Dynamic Programming and Optimal Control

Volume II

Approximate Dynamic Programming

FOURTH EDITION

Dimitri P. Bertsekas

Massachusetts Institute of Technology

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 Gallager, www.gallagerdesign.com
Cover photography: Dimitri and Melina Bertsekas

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 2012 519.703 01-75941

ISBN-10: 1-886529-44-2, ISBN-13: 978-1-886529-44-1 (Vol. II)
ISBN 1-886529-26-4 (Vol. I)
ISBN 1-886529-08-6 (Two-volume set – latest editions)

Updated 2nd Printing: 2018

Contents

1. Discounted Problems – Theory

2. Discounted Problems – Computational Methods

3. Stochastic Shortest Path Problems

4. Undiscounted Problems

5. Average Cost per Stage Problems

6. Approximate Dynamic Programming - Discounted Models

7. Approximate Dynamic Programming - Nondiscounted Models and Generalizations

Appendix A: Measure-Theoretic Issues in Dynamic Programming

CONTENTS OF VOLUME I
4TH EDITION, 2017

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 Department, Stanford University, and the Electrical Engineering Department 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 McAfee Professor of Engineering. In 2019, he joined the School of Computing, Informatics, and Decision Systems Engineering at the Arizona State University, Tempe, AZ, as Fulton Professor of Computational Decision Making.

Professor Bertsekas' teaching and research have spanned several fields, including deterministic optimization, dynamic programming and stochastic control, large-scale and distributed computation, artificial intelligence, and data communication networks. He has authored or coauthored numerous research papers and eighteen books, several of which are currently used as textbooks in MIT classes, including "Dynamic Programming and Optimal Control," "Data Networks," "Introduction to Probability," and "Nonlinear Programming."

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 2001 AACC John R. Ragazzini Education Award, the 2009 INFORMS Expository Writing Award, the 2014 AACC Richard Bellman Heritage Award, the 2014 INFORMS Khachiyan Prize for Life-Time Accomplishments in Optimization, the 2015 MOS/SIAM George B. Dantzig Prize, and the 2022 IEEE Control Systems Award. In 2018 he shared with his coauthor, John Tsitsiklis, the 2018 INFORMS John von Neumann Theory Prize for the contributions of the research monographs "Parallel and Distributed Computation" and "Neuro-Dynamic Programming." Professor Bertsekas was elected in 2001 to the United States National Academy of Engineering for "pioneering contributions to fundamental research, practice and education of optimization/control theory, and especially its application to data communication networks."

ATHENA SCIENTIFIC
OPTIMIZATION AND COMPUTATION SERIES

1. Rollout, Policy Iteration, and Distributed Reinforcement Learning, by Dimitri P. Bertsekas, 2020, ISBN 978-1-886529-07-6, 480 pages
2. Reinforcement Learning and Optimal Control, by Dimitri P. Bertsekas, 2019, ISBN 978-1-886529-39-7, 388 pages
3. Abstract Dynamic Programming, 2nd Edition, by Dimitri P. Bertsekas, 2018, ISBN 978-1-886529-46-5, 360 pages
4. Dynamic Programming and Optimal Control, Two-Volume Set, by Dimitri P. Bertsekas, 2017, ISBN 1-886529-08-6, 1270 pages
5. Nonlinear Programming, 3rd Edition, by Dimitri P. Bertsekas, 2016, ISBN 1-886529-05-1, 880 pages
6. Convex Optimization Algorithms, by Dimitri P. Bertsekas, 2015, ISBN 978-1-886529-28-1, 576 pages
7. Convex Optimization Theory, by Dimitri P. Bertsekas, 2009, ISBN 978-1-886529-31-1, 256 pages
8. Introduction to Probability, 2nd Edition, by Dimitri P. Bertsekas and John N. Tsitsiklis, 2008, ISBN 978-1-886529-23-6, 544 pages
9. Convex Analysis and Optimization, by Dimitri P. Bertsekas, Angelia Nedić, and Asuman E. Ozdaglar, 2003, ISBN 1-886529-45-0, 560 pages
10. Network Optimization: Continuous and Discrete Models, by Dimitri P. Bertsekas, 1998, ISBN 1-886529-02-7, 608 pages
11. Network Flows and Monotropic Optimization, by R. Tyrrell Rockafellar, 1998, ISBN 1-886529-06-X, 634 pages
12. Introduction to Linear Optimization, by Dimitris Bertsimas and John N. Tsitsiklis, 1997, ISBN 1-886529-19-1, 608 pages
13. Parallel and Distributed Computation: Numerical Methods, by Dimitri P. Bertsekas and John N. Tsitsiklis, 1997, ISBN 1-886529-01-9, 718 pages
14. Neuro-Dynamic Programming, by Dimitri P. Bertsekas and John N. Tsitsiklis, 1996, ISBN 1-886529-10-8, 512 pages
15. Constrained Optimization and Lagrange Multiplier Methods, by Dimitri P. Bertsekas, 1996, ISBN 1-886529-04-3, 410 pages
16. 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 Humblet, 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
November 1995

[†] 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 www

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

Dimitri P. Bertsekas
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 major 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 who 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
November 2006

Preface to the Fourth Edition

This is a major revision of Vol. II, and contains a substantial amount of new material, as well as a reorganization of old material. The length has increased by more than 60% from the third edition, and most of the old material has been restructured and/or revised. Volume II now numbers more than 700 pages and is larger in size than Vol. I. It can arguably be viewed as a new book!

Approximate DP has become the central focal point of Vol. II, and occupies more than half of the book (the last two chapters, and large parts of Chapters 1-3). Thus one may also view Vol. II as a followup to my 1996 book "Neuro-Dynamic Programming" (coauthored with John Tsitsiklis). The present book focuses to a great extent on new research that became available after 1996. On the other hand, the textbook style of the book has been preserved, and some material has been explained at an intuitive or informal level, while referring to the journal literature or the Neuro-Dynamic Programming book for a more mathematical treatment.

In the process of expansion and reorganization, the design of the book became more modular and suitable for classroom use. The core material, which can be covered in about a third to a half of one semester is Chapter 1 (except for the application-specific Sections 1.3 and 1.4), Chapter 2, and Chapter 6, which are self-contained when taken together. This material focuses on discounted problems, and may be supplemented by parts of Chapter 3 and Section 7.1 on stochastic shortest path problems. Indeed, this comprises half of what I cover in my MIT class (the remaining half comes from Volume I, including Chapter 6 of that volume that deals with finite horizon approximate DP). The material on average cost problems, given in Chapter 5, and Sections 7.2 and 7.4, and the advanced material on positive and negative DP models (Chapter 4), and Monte Carlo linear algebra (Section 7.3) are terminal subjects that may be covered at the instructor's discretion.

As the book's focus shifted, I placed increased emphasis on new or recent research in approximate DP and simulation-based methods, as well as on asynchronous iterative methods, in view of the central role of simulation, which is by nature asynchronous. A lot of this material is an outgrowth of my research and the research of my collaborators, conducted in the six years since the previous edition. Some of the highlights, in the order appearing in the book, are:

(a) Computational methods for generalized discounted DP (Sections 2.5 and 2.6), including error bounds for approximations in Section 2.5, and the asynchronous optimistic policy iteration methods of Sections 2.6.2 and 2.6.3, and their application to game and minimax problems, constrained policy iteration, and Q-learning.

(b) Policy iteration methods (including asynchronous optimistic versions) for stochastic shortest path problems that involve improper policies (Section 3.4).

(c) Extensive new material on various simulation-based, approximate value and policy iteration methods in Sections 6.3-6.6 (projected equation and aggregation methods).

(d) New reliable Q-learning algorithms for optimistic policy iteration (Sections 2.6.3 and 6.6.2).

(e) New simulation techniques for multistep methods, such as geometric and free-form sampling (Sections 6.4.1 and 7.3.3).

(f) Extensive new material on Monte Carlo linear algebra in Section 7.3 (primarily the simulation-based and approximate solution of large systems of linear equations), which extends the DP methodology of approximate policy evaluation.

Much of the research in (a)-(e) is based on my work with Janey (Huizhen) Yu, while most of the research in (f) is based on my work with Janey Yu and Mengdi Wang. My collaboration with Janey and Mengdi has had a strong impact on the book, and is greatly appreciated. Some of our work was presented in summary only, and was adapted to fit the style and purposes of this book; naturally, any shortcomings in its presentation are entirely my responsibility. The reader is referred to our joint and individual papers, which describe more fully our research, including material that could not be covered in this book.

I want to express my appreciation to colleagues and collaborators in approximate DP research, who contributed to the book in various ways, particularly Vivek Borkar, Angelia Nedić, and Ben Van Roy. A special thanks goes to John Tsitsiklis, with whom I have interacted extensively through collaboration and sharing of ideas on DP and asynchronous algorithms for more than 30 years. I also wish to acknowledge helpful interactions with many colleagues, including Vivek Farias, Eugene Feinberg, Warren Powell, Martin Puterman, Uriel Rothblum, and Bruno Scherrer. Finally, I want to thank the many students in my DP classes of the last decade, who patiently labored with a textbook under development, and contributed their ideas and experiences through their research projects from a broad variety of application fields.

Dimitri P. Bertsekas
May 2012

NOTE ABOUT THIS UPDATED PRINTING

In this 2nd printing of the 4th edition of Vol. II (2018) I have taken the opportunity to make a few changes, primarily to make better interconnections and provide cross-references to three complementary works, which share the style and notation of the present volume, but differ in the level of mathematical sophistication:

(a) The new (4th) edition of Vol. I, which appeared in 2017 and is mathematically less demanding than the present volume. It contains a lot of material on approximate DP that complements Chapters 6 and 7 of the present volume. Volume I also makes a connection with recent high profile advances in the field, including the AlphaZero program for Go and chess, and the use of deep reinforcement learning (or approximate dynamic programming using value and/or policy approximation with deep neural networks as an approximation architecture in the terminology of this book).

(b) My research monograph "Abstract Dynamic Programming," which appeared in 2018 and is more mathematical than the present volume. It contains an extensive unifying treatment of the discounted and undiscounted problems of Chapters 1-4, in the spirit of Section 1.6. It also contains a lot of advanced material on the stochastic shortest path and undiscounted problems of Chapters 3 and 4, respectively.

(c) The research monograph "Stochastic Optimal Control: The Discrete-Time Case," coauthored with Steven Shreve, which appeared in 1978, and is a mathematically advanced treatment of the measure-theoretic and other theoretical questions that arise in continuous-spaces stochastic optimal control. Appendix B provides a summary of this work and its connections to the present volume.

With these changes, the two volumes of "Dynamic Programming and Optimal Control," and the two research monographs above form a streamlined continuum, which covers the entire exact dynamic programming field, and the extraordinary progress that has occurred in its theory and applications over the last 50 years, since Bellman laid its foundations in the 1950s and early 60s.

These works, together with the 1996 "Neuro-Dynamic Programming" book, coauthored with John Tsitsiklis, also provide an entry point to the approximate dynamic programming/reinforcement learning field, and a mathematical counterpoint to the artificial intelligence approach towards the subject. This field is the focus of intensive research currently, and will undoubtedly undergo major developments in the coming years. I believe, however, that the principles laid out in the aforementioned books provide a solid foundation for future progress. I also believe that in view of the diversity and complexity of the dynamic programming problems currently being addressed, it is unlikely that a few dominant algorithms will emerge. Instead, a wide variety of techniques and combinations thereof will be needed. In particular, practical experience suggests that it is important to bring to bear the right mix of methodological ingredients into a given problem, and this requires a mathematical as well as an intuitive understanding of the properties of the broad range of available algorithmic approaches.

In addition to the connections and references to my other books, I have added in the 2nd printing notes and sources relating to some of the research progress that has occurred in the six years since the 2012 1st printing, and I have also corrected the typos that have been listed in the on-line errata sheet. Moreover, I replaced Section 4.5 on gambling strategies, which had outlived its usefulness, with a new section (4.1.4) on the relation between positive cost undiscounted problems and stochastic shortest path problems. Aside from these updates and a general polish of the text, the contents of this volume have not changed much.

Finally, let me note that during the period 2012-2018 the book has been supplemented by quite a few on-line extensions and instructional videos on exact, approximate, and abstract dynamic programming. This material together with the two aforementioned 1978 and 2018 research monographs, are freely accessible from my website:

http://web.mit.edu/dimitrib/www/home.html

Dimitri P. Bertsekas
January 2018

1

Discounted Problems – Theory

Contents

In this volume we consider stochastic optimal control problems with an infinite number of decision stages (an infinite horizon). We presented an introduction to these problems in Chapter 5 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.

As noted in Chapter 5 of Vol. I, we focus on 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.

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 4 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 typically cannot result in cost reduction, as we show in Section 1.1.4.

The first two chapters deal with discounted problems, covering the case of bounded cost per stage, but also setting the stage for the analytical and computational methodology to be used in other cases, both discounted and undiscounted. Chapters 3, 4, and 5 consider the other three major problem classes. The final two chapters discuss simulation-based methods that aim to compute approximations to the optimal cost-to-go function by using Monte-Carlo simulation and parametric architectures (such as feature-based architectures or neural networks, which we discussed in Chapter 6 of Vol. I). While this subject has been treated in several specialized books and monographs, the present volume includes a great deal of material that has not yet appeared in book form.

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.

In this chapter, after providing a broad introduction to infinite horizon problems involving minimization of total (discounted and undiscounted) cost, we focus on discounted problems with bounded cost per stage. We develop the basic theory of these problems in Section 1.2. We discuss multiarmed bandit problems (an important special case), and continuous-time variations in Sections 1.3 and 1.4, respectively. We discuss some extensions of the basic theory in Sections 1.5 and 1.6. After Section 1.2, the reader may proceed directly to the development of computational methods in Chapter 2, and return to the other sections of this chapter as necessary later.

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 three. 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

$$x_{k+1} = f(x_k, u_k, w_k), \qquad k = 0, 1, \ldots, \tag{1.1}$$

where for all k, the state x_k is an element of a space X, the control u_k is an element of a space U, and the random disturbance w_k is an element of a space W.† We assume that W is a countable set. The control u_k is constrained to take values in a given nonempty subset $U(x_k)$ of U, which depends on the current state x_k [$u_k \in U(x_k)$, for all $x_k \in X$]. The random disturbances $w_k, k = 0, 1, \ldots$, are characterized by probability distributions $P(\cdot \mid x_k, u_k)$ that are independent of k, where $P(w_k \mid x_k, u_k)$ 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 $w_{k-1}, \ldots, w_0$.

