) = \inf_{u_1} g(x_1, u_1)$$

need not be, because its level sets are defined in terms of projections of the level sets of g as

$$\{x_1 \mid J_1(x_1) < c\} = P\left(\{(x_1, u_1) \mid g(x_1, u_1) < c\}\right),$$

where c is a scalar and $P(\cdot)$ denotes projection on the space of x_1 . As an example, take g to be the indicator of a Borel subset of the unit square whose projection on the x_1 -axis is not Borel. Then J_1 is the indicator function of this projection, so it is not Borel measurable. This leads us to the notion of an analytic set.

A subset A of a Borel space Y is said to be *analytic* if there exists a Borel space Z and a Borel subset B of $Y \times Z$ such that $A = \text{proj}_Y(B)$, where proj_Y is the projection mapping from $Y \times Z$ to Y . It is clear that every Borel subset of a Borel space is analytic.

Analytic sets have many interesting properties, which are discussed in detail in [BeS78]. Some of these properties are particularly relevant to DP analysis. For example, let Y and Z be Borel spaces. Then:

- (i) If $A \subset Y$ is analytic and $h : Y \mapsto Z$ is Borel measurable, then $h(A)$ is analytic. In particular, if Y is a product of Borel spaces Y_1 and Y_2 , and $A \subset Y_1 \times Y_2$ is analytic, then $\text{proj}_{Y_1}(A)$ is analytic. Thus, the class of analytic sets is closed with respect to projection, a critical property for DP, which the class of Borel sets is lacking, as mentioned earlier.
- (ii) If $A \subset Z$ is analytic and $h : Y \mapsto Z$ is Borel measurable, then $h^{-1}(A)$ is analytic.

- (iii) If A_1, A_2, \dots are analytic subsets of Y , then $\cup_{k=1}^{\infty} A_k$ and $\cap_{k=1}^{\infty} A_k$ are analytic.

However, the complement of an analytic set need not be analytic, so the collection of analytic subsets of Y need not be a σ -algebra.

Lower Semianalytic Functions

Let Y be a Borel space and let $h : Y \mapsto [-\infty, \infty]$ be a function. We say that h is *lower semianalytic* if the level set

$$\{y \in Y \mid h(y) < c\}$$

is analytic for every $c \in \mathbb{R}$. The following proposition states that lower analyticity is preserved under partial minimization, a key result for our purposes. The proof follows from the preservation of analyticity of a subset of a product space under projection onto one of the component spaces, as in (i) above (see [BeS78], Prop. 7.47).

Proposition A.1: Let Y and Z be Borel spaces, and let $h : Y \times Z \mapsto [-\infty, \infty]$ be lower semianalytic. Then $h^* : Y \mapsto [-\infty, \infty]$ defined by

$$h^*(y) = \inf_{z \in Z} h(y, z)$$

is lower semianalytic.

By comparing the DP equation $J_1(x_1) = \inf_{u_1} g(x_1, u_1)$ [cf. Eq. (A.2)] and Prop. A.1, we see how lower semianalytic functions can arise in DP. In particular, J_1 is lower semianalytic if g is. Let us also give two additional properties of lower semianalytic functions that play an important role in DP (for a proof, see [BeS78], Lemma 7.40).

Proposition A.2: Let Y be a Borel space, and let $h : Y \mapsto [-\infty, \infty]$ and $l : Y \mapsto [-\infty, \infty]$ be lower semianalytic. Suppose that for every $y \in Y$, the sum $h(y) + l(y)$ is defined, i.e., is not of the form $\infty - \infty$. Then $h + l$ is lower semianalytic.

Proposition A.3: Let Y and Z be Borel spaces, let $h : Y \mapsto Z$ be Borel measurable, and let $l : Z \mapsto [-\infty, \infty]$ be lower semianalytic. Then the composition $l \circ h$ is lower semianalytic.

Universal Measurability

To address questions relating to the definition of the integrals appearing in the DP algorithm, we must discuss the measurability properties of lower semianalytic functions. In addition to the Borel σ -algebra \mathcal{B}_Y mentioned earlier, there is the *universal σ -algebra* \mathcal{U}_Y , which is the intersection of all completions of \mathcal{B}_Y with respect to all probability measures. Thus, $E \in \mathcal{U}_Y$ if and only if, given any probability measure p on (Y, \mathcal{B}_Y) , there is a Borel set B and a p -null set N such that $E = B \cup N$. Clearly, we have $\mathcal{B}_Y \subset \mathcal{U}_Y$. It is also true that every analytic set is universally measurable (for a proof, see [BeS78], Corollary 7.42.1), and hence the σ -algebra generated by the analytic sets, called the *analytic σ -algebra*, and denoted \mathcal{A}_Y , is contained in \mathcal{U}_Y :

$$\mathcal{B}_Y \subset \mathcal{A}_Y \subset \mathcal{U}_Y.$$

Let X , Y , and Z be Borel spaces, and consider a function $h : Y \mapsto Z$. We say that h is *universally measurable* if $h^{-1}(B) \in \mathcal{U}_Y$ for every $B \in \mathcal{B}_Z$. It can be shown that if $U \subset Z$ is universally measurable and h is universally measurable, then $h^{-1}(U)$ is also universally measurable. As a result, if $g : X \mapsto Y$, $h : Y \mapsto Z$ are universally measurable functions, then the composition $(g \circ h) : X \mapsto Z$ is universally measurable.

We say that $h : Y \mapsto Z$ is *analytically measurable* if $h^{-1}(B) \in \mathcal{A}_Y$ for every $B \in \mathcal{B}_Z$. It can be seen that *every lower semianalytic function is analytically measurable*, and in view of the inclusion $\mathcal{A}_Y \subset \mathcal{U}_Y$, it is also *universally measurable*.

Integration of Lower Semianalytic Functions

If p is a probability measure on (Y, \mathcal{B}_Y) , then p has a unique extension to a probability measure \bar{p} on (Y, \mathcal{U}_Y) . We write simply p instead of \bar{p} and $\int h d\bar{p}$ in place of $\int h d\bar{p}$. In particular, if h is lower semianalytic, then $\int h d\bar{p}$ is interpreted in this manner.

Let Y and Z be Borel spaces. A *stochastic kernel* $q(dz | y)$ on Z given Y is a collection of probability measures on (Z, \mathcal{B}_Z) parameterized by the elements of Y . If for each Borel set $B \in \mathcal{B}_Z$, the function $q(B | y)$ is Borel measurable (universally measurable) in y , the stochastic kernel $q(dz | y)$ is said to be *Borel measurable (universally measurable, respectively)*. The following proposition provides another basic property for the DP context (for a proof, see [BeS78], Props. 7.46 and 7.48).

Proposition A.4: Let Y and Z be Borel spaces, and let $q(dz | y)$ be a stochastic kernel on Z given Y . Let also $h : Y \times Z \mapsto [-\infty, \infty]$ be a function that is bounded either above or below.

- (a) If q is Borel measurable and h is lower semianalytic, then the function $l : Y \mapsto [-\infty, \infty]$ given by

$$l(y) = \int_Z h(y, z) q(dz \mid y)$$

is lower semianalytic.

- (b) If q is universally measurable and h is universally measurable, then the function $l : Y \mapsto [-\infty, \infty]$ given by

$$l(y) = \int_Z h(y, z) q(dz \mid y)$$

is universally measurable.

Note that the boundedness above or below assumption on h in the preceding proposition aims to ensure that $l(y)$ is well-defined for every y as an integral.[†]

Returning to the DP algorithm (A.2)-(A.3) of Section A.1, note that if the cost function g is lower semianalytic and bounded either above or below, then the partial minimum function J_1 given by the DP Eq. (A.2) is lower semianalytic (cf. Prop. A.1), and bounded either above or below, respectively. Furthermore, if the transition kernel $p(dx_1 \mid x_0, u_0)$ is Borel measurable, then the integral

$$\int J_1(x_1) p(dx_1 \mid x_0, u_0) \tag{A.7}$$

is a lower semianalytic function of (x_0, u_0) (cf. Prop. A.4), and in view of Prop. A.1, the same is true of the function J_0 given by the DP Eq. (A.3), which is the partial minimum over u_0 of the expression (A.7). Thus, with

[†] We use here the classical definition of integral, whereby for a probability measure p , the integral of an extended real-valued function f , with positive and negative parts f^+ and f^- , is defined as

$$\int f dp = \int f^+ dp - \int f^- dp,$$

provided $\int f^+ dp < \infty$ or $\int f^- dp < \infty$. The book [BeS78] (Section 7.4.4) uses a more general definition, which adopts the rule $\infty - \infty = \infty$ for the case where $\int f^+ dp = \infty$ and $\int f^- dp = \infty$. With this expanded definition of integral, there is no need for the boundedness assumption in Prop. A.4 (cf. [BeS78], Props. 7.46 and 7.48).

lower semianalytic g and Borel measurable p , the integrals appearing in the DP algorithm make sense.

Note that in the example of Section A.1, there is no cost incurred in the first stage of the system operation. When such a cost, call it $g_0(x_0, u_0)$, is introduced, the expression minimized in the DP Eq. (A.3) becomes

$$g_0(x_0, u_0) + \int J_1(x_1) p(dx_1 | x_0, u_0),$$

which is still a lower semianalytic function of (x_0, u_0) , provided g_0 is lower semianalytic and the sum above is not of the form $\infty - \infty$ for any (x_0, u_0) (Prop. A.2). Also, for alternative models defined in terms of a system function rather than a stochastic kernel (see the total cost model of Chapter 1), Prop. A.3 provides some of the necessary machinery to show that the functions generated by the DP algorithm are lower semianalytic.

Universally Measurable Selection

The preceding discussion has shown that if g is lower semianalytic and bounded either above or below, and p is Borel measurable, the DP algorithm (A.2)-(A.3) is well-defined and produces lower semianalytic functions J_1 and J_0 . However, this does not by itself imply that J_0 is equal to the optimal cost function J^* . For this it is necessary that the chosen class of policies has the ϵ -optimal selection property (A.5). It turns out that universally measurable policies have this property.

The following is the key selection theorem given in a general form, which also addresses the question of existence of optimal policies that can be obtained from the DP algorithm (for a proof, see [BeS78], Prop. 7.50). The theorem shows that if any functions $\bar{\mu}_1 : \mathcal{R} \rightarrow \mathcal{R}$ and $\bar{\mu}_0 : \mathcal{R} \rightarrow \mathcal{R}$ can be found such that $\bar{\mu}_1(x_1)$ and $\bar{\mu}_0(x_0)$ attain the respective minima in Eqs. (A.2) and (A.3), for every x_1 and x_0 , then $\bar{\mu}_1$ and $\bar{\mu}_0$ can be chosen to be universally measurable, the DP algorithm yields the optimal cost function and $\pi = (\bar{\mu}_0, \bar{\mu}_1)$ is optimal, provided that g is lower semianalytic and the integral in Eq. (A.3) is a lower semianalytic function of (x_0, u_0) .

Proposition A.5: (Measurable Selection Theorem) Let Y and Z be Borel spaces and let $h : Y \times Z \mapsto [-\infty, \infty]$ be lower semianalytic. Define $h^* : Y \mapsto [-\infty, \infty]$ by

$$h^*(y) = \inf_{z \in Z} h(y, z),$$

and let

$$I = \{y \in Y \mid \text{there exists a } z_y \in Z \text{ for which } h(y, z_y) = h^*(y)\},$$

i.e., I is the set of points y for which the infimum above is attained. For any $\epsilon > 0$, there exists a universally measurable function $\phi : Y \mapsto Z$ such that

$$h(y, \phi(y)) = h^*(y), \quad \forall y \in I,$$

$$h(y, \phi(y)) \leq \begin{cases} h^*(y) + \epsilon, & \forall y \notin I \text{ with } h^*(y) > -\infty, \\ -1/\epsilon, & \forall y \notin I \text{ with } h^*(y) = -\infty. \end{cases}$$

Universal Measurability Framework: A Summary

In conclusion, the preceding discussion shows that in the two-stage example of Section A.1, the measurability issues are resolved in the following sense: the DP algorithm (A.2)-(A.3) is well-defined, produces lower semianalytic functions J_1 and J_0 , and yields the optimal cost function (as in R.1), and furthermore there exist ϵ -optimal and possibly exactly optimal policies (as in R.2 and R.3), provided that:

- (a) *The stage cost function g is lower semianalytic and is bounded either above or below.* Lower analyticity is needed to show that the function J_1 of the DP Eq. (A.2) is lower semianalytic and hence also universally measurable (cf. Prop. A.1). Boundedness either above or below is needed to ensure the respective boundedness property for J_1 , which will be needed to guarantee that the integral of J_1 in Eq. (A.3) is defined (according to the classical definition). The more “natural” Borel measurability assumption on g implies lower analyticity of g , but will not keep the functions J_1 and J_0 produced by the DP algorithm within the domain of Borel measurability. This is because the partial minimum operation on Borel measurable functions takes us outside that domain (cf. Prop. A.1).
- (b) *The stochastic kernel p is Borel measurable.* This is needed in order for the integral in the DP Eq. (A.3) to be defined as a lower semianalytic function of (x_0, u_0) (cf. Prop. A.4). In turn, this is used to show that the function J_0 of the DP Eq. (A.3) is lower semianalytic (cf. Prop. A.1).
- (c) *The control functions μ_0 and μ_1 are allowed to be universally measurable, and we have $J_0(x_0) > -\infty$ for all x_0 and $J_1(x_1) > -\infty$ for all x_1 .* This is needed in order for the calculation of Eq. (A.4) to go through (using the measurable selection property of Prop. A.5), and

show that the DP algorithm produces the optimal cost function (cf. R.1). It is also needed (using again Prop. A.5) in order to show the associated existence of solutions results (cf. R.2 and R.3).

Extension to General Finite-Horizon DP

Let us now extend our analysis to an N -stage model with state x_k and control u_k that take values in Borel spaces X and U , respectively. We assume stochastic/transition kernels $p_k(dx_{k+1} | x_k, u_k)$, which are Borel measurable, and stage cost functions $g_k : X \times U \mapsto (-\infty, \infty]$, which are lower semianalytic and bounded either above or below.[†] Furthermore, we allow policies $\pi = \{\mu_0, \dots, \mu_{N-1}\}$ that are randomized: each component μ_k is a universally measurable stochastic kernel $\mu_k(du_k | x_k)$ from X to U . If for every x_k and k , $\mu_k(du_k | x_k)$ assigns probability 1 to a single control u_k , π is said to be *nonrandomized*.

Each policy π and initial state x_0 define a unique probability measure with respect to which $g_k(x_k, u_k)$ can be integrated to produce the expected value of g_k . The sum of these expected values for $k = 0, \dots, N-1$, is the cost $J_\pi(x_0)$. It is convenient to write this cost in terms of the following DP-like backwards recursion (see [BeS78], Section 8.1):

$$\begin{aligned} J_{\pi, N-1}(x_{N-1}) &= \int g_{N-1}(x_{N-1}, u_{N-1}) \mu_{N-1}(du_{N-1} | x_{N-1}), \\ J_{\pi, k}(x_k) &= \int \left(g_k(x_k, u_k) + \int J_{\pi, k+1}(x_{k+1}) p_k(dx_{k+1} | x_k, u_k) \right) \\ &\quad \mu_k(du_k | x_k), \quad k = 0, \dots, N-2. \end{aligned}$$

The function obtained at the last step is the cost of π starting at x_0 :

$$J_\pi(x_0) = J_{\pi, 0}(x_0).$$

We can interpret $J_{\pi, k}(x_k)$ as the expected cost-to-go starting from x_k at time k , and using π . Note that by Prop. A.4, the functions $J_{\pi, k}$ are all universally measurable.

The DP algorithm is given by

$$J_{N-1}(x_{N-1}) = \inf_{u_{N-1} \in U} g_{N-1}(x_{N-1}, u_{N-1}), \quad \forall x_{N-1},$$

$$J_k(x_k) = \inf_{u_k \in U} \left[g_k(x_k, u_k) + \int J_{k+1}(x_{k+1}) p_k(dx_{k+1} | x_k, u_k) \right], \quad \forall x_k, k.$$

[†] Note that since g_k may take the value ∞ , constraints of the form $u_k \in U_k(x_k)$ may be implicitly introduced by letting $g_k(x_k, u_k) = \infty$ when $u_k \notin U_k(x_k)$.

By essentially replicating the analysis of the two-stage example, we can show that the integrals in the above DP algorithm are well-defined, and that the functions J_{N-1}, \dots, J_0 are lower semianalytic.

It can be seen from the preceding expressions that we have for all policies π

$$J_k(x_k) \leq J_{\pi,k}(x_k), \quad \forall x_k, k = 0, \dots, N-1.$$

To show equality within $\epsilon \geq 0$ in the above relation, we may use the measurable selection theorem (Prop. A.5), assuming that

$$J_k(x_k) > -\infty, \quad \forall x_k, k,$$

so that ϵ -optimal universally measurable selection is possible in the DP algorithm. In particular, define $\bar{\pi} = \{\bar{\mu}_0, \dots, \bar{\mu}_{N-1}\}$ such that $\bar{\mu}_k : X \mapsto U$ is universally measurable, and for all x_k and k ,

$$g_k(x_k, \bar{\mu}_k(u_k)) + \int J_{k+1}(x_{k+1}) p_k(dx_{k+1} | x_k, \bar{\mu}_k(u_k)) \leq J_k(x_k) + \frac{\epsilon}{N}. \quad (\text{A.8})$$

Then, we can show by induction that

$$J_k(x_k) \leq J_{\bar{\pi},k}(x_k) \leq J_k(x_k) + \frac{(N-k)\epsilon}{N}, \quad \forall x_k, k = 0, \dots, N-1,$$

and in particular, for $k = 0$,

$$J_0(x_0) \leq J_{\bar{\pi}}(x_0) \leq J_0(x_0) + \epsilon, \quad \forall x_0.$$

and hence also

$$J^*(x_0) = \inf_{\pi} J_{\pi}(x_0) = J_0(x_0).$$

Thus, the DP algorithm produces the optimal cost function, and via the approximate minimization of Eq. (A.8), an ϵ -optimal policy. Similarly, if the infimum is attained for all x_k and k in the DP algorithm, then there exists an optimal policy. Note that both the ϵ -optimal and the exact optimal policies can be nonrandomized.

An interesting characteristic of the preceding line of development is that it decouples the issue of the definition of the DP algorithm from the question of whether it yields the optimal cost function and ϵ -optimal or nearly optimal policies. In the former question, the key fact is the preservation of lower semianalyticity under partial minimization and integration, while in the latter question, the key fact is whether ϵ -optimal selection is possible in the DP algorithm within the class of policies stipulated. To illustrate this point, suppose that we are interested in optimizing the cost $J_{\pi}(x_0)$ over a *restricted subset* Π of the randomized universally measurable policies. For example in problems with special structure, Π may be a class

of continuous functions, or linear functions, or functions with some special structural characteristics [e.g., (s, S) or other threshold policies in inventory control]. Then, Borel measurability of the stochastic kernels and lower semianalyticity of the costs per stage will guarantee that the functions J_k produced by the DP algorithm are well-defined and can be analyzed. If the analysis shows that the class of policies Π has the ϵ -selection property (A.8), then it follows that $J_0(x_0)$ is equal to the optimal cost over the restricted class Π , and that ϵ -optimal policies exist within this class.

The assumptions of Borel measurability of the stochastic kernels, lower semianalyticity of the costs per stage, and universally measurable policies, are the basis for the framework adopted by Bertsekas and Shreve [BeS78], which provides a comprehensive analysis of finite and infinite horizon total cost problems. The results obtained there using this framework closely parallel the results of Chapters 1 and 3 of the present volume, but apply to the more general case of uncountable disturbance spaces. There is also additional analysis in [BeS78] on problems of imperfect state information, as well as various refinements of the measurability framework just described. Among others, these refinements involve analytically measurable policies, and limit measurable policies (measurable with respect to the, so-called, limit σ -algebra, the smallest σ -algebra that has the properties necessary for a DP theory that is comparably powerful to the one for the universal σ -algebra).

References

- [ABB01] Abounadi, J., Bertsekas, B. P., and Borkar, V. S., 2001. “Learning Algorithms for Markov Decision Processes with Average Cost,” SIAM J. on Control and Optimization, Vol. 40, pp. 681-698.
- [ABB02] Abounadi, J., Bertsekas, B. P., and Borkar, V. S., 2002. “Stochastic Approximation for Non-Expansive Maps: Q-Learning Algorithms,” SIAM J. on Control and Optimization, Vol. 41, pp. 1-22.
- [AFB93] Arapostathis, A., Borkar, V., Fernandez-Gaucherand, E., Ghosh, M., and Marcus, S., 1993. “Discrete-Time Controlled Markov Processes with Average Cost Criterion: A Survey,” SIAM J. on Control and Optimization, Vol. 31, pp. 282-344.
- [AMT93] Archibald, T. W., McKinnon, K. I. M., and Thomas, L. C., 1993. “Serial and Parallel Value Iteration Algorithms for Discounted Markov Decision Processes,” Eur. J. Operations Research, Vol. 67, pp. 188-203.
- [AbB02] Aberdeen, D., and Baxter, J., 2002. “Scalable Internal-State Policy-Gradient Methods for POMDPs,” Proc. of the Nineteenth International Conference on Machine Learning, pp. 3 - 10.
- [Ama98] Amari, S., 1998. “Natural Gradient Works Efficiently in Learning,” Neural Computation, Vol. 10, pp. 251-276.
- [Ash70] Ash, R. B., 1970. Basic Probability Theory, Wiley, N. Y.
- [AyR91] Ayoun, S., and Rosberg, Z., 1991. “Optimal Routing to Two Parallel Heterogeneous Servers with Resequencing,” IEEE Trans. on Automatic Control, Vol. 36, pp. 1436-1449.
- [BBN04] Bertsekas, D. P., Borkar, V., and Nedić, A., 2004. “Improved Temporal Difference Methods with Linear Function Approximation,” in Learning and Approximate Dynamic Programming, by J. Si, A. Barto, W. Powell, (Eds.), IEEE Press, N. Y.
- [BBS95] Barto, A. G., Bradtke, S. J., and Singh, S. P., 1995. “Real-Time Learning and Control Using Asynchronous Dynamic Programming,” Artificial Intelligence, Vol. 72, pp. 81-138.
- [BDM83] Baras, J. S., Dorsey, A. J., and Makowski, A. M., 1983. “Two Competing Queues with Linear Costs: The μ -Rule is Often Optimal,” Report SRR

83-1, Department of Electrical Engineering, University of Maryland.