† We consider both problems with infinite state and control spaces [like the system of Eq. (1.1)], and problems with discrete (finite or countable) state space (in which case the underlying system is a Markov chain, like the ones of Chapter 5 of Vol. I). In this chapter, with a few exceptions, we place emphasis on the former case for generality. In the next chapter, we focus primarily on finite-state Markov chain problems (also referred to as *Markovian Decision Problems* or MDP for short), and we introduce compact Markov chain notation that is well-suited for such problems. Generally, to distinguish the infinite and finite state space cases, we denote an element of a continuous state space by x and an element of a discrete state space by i. Our notational system is consistent with the traditional optimal control notation that was established in the 1960s and 1970s.

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

$$J_\pi(x_0) = \lim_{N\to\infty} \mathop{E}_{\substack{w_k \\ k=0,1,\ldots}} \left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\}, \tag{1.2}$$

subject to the system equation constraint (1.1).† The cost per stage $g : X \times U \times W \mapsto \Re$ is given, and α is a positive scalar.

If $\alpha < 1$, the implication is that future costs are discounted, and then α is referred to as the *discount factor*. The other major possibility is $\alpha = 1$, in which case the problem is referred to as *undiscounted*. Such problems are considered in Chapters 3 and 4.‡

† In what follows we will generally impose appropriate assumptions on the cost per stage g and the scalar α, which 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

$$J_\pi(x_0) = \limsup_{N\to\infty} \mathop{E}_{\substack{w_k \\ k=0,1,\ldots}} \left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\}.$$

Note that the expected value of the N-stages cost of π is defined as a (possibly infinite) sum, since the disturbances w_k, $k = 0, 1, \ldots$, take values in a countable set. 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.

The cost $J_\pi(x_0)$ given by Eq. (1.2) represents the limit of expected finite horizon costs, which in all problems that we consider are assumed to be well defined and finite for all policies, in the sense discussed in Section 1.5 of Vol. I. Another possibility would be to minimize over π the expected infinite horizon cost

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

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 4.1) and other stochastic convergence theorems, which allow interchange of limit and expectation under appropriate conditions.

‡ We will occasionally consider a slightly more general form of discounting, where α may depend on the current state and control. The structure of these

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

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

[Note that we continue the convention of Vol. I to use $\min S$ (rather than $\inf S$) to denote the greatest lower bound of a set of numbers S, even if the minimum is not attained by an element of S.]

For a given initial state x, an optimal policy 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 $\pi = \{\mu, \mu, \ldots\}$, 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.

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

1.1.1 The Finite-Horizon DP Algorithm

For any admissible policy $\pi = \{\mu_0, \mu_1, \ldots\}$, 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 $J : X \mapsto \Re$ is some function. The total expected cost is

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

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 in Section 1.3 of Vol. I. This algorithm is given for $k = 1, \ldots, N$, by

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

with the initial condition

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

problems is not much different that the one of the present section. As we will show in Sections 1.5 and 1.6, fundamentally what is important is that the form of discounting used induces a contraction mapping structure in the associated DP equations.

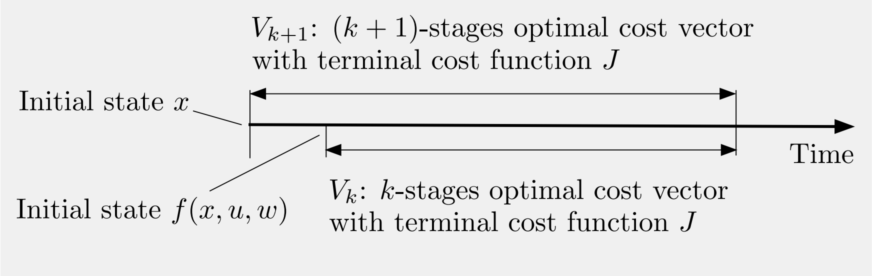

V_{k+1}: $(k+1)$-stages optimal cost vector with terminal cost function J
Initial state x
Time
Initial state $f(x,u,w)$
V_k: k-stages optimal cost vector with terminal cost function J

Figure 1.1.1 Interpretation of the DP recursion (1.4).

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 this 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 the DP recursion (1.3) becomes

$$V_{k+1}(x) = \min_{u\in U(x)} E\big\{g(x,u,w) + \alpha V_k\big(f(x,u,w)\big)\big\}, \tag{1.4}$$

with initial condition

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

Note the intuition here: to solve a $(k+1)$-stage problem, we minimize the sum of the first-stage cost plus the optimal cost of the future k stages, appropriately discounted to the present time by α (cf. Fig. 1.1.1).

The important feature of iteration (1.4) is that it can be used to calculate *all* the optimal finite horizon cost functions using a *single* DP recursion. With each iteration, we obtain the optimal cost function of some finite horizon problem, whose horizon is longer by one stage over the horizon of the preceding problem. 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 $J : X \mapsto \Re$, we consider the function obtained by applying the DP mapping to J, and we denote it by

$$(TJ)(x) = \min_{u \in U(x)} E\left\{ g(x,u,w) + \alpha J\big(f(x,u,w)\big)\right\}, \qquad x \in X, \qquad (1.5)$$

where $E\{\cdot\}$ denotes expected value over w with respect to the distribution $P(w \mid x,u)$.† Since $(TJ)(\cdot)$ is itself a function defined on the state space X, we view T as a mapping that transforms the function J on X into the function TJ on X.‡ 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 $J : X \mapsto \Re$ and any stationary policy μ, we denote

$$(T_\mu J)(x) = E\left\{ g\big(x,\mu(x),w\big) + \alpha J\big(f(x,\mu(x),w)\big)\right\}, \qquad x \in X. \qquad (1.6)$$

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 denote by T^k the composition of the mapping T with itself k times; i.e.,

$$(T^k J)(x) = \big(T(T^{k-1}J)\big)(x), \qquad x \in X, \quad k = 1,2,\ldots,$$

with

$$(T^0 J)(x) = J(x), \qquad x \in X.$$

Thus $T^k J$ is the function obtained by applying the mapping T to the function $T^{k-1}J$. Similarly, $T_\mu^k J$ is defined by

$$(T_\mu^k J)(x) = \big(T_\mu(T_\mu^{k-1}J)\big)(x), \qquad x \in X,$$

with

$$(T_\mu^0 J)(x) = J(x), \qquad x \in X.$$

It can be seen 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* $\alpha^k J$ [see Eq. (1.4); $T^k J$ is equal to V_k as given by this equation]. Similarly, $(T_\mu^k J)(x)$ *is the cost of a stationary policy* μ *for the same problem.*

Finally, consider a k-stage policy $\pi = \{\mu_0, \mu_1, \ldots, \mu_{k-1}\}$. Then, the expression $(T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{k-1}} J)(x)$ is defined sequentially by

$$(T_{\mu_i} T_{\mu_{i+1}} \cdots T_{\mu_{k-1}} J)(x) = \big(T_{\mu_i}(T_{\mu_{i+1}} \cdots T_{\mu_{k-1}} J)\big)(x), \quad i = 0,\ldots,k-2,$$

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* $\alpha^k J$.

† Whenever we use the mapping T, we will impose sufficient assumptions to guarantee that the expected value involved in Eq. (1.5) is well defined.

‡ To simplify notation, we try to avoid parentheses in function notation whenever there is no possibility of confusion, so for example we prefer to use TJ in place of $T(J)$ (which would also be correct), but we use $(TJ)(x)$ rather than TJx.

Example 1.1.1

The preceding abstract shorthand notation greatly simplifies long DP expressions that would be cumbersome to use with a more conventional notation. To illustrate the case where $k = 2$, let us write

$$
\begin{aligned}
(T^2J)(x) &= \min_{u\in U(x)} E\Big\{g(x,u,w) + \alpha(TJ)\big(f(x,u,w)\big)\Big\} \\
&= \min_{u_0\in U(x)} \mathop{E}_{w_0}\Bigg\{g(x,u_0,w_0) + \alpha \min_{u_1\in U(f(x,u_0,w_0))} \mathop{E}_{w_1}\bigg\{g\big(f(x,u_0,w_0),u_1,w_1\big) \\
&\qquad\qquad + \alpha J\Big(f\big(f(x,u_0,w_0),u_1,w_1\big)\Big)\bigg\}\Bigg\} \\
&= \min_{u_0\in U(x)} \mathop{E}_{w_0}\Bigg\{g(x,u_0,w_0) + \min_{u_1\in U(f(x,u_0,w_0))} \mathop{E}_{w_1}\bigg\{\alpha g\big(f(x,u_0,w_0),u_1,w_1\big) \\
&\qquad\qquad + \alpha^2 J\Big(f\big(f(x,u_0,w_0),u_1,w_1\big)\Big)\bigg\}\Bigg\}.
\end{aligned}
$$

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 $\alpha^2 J$.

Consider also the calculation of $(T_{\mu_0}T_{\mu_1}J)(x)$. We have

$$
\begin{aligned}
(T_{\mu_0}T_{\mu_1}J)(x) &= E\Big\{g\big(x,\mu_0(x),w\big) + \alpha(T_{\mu_1}J)\big(f\big(x,\mu_0(x),w\big)\big)\Big\} \\
&= \mathop{E}_{w_0}\Bigg\{g\big(x,\mu_0(x),w_0\big) + \alpha \mathop{E}_{w_1}\bigg\{g\big(f\big(x,\mu_0(x),w_0\big),\mu_1\big((x,\mu_0(x),w_0)\big),w_1\big) \\
&\qquad\qquad + \alpha J\bigg(f\Big(f(x,\mu_0(x),w_0),u_1,w_1\Big)\bigg)\bigg\}\Bigg\} \\
&= \mathop{E}_{w_0}\Bigg\{g(x,\mu_0(x),w_0) + \mathop{E}_{w_1}\bigg\{\alpha g\big(f(x,\mu_0(x),w_0),\mu_1\big(f(x,\mu_0(x),w_0)\big),w_1\big) \\
&\qquad\qquad + \alpha^2 J\bigg(f\Big(f\big(x,\mu_0(x),w_0\big),\mu_1\big(f(x,\mu_0(x),w_0)\big),w_1\Big)\bigg)\bigg\}\Bigg\}.
\end{aligned}
$$

Again this expression can be recognized as the cost of the 2-stage policy $\{\mu_0,\mu_1\}$ from initial state x and with terminal cost function $\alpha^2 J$.

The following monotonicity property plays a fundamental role in the subsequent developments.

Lemma 1.1.1: (Monotonicity Lemma) For any $J : X \mapsto \Re$ and $J' : X \mapsto \Re$, such that for all $x \in X$, $J(x) \leq J'(x)$, and any stationary policy $\mu : X \mapsto U$, we have

$$(T^k J)(x) \leq (T^k J')(x), \quad (T_\mu^k J)(x) \leq (T_\mu^k J')(x), \quad x \in X,\ k = 1, 2, \ldots.$$

In particular, if $J : X \mapsto \Re$ is such that for all $x \in X$, $J(x) \leq (TJ)(x)$,

$$(T^k J)(x) \leq (T^{k+1} J)(x), \qquad x \in X,\ k = 1, 2, \ldots.$$

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. (We may prove the lemma by using a straightforward induction argument.) **Q.E.D.**

For any two functions $J : X \mapsto \Re$ and $J' : X \mapsto \Re$, we write

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

With this notation, Lemma 1.1.1 is stated as

$$J \leq J' \qquad \Rightarrow \qquad T^k J \leq T^k J', \quad T_\mu^k J \leq T_\mu^k J', \qquad k = 1, 2, \ldots,$$

$$J \leq TJ \qquad \Rightarrow \qquad T^k J \leq T^{k+1} J, \qquad k = 1, 2, \ldots.$$

Let us also denote by $e : X \mapsto \Re$ the unit function that takes the value 1 identically on X:

$$e(x) \equiv 1, \qquad x \in X.$$

We have from the definitions (1.5) and (1.6) of T and T_μ, for any function $J : X \mapsto \Re$ and scalar r

$$\big(T(J + re)\big)(x) = (TJ)(x) + \alpha r, \qquad x \in X,$$

$$\big(T_\mu(J + re)\big)(x) = (T_\mu J)(x) + \alpha r, \qquad x \in X.$$

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

Lemma 1.1.2: (Constant Shift Lemma) For every k, function $J : X \mapsto \Re$, stationary policy μ, scalar r, and $x \in X$,

$$\big(T^k(J + re)\big)(x) = (T^k J)(x) + \alpha^k r,$$

$$\big(T_\mu^k(J + re)\big)(x) = (T_\mu^k J)(x) + \alpha^k r.$$

We introduce a final shorthand notation relating T and T_μ. Let us denote by $\mathcal{M}$ the set of all admissible stationary policies. Then, by viewing $\mathcal{M}$ as the Cartesian product $\Pi_{x\in X}U(x)$, we have for every $J : X \mapsto \Re$

$$(TJ)(x) = \min_{\mu\in\mathcal{M}} (T_\mu J)(x), \qquad x \in X,$$

or more compactly

$$TJ = \min_{\mu\in\mathcal{M}} (T_\mu J),$$

where the minimum is understood to be separate for each component of $T_\mu J$.

1.1.3 A Preview of Infinite Horizon Results

Let us speculate on the type of results that we will be aiming for, based also on the analysis of Chapter 5 of Vol. I.