- [BGM95] Bertsekas, D. P., Guerriero, F., and Musmanno, R., 1995. "Parallel Shortest Path Methods for Globally Optimal Trajectories," High Performance Computing: Technology, Methods, and Applications, (J. Dongarra et al., Eds.), Elsevier.
- [BNO03] Bertsekas, D. P., with Nedić, A., and Ozdaglar, A. E., 2003. Convex Analysis and Optimization, Athena Scientific, Belmont, MA.
- [BPT94a] Bertsimas, D., Paschalidis, I. C., and Tsitsiklis, J. N., 1994. "Optimization of Multiclass Queueing Networks: Polyhedral and Nonlinear Characterizations of Achievable Performance," Annals of Applied Probability, Vol. 4, pp. 43-75.
- [BPT94b] Bertsimas, D., Paschalidis, I. C., and Tsitsiklis, J. N., 1994. "Branching Bandits and Klimov's Problem: Achievable Region and Side Constraints," Proc. of the 1994 IEEE Conference on Decision and Control, pp. 174-180; also in IEEE Trans. on Automatic Control, Vol. 40, 1995, pp. 2063-2075.
- [BWN92] Blanc, J. P. C., de Waal, P. R., Nain, P., and Towsley, D., 1992. "Optimal Control of Admission to a Multiserver Queue with Two Arrival Streams," IEEE Trans. on Automatic Control, Vol. 37, pp. 785-797.
- [BaB01] Baxter, J., and Bartlett, P. L., 2001. "Infinite-Horizon Policy-Gradient Estimation," J. Artificial Intelligence Research, Vol. 15, pp. 319-350.
- [BarD81] Baras, J. S., and Dorsey, A. J., 1981. "Stochastic Control of Two Partially Observed Competing Queues," IEEE Trans. Automatic Control, Vol. AC-26, pp. 1106-1117.
- [Bai93] Baird, L. C., 1993. "Advantage Updating," Report WL-TR-93-1146, Wright Patterson AFB, OH.
- [Bai94] Baird, L. C., 1994. "Reinforcement Learning in Continuous Time: Advantage Updating," International Conf. on Neural Networks, Orlando, Fla.
- [Bai95] Baird, L. C., 1995. "Residual Algorithms: Reinforcement Learning with Function Approximation," Dept. of Computer Science Report, U.S. Air Force Academy, CO.
- [Bat73] Bather, J., 1973. "Optimal Decision Procedures for Finite Markov Chains," Advances in Appl. Probability, Vol. 5, pp. 328-339, pp. 521-540, 541-553.
- [BeC89] Bertsekas, D. P., and Castanon, D. A., 1989. "Adaptive Aggregation Methods for Infinite Horizon Dynamic Programming," IEEE Trans. on Automatic Control, Vol. AC-34, pp. 589-598.
- [BeC99] Bertsekas, D. P., and Castanon, D. A., 1999. "Rollout Algorithms for Stochastic Scheduling Problems," Heuristics, Vol. 5, pp. 89-108.
- [Bei96] Bertsekas, D. P., and Ioffe, S., 1996. "Temporal Differences-Based Policy Iteration and Applications in Neuro-Dynamic Programming," Lab. for Info. and Decision Systems Report LIDS-P-2349, Massachusetts Institute of Technology.
- [BeN93] Bertsimas, D., and Nino-Mora, J., 1993. "Conservation Laws, Extended Polymatroids, and the Multiarmed Bandit Problem: A Unified Polyhedral Ap-

- proach," *Mathematics of Operations Research*, Vol. 21, 1996, pp. 257-306.
- [BeR87] Beutler, F. J., and Ross, K. W., 1987. "Uniformization for Semi-Markov Decision Processes Under Stationary Policies," *J. Appl. Prob.*, Vol. 24, pp. 399-420.
- [BeS78] Bertsekas, D. P., and Shreve, S. E., 1978. *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, N. Y.; may be downloaded from <http://web.mit.edu/dimitrib/www/home.html>
- [BeS79] Bertsekas, D. P., and Shreve, S. E., 1979. "Existence of Optimal Stationary Policies in Deterministic Optimal Control," *J. Math. Anal. and Appl.*, Vol. 69, pp. 607-620.
- [BeT89] Bertsekas, D. P., and Tsitsiklis, J. N., 1989. *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, N. J.; may be downloaded from <http://web.mit.edu/dimitrib/www/home.html>
- [BeT91a] Bertsekas, D. P., and Tsitsiklis, J. N., 1991. "A Survey of Some Aspects of Parallel and Distributed Iterative Algorithms," *Automatica*, Vol. 27, pp. 3-21.
- [BeT91b] Bertsekas, D. P., and Tsitsiklis, J. N., 1991. "An Analysis of Stochastic Shortest Path Problems," *Math. Operations Research*, Vol. 16, pp. 580-595.
- [BeT96] Bertsekas, D. P., and Tsitsiklis, J. N., 1996. *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA.
- [BeT97] Bertsimas, D., and Tsitsiklis, J. N., 1997. *Introduction to Linear Optimization*, Athena Scientific, Belmont, MA.
- [BeT00] Bertsekas, D. P., and Tsitsiklis, J. N., 2000. "Gradient Convergence in Gradient Methods," *SIAM J. on Optimization*, Vol. 10, pp. 627-642.
- [BeT02] Bertsekas, D. P., and Tsitsiklis, J. N., 2002. *Introduction to Probability*, Athena Scientific, Belmont, MA.
- [Bel57] Bellman, R., 1957. *Applied Dynamic Programming*, Princeton University Press, Princeton, N. J.
- [Ber71] Bertsekas, D. P., 1971. "Control of Uncertain Systems With a Set-Membership Description of the Uncertainty," Ph.D. Thesis, Dept. of EECS, MIT; may be downloaded from <http://web.mit.edu/dimitrib/www/publ.html>.
- [Ber72] Bertsekas, D. P., 1972. "Infinite Time Reachability of State Space Regions by Using Feedback Control," *IEEE Trans. Automatic Control*, Vol. AC-17, pp. 604-613.
- [Ber73a] Bertsekas, D. P., 1973. "Stochastic Optimization Problems with Non-differentiable Cost Functionals," *J. Optimization Theory Appl.*, Vol. 12, pp. 218-231.
- [Ber73b] Bertsekas, D. P., 1973. "Linear Convex Stochastic Control Problems Over an Infinite Time Horizon," *IEEE Trans. Automatic Control*, Vol. AC-18, pp. 314-315.

- [Ber75] Bertsekas, D. P., 1975. "Convergence of Discretization Procedures in Dynamic Programming," IEEE Trans. Automatic Control, Vol. AC-20, pp. 415-419.
- [Ber76] Bertsekas, D. P., 1976. "On Error Bounds for Successive Approximation Methods," IEEE Trans. Automatic Control, Vol. AC-21, pp. 394-396.
- [Ber77] Bertsekas, D. P., 1977. "Monotone Mappings with Application in Dynamic Programming," SIAM J. on Control and Optimization, Vol. 15, pp. 438-464.
- [Ber82a] Bertsekas, D. P., 1982. "Distributed Dynamic Programming," IEEE Trans. Automatic Control, Vol. AC-27, pp. 610-616.
- [Ber82b] Bertsekas, D. P., 1982. Constrained Optimization and Lagrange Multiplier Methods, Academic Press, N. Y.
- [Ber83] Bertsekas, D. P., 1983. "Asynchronous Distributed Computation of Fixed Points," Math. Programming, Vol. 27, pp. 107-120.
- [Ber95a] Bertsekas, D. P., 1995. "A Generic Rank One Correction Algorithm for Markovian Decision Problems," Operations Research Letters, Vol. 17, pp. 111-119.
- [Ber95b] Bertsekas, D. P., 1995. "A Counterexample to Temporal Differences Learning," Neural Computation, Vol. 7, pp. 270-279.
- [Ber98] Bertsekas, D. P., 1998. "A New Value Iteration Method for the Average Cost Dynamic Programming Problem," SIAM J. on Control and Optimization, Vol. 36, pp. 742-759.
- [Ber99] Bertsekas, D. P., 1999. Nonlinear Programming: 2nd Edition, Athena Scientific, Belmont, MA.
- [Ber05a] Bertsekas, D. P., 2005. "Dynamic Programming and Suboptimal Control: A Survey from ADP to MPC," Fundamental Issues in Control, Special Issue for the CDC-ECC 05, European J. of Control, Vol. 11, Nos. 4-5.
- [Ber05b] Bertsekas, D. P., 2005. "Rollout Algorithms for Constrained Dynamic Programming," Lab. for Information and Decision Systems Report 2646, MIT.
- [BhE91] Bhattacharya, P. P., and Ephremides, A., 1991. "Optimal Allocations of a Server Between Two Queues with Due Times," IEEE Trans. on Automatic Control, Vol. 36, pp. 1417-1423.
- [Bil83] Billingsley, P., 1983. "The Singular Function of Bold Play," American Scientist, Vol. 71, pp. 392-397.
- [Bla62] Blackwell, D., 1962. "Discrete Dynamic Programming," Ann. Math. Statist., Vol. 33, pp. 719-726.
- [Bla65] Blackwell, D., 1965. "Discounted Dynamic Programming," Ann. Math. Statist., Vol. 36, pp. 226-235.
- [Bla70] Blackwell, D., 1970. "On Stationary Policies," J. Roy. Statist. Soc. Ser. A, Vol. 133, pp. 33-38.

- [Bor88] Borkar, V. S., 1988. "A Convex Analytic Approach to Markov Decision Processes," *Prob. Theory and Related Fields*, Vol. 78, pp. 583-602.
- [Bor89] Borkar, V. S., 1989. "Control of Markov Chains with Long-Run Average Cost Criterion: The Dynamic Programming Equations," *SIAM J. on Control and Optimization*, Vol. 27, pp. 642-657.
- [Bor91] Borkar, V. S., 1991. *Topics in Controlled Markov Chains*, Pitman Research Notes in Math. No. 240, Longman Scientific and Technical, Harlow.
- [Boy02] Boyan, J. A., 2002. "Technical Update: Least-Squares Temporal Difference Learning," *Machine Learning*, Vol. 49, pp. 1-15.
- [BrB96] Bradtke, S. J., and Barto, A. G., 1996. "Linear Least-Squares Algorithms for Temporal Difference Learning," *Machine Learning*, Vol. 22, pp. 33-57.
- [Bro65] Brown, B. W., 1965. "On the Iterative Method of Dynamic Programming on a Finite Space Discrete Markov Process," *Ann. Math. Statist.*, Vol. 36, pp. 1279-1286.
- [CaC97] Cao, X. R., and Chen, H. F., 1997. "Perturbation Realization Potentials and Sensitivity Analysis of Markov Processes," *IEEE Transactions on Automatic Control*, Vol. 32, pp. 1382-1393.
- [CaS92] Cavazos-Cadena, R., and Sennott, L. I., 1992. "Comparing Recent Assumptions for the Existence of Optimal Stationary Policies," *Operations Research Letters*, Vol. 11, pp. 33-37.
- [CaW98] Cao, X. R., and Wan, Y. W., 1998. "Algorithms for Sensitivity Analysis of Markov Systems Through Potentials and Perturbation Realization," *IEEE Transactions Control Systems Technology*, Vol. 6, pp. 482-494.
- [Cao99] Cao, X. R., 1999. "Single Sample Path Based Optimization of Markov Chains," *J. of Optimization Theory and Applicationa*, Vol. 100, pp. 527-548.
- [Cao04] Cao, X. R., 2004. "Learning and Optimization from a System Theoretic Perspective," in *Learning and Approximate Dynamic Programming*, by J. Si, A. Barto, W. Powell, (Eds.), IEEE Press, N. Y.
- [Cao05] Cao, X. R., 2005. "A Basic Formula for Online Policy Gradient Algorithms," *IEEE Transactions on Automatic Control*, Vol. 50, pp. 696-699.
- [Cav86] Cavazos-Cadena, R., 1986. "Finite-State Approximations for Denumerable State Discounted Markov Decision Processes," *Appl. Math. Opt.*, Vol. 14, pp. 1-26.
- [Cav89a] Cavazos-Cadena, R., 1989. "Necessary Conditions for the Optimality Equations in Average-Reward Markov Decision Processes," *Sys. Control Letters*, Vol. 11, pp. 65-71.
- [Cav89b] Cavazos-Cadena, R., 1989. "Weak Conditions for the Existence of Optimal Stationary Policies in Average Markov Decisions Chains with Unbounded Costs," *Kybernetika*, Vol. 25, pp. 145-156.
- [Cav91] Cavazos-Cadena, R., 1991. "Recent Results on Conditions for the Existence of Average Optimal Stationary Policies," *Annals of Operations Research*, Vol. 28, pp. 3-28.

- [ChT89] Chow, C.-S., and Tsitsiklis, J. N., 1989. "The Complexity of Dynamic Programming," *Journal of Complexity*, Vol. 5, pp. 466-488.
- [ChT91] Chow, C.-S., and Tsitsiklis, J. N., 1991. "An Optimal One-Way Multi-grid Algorithm for Discrete-Time Stochastic Control," *IEEE Trans. on Automatic Control*, Vol. AC-36, pp. 898-914.
- [ChV06] Choi, D. S., and Van Roy, B., 2006. "A Generalized Kalman Filter for Fixed Point Approximation and Efficient Temporal-Difference Learning," *Discrete Event Dynamic Systems*, Vol. 16, pp. 207-239.
- [CoR87] Courcoubetis, C. A., and Reiman, M. I., 1987. "Optimal Control of a Queueing System with Simultaneous Service Requirements," *IEEE Trans. on Automatic Control*, Vol. AC-32, pp. 717-727.
- [CoV84] Courcoubetis, C., and Varaiya, P. P., 1984. "The Service Process with Least Thinking Time Maximizes Resource Utilization," *IEEE Trans. Automatic Control*, Vol. AC-29, pp. 1005-1008.
- [CrC91] Cruz, R. L., and Chuah, M. C., 1991. "A Minimax Approach to a Simple Routing Problem," *IEEE Trans. on Automatic Control*, Vol. 36, pp. 1424-1435.
- [D'Ep60] D'Epenoux, F., 1960. "Sur un Probleme de Production et de Stockage Dans l'Aleatoire," *Rev. Francaise Aut. Infor. Recherche Operationnelle*, Vol. 14, (English Transl.: *Management Sci.*, Vol. 10, 1963, pp. 98-108).
- [DFV00] de Farias, D. P., and Van Roy, B., 2000. "On the Existence of Fixed Points for Approximate Value Iteration and Temporal-Difference Learning," *J. of Optimization Theory and Applications*, Vol. 105.
- [DFV03] de Farias, D. P., and Van Roy, B., 2003. "The Linear Programming Approach to Approximate Dynamic Programming," *Operations Research*, Vol. 51, pp. 850-865.
- [DFV04a] de Farias, D. P., and Van Roy, B., 2004. "On Constraint Sampling in the Linear Programming Approach to Approximate Dynamic Programming," *Mathematics of Operations Research*, Vol. 29, pp. 462-478.
- [Dan63] Dantzig, G. B., 1963. *Linear Programming and Extensions*, Princeton Univ. Press, Princeton, N. J.
- [Day92] Dayan, P., 1992. "The Convergence of TD(λ) for General λ ," *Machine Learning*, Vol. 8, pp. 341-362.
- [DeF68] Denardo, E. V., and Fox, B., 1968. "Multichain Markov Renewal Programs," *SIAM J. of Applied Math.*, Vol. 16, pp. 468-487.
- [DeF04] De Farias, D. P., 2004. "The Linear Programming Approach to Approximate Dynamic Programming," in *Learning and Approximate Dynamic Programming*, by J. Si, A. Barto, W. Powell, (Eds.), IEEE Press, N. Y.
- [DeG60] De Ghellinck, G. T., 1960. "Les Problemes de Decisions Sequentielles," *Cah. Centre d'Etudes Rec. Oper.*, Vol. 2, pp. 161-179.
- [DeV67] Derman, C., and Veinott, A. F., Jr., 1967. "A Solution to a Countable System of Equations Arising in Markovian Decision Processes," *Ann. Math. Statist.*, Vol. 37, pp. 582-584.

- [Dek87] Dekker, R., 1987. "Counter Examples for Compact Action Markov Decision Chains with Average Reward Criteria," *Communications in Statistics: Stochastic Models*, Vol. 3, pp. 357-368.
- [Den67] Denardo, E. V., 1967. "Contraction Mappings in the Theory Underlying Dynamic Programming," *SIAM Review*, Vol. 9, pp. 165-177.
- [Der62] Derman, C., 1962. "On Sequential Decisions and Markov Chains," *Management Sci.*, Vol. 9, pp. 16-24.
- [Der70] Derman, C., 1970. *Finite State Markovian Decision Processes*, Academic Press, N. Y.
- [DuS65] Dubins, L., and Savage, L. M., 1965. *How to Gamble If You Must*, McGraw-Hill, N. Y.
- [DyY79] Dynkin, E. B., and Yuskevich, A. A., 1979. *Controlled Markov Processes*, Springer-Verlag, N. Y.
- [EVW80] Ephremides, A., Varaiya, P. P., and Walrand, J. C., 1980. "A Simple Dynamic Routing Problem," *IEEE Trans. Automatic Control*, Vol. AC-25, pp. 690-693.
- [EaZ62] Eaton, J. H., and Zadeh, L. A., 1962. "Optimal Pursuit Strategies in Discrete State Probabilistic Systems," *Trans. ASME Ser. D. J. Basic Eng.*, Vol. 84, pp. 23-29.
- [EpV89] Ephremides, A., and Verd'u, S., 1989. "Control and Optimization Methods in Communication Network Problems," *IEEE Trans. Automatic Control*, Vol. AC-34, pp. 930-942.
- [FAM90] Fernández-Gaucherand, E., Arapostathis, A., and Marcus, S. I., 1990. "Remarks on the Existence of Solutions to the Average Cost Optimality Equation in Markov Decision Processes," *Systems and Control Letters*, Vol. 15, pp. 425-432.
- [FAM91] Fernández-Gaucherand, E., Arapostathis, A., and Marcus, S. I., 1991. "On the Average Cost Optimality Equation and the Structure of Optimal Policies for Partially Observable Markov Decision Processes," *Annals of Operations Research*, Vol. 29, pp. 439-470.
- [FHT79] Federgruen, A., Hordijk, A., and Tijms, H. C., 1979. "Denumerable State Semi-Markov Decision Processes with Unbounded Costs, Average Cost Criterion," *Stochastic Processes and their Applications*, Vol. 9, pp. 223-235.
- [FST78] Federgruen, A., Schweitzer, P. J., and Tijms, H. C., 1978. "Contraction Mappings Underlying Undiscounted Markov Decision Problems," *J. of Math. Analysis and Applications*, Vol. 65, pp. 711-730.
- [FeS94] Feinberg, E. A., and Shwartz, A., 1994. "Markov Decision Models with Weighted Discounted Criteria," *Mathematics of Operations Research*, Vol. 19, pp. 1-17.
- [FeS02] Feinberg, E. A., and Shwartz, A., 2002. *Handbook of Markov Decision Processes: Methods and Applications*, Kluwer, N. Y.
- [FeS04] Ferrari, S., and Stengel, R. F., 2004. "Model-Based Adaptive Critic De-

- signs," in Learning and Approximate Dynamic Programming, by J. Si, A. Barto, W. Powell, (Eds.), IEEE Press, N. Y.
- [Fei78] Feinberg, E. A., 1978. "The Existence of a Stationary ϵ -Optimal Policy for a Finite-State Markov Chain," *Theor. Prob. Appl.*, Vol. 23, pp. 297-313.
- [Fei92a] Feinberg, E. A., 1992. "Stationary Strategies in Borel Dynamic Programming," *Mathematics of Operations Research*, Vol. 125, pp. 87-96.
- [Fei92b] Feinberg, E. A., 1992. "A Markov Decision Model of a Search Process," *Comtemporary Mathematics*, Vol. 125, pp. 87-96.
- [FiV96] Filar, J., and Vrieze, K., 1996. Competitive Markov Decision Processes, Springer, N. Y.
- [Fox71] Fox, B. L., 1971. "Finite State Approximations to Denumerable State Dynamic Progams," *J. Math. Anal. Appl.*, Vol. 34, pp. 665-670.
- [FuH94] Fu, M. C., and Hu, 1994. "Smoothed Perurbation Analysis Derivative Estimation for Markov Chains," *Oper. Res. Letters*, Vol. 41, pp. 241-251.
- [GKP03] Guestrin, C. E., Koller, D., Parr, R., and Venkataraman, S., 2003. "Efficient Solution Algorithms for Factored MDPs," *J. of Artificial Intelligence Research*, Vol. 19, pp. 399-468.
- [GLH94] Gurvits, L., Lin, L. J., and Hanson, S. J., 1994. "Incremental Learning of Evaluation Functions for Absorbing Markov Chains: New Methods and Theorems," Preprint.
- [Gal95] Gallager, R. G., 1995. Discrete Stochastic Processes, Kluwer, N. Y.
- [Gho90] Ghosh, M. K., 1990. "Markov Decision Processes with Multiple Costs," *Operations Research Letters*, Vol. 9, pp. 257-260.
- [GiJ74] Gittins, J. C., and Jones, D. M., 1974. "A Dynamic Allocation Index for the Sequential Design of Experiments," *Progress in Statistics* (J. Gani, ed.), North-Holland, Amsterdam, pp. 241-266.
- [Gil57] Gillette, D., 1957. "Stochastic Games with Zero Stop Probabilities," in Contributions to the Theory of Games, III, Princeton Univ. Press, Princeton, N. J., *Annals of Math. Studies*, Vol. 39, pp. 71-187.
- [Git79] Gittins, J. C., 1979. "Bandit Processes and Dynamic Allocation Indices," *J. Roy. Statist. Soc.*, Vol. B, No. 41, pp. 148-164.
- [Gly87] Glynn, P. W., 1987. "Likelihood Ratio Gradient Estimation: An Overview," Proc. of the 1987 Winter Simulation Conference, pp. 366-375.
- [Gol03] Golubin, A. Y., 2003. "A Note on the Convergence of Policy Iteration in Markov Decision Processes with Compact Action Spaces," *Math. Operations Research*, Vol. 28, pp. 194-200.
- [Gos04] Gosavi, A., 2004. "Reinforcement Learning for Long-Run Average Cost," *European J. of Operational Research*, Vol. 155, pp. 654-674.
- [GrU04] Grudic, G., and Ungar, L., 2004. "Reinforcement Learning in Large, High-Dimensional State Spaces," in Learning and Approximate Dynamic Programming, by J. Si, A. Barto, W. Powell, (Eds.), IEEE Press, N. Y.