(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 $k \to \infty$, it is reasonable to speculate that the optimal infinite horizon cost is equal to the limit of the optimal k-stage costs; i.e.,

$$J^*(x) = \lim_{k\to\infty} (T^k J_0)(x), \qquad x \in X.$$

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 $\alpha < 1$, the convergence property

$$J^*(x) = \lim_{k\to\infty} (T^k J)(x), \qquad x \in X,$$

holds *regardless of the choice of J.*

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

$$(T^{k+1}J_0)(x) = \min_{u\in U(x)} \mathop{E}_{w} \left\{ g(x,u,w) + \alpha (T^k J_0)\big(f(x,u,w)\big) \right\},$$

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

$$J^*(x) = \min_{u\in U(x)} \mathop{E}_{w} \big\{ g(x,u,w) + \alpha J^*\big(f(x,u,w)\big) \big\}, \qquad x \in X,$$

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 $k \to \infty$, 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 3, the preceding conjectures are correct. For problems with unbounded costs per stage and for stochastic shortest path problems where our assumptions of Chapter 3 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

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

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* $\pi = \{\mu_0, \mu_1, \ldots\}$, where μ_k is a function that maps a history h_k into a probability distribution $\mu_k(u_k \mid h_k)$ 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

$\mu_k(u_k \mid h_k)$ 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* $\pi = \{\mu_0, \mu_1, \ldots\}$, where μ_k is a function that maps the state x_k into a probability distribution $\mu_k(u_k \mid x_k)$ 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 $\pi = \{\mu_0, \mu_1, \ldots\}$ 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 $\overline{\pi} = \{\overline{\mu}_0, \overline{\mu}_1, \ldots\}$, where $\overline{\mu}_k$ is defined for all x_k with $\xi_k(x_k) > 0$ by

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

Let $\overline{\xi}_k(x_k)$ and $\overline{\zeta}_k(x_k, u_k)$ 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

$$\xi_k(x_k) = \overline{\xi}_k(x_k), \qquad \zeta_k(x_k, u_k) = \overline{\zeta}_k(x_k, u_k). \tag{1.7}$$

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 $\overline{\xi}_0(x_0)$ are both equal to the distribution of the initial state, while

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

Suppose that Eq. (1.7) holds for some k. Then, we have

$$\overline{\xi}_{k+1}(x_{k+1}) = \sum_{x_k, u_k} \overline{\zeta}_k(x_k, u_k)\, p_{x_{k+1}x_k}(u_k)$$

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

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,

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

thereby completing the induction. Thus π and $\overline{\pi}$ 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 $N \to \infty$. In particular this is true for the finite state and control spaces versions of the total cost problems that we will discuss: discounted problems with bounded cost per stage (the present chapter), problems with nonnegative cost per stage (Chapter 4), and problems with nonpositive cost per stage (Chapter 4).

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 discounted cost problems with bounded cost per stage of the present chapter, and the finite state and control spaces models of Chapters 3 and 5; 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 4, and also in some models not considered in this book, such as constrained DP problems where policies are required to satisfy additional constraint inequalities (see e.g., FeS96], [FeS02]). The monograph [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 - BOUNDED COST PER STAGE

In this section we develop the theory of the most well-behaved type of infinite horizon problem, characterized by the following assumption.

Assumption D (Discounted Cost – Bounded Cost per Stage): The cost per stage g satisfies for all $(x,u) \in X \times U$,

$$\big|E\big\{g(x,u,w)\big\}\big| \leq M,$$

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

The preceding boundedness assumption is not as restrictive as might appear. It holds when the spaces X, U, and W either are finite sets, or they are approximated by finite sets for the purposes of computation. Also, it is often possible to reformulate the problem so that X, U, and W are bounded subsets of Euclidean spaces, 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 the preceding Assumption D, which implies that the "tail" of the cost after stage N,

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

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

Proposition 1.2.1: (Convergence of the DP Algorithm) For any bounded function $J : X \mapsto \Re$, we have for all $x \in X$,

$$J^*(x) = \lim_{N\to\infty} (T^N J)(x).$$

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

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

Since by Assumption D, $\big|g\big(x_k, \mu_k(x_k), w_k\big)\big| \le M$, we obtain

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

Using the above relations, it follows that

$$\begin{aligned} J_\pi(x_0) - \frac{\alpha^N M}{1-\alpha} - \alpha^N \max_{x\in X}\big|J(x)\big| & \\ \le E\left\{\alpha^N J(x_N) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} & \\ \le J_\pi(x_0) + \frac{\alpha^N M}{1-\alpha} + \alpha^N \max_{x\in X}\big|J(x)\big|. & \end{aligned}$$

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

$$\begin{aligned} J^*(x_0) - \frac{\alpha^N M}{1-\alpha} - \alpha^N \max_{x\in X}\big|J(x)\big| & \\ \le (T^N J)(x_0) & \\ \le J^*(x_0) + \frac{\alpha^N M}{1-\alpha} + \alpha^N \max_{x\in X}\big|J(x)\big|, & \end{aligned} \tag{1.8}$$

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

Note that based on the preceding proposition, the DP algorithm may be used to compute an approximation to J^*. This computational method, called *value iteration* (cf. Chapter 5 of Vol. I), together with some additional methods will be examined in the next chapter.

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 $\tilde{U}(x) = \big\{\mu(x)\big\}$ instead of $U(x)$. Proposition 1.2.1 applies to this modified problem and yields the following:

Proposition 1.2.2: For every stationary policy μ, the associated cost function satisfies for all $x \in X$,

$$J_\mu(x) = \lim_{N\to\infty} (T_\mu^N J)(x).$$

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

Proposition 1.2.3: (Bellman's Equation) The optimal cost function J^* satisfies for all $x \in X$,

$$J^*(x) = \min_{u\in U(x)} \mathop{E}_{w}\Big\{g(x,u,w) + \alpha J^*\big(f(x,u,w)\big)\Big\}, \tag{1.9}$$

or, equivalently,

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

Furthermore, J^* is the unique solution of this equation within the class of bounded functions. Moreover, for any bounded function J with $J \geq TJ$ (or $J \leq TJ$), we have $J \geq J^*$ (or $J \leq J^*$, respectively).

Proof: From Eq. (1.8), we have for all $x \in X$ 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 $x \in X$]. Applying the mapping T to this relation and using the Monotonicity and Constant Shift Lemmas 1.1.1 and 1.1.2, we obtain for all $x \in X$ 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 $N \to \infty$ in the preceding relation and using the fact

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

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

To show uniqueness, note that if J is bounded and satisfies $J = TJ$, then $J = \lim_{N\to\infty} T^N J$, so by Prop. 1.2.1, we have $J = J^*$. Finally, for any bounded J with $J \geq TJ$, using the Monotonicity Lemma 1.1.1, we have for all k,

$$J \geq TJ \geq \cdots \geq T^k J \geq T^{k+1} J \geq \cdots$$

and by taking limit as $k \to \infty$ and by using Prop. 1.2.1, we obtain $J \geq \lim_{k\to\infty} T^k J = J^*$. The proof for the case $J \leq TJ$ is similar. **Q.E.D.**

Based on the same reasoning we used to obtain Prop. 1.2.2 from Prop. 1.2.1, we have the following.

Proposition 1.2.4: For every stationary policy μ, the associated cost function satisfies for all $x \in X$,

$$J_\mu(x) = \mathop{E}_{w}\Big\{g\big(x, \mu(x), w\big) + \alpha J_\mu\big(f(x, \mu(x), w)\big)\Big\},$$

or, equivalently,

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

Furthermore, J_μ is the unique solution of this equation within the class of bounded functions. Moreover, for any bounded function J with $J \geq T_\mu J$ (or $J \leq T_\mu J$), we have $J \geq J_\mu$ (or $J \leq J_\mu$, respectively).

The next proposition characterizes stationary optimal policies.

Proposition 1.2.5: (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 (1.9) for each $x \in X$; 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 Prop. 1.2.4, 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 Prop. 1.2.4, 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. 1.2.5 implies the existence of an optimal stationary policy when the minimum in the right-hand side of Bellman's equation is attained for all $x \in X$. In particular, when $U(x)$ is finite for each $x \in X$, an optimal stationary policy is guaranteed to exist.

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

$$\max_{x\in X}\big|(T^kJ)(x) - J^*(x)\big| \le \alpha^k \max_{x\in X}\big|J(x) - J^*(x)\big|, \qquad k = 0, 1, \ldots$$

This relation is obtained by using the fact $T^kJ^* = J^*$ (which follows from Bellman's equation) and the following proposition, which is a fundamental contraction property of T that we will revisit in Section 1.5.

Proposition 1.2.6: (Convergence Rate) For any two bounded functions $J : X \mapsto \Re$, $J' : X \mapsto \Re$, and for all $k = 0, 1, \ldots$, there holds

$$\max_{x\in X}\big|(T^kJ)(x) - (T^kJ')(x)\big| \le \alpha^k \max_{x\in X}\big|J(x) - J'(x)\big|.$$

Proof: Denote

$$c = \max_{x\in X}\big|J(x) - J'(x)\big|,$$

so that for all $x \in X$,

$$J(x) - c \le J'(x) \le J(x) + c.$$

Applying T^k in this relation and using the Monotonicity and Constant Shift Lemmas 1.1.1 and 1.1.2, we obtain for all $x \in X$,

$$(T^kJ)(x) - \alpha^k c \le (T^kJ')(x) \le (T^kJ)(x) + \alpha^k c.$$

It follows that for all $x \in X$,

$$\big|(T^kJ)(x) - (T^kJ')(x)\big| \le \alpha^k c,$$

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

As earlier, by specializing Prop. 1.2.6 we obtain the following.

Proposition 1.2.7: For any two bounded functions $J : X \mapsto \Re$, $J' : X \mapsto \Re$, and any stationary policy μ, we have

$$\max_{x\in X}\big|(T_\mu^kJ)(x) - (T_\mu^kJ')(x)\big| \le \alpha^k \max_{x\in X}\big|J(x) - J'(x)\big|, \qquad k = 0, 1, \ldots$$

Markov Chain Notation

Let us now describe the preceding results with a different notation, for the case where the state space X is finite or countable. Then, similar to Chapter 5 in Vol. I, states may be denoted by $i = 1, 2, \dots,$ and the system may be described in terms of transition probabilities $p_{ij}(u)$:

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

These 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

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

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

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

The mappings T and T_μ are written in terms of transition probabilities as

$$(TJ)(i) = \min_{u \in U(i)} \sum_{j \in X} p_{ij}(u)\big(g(i, u, j) + \alpha J(j)\big), \qquad i \in X,$$

$$(T_\mu J)(i) = \sum_{j \in X} p_{ij}\big(\mu(i)\big)\big(g(i, \mu(i), j) + \alpha J(j)\big), \qquad i \in X,$$

and the results of this section can be translated in the new notation. For example, Bellman's equation takes the form

$$J^*(i) = \min_{u \in U(i)} \sum_{j \in X} p_{ij}(u)\big(g(i, u, j) + \alpha J^*(j)\big), \qquad i \in X.$$

The following example illustrates, among others, this notation.

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 $1, \dots, n$. 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. 1.2.3) takes the form

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

By Prop. 1.2.5, 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

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

Based on the convergence of the DP algorithm (cf. Prop. 1.2.1), we can 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

$$J_0(i) = 0,$$

$$(TJ_0)(i) = \min\big[R + g(1), g(i)\big],$$

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

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

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

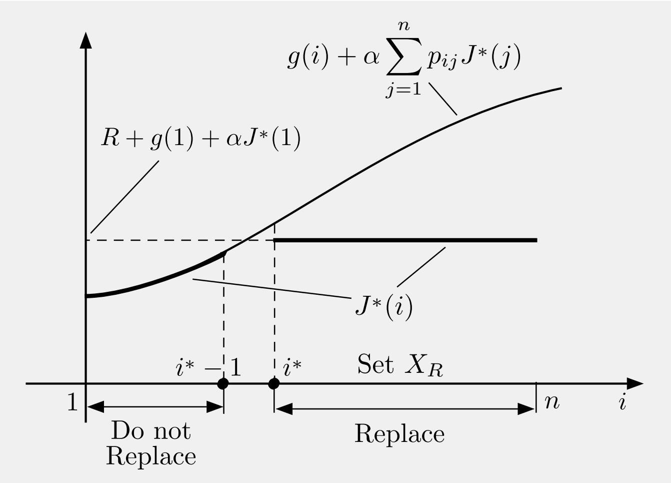
for all functions $J(i)$, which are monotonically nondecreasing in i. This assumption is satisfied if

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

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

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

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



$g(i) + \alpha \sum_{j=1}^{n} p_{ij} J^*(j)$
$R + g(1) + \alpha J^*(1)$
$J^*(i)$
$i^* - 1$
i^*
Set X_R
1
n
i
Do not Replace
Replace

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

$(TJ_0)(i)$ 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

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

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

$$X_R = \left\{ i \;\middle|\; R + g(1) + \alpha J^*(1) \le g(i) + \alpha \sum_{j=1}^{n} p_{ij} J^*(j) \right\},$$

and let

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

Then, an optimal policy takes the form

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

as shown in Fig. 1.2.1.

In the next two sections we illustrate the theory of the present section with two interesting classes of problems: multiarmed bandit problems in Section 1.3, and continuous-time semi-Markov problems in Section 1.4. The material of these sections will not be used later and the reader may skip them with no loss of continuity. Then, in Sections 1.5 and 1.6, we will provide some extensions of the basic theory, which will be used sparingly in Chapters 2-5 (mostly for Sections 2.5 and 2.6). Thus the reader may proceed directly to the development of computational methods in Chapter 2 at this point, and return to this chapter when needed later.

1.3 SCHEDULING AND MULTIARMED BANDIT PROBLEMS

In this section we will discuss an important class of discounted cost problems with bounded cost per stage. There are n projects (or activities) of which only one can be worked on at any time period. Each project ℓ is characterized at time k by its state x_k^ℓ. If project ℓ is worked on at time k, one receives an expected reward $\alpha^k R^\ell(x_k^\ell)$, where $\alpha \in (0,1)$ is a discount factor; the state x_k^ℓ then evolves according to the equation

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

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

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

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

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 analytically. 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 ℓ, there is a function $m^\ell(x^\ell)$ such that an optimal policy at time k is to

$$\begin{array}{lll} \text{retire} & \text{if} & M > \max\limits_{\bar{\ell}}\bigl\{m^{\bar{\ell}}(x_k^{\bar{\ell}})\bigr\}, \\ \text{work on project } \ell & \text{if} & m^\ell(x_k^\ell) = \max\limits_{\bar{\ell}}\bigl\{m^{\bar{\ell}}(x_k^{\bar{\ell}})\bigr\} \geq M. \end{array} \tag{1.11}$$

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

The problem 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 $x = (x^1, \ldots, x^n)$ 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

$$J(x, M) = \max\left[M,\ \max_\ell L^\ell(x, M, J)\right], \tag{1.12}$$

where L^ℓ is defined by

$$L^\ell(x, M, J) = R^\ell(x^\ell) + \alpha \mathop{E}_{w^\ell}\bigl\{J\bigl(x^1, \ldots, x^{\ell-1}, f^\ell(x^\ell, w^\ell), x^{\ell+1}, \ldots, x^n, M\bigr)\bigr\}. \tag{1.13}$$

The next proposition gives some useful properties of J.