- [GuR06] Guo, X., and Rieder, U., 2006. "Average Optimality for Continuous-Time Markov Decision Processes in Polish Spaces," *Ann. Appl. Probability*, Vol. 16, pp. 730756.
- [HBK94] Harmon, M. E., Baird, L. C., and Klopf, A. H., 1994. "Advantage Updating Applied to a Differential Game," Presented at NIPS Conf., Denver, Colo.
- [HCP99] Hernandez-Lerma, O., Carrasco, O., and Perez-Hernandez, 1999. "Markov Control Processes with the Expected Total Cost Criterion: Optimality, Stability, and Transient Models," *Acta Appl. Math.*, Vol. 59, pp. 229-269.
- [HFM05] He, Y., Fu, M. C., and Marcus, S. I., 2005. "A Two-Timescale Simulation-Based Gradient Algorithm for Weighted Cost Markov Decision Processes," Proc. of the 2005 Conf. on Decision and Control, Seville, Spain, pp. 8022-8027.
- [HHL91] Hernandez-Lerma, O., Hennet, J. C., and Lasserre, J. B., 1991. "Average Cost Markov Decision Processes: Optimality Conditions," *J. Math. Anal. Appl.*, Vol. 158, pp. 396-406.
- [HPC96] Helmsen, J., Puckett, E. G., Colella, P., and Dorr, M., 1996. "Two New Methods for Simulating Photolithography Development," SPIE, Vol. 2726, pp. 253-261.
- [HaL86] Haurie, A., and L'Ecuyer, P., 1986. "Approximation and Bounds in Discrete Event Dynamic Programming," *IEEE Trans. Automatic Control*, Vol. AC-31, pp. 227-235.
- [Haj84] Hajek, B., 1984. "Optimal Control of Two Interacting Service Stations," *IEEE Trans. Automatic Control*, Vol. AC-29, pp. 491-499.
- [Har72] Harrison, J. M., 1972. "Discrete Dynamic Programming with Unbounded Rewards," *Ann. Math. Stat.*, Vol. 43, pp. 636-644.
- [Har75a] Harrison, J. M., 1975. "A Priority Queue with Discounted Linear Costs," *Operations Research*, Vol. 23, pp. 260-269.
- [Har75b] Harrison, J. M., 1975. "Dynamic Scheduling of a Multiclass Queue: Discount Optimality," *Operations Research*, Vol. 23, pp. 270-282.
- [Has68] Hastings, N. A. J., 1968. "Some Notes on Dynamic Programming and Replacement," *Operational Research Quart.*, Vol. 19, pp. 453-464.
- [He02] He, Y., 2002. *Simulation-Based Algorithms for Markov Decision Processes*, Ph.D. Thesis, University of Maryland.
- [HeL96] Hernandez-Lerma, O., and Lasserre, J. B., 1996. *Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, N. Y.
- [HeL97] Hernandez-Lerma, O., and Lasserre, J. B., 1997. "Policy Iteration for Average Cost Markov Control Processes on Borel Spaces," *Acta Applicandae Mathematicae*, Vol. 47, pp. 125-154.
- [HeL99] Hernandez-Lerma, O., and Lasserre, J. B., 1999. *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, N. Y.

- [HeS84] Heyman, D. P., and Sobel, M. J., 1984. Stochastic Models in Operations Research, Vol. II, McGraw-Hill, N. Y.
- [Her89] Hernandez-Lerma, O., 1989. Adaptive Markov Control Processes, Springer-Verlag, N. Y.
- [HiW05] Hinderer, K., and Waldmann, K.-H., 2005. "Algorithms for Countable State Markov Decision Models with an Absorbing Set," SIAM J. of Control and Optimization, Vol. 43, pp. 2109-2131.
- [Hin70] Hinderer, K., 1970. Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter, Springer-Verlag, N. Y.
- [HoP87] Hordijk, A., and Puterman, M. 1987. "On the Convergence of Policy Iteration in Finite State Undiscounted Markov Decision Processes: the Unichain Case," Math. of Operations Research, Vol. 12, pp. 163-176.
- [How60] Howard, R., 1960. Dynamic Programming and Markov Processes, MIT Press, Cambridge, MA.
- [JJS94] Jaakkola, T., Jordan, M. I., and Singh, S. P., 1994. "On the Convergence of Stochastic Iterative Dynamic Programming Algorithms," Neural Computation, Vol. 6, pp. 1185-1201.
- [JSJ95] Jaakkola, T., Singh, S. P., and Jordan, M. I., 1995. "Reinforcement Learning Algorithm for Partially Observable Markov Decision Problems," Advances in Neural Information Processing Systems, Vol. 7, pp. 345-352.
- [JaC06] James, H. W., and Collins, E. J., 2006. "An Analysis of Transient Markov Decision Processes," J. Appl. Prob., Vol. 43, pp. 603-621.
- [Jew63] Jewell, W., 1963. "Markov Renewal Programming I and II," Operations Research, Vol. 2, pp. 938-971.
- [KaV87] Katehakis, M., and Veinott, A. F., 1987. "The Multi-Armed Bandit Problem: Decomposition and Computation," Math. of Operations Research, Vol. 12, pp. 262-268.
- [Kak01] Kakade, S., 2001. "A Natural Policy Gradient," Proc. Advances in Neural Information Processing Systems, Vancouver, BC, Vol. 14, pp. 1531-1538.
- [Kal83] Kallenberg, L. C. M., 1983. Linear Programming and Finite Markov Control Problems, Mathematical Centre Report, Amsterdam.
- [Kal94a] Kallenberg, L. C. M., 1994. "Survey of linear programming for standard and nonstandard Markovian control problems. Part I: Theory," J. Math. Methods of Operations Research (ZOR), Vol. 40.
- [Kal94b] Kallenberg, L. C. M., 1994. "Survey of linear programming for standard and nonstandard Markovian control problems. Part II: Applications," J. Math. Methods of Operations Research (ZOR), Vol. 40.
- [Kel81] Kelly, F. P., 1981. "Multi-Armed Bandits with Discount Factor Near One: The Bernoulli Case," The Annals of Statistics, Vol. 9, pp. 987-1001.
- [Kle68] Kleinman, D. L., 1968. "On an Iterative Technique for Riccati Equation Computations," IEEE Trans. Automatic Control, Vol. AC-13, pp. 114-115.

- [KoB99] Konda, V. R., and Borkar, V. S., 1999. "Actor-Critic Like Learning Algorithms for Markov Decision Processes," SIAM J. on Control and Optimization, Vol. 38, pp. 94-123.
- [KoP00] Koller, K., and Parr, R., 2000. "Policy Iteration for Factored MDPs," Proc. of the 16th Annual Conference on Uncertainty in AI, pp. 326-334.
- [KoT99] Konda, V. R., and Tsitsiklis, J. N., 1999. "Actor-Critic Algorithms," Proc. 1999 Neural Information Processing Systems Conference, Denver, Colorado, pp. 1008-1014.
- [KoT03] Konda, V. R., and Tsitsiklis, J. N., 2003. "Actor-Critic Algorithms," SIAM J. on Control and Optimization, Vol. 42, pp. 1143-1166.
- [Kon02] Konda, V. R., 2002. Actor-Critic Algorithms, Ph.D. Thesis, Dept. of EECS, M.I.T., Cambridge, MA.
- [KuV86] Kumar, P. R., and Varaiya, P. P., 1986. Stochastic Systems: Estimation, Identification, and Adaptive Control, Prentice-Hall, Englewood Cliffs, N. J.
- [KuY97] Kushner, H. J., and Yin, G., 1997. Stochastic Approximation Algorithms and Applications, Springer-Verlag, New York.
- [Kum85] Kumar, P. R., 1985. "A Survey of Some Results in Stochastic Adaptive Control," SIAM J. on Control and Optimization, Vol. 23, pp. 329-380.
- [L'Ec91] L'Ecuyer, P., 1991. "An overview of derivative estimation, Proceedings of the 1991 Winter Simulation Conference, pp. 207-217.
- [LaP03] Lagoudakis, M. G., and Parr, R., 2003. "Least-Squares Policy Iteration," J. of Machine Learning Research, Vol. 4, pp. 1107-1149.
- [Las88] Lasserre, J. B., 1988. Conditions for Existence of Average and Blackwell Optimal Stationary Policies in Denumerable Markov Decision Processes," J. Math. Anal. Appl., Vol. 136, pp. 479-490.
- [LiK84] Lin, W., and Kumar, P. R., 1984. "Optimal Control of a Queueing System with Two Heterogeneous Servers," IEEE Trans. Automatic Control, Vol. AC-29, pp. 696-703.
- [LiR71] Lippman, S. A., and Ross, S. M., 1971. "The Streetwalker's Dilemma: A Job-Shop Model," SIAM J. of Appl. Math., Vol. 20, pp. 336-342.
- [LiS61] Liusternik, L., and Sobolev, V., 1961. Elements of Functional Analysis, Ungar, N. Y.
- [Lip73] Lippman, S. A., 1973. "Semi-Markov Decision Processes with Unbounded Rewards," Management Sci., Vol. 21, pp. 717-731.
- [Lip75a] Lippman, S. A., 1975. "On Dynamic Programming with Unbounded Rewards," Management Sci., Vol. 19, pp. 1225-1233.
- [Lip75b] Lippman, S. A., 1975. "Applying a New Device in the Optimization of Exponential Queuing Systems," Operations Research, Vol. 23, pp. 687-710.
- [Lit96] Littman, M. L., 1996. "Algorithms for Sequential Decision Making," Ph. D. thesis, Brown University, Providence, R. I.

- [LjS83] Ljung, L., and Soderstrom, T., 1983. Theory and Practice of Recursive Identification, MIT Press, Cambridge, MA.
- [LoS01] Longstaff, F. A., and Schwartz, E. S., 2001. "Valuing American Options by Simulation: A Simple Least-Squares Approach," *Review of Financial Studies*, Vol. 14, pp. 113-147.
- [Lue69] Luenberger, D. G., 1969. Optimization by Vector Space Methods, Wiley, N. Y.
- [MaT01] Marbach, P., and Tsitsiklis, J. N., 2001. "Simulation-Based Optimization of Markov Reward Processes," *IEEE Transactions on Automatic Control*, Vol. 46, pp. 191-209.
- [MaT03] Marbach, P., and Tsitsiklis, J. N., 2003. "Approximate Gradient Methods in Policy-Space Optimization of Markov Reward Processes," *J. Discrete Event Dynamic Systems*, Vol. 13, pp. 111-148.
- [Mah96] Mahadevan, S., 1996. "Average Reward Reinforcement Learning: Foundations, Algorithms, and Empirical Results," *Machine Learning*, Vol. 22, pp. 1-38.
- [Man60] Manne A., 1960. "Linear Programming and Sequential Decisions," *Man. Science*, Vol. 6, pp. 259-267.
- [McQ66] MacQueen, J., 1966. "A Modified Dynamic Programming Method for Markovian Decision Problems," *J. Math. Anal. Appl.*, Vol. 14, pp. 38-43.
- [Mey97] Meyn, S., 1997. "The Policy Iteration Algorithm for Average Reward Markov Decision Processes with General State Space," *IEEE Trans. on Automatic Control*, Vol. 42, pp. 1663-1680.
- [Mey99] Meyn, S., 1999. "Algorithms for Optimization and Stabilization of Controlled Markov Chains," *Sadhana*, Vol. 24, pp. 339-367.
- [MoW77] Morton, T. E., and Wecker, W., 1977. "Discounting, Ergodicity and Convergence for Markov Decision Processes," *Management Sci.*, Vol. 23, pp. 890-900.
- [Mor71] Morton, T. E., 1971. "On the Asymptotic Convergence Rate of Cost Differences for Markovian Decision Processes," *Operations Research*, Vol. 19, pp. 244-248.
- [NTW89] Nain, P., Tsoucas, P., and Walrand, J., 1989. "Interchange Arguments in Stochastic Scheduling," *J. of Appl. Prob.*, Vol. 27, pp. 815-826.
- [NeB03] Nedić, A., and Bertsekas, D. P., 2003. "Least-Squares Policy Evaluation Algorithms with Linear Function Approximation," *J. of Discrete Event Systems*, Vol. 13, pp. 79-110.
- [NgP86] Nguyen, S., and Pallottino, S., 1986. "Hyperpaths and Shortest Hyperpaths," in Combinatorial Optimization by B. Simeone (ed.), Springer-Verlag, N. Y., pp. 258-271.
- [OMK84] Ohnishi, M., Mine, H., and Kawai, H., 1984. "An Optimal Inspection and Replacement Policy Under Incomplete State Information: Average Cost Criterion," in Stochastic Models in Reliability Theory (S. Osaki and Y. Hatoyama,

- Eds.), Lect. Notes Econ. Math. Systems, Vol. 135, Springer-Verlag, Berlin, pp. 187-197.
- [Odo69] Odoni, A. R., 1969. "On Finding the Maximal Gain for Markov Decision Processes," *Operations Research*, Vol. 17, pp. 857-860.
- [OrR70] Ortega, J. M., and Rheinboldt, W. C., 1970. *Iterative Solution of Non-linear Equations in Several Variables*, Academic Press, N. Y.
- [Orn69] Ornstein, D., 1969. "On the Existence of Stationary Optimal Strategies," *Proc. Amer. Math. Soc.*, Vol. 20, pp. 563-569.
- [PBT98] Polymenakos, L. C., Bertsekas, D. P., and Tsitsiklis, J. N., 1998. "Efficient Algorithms for Continuous-Space Shortest Path Problems," *IEEE Trans. on Automatic Control*, Vol. 43, pp. 278-283.
- [PBW79] Popyack, J. L., Brown, R. L., and White, C. C., III, 1969. "Discrete Versions of an Algorithm due to Varaiya," *IEEE Trans. Aut. Control*, Vol. 24, pp. 503-504.
- [PaB99] Patek, S. D., and Bertsekas, D. P., 1999. "Stochastic Shortest Path Games," *SIAM J. on Control and Optimization*, Vol. 36, pp. 804-824.
- [PaK81] Pattipati, K. R., and Kleinman, D. L., 1981. "Priority Assignment Using Dynamic Programming for a Class of Queueing Systems," *IEEE Trans. on Automatic Control*, Vol. AC-26, pp. 1095-1106.
- [PaT87] Papadimitriou, C. H., and Tsitsiklis, J. N., 1987. "The Complexity of Markov Decision Processes," *Math. Operations Research*, Vol. 12, pp. 441-450.
- [PaT00] Paschalidis, I. C., and Tsitsiklis, J. N., 2000. "Congestion-Dependent Pricing of Network Services," *IEEE/ACM Transactions on Networking*, Vol. 8, pp. 171-184.
- [Pat04] Patek, S. D., 2004. "Policy Iteration Type Algorithms for Recurrent State Markov Decision Processes," *Computers and Operations Research*, Vol. 31, pp. 2333-2347.
- [Pin97] Pineda, F., 1997. "Mean-Field Analysis for Batched TD(λ)," *Neural Computation*, pp. 1403-1419.
- [Pla77a] Platzman, L., 1977. Finite Memory Estimation and Control of Finite Probabilistic Systems, Ph.D. Thesis, Dept. of EECS, MIT, Cambridge, MA.
- [Pla77b] Platzman, L., 1977. "Improved Conditions for Convergence in Undiscounted Markov Renewal Programming," *Operations Research*, Vol. 25, pp. 529-533.
- [Pla80] Platzman, L., 1980. "Optimal Infinite Horizon Undiscounted Control of Finite Probabilistic Systems," *SIAM J. Control and Opt.*, Vol. 18, pp. 362-380.
- [Pli78] Pliska, S. R., 1978. "On the Transient Case for Markov Decision Chains with General State Spaces," in *Dynamic Programming and Its Applications*, M. L. Puterman (ed.), Academic Press, N. Y.
- [PoA69] Pollatschek, M., and Avi-Itzhak, B., 1969. "Algorithms for Stochastic Games with Geometrical Interpretation," *Management Sci.*, Vol. 15, pp. 399-413.

- [PoT78] Porteus, E., and Totten, J., 1978. "Accelerated Computation of the Expected Discounted Return in a Markov Chain," *Operations Research*, Vol. 26, pp. 350-358.
- [PoT96] Polychronopoulos, G. H., and Tsitsiklis, J. N., 1996. "Stochastic Shortest Path Problems with Recourse," *Networks*, Vol. 27, pp. 133-143.
- [PoV04] Powell, W. B., and Van Roy, B., 2004. "Approximate Dynamic Programming for High-Dimensional Resource Allocation Problems," in *Learning and Approximate Dynamic Programming*, by J. Si, A. Barto, W. Powell, (Eds.), IEEE Press, N. Y.
- [Por71] Porteus, E., 1971. "Some Bounds for Discounted Sequential Decision Processes," *Management Sci.*, Vol. 18, pp. 7-11.
- [Por75] Porteus, E., 1975. "Bounds and Transformations for Finite Markov Decision Chains," *Operations Research*, Vol. 23, pp. 761-784.
- [Por81] Porteus, E., 1981. "Improved Conditions for Convergence in Undiscounted Markov Renewal Programming," *Operations Research*, Vol. 25, pp. 529-533.
- [PsT93] Psaraftis, H. N., and Tsitsiklis, J. N., 1993. "Dynamic Shortest Paths in Acyclic Networks with Markovian Arc Costs," *Operations Research*, Vol. 41, pp. 91-101.
- [PuB78] Puterman, M. L., and Brumelle, S. L., 1978. "The Analytic Theory of Policy Iteration," in *Dynamic Programming and Its Applications*, M. L. Puterman (ed.), Academic Press, N. Y.
- [PuS78] Puterman, M. L., and Shin, M. C., 1978. "Modified Policy Iteration Algorithms for Discounted Markov Decision Problems," *Management Sci.*, Vol. 24, pp. 1127-1137.
- [PuS82] Puterman, M. L., and Shin, M. C., 1982. "Action Elimination Procedures for Modified Policy Iteration Algorithms," *Operations Research*, Vol. 30, pp. 301-318.
- [Put78] Puterman, M. L. (ed.), 1978. *Dynamic Programming and its Applications*, Academic Press, N. Y.
- [Put94] Puterman, M. L., 1994. *Markovian Decision Problems*, J. Wiley, N. Y.
- [RVW82] Rosberg, Z., Varaiya, P. P., and Walrand, J. C., 1982. "Optimal Control of Service in Tandem Queues," *IEEE Trans. Automatic Control*, Vol. AC-27, pp. 600-609.
- [RaF91] Raghavan, T. E. S., and Filar, J. A., 1991. "Algorithms for Stochastic Games - A Survey," *ZOR - Methods and Models of Operations Research*, Vol. 35, pp. 437-472.
- [RiS92] Ritt, R. K., and Sennott, L. I., 1992. "Optimal Stationary Policies in General State Markov Decision Chains with Finite Action Set," *Math. Operations Research*, Vol. 17, pp. 901-909.
- [Roc70] Rockafellar, R. T., 1970. *Convex Analysis*, Princeton University Press, Princeton, N. J.

- [Ros70] Ross, S. M., 1970. *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, CA.
- [Ros71] Ross, S. M., 1971. "On the Nonexistence of ϵ -Optimal Randomized Stationary Policies in Average Cost Markov Decision Models," *The Annals of Math. Statistics*, Vol. 42, pp. 1767-1768.
- [Ros83a] Ross, S. M., 1983. *Introduction to Stochastic Dynamic Programming*, Academic Press, N. Y.
- [Ros83b] Ross, S. M., 1983. *Stochastic Processes*, Wiley, N. Y.
- [Ros89] Ross, K. W., 1989. "Randomized and Past-Dependent Policies for Markov Decision Processes with Multiple Constraints," *Operations Research*, Vol. 37, pp. 474-477.
- [Roy88] Royden, H. L., 1988. *Principles of Mathematical Analysis*, (3rd Ed.), McGraw-Hill, N. Y.
- [RuS94] Runggaldier, W. J., and Stettner, L., 1994. *Approximations of Discrete Time Partially Observed Control Problems*, Applied Math. Monographs 6, Giardini Editori e Stampatori, Pisa.
- [Rud76] Rudin, W., 1976. *Real Analysis*, (3rd Ed.), McGraw-Hill, N. Y.
- [Rus95] Rust, J., 1995. "Numerical Dynamic Programming in Economics," in *Handbook of Computational Economics*, H. Amman, D. Kendrick, and J. Rust (eds.).
- [Rus97] Rust, J., 1997. "Using Randomization to Break the Curse of Dimensionality," *Econometrica*, Vol. 65, pp. 487-516.
- [SBP04] Si, J., Barto, A., Powell, W., and Wunsch, D., (Eds.) 2004. *Learning and Approximate Dynamic Programming*, IEEE Press, N. Y.
- [SJ94] Singh, T. S., Jaakkola, T., and Jordan, M. I., 1994. "Learning Without State-Estimation in Partially Observable Markovian Decision Processes," Proc. 11th Conf. Machine Learning.
- [SMS99] Sutton, R. S., McAllester, D., Singh, S. P., and Mansour, Y., 1999. "Policy Gradient Methods for Reinforcement Learning with Function Approximation," Proc. 1999 Neural Information Processing Systems Conference, Denver, Colorado.
- [SYL04] Si, J., Yang, L., and Liu, D., 2004. "Direct Neural Dynamic Programming," in *Learning and Approximate Dynamic Programming*, by J. Si, A. Barto, W. Powell, (Eds.), IEEE Press, N. Y.
- [ScF77] Schweitzer, P. J., and Federgruen, A., 1977. "The Asymptotic Behavior of Value Iteration in Markov Decision Problems," *Math. Operations Research*, Vol. 2, pp. 360-381.
- [ScF78] Schweitzer, P. J., and Federgruen, A., 1978. "The Functional Equations of Undiscounted Markov Renewal Programming," *Math. Operations Research*, Vol. 3, pp. 308-321.
- [ScS85] Schweitzer, P. J., and Seidman, A., 1985. "Generalized Polynomial Ap-

- proximations in Markovian Decision Problems," *J. Math. Anal. and Appl.*, Vol. 110, pp. 568-582.
- [Sch68] Schweitzer, P. J., 1968. "Perturbation Theory and Finite Markov Chains," *J. Appl. Prob.*, Vol. 5, pp. 401-413.
- [Sch71] Schweitzer, P. J., 1971. "Iterative Solution of the Functional Equations of Undiscounted Markov Renewal Programming," *J. Math. Anal. Appl.*, Vol. 34, pp. 495-501.
- [Sch72] Schweitzer, P. J., 1972. "Data Transformations for Markov Renewal Programming," talk at National ORSA Meeting, Atlantic City, N. J.
- [Sch75] Schal, M., 1975. "Conditions for Optimality in Dynamic Programming and for the Limit of n -Stage Optimal Policies to be Optimal," *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, Vol. 32, pp. 179-196.
- [Sch81] Schweitzer, P. J., 1981. "Bottleneck Determination in a Network of Queues," Graduate School of Management Working Paper No. 8107, University of Rochester, Rochester, N. Y.
- [Sch93a] Schal, M., 1993. "Average Optimality in Dynamic Programming with General State Space," *Math. of Operations Research*, Vol. 18, pp. 163-172.
- [Sch93b] Schwartz, A., 1993. "A Reinforcement Learning Method for Maximizing Undiscounted Rewards," *Proc. of the 10th Machine Learning Conference*.
- [Sen86] Sennott, L. I., 1986. "A New Condition for the Existence of Optimum Stationary Policies in Average Cost Markov Decision Processes," *Operations Research Lett.*, Vol. 5, pp. 17-23.
- [Sen89a] Sennott, L. I., 1989. "Average Cost Optimal Stationary Policies in Infinite State Markov Decision Processes with Unbounded Costs," *Operations Research*, Vol. 37, pp. 626-633.
- [Sen89b] Sennott, L. I., 1989. "Average Cost Semi-Markov Decision Processes and the Control of Queueing Systems," *Prob. Eng. Info. Sci.*, Vol. 3, pp. 247-272.
- [Sen91] Sennott, L. I., 1991. "Value Iteration in Countable State Average Cost Markov Decision Processes with Unbounded Cost," *Annals of Operations Research*, Vol. 28, pp. 261-272.
- [Sen93a] Sennott, L. I., 1993. "The Average Cost Optimality Equation and Critical Number Policies," *Prob. Eng. Info. Sci.*, Vol. 7, pp. 47-67.
- [Sen93b] Sennott, L. I., 1993. "Constrained Average Cost Markov Decision Chains," *Prob. Eng. Info. Sci.*, Vol. 7, pp. 69-83.
- [Sen98] Sennott, L. I., 1998. *Stochastic Dynamic Programming and the Control of Queueing Systems*, Wiley, N. Y.
- [Ser79] Serfozo, R., 1979. "An Equivalence Between Discrete and Continuous Time Markov Decision Processes," *Operations Research*, Vol. 27, pp. 616-620.
- [Set99a] Sethian, J. A., 1999. *Level Set Methods and Fast Marching Methods Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, Cambridge University Press, N. Y.