Proposition 1.3.1: Let $B = \max_\ell \max_{x^\ell}\bigl|R^\ell(x^\ell)\bigr|$. 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 a DP iteration starting with the function

$$J_0(x, M) = \max[0, M].$$

It has the form

$$J_{k+1}(x, M) = \max\left[M, \max_\ell L^\ell(x, M, J_k)\right], \qquad k = 0, 1, \ldots, \qquad (1.14)$$

and from Prop. 1.2.1, we know that for all x and M,

$$\lim_{k\to\infty} J_k(x, M) = J(x, M).$$

We show inductively that $J_k(x, M)$ has the properties (a)-(c) stated in the proposition and by taking the limit as $k \to \infty$, we establish the same properties for J. Indeed, clearly $J_0(x, M)$ satisfies properties (a)-(c). Assume that $J_k(x, M)$ satisfies (a)-(c). Then from Eqs. (1.12) and (1.14), $J_{k+1}(x, M)$ is convex and monotonically nondecreasing in M, since the expectation and maximization operations preserve these properties. Hence property (a) follows. Verification of (b) and (c) is similarly 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 ℓ. The optimal reward function for this problem is denoted $J^\ell(x^\ell, M)$ and has the properties indicated in Prop. 1.3.1. A typical form for $J^\ell(x^\ell, M)$, viewed as a function of M for fixed x^ℓ, is shown in Fig. 1.3.1. Clearly, there is a minimal value $m^\ell(x^\ell)$ of M for which $J^\ell(x^\ell, M) = M$; i.e., for all x^ℓ,

$$m^\ell(x^\ell) = \min\bigl\{M \mid J^\ell(x^\ell, M) = M\bigr\}. \qquad (1.15)$$

The function $m^\ell(x^\ell)$ is called the *index function* (or simply index) of project ℓ. It provides an indifference threshold at each state; i.e., $m^\ell(x^\ell)$ is the retirement reward for which we are indifferent between retiring and operating the project when at state x^ℓ.

Our objective is to show the optimality of the index rule (1.11) for the index function defined by Eq. (1.15).

Project-by-Project Retirement Policies

Consider first a problem with a single project, say project ℓ, and a fixed retirement reward M. Then by the definition (1.15) of the index, an optimal policy is to

$$\begin{aligned} \text{retire project } \ell \text{ if} \quad & m^\ell(x^\ell) < M, \\ \text{work on project } \ell \text{ if} \quad & m^\ell(x^\ell) \geq M. \end{aligned} \qquad (1.16)$$

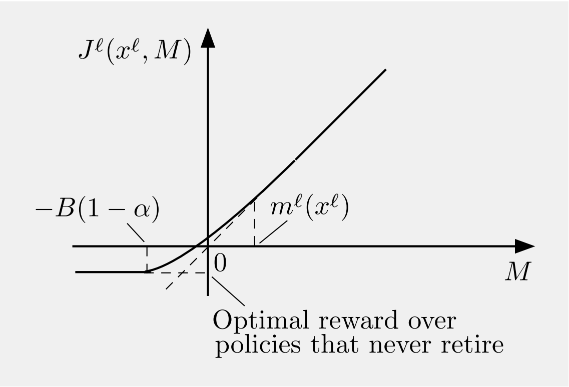


Figure 1.3.1 Form of the ℓth project reward function $J^\ell(x^\ell, M)$ for fixed x^ℓ and definition of the index $m^\ell(x^\ell)$.

In other words, the project is operated continuously up to the time that its state falls into the *retirement set*

$$X^\ell = \big\{x^\ell \mid m^\ell(x^\ell) < M\big\}. \tag{1.17}$$

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 $x = (x^1, \ldots, x^n)$. Let us ask two questions:

1. Does it make sense to retire (from all projects) when there is still a project ℓ with state x^ℓ such that $m^\ell(x^\ell) > M$? The answer is negative. Retiring when $m^\ell(x^\ell) > M$ cannot be optimal, since if we operate project ℓ exclusively up to the time that its state x^ℓ falls within the retirement set X^ℓ of Eq. (1.17) and then retire, we will gain a higher expected reward. [This follows from the definition (1.15) of the index and the nature of the optimal policy (1.16) for the single-project problem.]

2. Does it ever make sense to work on a project ℓ with state in the retirement set X^ℓ of Eq. (1.17)? 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 $x = (x^1, \ldots, x^n)$,

$$\begin{array}{rll} \text{permanently retires project } \ell & \text{if} & x^\ell \in X^\ell, \\ \text{works on some project} & \text{if} & x^j \notin X^j \text{ for some } j, \end{array} \qquad (1.18)$$

where X^ℓ is the ℓth project retirement set of Eq. (1.17). 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 1.3.2: There exists an optimal PPR policy.

Proof: In view of Eqs. (1.12), and (1.18), existence of a PPR policy is equivalent to having, for all ℓ,

$$\max\left[M, \max_{\bar{\ell}\neq\ell} L^{\bar{\ell}}(x, M, J)\right] \geq L^\ell(x, M, J), \qquad \text{for all } x \text{ with } x^\ell \in X^\ell, \qquad (1.19)$$

$$M \leq L^\ell(x, M, J), \qquad \text{for all } x \text{ with } x^\ell \notin X^\ell, \qquad (1.20)$$

where L^ℓ is given by

$$L^\ell(x, M, J) = R^\ell(x^\ell) + \alpha \mathop{E}_{w^\ell}\Big\{J\big(x^1, \ldots, x^{\ell-1}, f^\ell(x^\ell, w^\ell), x^{\ell+1}, \ldots, x^n, M\big)\Big\}, \qquad (1.21)$$

and $J(x, M)$ is the optimal reward function corresponding to x and M.

The ℓth single-project optimal reward function J^ℓ clearly satisfies, for all x^ℓ,

$$J^\ell(x^\ell, M) \leq J(x^1, \ldots, x^{\ell-1}, x^\ell, x^{\ell+1}, \ldots, x^n, M), \qquad (1.22)$$

since having the option of working at projects other than ℓ cannot decrease the optimal reward. Furthermore, from the definition of the retirement set X^ℓ [cf. Eq. (1.17)],

$$x^\ell \notin X^\ell, \qquad \text{if } M \leq R^\ell(x^\ell) + \alpha \mathop{E}_{w^\ell}\Big\{J^\ell\big(f^\ell(x^\ell, w^\ell), M\big)\Big\}. \qquad (1.23)$$

Using Eqs. (1.21)-(1.23), we obtain Eq. (1.20).

It will suffice to show Eq. (1.19) for $\ell = 1$. Denote:

$\underline{x} = (x^2, \ldots, x^n)$: The state of all projects other than project 1.

$\underline{J}(\underline{x}, M)$: 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 $x = (x^1, \underline{x})$:

$$\underline{J}(\underline{x}, M) \le J(x^1, \underline{x}, M) \le \underline{J}(\underline{x}, M) + \big(J^1(x^1, M) - M\big). \tag{1.24}$$

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. (1.19) for $\ell = 1$, it will suffice to show Eq. (1.24). Indeed, when $x^1 \in X^1$, then $J^1(x^1, M) = M$, so from Eq. (1.24) we obtain $J(x^1, \underline{x}, M) = \underline{J}(\underline{x}, M)$, which is in turn equivalent to Eq. (1.19) for $\ell = 1$.

We now turn to the proof of Eq. (1.24). Its left side is evident. To show the right side, we proceed by induction on the DP recursions

$$\begin{aligned} J_{k+1}(x^1, \underline{x}) = \max \Big[& M, R^1(x^1) + \alpha E\big\{J_k\big(f^1(x^1, w^1), x\big)\big\}, \\ & \max_{\ell \ne 1} \big[R^\ell(x^\ell) + \alpha E\big\{J_k\big(x^1, F^\ell(\underline{x}, w^\ell)\big)\big\}\big]\Big], \end{aligned} \tag{1.25}$$

$$\underline{J}_{k+1}(\underline{x}) = \max \left[M, \max_{\ell \ne 1} \big[R^\ell(x^\ell) + \alpha E\big\{\underline{J}_k\big(F^\ell(\underline{x}, w^\ell)\big)\big\}\big] \right], \tag{1.26}$$

$$J^1_{k+1}(x^1) = \max \big[M, R^1(x^1) + \alpha E\big\{J^1_k\big(f^1(x^1, w^1)\big)\big\}\big], \tag{1.27}$$

where, for all $\ell \ne 1$ and $\underline{x} = (x^2, \ldots, x^n)$,

$$F^\ell(\underline{x}, w^\ell) = \big(x^2, \ldots, x^{\ell-1}, f^\ell(x^\ell, w^\ell), x^{\ell+1}, \ldots, x^n\big).$$

The initial conditions for the recursions (1.25)-(1.27) are

$$J_0(x^1, \underline{x}) = M, \qquad \text{for all } (x^1, \underline{x}), \tag{1.28}$$

$$\underline{J}_0(\underline{x}) = M, \qquad \text{for all } \underline{x}, \tag{1.29}$$

$$J^1_0(x^1) = M, \qquad \text{for all } x^1. \tag{1.30}$$

We know that $J_k(x^1, \underline{x}) \to J(x^1, \underline{x}, M)$, $\underline{J}_k(\underline{x}) \to \underline{J}(\underline{x}, M)$, and $J^1_k(x^1) \to J^1(x^1, M)$, so to show Eq. (1.24) it will suffice to show that for all k and $x = (x^1, \underline{x})$ we have

$$J_k(x^1, \underline{x}) \le \underline{J}_k(\underline{x}) + \big(J^1_k(x^1) - M\big). \tag{1.31}$$

In view of the definitions (1.28)-(1.30), we see that Eq. (1.31) holds for $k = 0$. Assume that it holds for some k. We will show that it holds for $k+1$. From Eqs. (1.25)-(1.27) and the induction hypothesis (1.31), we have

$$\begin{aligned} J_{k+1}(x^1, \underline{x}) \le \max \Big[& M,\ R^1(x^1) + \alpha E\big\{\underline{J}_k(\underline{x}) + J^1_k\big(f^1(x^1, w^1)\big) - M\big\}, \\ & \max_{\ell \ne 1} \big[R^\ell(x^\ell) + \alpha E\big\{\underline{J}_k\big(F^\ell(\underline{x}, w^\ell)\big) + J^1_k(x^1) - M\big\}\big]\Big]. \end{aligned}$$

Using the facts $\underline{J}_k(\underline{x}) \geq M$ and $J_k^1(x^1) \geq M$ [cf. Eqs. (1.25)-(1.27)], and the preceding equation, we see that

$$J_{k+1}(x^1, \underline{x}) \leq \max[\beta_1, \beta_2],$$

where

$$\beta_1 = \max\Big[M,\, R^1(x^1) + \alpha E\big\{J_k^1\big(f^1(x^1, w^1)\big)\big\}\Big] + \alpha\big(\underline{J}_k(\underline{x}) - M\big),$$

$$\beta_2 = \max\Bigg[M,\, \max_{\ell \neq 1}\big[R^\ell(x^\ell) + \alpha E\big\{\underline{J}_k\big(F^\ell(\underline{x}, w^\ell)\big)\big\}\big]\Bigg] + \alpha(J_k^1(x^1) - M).$$

Using Eqs. (1.26), (1.27), and the preceding equations, we see that

$$J_{k+1}(x^1, \underline{x}) \leq \max\big[J_{k+1}^1(x^1) + \underline{J}_k(\underline{x}) - M,\, \underline{J}_{k+1}(\underline{x}) + J_k^1(x^1) - M\big]. \quad (1.32)$$

It can be seen from Eqs. (1.25)-(1.27) and (1.28)-(1.30) 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. (1.32) we obtain that Eq. (1.31) 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.

Proposition 1.3.3: 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

$$\frac{\partial J(x, M)}{\partial M} = E\big\{\alpha^{K_M} \mid x_0 = x\big\}.$$

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

$$\frac{J(x,M) - J(x,M-\epsilon)}{\epsilon} \le E\{\alpha^{K_M}\} \le \frac{J(x,M+\epsilon) - J(x,M)}{\epsilon}.$$

The result follows by taking $\epsilon \to 0$. **Q.E.D.**

Note that the convexity of $J(x,\cdot)$ with respect to M (Prop. 1.3.1) implies that the derivative $\partial J(x,M)/\partial M$ exists almost everywhere with respect to Lebesgue measure. 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_ℓ be the retirement time of project ℓ if it were the only project available, let T be the retirement time for the multiproject problem. Both T_ℓ 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 $T_1, \ldots, T_n$ are independent random variables. Therefore,

$$E\{\alpha^T\} = E\left\{\alpha^{T_1+\cdots+T_n}\right\} = \prod_{\ell=1}^{n} E\{\alpha^{T_\ell}\}.$$

Using Prop. 1.3.3, we obtain

$$\frac{\partial J(x,M)}{\partial M} = \prod_{\ell=1}^{n} \frac{\partial J^\ell(x^\ell,M)}{\partial M}. \tag{1.33}$$

Optimality of the Index Rule

We are now ready to show our main result.

Proposition 1.3.4: The index rule (1.11) is an optimal stationary policy.