- [Set99b] Sethian, J. A., 1999. "Fast Marching Methods," SIAM Review, Vol. 41.
- [Sha53] Shapley, L. S., 1953. "Stochastic Games," Proc. Nat. Acad. Sci. U.S.A., Vol. 39.
- [Sin94] Singh, S. P., 1994. "Reinforcement Learning Algorithms for Average-Payoff Markovian Decision Processes," Proc. of 12th National Conference on Artificial Intelligence, pp. 202-207.
- [Sob82] Sobel, M. J., 1982. "The Optimality of Full-Service Policies," Operations Research, Vol. 30, pp. 636-649.
- [StP74] Stidham, S., and Prabhu, N. U., 1974. "Optimal Control of Queueing Systems," in Mathematical Methods in Queueing Theory (Lecture Notes in Economics and Math. Syst., Vol. 98), A. B. Clarke (Ed.), Springer-Verlag, N. Y., pp. 263-294.
- [Ste93] Stettner, L., 1993. "Ergodic Control of Partially Observed Markov Processes with Equivalent Transition Probabilities," Applicationes Math. (Warsaw), Vol. 22, pp. 25-38.
- [Sti85] Stidham, S. S., 1985. "Optimal Control of Admission to a Queueing System," IEEE Trans. Automatic Control, Vol. AC-30, pp. 705-713.
- [Str66] Strauch, R., 1966. "Negative Dynamic Programming," Ann. Math. Statist., Vol. 37, pp. 871-890.
- [SuB98] Sutton, R. S., and Barto, A. G., 1998. Reinforcement Learning, MIT Press, Cambridge, MA.
- [SuC91] Suk, J.-B., and Cassandras, C. G., 1991. "Optimal Scheduling of Two Competing Queues with Blocking," IEEE Trans. on Automatic Control, Vol. 36, pp. 1086-1091.
- [Sut88] Sutton, R. S., 1988. "Learning to Predict by the Methods of Temporal Differences," Machine Learning, Vol. 3, pp. 9-44.
- [TSC92] Towsley, D., Sparaggis, P. D., and Cassandras, C. G., 1992. "Optimal Routing and Buffer Allocation for a Class of Finite Capacity Queueing Systems," IEEE Trans. on Automatic Control, Vol. 37, pp. 1446-1451.
- [Tes92] Tesauro, G., 1992. "Practical Issues in Temporal Difference Learning," Machine Learning, Vol. 8, pp. 257-277.
- [Tho80] Thomas, L. C., 1980. "Connectedness Conditions for Denumerable State Markov Decision Processes," in Recent Developments in Markov Decision Processes, by R. Hartley, L. C. Thomas, and D. F. White (Eds.), Academic Press, N. Y., pp. 181-204.
- [TsV96] Tsitsiklis, J. N., and Van Roy, B., 1996. "Feature-Based Methods for Large-Scale Dynamic Programming," Machine Learning, Vol. 22, pp. 59-94.
- [TsV97] Tsitsiklis, J. N., and Van Roy, B., 1997. "An Analysis of Temporal-Difference Learning with Function Approximation," IEEE Transactions on Automatic Control, Vol. 42, pp. 674-690.

- [TsV99a] Tsitsiklis, J. N., and Van Roy, B., 1999. "Average Cost Temporal-Difference Learning," *Automatica*, Vol. 35, pp. 1799-1808.
- [TsV99b] Tsitsiklis, J. N., and Van Roy, B., 1999. "Optimal Stopping of Markov Processes: Hilbert Space Theory, Approximation Algorithms, and an Application to Pricing Financial Derivatives", *IEEE Transactions on Automatic Control*, Vol. 44, pp. 1840-1851.
- [TsV02] Tsitsiklis, J. N., and Van Roy, B., 2002. "On Average Versus Discounted Reward Temporal-Difference Learning," *Machine Learning*, Vol. 49, pp. 179-191.
- [Tse90] Tseng, P., 1990. "Solving H -Horizon, Stationary Markov Decision Problems in Time Proportional to $\log(H)$," *Operations Research Letters*, Vol. 9, pp. 287-297.
- [Tsi86] Tsitsiklis, J. N., 1986. "A Lemma on the Multiarmed Bandit Problem," *IEEE Trans. Automatic Control*, Vol. AC-31, pp. 576-577.
- [Tsi89] Tsitsiklis, J. N., 1989. "A Comparison of Jacobi and Gauss-Seidel Parallel Iterations," *Applied Math. Lett.*, Vol. 2, pp. 167-170.
- [Tsi94a] Tsitsiklis, J. N., 1994. "A Short Proof of the Gittins Index Theorem," *Annals of Applied Probability*, Vol. 4, pp. 194-199.
- [Tsi94b] Tsitsiklis, J. N., 1994. "Asynchronous Stochastic Approximation and Q-Learning," *Machine Learning*, Vol. 16, pp. 185-202.
- [Tsi95] Tsitsiklis, J. N., 1995. "Efficient Algorithms for Globally Optimal Trajectories," *IEEE Trans. Automatic Control*, Vol. AC-40, pp. 1528-1538.
- [Tsi06] Tsitsiklis, J. N., 2006. "NP-Hardness of Checking the Unichain Condition in Average Cost MDPs," unpublished report, MIT.
- [VWB85] Varaiya, P. P., Walrand, J. C., and Buyukkoc, C., 1985. "Extensions of the Multiarmed Bandit Problem: The Discounted Case," *IEEE Trans. Automatic Control*, Vol. AC-30, pp. 426-439.
- [VaW78] Van Nunen, J. A., and Wessels, J., 1978. "A Note on Dynamic Programming with Unbounded Rewards," *Management Sci.*, Vol. 24, pp. 576-580.
- [Van76] Van Nunen, J. A., 1976. Contracting Markov Decision Processes, Mathematical Centre Report, Amsterdam.
- [Van98] Van Roy, B., 1998. Learning and Value Function Approximation in Complex Decision Processes, Ph.D. Thesis, Dept. of EECS, MIT, Cambridge, MA.
- [Var78] Varaiya, P. P., 1978. "Optimal and Suboptimal Stationary Controls of Markov Chains," *IEEE Trans. Automatic Control*, Vol. AC-23, pp. 388-394.
- [VeP84] Verd'u, S., and Poor, H. V., 1984. "Backward, Forward, and Backward-Forward Dynamic Programming Models under Commutativity Conditions," Proc. 1984 IEEE Decision and Control Conference, Las Vegas, NE, pp. 1081-1086.
- [VeP87] Verd'u, S., and Poor, H. V., 1987. "Abstract Dynamic Programming Models under Commutativity Conditions," *SIAM J. on Control and Optimization*, Vol. 25, pp. 990-1006.
- [Wei66] Veinott, A. F., Jr., 1966. "On Finding Optimal Policies in Discrete Dy-

- namic Programming with no Discounting," *Ann. Math. Statist.*, Vol. 37, pp. 1284-1294.
- [Wei69] Veinott, A. F., Jr., 1969. "Discrete Dynamic Programming with Sensitive Discount Optimality Criteria," *Ann. Math. Statist.*, Vol. 40, pp. 1635-1660.
- [ViE88] Viniotis, I., and Ephremides, A., 1988. "Extension of the Optimality of the Threshold Policy in Heterogeneous Multiserver Queueing Systems," *IEEE Trans. on Automatic Control*, Vol. 33, pp. 104-109.
- [WaB92] Watkins, C. J. C. H., and Dayan, P., 1992. "Q-Learning," *Machine Learning*, Vol. 8, pp. 279-292.
- [Wat89] Watkins, C. J. C. H., *Learning from Delayed Rewards*, Ph.D. Thesis, Cambridge Univ., England.
- [Web93] Weber, R., 1993. "On the Gittins Index for Multiarmed Bandits," *Annals of Applied Probability*, Vol. 3.
- [Wes77] Wessels, J., 1977. "Markov Programming by Successive Approximations with Respect to Weighted Supremum Norms," *J. Math. Anal. Appl.*, Vol. 58, pp. 326-335.
- [WhK80] White, C. C., and Kim, K., 1980. "Solution Procedures for Partially Observed Markov Decision Processes," *J. Large Scale Systems*, Vol. 1, pp. 129-140.
- [Whi63] White, D. J., 1963. "Dynamic Programming, Markov Chains, and the Method of Successive Approximations," *J. Math. Anal. and Appl.*, Vol. 6, pp. 373-376.
- [Whi78] Whitt, W., 1978. "Approximations of Dynamic Programs I," *Math. Operations Research*, Vol. 3, pp. 231-243.
- [Whi79] Whitt, W., 1979. "Approximations of Dynamic Programs II," *Math. Operations Research*, Vol. 4, pp. 179-185.
- [Whi80a] White, D. J., 1980. "Finite State Approximations for Denumerable State Infinite Horizon Discounted Markov Decision Processes: The Method of Successive Approximations," in *Recent Developments in Markov Decision Processes*, Hartley, R., Thomas, L. C., and White, D. J. (eds.), Academic Press, N. Y., pp. 57-72.
- [Whi80b] Whittle, P., 1980. "Multi-Armed Bandits and the Gittins Index," *J. Roy. Statist. Soc. Ser. B*, Vol. 42, pp. 143-149.
- [Whi81] Whittle, P., 1981. "Arm-Acquiring-Bandits," *The Annals of Probability*, Vol. 9, pp. 284-292.
- [Whi82] Whittle, P., 1982. *Optimization Over Time*, Wiley, N. Y., Vol. 1, 1982, Vol. 2, 1983.
- [WiB93] Williams, R. J., and Baird, L. C., 1993. "Analysis of Some Incremental Variants of Policy Iteration: First Steps Toward Understanding Actor-Critic Learning Systems," Report NU-CCS-93-11, College of Computer Science, Northeastern University, Boston, MA.

- [Wil92] Williams, R. J., 1992. "Simple Statistical Gradient Following Algorithms for Connectionist Reinforcement Learning," *Machine Learning*, Vol. 8, pp. 229-256,
- [YuB04] Yu, H., and Bertsekas, D. P., 2004. "Discretized Approximations for POMDP with Average Cost," Proc. of the 20th Conference on Uncertainty in Artificial Intelligence, Banff, Canada.
- [YuB06a] Yu, H., and Bertsekas, D. P., 2006. "On Near-Optimality of the Set of Finite-State Controllers for Average Cost POMDP," Lab. for Information and Decision Systems Report 2689, MIT.
- [YuB06b] Yu, H., and Bertsekas, D. P., 2006. "Convergence Results for Some Temporal Difference Methods Based on Least Squares," Lab. for Information and Decision Systems Report 2697, MIT.
- [YuB06c] Yu, H., and Bertsekas, D. P., 2006. "A Least Squares Q-Learning Algorithm for Optimal Stopping Problems," Lab. for Information and Decision Systems Report 2731, MIT.
- [Yu05] Yu, H., 2005. "A Function Approximation Approach to Estimation of Policy Gradient for POMDP with Structured Policies," Proc. of the 21st Conference on Uncertainty in Artificial Intelligence, Edinburgh, Scotland.
- [Yu06] Yu, H., 2006. Approximate Solution Methods for Partially Observable Markov and Semi-Markov Decision Processes, Ph.D. Thesis, Dept. of EECS, M.I.T., Cambridge, MA.

INDEX

A

- Accessibility condition, 199
- Admissible policy, 4
- Advantage updating, 401
- Aggregation, 46, 111, 232
- Analytic set, 413
- Analytically measurable, 413
- Aperiodic Markov chain, 177
- Approximation in policy space, 392
- Asset selling, 147, 319
- Asynchronous algorithms, 33, 44, 91
- Asynchronous policy iteration, 44, 73
- Average cost problem, 174, 310

B

- Basis functions, 52, 79, 326
- Bellman's equation, 9, 15, 97, 127, 176, 198, 254, 262, 278, 291, 294, 317, 323
- Bias, 183
- Blackwell optimal policy, 176, 184, 297
- Bold strategy, 151
- Borel space, 413
- Borel measurable, 413

C

- Chattering, 338, 362, 399
- Column reduction, 82
- Contracting value iteration, 217, 222, 391
- Contraction mappings, 56–66, 80, 102, 113, 120, 121
- Consistently improving policies, 108, 114, 119
- Controllability, 141, 272
- Cost approximation, 325

D

- Data transformations, 87
- Differential cost, 176
- Dijkstra's algorithm, 108, 119
- Direct policy evaluation, 323, 329
- Discounted cost, 12, 289, 306

Distributed computation, 91
Duality, 240, 243

E

- ϵ -optimal policy, 162
- Error bounds, 22, 84, 213

F

- Feature extraction, 327
- Feature vectors, 327

G

- Gain, 183
- Gambling, 150, 163, 170
- Gauss-Seidel method, 31, 108, 117, 222

H

- History-dependent policy, 10

I

- Improper policy, 96
- Index function, 69
- Index of a project, 67
- Index rule, 67
- Indirect policy evaluation, 323, 340
- Inventory control, 142
- Irreducible Markov chain, 230

J

- Jacobi method, 83

L

- LSPE(λ), 324, 348
- LSTD(λ), 324, 355
- LLL strategy, 108
- Label correcting method, 108
- Laurent series, 182, 183, 275
- Limited lookahead policy, 53
- Linear programming, 51, 140, 239
- Linear quadratic problems, 140, 166–168, 272, 285
- Lower semianalytic, 414

M

Measurability issues, 3, 79, 407
 Measurable selection, 417
 Minimax problems, 86
 Monotone convergence theorem, 126
 Monte-Carlo simulation, 322, 402
 Multiarmed bandit problem, 66, 301
 Multiple-rank corrections, 36, 301

N

Negative DP model, 124
 Neuro-dynamic programming, 322, 399
 Newton's method, 85
 Nonstationary problems, 157
 Normalization rule, 230

O

Observability, 141, 272
 One-step-lookahead rule, 146, 147, 149, 150
 Optimistic policy iteration, 337, 347

P

$PVI(\lambda)$, 342, 348
 Parallel computation, 79, 91
 Periodic Markov chain, 177
 Periodic problems, 157, 161, 167, 169
 Policy, 3
 Policy evaluation, 39, 229, 235
 Policy improvement, 40, 230
 Policy iteration, 38, 85, 88, 108, 140, 229
 Policy iteration, approximate, 47, 109, 111, 329
 Policy iteration, modified, 43, 88, 109
 Policy gradient method, 406
 Polynomial approximations, 327
 Positive DP model, 124
 Priority assignment, 299
 Projected Bellman equation, 323, 341
 Projected value iteration, 342
 Proper policy, 96

Q

Q-factor, 325, 326
 Q-learning, 324, 363, 389, 401
 Quadratic cost, 140, 166–168, 272, 285
 Queueing control, 295, 309

R

Randomized policy, 10, 244
 Rank-one correction, 33, 83
 Reachability, 171, 172
 Recurrent class, 177
 Recurrent state, 177
 Reinforcement learning, 322, 399
 Relative cost, 176, 184
 Replacement problems, 17, 200
 Riccati equation, 141, 272
 Rollout, 53
 Routing, 302

S

SLF strategy, 108
 Scheduling problems, 66
 Semi-Markov problems, 306
 Semicontinuous model, 412
 Sequential hypothesis testing, 147
 Sequential probability ratio, 148
 Sequential space decomposition, 117
 Shortest path problem, 94, 105, 118
 Span seminorm, 258
 State-action frequencies, 244
 Stationary policy, 4
 Stochastic shortest paths, 93, 369
 Stopping problems, 105, 145
 Successive approximation, 22
 Sup-norm contraction, 57

T

$TD(\lambda)$, 324, 357
 Temporal differences, 324, 335, 352, 378
 Threshold policies, 88

U

Unbounded costs per stage, 124
 Undiscounted problems, 124
 Unichain policy, 201
 Uniformization, 288, 317
 Universally measurable, 407, 415

V

Value iteration, 22, 105, 139, 204
 Value iteration, approximate, 37
 Value iteration, contracting, 217, 222, 391

Value iteration, relative, 207, 222, 279,

390

Value iteration, termination, 26, 106

W

Weak accessibility condition, 198

Weighted sup norm, 57

Weighted sup-norm contraction, 57

**ATHENA SCIENTIFIC
OPTIMIZATION AND COMPUTATION SERIES**

1. Convex Analysis and Optimization, by Dimitri P. Bertsekas, with Angelia Nedić and Asuman E. Ozdaglar, 2003, ISBN 1-886529-45-0, 560 pages
2. Introduction to Probability, by Dimitri P. Bertsekas and John N. Tsitsiklis, 2002, ISBN 1-886529-40-X, 430 pages
3. Dynamic Programming and Optimal Control, Two-Volume Set, by Dimitri P. Bertsekas, 2007, ISBN 1-886529-08-6, 1020 pages
4. Nonlinear Programming, 2nd Edition, by Dimitri P. Bertsekas, 1999, ISBN 1-886529-00-0, 791 pages
5. Network Optimization: Continuous and Discrete Models, by Dimitri P. Bertsekas, 1998, ISBN 1-886529-02-7, 608 pages
6. Network Flows and Monotropic Optimization, by R. Tyrrell Rockafellar, 1998, ISBN 1-886529-06-X, 634 pages
7. Introduction to Linear Optimization, by Dimitris Bertsimas and John N. Tsitsiklis, 1997, ISBN 1-886529-19-1, 608 pages
8. Parallel and Distributed Computation: Numerical Methods, by Dimitri P. Bertsekas and John N. Tsitsiklis, 1997, ISBN 1-886529-01-9, 718 pages
9. Neuro-Dynamic Programming, by Dimitri P. Bertsekas and John N. Tsitsiklis, 1996, ISBN 1-886529-10-8, 512 pages
10. Constrained Optimization and Lagrange Multiplier Methods, by Dimitri P. Bertsekas, 1996, ISBN 1-886529-04-3, 410 pages
11. Stochastic Optimal Control: The Discrete-Time Case, by Dimitri P. Bertsekas and Steven E. Shreve, 1996, ISBN 1-886529-03-5, 330 pages

Neuro-Dynamic Programming

Dimitri P. Bertsekas and John N. Tsitsiklis
Massachusetts Institute of Technology

This is the first textbook that fully explains the neuro-dynamic programming/reinforcement learning methodology, which is a recent breakthrough in the practical application of neural networks and dynamic programming to complex problems of planning, optimal decision making, and intelligent control.

From the review by George Cybenko for IEEE Computational Science and Engineering, May 1998:

"Neurodynamic Programming is a remarkable monograph that integrates a sweeping mathematical and computational landscape into a coherent body of rigorous knowledge. The topics are current, the writing is clear and to the point, the examples are comprehensive and the historical notes and comments are scholarly."

"In this monograph, Bertsekas and Tsitsiklis have performed a Herculean task that will be studied and appreciated by generations to come. I strongly recommend it to scientists and engineers eager to seriously understand the mathematics and computations behind modern behavioral machine learning."

Among its special features, the book:

- Describes and unifies a large number of NDP methods, including several that are new
- Describes new approaches to formulation and solution of important problems in stochastic optimal control, sequential decision making, and discrete optimization
- Rigorously explains the mathematical principles behind NDP
- Illustrates through examples and case studies the practical application of NDP to complex problems from optimal resource allocation, optimal feedback control, data communications, game playing, and combinatorial optimization
- Presents extensive background and new research material on dynamic programming and neural network training

Neuro-Dynamic Programming is the winner of the 1997 INFORMS CSTS prize for research excellence in the interface between Operations Research and Computer Science

ISBN 1-886529-10-8, 512 pp., hardcover, 1996

This is a substantially expanded and improved edition of the best-selling book by Bertsekas on dynamic programming, a central algorithmic method for optimal control, sequential decision making under uncertainty, and combinatorial optimization. The treatment focuses on basic unifying themes and conceptual foundations. It illustrates the versatility, power, and generality of the method with many examples and applications from engineering, operations research, and economics.

“Here is a tour-de-force in its field.” — from the review of the first edition by David K. Smith, *Journal of Operational Research Society*

“In conclusion, this book is an excellent source of reference . . . The main strengths of the book are the clarity of the exposition, the quality and variety of the examples, and its coverage of the most recent advances.” — from the review of the first edition by Thomas W. Archibald, *IMA Journal of Mathematics Applied in Business and Industry*

This new edition of the second volume contains:

- new** A comprehensive and mathematically rigorous treatment of infinite horizon problems
- new** Extensive coverage of recent research on simulation-based approximation techniques (neuro-dynamic programming), which allow the practical application of dynamic programming to complex problems
- new** An in-depth development of the average cost problem, including a full analysis of multichain problems, and an extensive analysis of infinite-spaces problems
- new** Highlighting of the role of contraction mappings in infinite state space problems and in approximate dynamic programming
- new** A summary and orientation appendix on the mathematical measure-theoretic issues that must be addressed for a rigorous theory of stochastic dynamic programming.

DIMITRI P. BERTSEKAS, a member of the National Academy of Engineering, is Professor of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology.



Related books of interest by
Athena Scientific:

Neuro-Dynamic Programming, Dimitri P. Bertsekas and John N. Tsitsiklis, 1996

Stochastic Optimal Control: The Discrete-Time Case, Dimitri P. Bertsekas and Steven E. Shreve, 1996

Introduction to Probability, Dimitri P. Bertsekas and John N. Tsitsiklis, 2002

ISBN 1-886529-30-2

9 0000