Proof: Fix $x = (x^1, \ldots, x^n)$, denote

$$m(x) = \max_{\bar\ell}\bigl\{m^{\bar\ell}(x^{\bar\ell})\bigr\},$$

and let ℓ attain the maximum above, i.e.,

$$m^\ell(x^\ell) = \max_{\bar\ell}\bigl\{m^{\bar\ell}(x^{\bar\ell})\bigr\}.$$

If $m(x) < M$ the optimality of the index rule (1.11) at state x follows from the existence of an optimal PPR policy. If $m(x) \geq M$, we note that

$$J^\ell(x^\ell, M) = R^\ell(x^\ell) + \alpha E\big\{J^\ell\big(f^\ell(x^\ell, w^\ell), M\big)\big\}$$

and then use this relation together with Eq. (1.33) to write

$$\begin{aligned}
\frac{\partial J(x,M)}{\partial M} &= \frac{\partial J^\ell(x^\ell, M)}{\partial M} \cdot \prod_{j\neq\ell} \frac{\partial J^j(x^j, M)}{\partial M} \\
&= \alpha \frac{\partial}{\partial M} E\left\{ J^\ell\big(f^\ell(x^\ell, w^\ell), M\big) \cdot \prod_{j\neq\ell} \frac{\partial J^j(x^j, M)}{\partial M} \right\} \\
&= \alpha E\left\{ \frac{\partial}{\partial M} J^\ell\big(f^\ell(x^\ell, w^\ell), M\big) \cdot \prod_{j\neq\ell} \frac{\partial J^j(x^j, M)}{\partial M} \right\} \\
&= \alpha E\left\{ \frac{\partial}{\partial M} J\big(x^1, \ldots, x^{\ell-1}, f^\ell(x^\ell, w^\ell), x^{\ell+1}, \ldots, x^n, M\big) \right\} \\
&= \alpha \frac{\partial}{\partial M} E\big\{ J\big(x^1, \ldots, x^{\ell-1}, f^\ell(x^\ell, w^\ell), x^{\ell+1}, \ldots, x^n, M\big)\big\},
\end{aligned}$$

and finally

$$\frac{\partial J(x,M)}{\partial M} = \frac{\partial}{\partial M} L^\ell(x, M, J),$$

where

$$L^\ell(x, M, J) = R^\ell(x^\ell) + \alpha E\big\{J\big(x^1, \ldots, x^{\ell-1}, f^\ell(x^\ell, w^\ell), x^{\ell+1}, \ldots, x^n, M\big)\big\}.$$

(The interchange of differentiation and expectation can be justified for almost all M; see [Ber73a].) By the existence of an optimal PPR policy, we also have

$$J\big(x, m(x)\big) = L^\ell\big(x, m(x), J\big).$$

Therefore, the convex functions $J(x,M)$ and $L^\ell(x,M,J)$ 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

$$J(x, M) = L^\ell(x, M, J).$$

This implies that the index rule (1.11) 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 (1.11), 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 ℓ, x^ℓ, and w^ℓ that can occur with positive probability, we have either

$$m^\ell(x^\ell) \leq m^\ell\big(f^\ell(x^\ell, w^\ell)\big) \tag{1.34}$$

or

$$m^\ell(x^\ell) \geq m^\ell\big(f^\ell(x^\ell, w^\ell)\big). \tag{1.35}$$

Under Eq. (1.34) [or Eq. (1.35)] 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. (1.35) implies that if retirement is optimal when at state x^ℓ then it is also optimal at each state $f^\ell(x^\ell, w^\ell)$. Therefore, for all x^ℓ such that $M = m^\ell(x^\ell)$ we have, for all w^ℓ,

$$J^\ell(x^\ell, M) = M, \qquad J^\ell\big(f^\ell(x^\ell, w^\ell), M\big) = M.$$

From Bellman's equation

$$J^\ell(x^\ell, M) = \max\Big[M,\ R^\ell(x^\ell) + \alpha E\Big\{J^\ell\big(f^\ell(x^\ell, w^\ell), M\big)\Big\}\Big]$$

we obtain

$$m^\ell(x^\ell) = R^\ell(x^\ell) + \alpha m^\ell(x^\ell)$$

or

$$m^\ell(x^\ell) = \frac{R^\ell(x^\ell)}{1-\alpha}.$$

Thus the optimal policy in the deteriorating case is

retire if $M > \max_\ell \frac{R^\ell(x^\ell)}{1-\alpha}$

engage the project ℓ with maximal one-step reward $R^\ell(x^\ell)$ otherwise.

The following is an example of the deteriorating case.

Example 1.3.1 (Treasure Hunting)

Consider a search problem involving N sites. Each site ℓ may contain a treasure with expected value v_ℓ. A search at site ℓ costs $c_\ell > 0$ and reveals the treasure with probability β_ℓ (assuming a treasure is there). Let P_ℓ be the probability that there is a treasure at site ℓ. We take P_ℓ as the state of the project corresponding to searching site ℓ. Then the corresponding one-step reward is

$$R^\ell(P_\ell) = \beta_\ell P_\ell v_\ell - c_\ell, \tag{1.36}$$

with the retirement reward being $M = 0$. If a search at site ℓ does not reveal the treasure, the probability P_ℓ drops to

$$\overline{P}_\ell = \frac{P_\ell(1-\beta_\ell)}{P_\ell(1-\beta_\ell) + 1 - P_\ell},$$

as can be verified using Bayes' rule. If the search finds the treasure, the probability P_ℓ drops to zero, since the treasure is removed from the site. Based on this and the fact that $R^\ell(P_\ell)$ is increasing with P_ℓ [cf. Eq. (1.36)], it is seen that the deteriorating condition (1.35) holds. Therefore, it is optimal to search the site ℓ for which the expression $R^\ell(P_\ell)$ of Eq. (1.36) is maximal, provided $\max_\ell R^\ell(P_\ell) \geq 0$, and to retire if $R^\ell(P_\ell) < 0$ for all ℓ.

1.4 DISCOUNTED CONTINUOUS-TIME PROBLEMS

In this section, we consider continuous-time semi-Markov problems of the type that we discussed in Section 5.6 of Vol. I. We saw there that they are closely connected to discrete-time problems, the main difference being that the discount factor depends on the state and the control. In this section, we reexamine these problems in the light of the framework of this chapter, and we discuss some interesting special cases. We restrict attention to the case of a finite or countable number of states.

We first focus on an important special case, where the times between successive transitions have an *exponential probability distribution*. We show that by using a conversion process called *uniformization*, the analysis of these models and the relation to the discrete-time framework can be simplified. Many of the practical systems of this type involve the Poisson process, so for most of the examples discussed here and later in Section 4.6, we assume that the reader is familiar with this process at the level of textbooks such as [Ros83b], [Gal95], and [BeT08].

We then discuss problems where the times between successive transitions need not have an exponential distribution. This is the case discussed for a finite number of states in Section 5.6 of Vol. I. Here we extend the analysis to the case where the number of states is countably infinite.

Uniformization

We consider a continuous-time system with a finite or a countable number of states. Accordingly, states are denoted by $i = 1, 2, \ldots,$ and the system is described by transition distributions. In particular, state transitions and control selections take place at discrete times, but the time from one transition to the next is random. We first assume the following:

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 transition to state i and transition to the next state is exponentially distributed with parameter $\nu_i(u) > 0$; 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

   $$p(\tau) = \nu_i(u)e^{-\nu_i(u)\tau}, \qquad \tau \geq 0.$$

   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

   $$\nu_i(u) \leq \nu, \qquad \text{for all } i,\ u \in U(i).$$

The parameter $\nu_i(u)$ is referred to as the *transition rate* 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 $i(t)$ and $u(t)$, respectively, and stay constant between transitions. We use the following notation:

t_k: The time of occurrence of the kth transition. By convention, we denote $t_0 = 0$.

$\tau_k = t_k - t_{k-1}$: The kth transition time interval.

$i_k = i(t_k)$: We have $i(t) = i_k$ 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

$$\lim_{N\to\infty} E\left\{\int_0^{t_N} e^{-\beta t}g\big(i(t), u(t)\big)dt\right\}, \tag{1.37}$$

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 $\pi = \{\mu_0, \mu_1, \ldots\}$, 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 $\mu_k(i_k)$. Because states stay constant between transitions, the cost function of π is given by

$$J_\pi(i_0) = \sum_{k=0}^{\infty} E\left\{ \int_{t_k}^{t_{k+1}} e^{-\beta t} g\big(i_k, \mu_k(i_k)\big) \,\Big|\, i_0 \right\}.$$

We first consider the case where *the transition rate is the same for all states and controls*; i.e., for all i and u,

$$\nu_i(u) = \nu.$$

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 (1.37) corresponding to a sequence $\big\{(i_k, u_k)\big\}$ can be expressed as

$$\sum_{k=0}^{\infty} E\left\{ \int_{t_k}^{t_{k+1}} e^{-\beta t} g\big(i(t), u(t)\big) dt \right\} = \sum_{k=0}^{\infty} E\left\{ \int_{t_k}^{t_{k+1}} e^{-\beta t} dt \right\} E\big\{g(i_k, u_k)\big\}. \tag{1.38}$$

We have (using the independence of the transition time intervals)

$$\begin{aligned} E\left\{ \int_{t_k}^{t_{k+1}} e^{-\beta t} dt \right\} &= \frac{E\{e^{-\beta t_k}\}\big(1 - E\{e^{-\beta \tau_{k+1}}\}\big)}{\beta} \\ &= \frac{E\{e^{-\beta(\tau_1 + \ldots + \tau_k)}\}\big(1 - E\{e^{-\beta \tau_{k+1}}\}\big)}{\beta} \\ &= \frac{\alpha^k (1 - \alpha)}{\beta}, \end{aligned} \tag{1.39}$$

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. (1.39), 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. (1.38) it follows that the cost of the problem can be expressed as

$$\frac{1}{\beta+\nu}\sum_{k=0}^{\infty}\alpha^k E\big\{g(i_k,u_k)\big\}.$$

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

$$\lim_{N\to\infty} E\left\{\int_0^{t_N} e^{-\beta t} g\big(i(t),u(t)\big)dt\right\},$$

and transition rate ν that is independent of state and control is equivalent to a discrete-time Markov chain problem with discount factor

$$\alpha=\frac{\nu}{\beta+\nu},$$

and cost per stage given by

$$\tilde g(i,u)=\frac{g(i,u)}{\beta+\nu}. \tag{1.40}$$

In particular, Bellman's equation takes the form

$$J(i)=\min_{u\in U(i)}\left[\tilde g(i,u)+\alpha\sum_j p_{ij}(u)J(j)\right]. \tag{1.41}$$

In some problems, in addition to the cost (1.37), 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 (1.40) should be changed to $\hat g(i,u)+\tilde g(i,u)$, and Bellman's equation (1.41) becomes

$$J(i)=\min_{u\in U(i)}\left[\hat g(i,u)+\tilde g(i,u)+\alpha\sum_j p_{ij}(u)J(j)\right]. \tag{1.42}$$

Example 1.4.1

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. There is a cost $c(i)$ per unit time that i orders remain unfilled. We

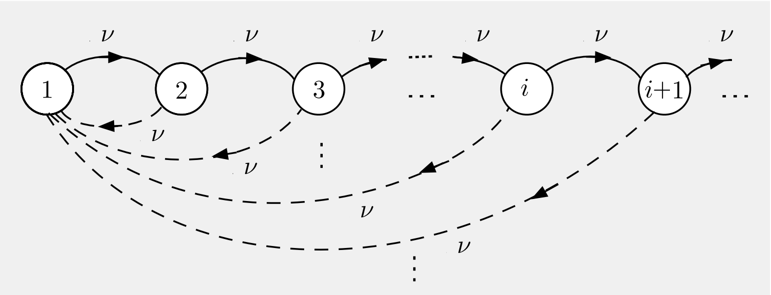


Figure 1.4.1 Transition diagram for the continuous-time Markov chain of Example 1.4.1. The transitions associated with the first control (do not fill the orders) are shown with solid lines, and the transitions associated with the second control (fill the orders) are shown with broken lines.

assume that $c(i)$ is bounded and monotonically nondecreasing with i. 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 expected cost $c(i)/(\beta + \nu)$ up to the transition to the next state $i + 1$ [cf. Eq. (1.40)], as shown in Fig. 1.4.1. We are in effect faced with a discounted discrete-time problem with bounded cost per stage.

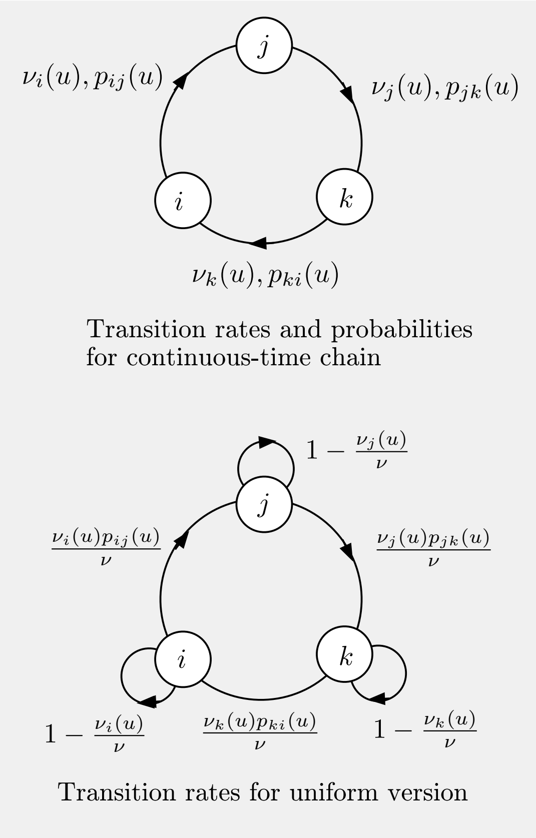
Bellman's equation takes the form

$$J(i) = \min \left[K + \alpha J(1), \frac{c(i)}{\beta + \nu} + \alpha J(i+1) \right], \qquad i = 1, 2, \ldots,$$

where $\alpha = \nu/(\beta + \nu)$ is the effective discount factor [cf. Eq. (1.5)]. [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. (1.42).] Reasoning from first principles (or using the value iteration algorithm and induction), 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^*.

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


j
$\nu_i(u), p_{ij}(u)$
$\nu_j(u), p_{jk}(u)$
i
k
$\nu_k(u), p_{ki}(u)$
$1 - \frac{\nu_j(u)}{\nu}$
j
$\frac{\nu_i(u)p_{ij}(u)}{\nu}$
$\frac{\nu_j(u)p_{jk}(u)}{\nu}$
i
k
$1 - \frac{\nu_i(u)}{\nu}$
$\frac{\nu_k(u)p_{ki}(u)}{\nu}$
$1 - \frac{\nu_k(u)}{\nu}$

Transition rates and probabilities
for continuous-time chain

Transition rates for uniform version

Figure 1.4.2 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

$$\tilde{p}_{ij}(u) = \big(\nu_i(u)/\nu\big)p_{ij}(u), \qquad \text{for } i \neq j,$$

and

$$\tilde{p}_{ii}(u) = \big(\nu_i(u)/\nu\big)p_{ii}(u) + 1 - \nu_i(u)/\nu, \qquad \text{for } j = i.$$

In the example of the figure we have $p_{ii}(u) = 0$ for all i and u.

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. 1.4.2). We argue now that leaving state i at a rate $\nu_i(u)$ in the original process is probabilistically 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 probabilistic equivalence, we mean that for any given policy π, initial state i_0, and time t, the probabilities $P\{i(t) = i \mid i_0, \pi\}$ are identical for the original process and its uniform version, for all i. We give a proof of this fact in Exercise 1.10 for the case of a finite number of states (see [Lip75b] and [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

$$\lim_{N\to\infty} E\left\{\int_0^{t_N} e^{-\beta t} g\big(i(t), u(t)\big)\,dt\right\},$$

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 for all i and u,

$$\nu_i(u) \le \nu.$$

The transition probabilities are

$$\tilde p_{ij}(u) = \begin{cases} \frac{\nu_i(u)}{\nu} p_{ij}(u) & \text{if } i \ne j, \\ \frac{\nu_i(u)}{\nu} p_{ii}(u) + 1 - \frac{\nu_i(u)}{\nu} & \text{if } i = j, \end{cases}$$

and the cost per stage is for all i and u,

$$\tilde g(i,u) = \frac{g(i,u)}{\beta + \nu}.$$

In particular, Bellman's equation takes the form

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

which, after some calculation using the preceding definitions, can be written as

$$J(i) = \frac{1}{\beta+\nu} \min_{u\in U(i)} \left[g(i,u) + \big(\nu - \nu_i(u)\big) J(i) + \nu_i(u) \sum_j p_{ij}(u) J(j)\right].$$

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. (1.42)]

$$J(i) = \frac{1}{\beta+\nu} \min_{u\in U(i)} \Biggl[(\beta+\nu)\hat{g}(i,u) + g(i,u) \\ + \bigl(\nu - \nu_i(u)\bigr)J(i) + \nu_i(u)\sum_j p_{ij}(u)J(j) \Biggr].$$

Example 1.4.2 (Priority Assignment and the μc Rule)

Consider r queues that share a single server. There is a positive cost c_ℓ per unit time and per customer in each queue $\ell = 1,\ldots,r$. The service time of a customer of queue ℓ is exponentially distributed with parameter μ_ℓ, 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

$$\lim_{N\to\infty} E\left\{ \int_0^{t_N} e^{-\beta t} \sum_{\ell=1}^{r} c_\ell \, x_\ell(t) dt \right\},$$

where $x_\ell(t)$ is the number of customers in the ℓth 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. 1.4.3. The discount factor is

$$\alpha = \frac{\mu}{\beta+\mu},$$

where

$$\mu = \max_\ell\{\mu_\ell\},$$

and the corresponding cost is

$$\frac{1}{\beta+\mu}\sum_{k=0}^{\infty} \alpha^k E\left\{ \sum_{\ell=1}^{r} c_\ell \, x_k^\ell \right\}, \tag{1.43}$$

where x_k^ℓ is the number of customers in the ℓth queue after the kth transition (real or fictitious).

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. For $k = 0,1,\ldots$, define

$$\ell_k = \begin{cases} \ell & \text{if the } k\text{th transition corresponds to a departure from queue } \ell, \\ 0 & \text{if the } k\text{th transition is fictitious.} \end{cases}$$

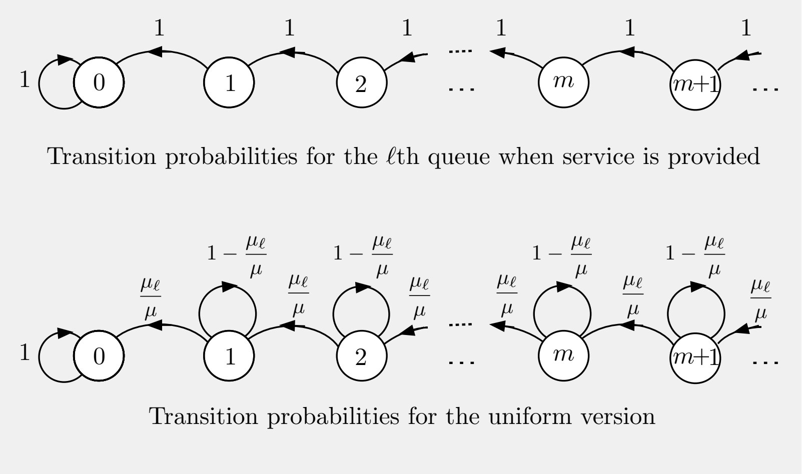

1
1
1
1
1
1
1
0
1
2
...
m
m+1
...
Transition probabilities for the ℓth queue when service is provided
$1-\frac{\mu_\ell}{\mu}$
$\frac{\mu_\ell}{\mu}$
1
0
1
2
...
m
m+1
...
Transition probabilities for the uniform version

Figure 1.4.3 Continuous-time Markov chain and uniform version for the ℓth queue of Example 1.4.2 when service is provided. The transition rate for the uniform version is $\mu = \max_\ell\{\mu_\ell\}$.

Denote also

$$\begin{aligned} &c_{\ell_0} = 0, \\ &x_0^\ell : \text{the initial number of customers in queue } \ell. \end{aligned}$$

Then the cost (1.43) can also be written as

$$\begin{aligned} \frac{1}{\beta+\mu}&\left[\sum_{\ell=1}^{r} c_\ell\, x_0^\ell + \sum_{k=1}^{\infty}\alpha^k E\left\{\sum_{\ell=1}^{r} c_\ell\, x_0^\ell - \sum_{m=0}^{k-1} c_{\ell_m}\right\}\right] \\ &= \frac{1}{\beta+\mu}\left[\sum_{k=0}^{\infty}\alpha^k\left(\sum_{\ell=1}^{r} c_\ell\, x_0^\ell\right) - E\left\{\sum_{m=0}^{\infty}\sum_{k=m+1}^{\infty}\alpha^k c_{\ell_m}\right\}\right] \\ &= \frac{1}{(\beta+\mu)(1-\alpha)}\sum_{\ell=1}^{r} c_\ell\, x_0^\ell - \frac{\alpha}{(\beta+\mu)(1-\alpha)}\sum_{k=0}^{\infty}\alpha^k E\{c_{\ell_k}\} \\ &= \frac{1}{\beta}\sum_{\ell=1}^{r} c_\ell\, x_0^\ell - \frac{\alpha}{\beta}\sum_{k=0}^{\infty}\alpha^k E\{c_{\ell_k}\}. \end{aligned}$$

Therefore, instead of minimizing the cost (1.43), we can equivalently

$$\text{maximize } \sum_{k=0}^{\infty}\alpha^k E\{c_{\ell_k}\}, \tag{1.44}$$

where c_{ℓ_k} can be viewed as the *savings in waiting cost rate* obtained from the kth transition.

We now recognize problem (1.44) as a multiarmed bandit problem. The r queues can be viewed as separate projects. At each time, a nonempty queue, say ℓ, is selected and served. Since a customer departure occurs with probability μ_ℓ/μ, and a fictitious transition that leaves the state unchanged occurs with probability $1 - \mu_\ell/\mu$, the corresponding expected reward is

$$\frac{\mu_\ell}{\mu} c_\ell. \tag{1.45}$$

Note that the reward per stage is bounded, so we may use the framework and results of Sections 1.2 and 1.3. It is evident that the problem falls in the deteriorating case examined at the end of Section 1.3. 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 1.3). Equivalently [cf. Eq. (1.45)], *it is optimal to serve the nonempty queue ℓ for which $\mu_\ell c_\ell$ is maximum*. This policy is known as the *μc rule*. It plays an important role in several other formulations of the priority assignment problem. We can view $\mu_\ell c_\ell$ as the ratio of the waiting cost rate c_ℓ by the average time $1/\mu_\ell$ 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.

Discounted Semi-Markov Problems

We now consider a more general version of the continuous-time problem where we cannot use uniformization, and as a result we cannot pose the problem as a discounted problem that fits the framework of Section 1.2. We instead formulate the problem as a discrete-time problem that nearly fits that framework. The only difference is that the discount factor may depend on the state and/or the control.

We continue to assume a finite or a countable number of states, but we replace transition probabilities with *transition distributions* $Q_{ij}(\tau, u)$, which for a given pair (i, u), specify the joint distribution of the transition interval and the next state:

$$Q_{ij}(\tau, u) = P\{t_{k+1} - t_k \le \tau,\ i_{k+1} = j \mid i_k = i,\ u_k = u\}.$$

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:

$$\int_0^\infty \tau Q_{ij}(\tau, u) < \infty.$$

Note that the transition distributions specify the ordinary transition probabilities via

$$p_{ij}(u) = P\{i_{k+1} = j \mid i_k = i,\ u_k = u\} = \lim_{\tau \to \infty} Q_{ij}(\tau, u).$$

Thus, contrary to the case where uniformization applies, $Q_{ij}(\tau, u)$ need not be an exponential distribution.

As earlier, we consider a cost function of the form

$$\lim_{N\to\infty} E\left\{\int_0^{t_N} e^{-\beta t} g\big(i(t), u(t)\big)dt\right\}, \tag{1.46}$$

where t_N is the completion time of the Nth transition, and the function g and the positive discount parameter β are given. The cost function of an admissible N-stage policy $\pi = \{\mu_0, \mu_1, \ldots, \mu_{N-1}\}$ is given by

$$J_\pi^N(i) = \sum_{k=0}^{N-1} E\left\{\int_{t_k}^{t_{k+1}} e^{-\beta t} g\big(i_k, \mu_k(i_k)\big)dt \;\Big|\; i_0 = i\right\}.$$

We see that for all states i we have

$$J_\pi^N(i) = G\big(i, \mu_0(i)\big) + \sum_j \int_0^\infty e^{-\beta\tau} Q_{ij}\big(d\tau, \mu(i)\big) J_{\pi_1}^{N-1}(j), \tag{1.47}$$

where $J_{\pi_1}^{N-1}(j)$ is the $(N-1)$-stage cost of the policy $\pi_1 = \{\mu_1, \mu_2, \ldots, \mu_{N-1}\}$ 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 $\int_0^\tau e^{-\beta t}dt = (1 - e^{-\beta\tau})/\beta$,

$$G(i, u) = g(i, u) \sum_j \int_0^\infty \frac{1 - e^{-\beta\tau}}{\beta} Q_{ij}(d\tau, u). \tag{1.48}$$

If we denote

$$m_{ij}(u) = \int_0^\infty e^{-\beta\tau} Q_{ij}(d\tau, u),$$

we see that Eq. (1.47) can be written in the form

$$J_\pi^N(i) = G\big(i, \mu_0(i)\big) + \sum_j m_{ij}\big(\mu_0(i)\big) J_{\pi_1}^{N-1}(j), \tag{1.49}$$

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 (1.49) motivates the use of mappings T and T_μ that are similar to those introduced in Section 1.1.2. For a function J and a stationary policy μ, let us define

$$(T_\mu J)(i) = G\big(i, \mu(i)\big) + \sum_j m_{ij}\big(\mu(i)\big) J(j), \tag{1.50}$$

$$(TJ)(i) = \min_{u \in U(i)} \left[G(i,u) + \sum_j m_{ij}(u) J(j) \right]. \tag{1.51}$$

Then by using Eq. (1.49), it can be seen that the cost function J_π of an infinite horizon policy $\pi = \{\mu_0, \mu_1, \ldots\}$ can be expressed as

$$J_\pi(i) = \lim_{N \to \infty} J_\pi^N(i) = \lim_{N \to \infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} J_0)(i),$$

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

$$J_\mu(i) = \lim_{N \to \infty} (T_\mu^N J_0)(i).$$

These expressions implicitly assume that the corresponding limits exist, something that we will verify shortly under suitable conditions.

The discounted cost analysis of Section 1.2 carries through in its entirety (see also Section 5.6 of Vol. I), 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.,

$$\rho = \max_{i,\, u \in U(i)} \sum_j m_{ij}(u) < 1. \tag{1.52}$$

Under these circumstances, analogs of the results of Section 1.2 can be readily shown. In particular, the optimal cost function J^* is the unique bounded solution of Bellman's equation $J = TJ$ or

$$J(i) = \min_{u \in U(i)} \left[G(i,u) + \sum_j m_{ij}(u) J(j) \right].$$

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.

We note that for the property $\rho < 1$ [cf. Eq. (1.52)] to hold, it is sufficient that there exist $\overline{\tau} > 0$ and $\epsilon > 0$ such that the transition time is greater than $\overline{\tau}$ with probability greater than $\epsilon > 0$; i.e., we have for all i and $u \in U(i)$,

$$1 - \sum_j Q_{ij}(\overline{\tau}, u) = \sum_j P\{\tau \geq \overline{\tau} \mid i,\, u,\, j\} \geq \epsilon.$$

We finally note that in some problems, in addition to the cost (1.46), 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

$$(T_\mu J)(i) = \hat{g}\big(i, \mu(i)\big) + G\big(i, \mu(i)\big) + \sum_j m_{ij}\big(\mu(i)\big) J(j),$$

$$(TJ)(i) = \min_{u \in U(i)} \left[\hat{g}(i,u) + G(i,u) + \sum_j m_{ij}(u) J(j) \right].$$

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 $Q_{ij}(\tau,u)/p_{ij}(u)$. Then $G(i,u)$ 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. (1.48)] and the preceding development goes through without modification.

Example 1.4.3 (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 discounted cost.

Note that 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

$$Q_{0j}(\tau, \lambda, \mu) = \begin{cases} 1 - e^{-\lambda\tau} & \text{if } j = 1, \\ 0 & \text{otherwise,} \end{cases}$$

$$Q_{ij}(\tau, \lambda, \mu) = \begin{cases} p_{ij}(\lambda, \mu) & \text{if } 1/\mu \le \tau, \\ 0 & \text{otherwise,} \end{cases} \qquad i \ge 1,$$

where $p_{ij}(\lambda,\mu)$ are the state transition probabilities. 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 $[0,1/\mu]$, which is given by the Poisson distribution (see e.g., [BeT08]). In particular, we have

$$p_{ij}(\lambda,\mu) = \begin{cases} \frac{e^{-\lambda/\mu}(\lambda/\mu)^{(j-i+1)}}{(j-i+1)!} & \text{if } j \geq i-1, \\ 0 & \text{otherwise,} \end{cases} \qquad i \geq 1.$$

Using the above formulas, one can write Bellman's equation and solve the problem as if it were essentially a discrete-time discounted problem.

1.5 THE ROLE OF CONTRACTION MAPPINGS

Two key structural properties of 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 plays an important role in the stochastic shortest path models of Chapter 3, and it forms the basis for the analysis of positive and negative DP models, given in Chapter 4.

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 for all $y \in Y$, $\|y\| \geq 0$, $\|y\| = 0$ if and only if $y = 0$, $\|ay\| = |a|\|y\|$ for all scalars a, 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

$$\|Fy - Fz\| \leq \rho\|y - z\|, \qquad \forall\, y, z \in Y.$$

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, in the sense that for some $\bar{y} \in Y$, we have $\|y_k - \bar{y}\| \to 0$.† When

† In this section we will use some introductory material from real analysis; we refer to textbooks such as [LiS61], [Roy88], [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 $\|y_m - y_n\| \to 0$ as $m, n \to \infty$, 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, implying that the real line is a complete space and so is every real finite-dimensional vector space. On the other hand, an infinite dimensional space may not be complete under some norms, while it may be complete under other norms.

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

$$y = Fy$$

has a unique solution y^*, called the *fixed point of* F. Furthermore, the sequence $\{y_k\}$ generated by the iteration

$$y_{k+1} = Fy_k$$

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.

Example 1.5.1 (Linear Contraction Mappings in $\Re^n$)

Consider the case of a linear mapping $F : \Re^n \mapsto \Re^n$ of the form

$$Fy = b + Ay,$$

where A is an $n \times n$ matrix and b is a vector in $\Re^n$. Let $\sigma(A)$ denote the spectral radius of A (the largest modulus among the moduli of the eigenvalues of A). Then it can be shown that A *is a contraction mapping with respect to some norm if and only if* $\sigma(A) < 1$.

Specifically, given $\epsilon > 0$, there exists a norm $\|\cdot\|_s$ such that

$$\|Ay\|_s \le \big(\sigma(A) + \epsilon\big)\|y\|_s, \qquad \forall\ y \in \Re^n. \tag{1.53}$$

Thus, if $\sigma(A) < 1$ we may select $\epsilon > 0$ such that $\rho = \sigma(A) + \epsilon < 1$, and obtain the contraction relation

$$\|Fy - Fz\|_s = \big\|A(y - z)\big\|_s \le \rho\|y - z\|_s, \qquad \forall\ y, z \in \Re^n. \tag{1.54}$$

The norm $\|\cdot\|_s$ can be taken to be a weighted Euclidean norm, i.e., it may have the form $\|y\|_s = \|My\|$, where M is a square invertible matrix, and $\|\cdot\|$ is the standard Euclidean norm, i.e., $\|x\| = \sqrt{x'x}$. †

† We may show Eq. (1.53) by using the Jordan canonical form of A, which is denoted by J. In particular, if P is a nonsingular matrix such that $P^{-1}AP = J$ and D is the diagonal matrix with $1, \delta, \ldots, \delta^{n-1}$ along the diagonal, where $\delta > 0$, it is straightforward to verify that $D^{-1}P^{-1}APD = \hat{J}$, where $\hat{J}$ is the matrix that is identical to J except that each nonzero off-diagonal term is replaced by δ. Defining $\hat{P} = PD$, we have $A = \hat{P}\hat{J}\hat{P}^{-1}$. Now if $\|\cdot\|$ is the standard Euclidean norm, we note that for some $\beta > 0$, we have $\|\hat{J}z\| \le \big(\sigma(A) + \beta\delta\big)\|z\|$ for all $z \in \Re^n$ and $\delta \in (0, 1]$. For a given $\delta \in (0, 1]$, consider the weighted Euclidean norm $\|\cdot\|_s$ defined by $\|y\|_s = \|\hat{P}^{-1}y\|$. Then we have for all $y \in \Re^n$,

$$\|Ay\|_s = \|\hat{P}^{-1}Ay\| = \|\hat{P}^{-1}\hat{P}\hat{J}\hat{P}^{-1}y\| = \|\hat{J}\hat{P}^{-1}y\| \le \big(\sigma(A) + \beta\delta\big)\|\hat{P}^{-1}y\|,$$

so that $\|Ay\|_s \le \big(\sigma(A) + \beta\delta\big)\|y\|_s$, for all $y \in \Re^n$. For a given $\epsilon > 0$, we choose $\delta = \epsilon/\beta$, so the preceding relation yields Eq. (1.53).

Conversely, if Eq. (1.54) holds for some norm $\|\cdot\|_s$ and all real vectors y, z, it also holds for all complex vectors y, z with the squared norm $\|c\|_s^2$ of a complex vector c defined as the sum of the squares of the norms of the real and the imaginary components. Thus from Eq. (1.54), by taking $y - z = u$, where u is an eigenvector corresponding to an eigenvalue λ with $|\lambda| = \sigma(A)$, we have $\sigma(A)\|u\|_s = \|Au\|_s \leq \rho\|u\|_s$. Hence $\sigma(A) \leq \rho$, and it follows that if F is a contraction with respect to a given norm, we must have $\sigma(A) < 1$.

1.5.1 Sup-Norm Contractions

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

$$v(x) > 0, \qquad \forall\, x \in X.$$

Let $B(X)$ denote the set of all functions $J : X \mapsto \Re$ such that $J(x)/v(x)$ is bounded as x ranges over X. We define a norm on $B(X)$, called the *weighted sup-norm*, by

$$\|J\| = \max_{x \in X} \frac{|J(x)|}{v(x)}. \tag{1.55}$$

(The maximum in the above equation need not be attained. We are still following the convention of denoting by "max" the least upper bound of the 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, $B(X)$ is complete under this norm. †

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

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

for all $x \in X$ and $k \geq K$; this will imply that

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

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

$$\epsilon < \big|J_{m_i}(x_{m_i}) - J^*(x_{m_i})\big|/v(x_{m_i}), \qquad \forall\, i \geq 1.$$

The right-hand side above is less or equal to

$$\big|J_{m_i}(x_{m_i}) - J_n(x_{m_i})\big|/v(x_{m_i}) + \big|J_n(x_{m_i}) - J^*(x_{m_i})\big|/v(x_{m_i}),\ \forall\, n \geq 1,\ i \geq 1.$$

In what follows, we will always assume that $B(X)$ is equipped with the weighted sup-norm above, where the weight function v will be clear from the context. There will be frequent occasions where the norm will be unweighted, i.e., $v(x) \equiv 1$ and $\|J\| = \max_{x\in X} \big|J(x)\big|$, in which case we will explicitly state so.

For a mapping $F : B(X) \mapsto B(X)$ and a function $J \in B(X)$, the function F^kJ obtained by applying F to J successively k times belongs to $B(X)$. The following is the central result regarding contraction mappings, specialized to the case of $B(X)$. Assuming that F is a contraction mapping, it guarantees the convergence of F^kJ to the unique fixed point of F, and provides the basis for an important algorithm for computing fixed points.

Proposition 1.5.1: (Contraction Mapping Fixed-Point Theorem) If $F : B(X) \mapsto B(X)$ is a contraction mapping with modulus $\rho \in (0,1)$, then there exists a unique $J^* \in B(X)$ such that

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

Furthermore, $\{F^kJ\}$ converges to J^* for any $J \in B(X)$, and we have

$$\|F^kJ - J^*\| \le \rho^k\|J - J^*\|, \qquad k = 1, 2, \ldots .$$

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

$$\|J_{k+1} - J_k\| \le \rho\|J_k - J_{k-1}\|, \qquad k = 1, 2, \ldots,$$

which implies that

$$\|J_{k+1} - J_k\| \le \rho^k\|J_1 - J_0\|, \qquad k = 1, 2, \ldots .$$

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

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

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.

Therefore, $\{J_k\}$ is a Cauchy sequence and must converge to a limit $J^* \in B(X)$, since $B(X)$ 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 convergence rate bound of the last part, note that

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

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

The convergence rate exhibited by F^kJ in the preceding proposition is said to be *geometric*, and F^kJ is said to converge to its limit J^* *geometrically*. This is in reference to the fact that the error $\|F^kJ - J^*\|$ converges to 0 faster than some geometric progression.

Consider now the mappings T and T_μ associated with the discounted cost problem with bounded cost per stage [cf. Eqs. (1.5) and (1.40)]. Propositions 1.2.6 and 1.2.7 show that T and T_μ are contraction mappings ($\rho = \alpha$) with respect to the unweighted sup-norm, where $v(x) \equiv 1$. Their unique fixed points are J^* (the optimal cost function) and J_μ, respectively. Furthermore, the convergence of the DP recursion/value iteration to J^* follows from the general convergence result of Prop. 1.5.1. The same is true for the mappings T and T_μ corresponding to semi-Markov decision problems [cf. Eqs. (1.50) and (1.51)]. Later 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. An important such case arises in stochastic shortest path problems (see Chapter 3).

Let us now focus on the finite-dimensional case $X = \{1, \ldots, n\}$. Consider a linear mapping $F : \Re^n \mapsto \Re^n$ of the form

$$Fy = b + Ay,$$

where A is an $n \times n$ matrix with components a_{ij}, and b is a vector in $\Re^n$ (cf. Example 1.5.1). Then it is straightforward to verify (see the following proposition) that F *is a contraction with respect to the weighted sup-norm* $\|y\| = \max_{i=1,\ldots,n} |y_i|/v(i)$ *if and only if*

$$\frac{\sum_{j=1}^n |a_{ij}|\, v(j)}{v(i)} < 1, \qquad i = 1, \ldots, n.$$

Let us also denote by $|A|$ the matrix whose components are the absolute values of the components of A and let $\sigma\big(|A|\big)$ denote the spectral radius of $|A|$. Then it can be shown that F *is a contraction with respect to some weighted sup-norm if and only if* $\sigma\big(|A|\big) < 1$. A proof of this may be found in [BeT89], Ch. 2, Cor. 6.2. A proof may also be constructed using the analysis of SSP problems in Chapter 3 (see Prop. 3.2.3), and the fact that A is a weighted sup-norm contraction if and only if $|A|$ is. Thus any substochastic matrix P ($p_{ij} \geq 0$ for all i, j, and $\sum_{j=1}^n p_{ij} \leq 1$, for all i) is a contraction with respect to some weighted sup-norm if and only if $\sigma(P) < 1$.

Finally, let us consider a nonlinear mapping $F : \Re^n \mapsto \Re^n$ that has the property

$$|Fy - Fz| \leq P\,|y - z|, \qquad \forall\ y, z \in \Re^n,$$

for some matrix P with nonnegative components and $\sigma(P) < 1$. Here, we generically denote by $|w|$ the vector whose components are the absolute values of the components of w, and the inequality is componentwise. Then we claim that F is a contraction with respect to some weighted sup-norm. To see this note that by the preceding discussion, P is a contraction with respect to some weighted sup-norm $\|w\| = \max_{i=1,\ldots,n} |w(i)|/v(i)$, and we have

$$\frac{\big(|Fy - Fz|\big)(i)}{v(i)} \leq \frac{\big(P\,|y - z|\big)(i)}{v(i)} \leq \alpha\,\|y - z\|, \qquad \forall\ i = 1, \ldots, n,$$

for some $\alpha \in (0, 1)$, so that F is a contraction with respect to $\|\cdot\|$. For additional discussion of linear and nonlinear contraction mapping properties and characterizations such as the one above, see the book [OrR70].

Some Special Cases for Countable State Space

The case where X 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 1.5.2: Let $X = \{1, 2, \ldots\}$.

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

$$(FJ)(i) = b_i + \sum_{j\in X} a_{ij} J(j), \qquad i \in X,$$

where b_i and a_{ij} are some scalars. Then F is a contraction with modulus ρ with respect to the weighted sup-norm (1.55) if and only if

$$\frac{\sum_{j\in X} |a_{ij}|\, v(j)}{v(i)} \le \rho, \qquad i \in X. \tag{1.56}$$

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

$$(FJ)(i) = \min_{\mu\in M} (F_\mu J)(i), \qquad i \in X,$$

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

Proof: (a) Assume that Eq. (1.56) holds. For any $J, J' \in B(X)$, we have

$$\begin{aligned}\|FJ - FJ'\| &= \max_{i\in X} \frac{\left|\sum_{j\in X} a_{ij}\big(J(j) - J'(j)\big)\right|}{v(i)}\\ &\le \max_{i\in X} \frac{\sum_{j\in X} \big|a_{ij}\big|\, v(j)\Big(\big|J(j) - J'(j)\big|/v(j)\Big)}{v(i)}\\ &\le \max_{i\in X} \frac{\sum_{j\in X} \big|a_{ij}\big|\, v(j)}{v(i)}\, \big\|J - J'\big\|\\ &\le \rho\, \|J - J'\|,\end{aligned}$$

where the last inequality follows from the hypothesis.

Conversely, arguing by contradiction, let's assume that Eq. (1.56) is violated for some $i \in X$. Define $J(j) = v(j)\,\mathrm{sgn}(a_{ij})$ and $J'(j) = 0$ for all $j \in X$. Then we have $\|J - J'\| = \|J\| = 1$, and

$$\frac{\big|(FJ)(i) - (FJ')(i)\big|}{v(i)} = \frac{\sum_{j\in X} |a_{ij}|\, v(j)}{v(i)} > \rho = \rho\, \|J - J'\|,$$

showing that F is not a contraction of modulus ρ.

(b) Since F_μ is a contraction of modulus ρ, we have for any $J, J' \in B(X)$,

$$\frac{(F_\mu J)(i)}{v(i)} \le \frac{(F_\mu J')(i)}{v(i)} + \rho\, \|J - J'\|, \qquad i \in X,$$

so by taking the minimum over $\mu \in M$,

$$\frac{(FJ)(i)}{v(i)} \le \frac{(FJ')(i)}{v(i)} + \rho\, \|J - J'\|, \qquad i \in X.$$

Reversing the roles of J and J', we obtain

$$\frac{\big|(FJ)(i) - (FJ')(i)\big|}{v(i)} \le \rho\, \|J - J'\|, \qquad i \in X,$$

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

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

Proposition 1.5.3: Let $X = \{1, 2, \ldots\}$, let M be a parameter set, and for each $\mu \in M$, let F_μ be a linear mapping of the form

$$(F_\mu J)(i) = b_i(\mu) + \sum_{j\in X} a_{ij}(\mu)\, J(j), \qquad i \in X.$$

(a) We have $F_\mu J \in B(X)$ for all $J \in B(X)$ provided $b(\mu) \in B(X)$ and $V(\mu) \in B(X)$, where

$$b(\mu) = \big\{b_1(\mu), b_2(\mu), \ldots\big\}, \qquad V(\mu) = \big\{V_1(\mu), V_2(\mu), \ldots\big\},$$

with

$$V_i(\mu) = \sum_{j\in X} \big|a_{ij}(\mu)\big|\, v(j), \qquad i \in X.$$

(b) Consider the mapping F

$$(FJ)(i) = \min_{\mu\in M} (F_\mu J)(i), \qquad i \in X.$$

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

$$b = \big\{b_1, b_2, \ldots\big\}, \qquad V = \{V_1, V_2, \ldots\},$$

with $b_i = \max_{\mu\in M} b_i(\mu)$ and $V_i = \max_{\mu\in M} V_i(\mu)$.

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

$$\begin{aligned}(F_\mu J)(i) &\le \big|b_i(\mu)\big| + \sum_{j\in X} \big|a_{ij}(\mu)\big|\, \big|J(j)/v(j)\big|\, v(j)\\ &\le \big|b_i(\mu)\big| + \|J\| \sum_{j\in X} \big|a_{ij}(\mu)\big|\, v(j)\end{aligned}$$

$$= \big|b_i(\mu)\big| + \|J\|\, V_i(\mu),$$

and similarly $(F_\mu J)(i) \geq -\big|b_i(\mu)\big| - \|J\|\, V_i(\mu)$. Thus

$$\big|(F_\mu J)(i)\big| \leq \big|b_i(\mu)\big| + \|J\|\, V_i(\mu), \qquad i \in X.$$

By dividing this inequality with $v(i)$ and by taking the maximum over $i \in X$, 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.

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

$$\big\|F^m J - F^m J'\big\| \leq \rho \|J - J'\|, \qquad \forall\ J, J' \in B(X),$$

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. 1.5.1.

Proposition 1.5.4: (m-Stage Contraction Mapping Fixed-Point Theorem) If $F : B(X) \mapsto B(X)$ is an m-stage contraction mapping with modulus $\rho \in (0, 1)$, then there exists a unique $J^* \in B(X)$ such that

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

Furthermore, $\{F^k J\}$ converges to J^* for any $J \in B(X)$.

Proof: Since F^m maps $B(X)$ into $B(X)$ and is a contraction mapping, by Prop. 1.5.1, it has a unique fixed point in $B(X)$, 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. 1.5.1, we have for all $J \in B(X)$,

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

Using $F^\ell J$ in place of J, we obtain

$$\lim_{k\to\infty} \|F^{mk+\ell} J - J^*\| = 0, \qquad \ell = 0, 1, \ldots, m-1,$$

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

In the next subsection, we discuss an interesting discounted problem that cannot be analyzed with the theory of Sections 1.1 and 1.2, but can be addressed with m-stage contraction mapping theory.

1.5.2 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 and unbounded cost per stage there is a method of analysis that relies on weighted sup-norm contractions.

Consider a problem where the state space is $X = \{1, 2, \ldots\}$, the discount factor is $\alpha \in (0, 1)$, the transition probabilities are denoted $p_{ij}(u)$ for $i, j \in X$ and $u \in U(i)$, and the expected cost per stage is denoted by $g(i, u)$, $i \in X$, $u \in U(i)$. The constraint set $U(i)$ may be infinite. For a positive weight sequence $v = \big\{v(1), v(2), \ldots\big\}$, we consider the space $B(X)$ of sequences $J = \big\{J(1), J(2), \ldots\big\}$ such that $\|J\| < \infty$, where $\|\cdot\|$ is the weighted sup-norm

$$\|J\| = \max_{i\in X} \frac{|J(i)|}{v(i)}.$$

The following assumption will allow the use of Props. 1.5.2 and 1.5.3 for the purpose of showing that the DP mappings T and T_μ are m-stage contraction mappings. We assume the following.

Assumption 1.5.1:

(a) The sequence $G = \big\{G_1, G_2, \ldots\big\}$, where

$$G_i = \max_{u \in U(i)} \big|g(i,u)\big|, \qquad i \in X,$$

belongs to $B(X)$.

(b) The sequence $V = \big\{V_1, V_2, \ldots\big\}$, where

$$V_i = \max_{u \in U(i)} \sum_{j \in X} p_{ij}(u)\, v(j), \qquad i \in X,$$

belongs to $B(X)$.

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

$$\alpha^m \frac{\sum_{j \in X} P(x_m = j \mid x_0 = i, \pi)\, v(j)}{v(i)} \leq \rho, \qquad i \in X.$$

Assumption 1.5.1(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

$$v(i) = \max\left\{1,\ \max_{u \in U(i)} \big|g(i,u)\big|\right\}, \qquad i \in X.$$

Assumption 1.5.1(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 1.5.1(c) is satisfied if the expression

$$\frac{\sum_{j \in X} P(x_m = j \mid x_0 = i, \pi)\, v(j)}{v(i)}$$

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 1.5.2

Let

$$v(i) = i, \qquad i \in X.$$

Then Assumption 1.5.1(a) is satisfied if the maximum expected absolute cost per stage at state i grows no faster than linearly with i. Assumption 1.5.1(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 1.5.1(c) is satisfied if

$$\alpha^m \sum_{j\in X} P(x_m = j \mid x_0 = i, \pi)\, j \le \rho\, i, \qquad i \in X.$$

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 1.5.1(c) is satisfied. Similar interpretations are possible for other choices of $v(i)$, such as

$$v(i) = i^t, \qquad i \in X,$$

for some positive integer t.

We now consider the DP mappings T_μ and T,

$$(T_\mu J)(i) = g\big(i,\mu(i)\big) + \alpha \sum_{j\in X} p_{ij}\big(\mu(i)\big)J(j), \qquad i \in X,$$

$$(TJ)(i) = \min_{u\in U(i)} \left[g(i,u) + \alpha \sum_{j\in X} p_{ij}(u)J(j) \right], \qquad i \in X,$$

and show their contraction property.

Proposition 1.5.5: Under Assumption 1.5.1, the mappings T and T_μ map $B(X)$ into $B(X)$, and are m-stage contraction mappings with modulus ρ.

Proof: Assumptions 1.5.1(a) and 1.5.1(b), together with Prop. 1.5.3, show that if $J \in B(X)$, then $TJ \in B(X)$ and $T_\mu J \in B(X)$ for all μ. For any $J \in B(X)$, and any policy $\pi = \{\mu_0, \mu_1, \ldots\}$, we have for all $i \in X$,

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

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. 1.5.2(a) in conjunction with Assumption 1.5.1(c), it follows that $T_{\mu_0} \cdots T_{\mu_{m-1}}$ is a contraction of modulus ρ, and then using Prop. 1.5.2(b), it follows that the same is true for T^m. **Q.E.D.**

The m-stage contraction property of T and Prop. 1.5.4 can now be used to essentially replicate the analysis of Section 1.2, and to show the standard results:

(a) The DP iteration $J_{k+1} = TJ_k$ 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 $T_\mu J^* = TJ^*$.

Let us finally note that the preceding analysis generalizes to the undiscounted case where $\alpha = 1$ (under some additional conditions). Indeed, we will revisit the corresponding contraction property of T in Section 3.6, in the context of stochastic shortest path problems with a countable number of states.

1.6 ABSTRACT FORMS OF DISCOUNTED DYNAMIC PROGRAMMING

In the preceding sections we have investigated several analytical issues for discounted problems, relating to:

(a) The existence of a unique solution of Bellman's equation.

(b) The convergence of the DP recursion/value iteration.

(c) Conditions for optimality of stationary policies.

Taking an abstract point of view, these results revolve around the mappings T_μ and T introduced in Section 1.1.2, and their variants for semi-Markov problems discussed in Section 1.4. In Section 1.5 we discussed how these results derive their validity from a central characteristic of T_μ and T (in addition to their monotonicity property), namely their contraction property. This has motivated a powerful unifying analytical approach, whereby for a given DP problem, we may investigate whether there is an underlying contraction mapping and if so, address the issues (a)-(c) above using the theory of Section 1.5. We saw an example of the utility of this approach in the context of some discounted problems with unbounded cost per stage; cf. Section 1.5.2.

In this section we develop further this abstract view, aiming at unification and generalization of the theory of discounted problems (in this chapter) and the associated computational methods (in the next chapter).[†] We consider a general class of mappings that are patterned after those appearing in stochastic DP, but are more general: for example they apply to minimax problems, game theoretic problems, undiscounted DP problems, and even to important problems beyond DP. We discuss the properties of these mappings with the issues (a)-(c) above in mind, and thus address interesting DP questions in substantially greater generality than heretofore.

While in this section, we will focus on a connection with the contraction mapping material of Section 1.5, our viewpoint transcends contraction mappings and applies to problems that are undiscounted, such as the stochastic shortest path problems of Chapter 3, and the undiscounted problems of Chapter 4. Moreover, the abstract viewpoint of this section has an algorithmic value, and will serve to unify and enhance the development of algorithms in Chapters 2 and 3.

Let X and U be two sets, which in view of connections to DP that will become apparent shortly, we will loosely refer to as a set of "states" and a set of "controls." For each $x \in X$, let $U(x) \subset U$ be a nonempty subset of controls that are feasible at state x. Consistent with the DP context, we refer to a function $\mu : X \mapsto U$ with $\mu(x) \in U(x)$, for all $x \in X$, as a "policy." We denote by $\mathcal{M}$ the set of all policies.

Let $R(X)$ be the set of real-valued functions $J : X \mapsto \Re$, and let $H : X \times U \times R(X) \mapsto \Re$ be a given mapping. We consider the mapping T defined by

$$(TJ)(x) = \min_{u \in U(x)} H(x, u, J), \qquad \forall\, x \in X.$$

We assume that $(TJ)(x) > -\infty$ for all $x \in X$, so that T maps $R(X)$ into $R(X)$. For each policy $\mu \in \mathcal{M}$, we consider the mapping $T_\mu : R(X) \mapsto R(X)$ defined by

$$(T_\mu J)(x) = H\big(x, \mu(x), J\big), \qquad \forall\, x \in X.$$

We want to find a function $J^* \in R(X)$ such that

$$J^*(x) = \min_{u \in U(x)} H(x, u, J^*), \qquad \forall\, x \in X,$$

i.e., find a fixed point of T. We also want to obtain a policy μ^* such that $T_{\mu^*} J^* = TJ^*$.

We give a few special cases. Additional examples will arise in the development of DP models and algorithms with special structure, which we will encounter later.

[†] This section, the corresponding algorithmic Sections 2.5 and 2.6 in Chapter 2, and Sections 3.3.2 and 3.4.1 of Chapter 3 contain relatively advanced topics that deal with abstract DP models, and may be skipped at first reading. This material will be used sparingly in Chapter 6, and will not be used in Chapters 4, 5, and 7. The abstract point of view is developed in much greater detail in the author's monograph [Ber18], and several applications to stochastic optimal control problems are discussed in greater depth than in the present volume.

Example 1.6.1 (Discounted Problems)

Consider the α-discounted total cost problem of Section 1.1. For

$$H(x,u,J) = E\big\{g(x,u,w) + \alpha J\big(f(x,u,w)\big)\big\},$$

the equation $J = TJ$, i.e.,

$$J(x) = \min_{u \in U(x)} H(x,u,J) = \min_{u \in U(x)} E\big\{g(x,u,w) + \alpha J\big(f(x,u,w)\big)\big\}, \quad \forall\, x \in X,$$

is Bellman's equation. In the special case of an MDP involving states $x = 1,\ldots,n$, controls $u \in U(x)$ at state x, transition probabilities $p_{xy}(u)$, and cost per stage $g(x,u,y)$, H takes the form

$$H(x,u,J) = \sum_{y=1}^{n} p_{xy}(u)\big(g(x,u,y) + \alpha J(y)\big),$$

and again the equation $J = TJ$ is Bellman's equation for the MDP.

Example 1.6.2 (Discounted Semi-Markov Problems)

With x, y, u as in Example 1.6.1, consider the mapping

$$H(x,u,J) = G(x,u) + \sum_{y=1}^{n} m_{xy}(u)J(y),$$

where G is some function representing cost per stage, and $m_{xy}(u)$ are nonnegative numbers with $\sum_{y=1}^{n} m_{xy}(u) < 1$ for all $x \in X$ and $u \in U(x)$. The equation $J = TJ$ is Bellman's equation for a continuous-time semi-Markov decision problem, after it is converted into an equivalent discrete-time problem (cf. Section 1.4).

Example 1.6.3 (Minimax Problems)

Consider a minimax version of Example 1.6.1, where an antagonistic player chooses w from a set $W(x,u)$, and let

$$H(x,u,J) = \max_{w \in W(x,u)} \big[g(x,u,w) + \alpha J\big(f(x,u,w)\big)\big].$$

Then the equation $J = TJ$ is Bellman's equation for an infinite horizon version of the minimax control problem discussed in Section 1.6 of Vol. I.

Example 1.6.4 (Deterministic and Stochastic Shortest Path Problems)

Consider a classical deterministic shortest path problem (cf. Vol. I, Chapter 2) involving a graph of n nodes $x = 1, \ldots, n$, plus a destination t, an arc length a_{xu} for each arc (x, u), and the mapping

$$H(x, u, J) = \begin{cases} a_{xu} + J(u) & \text{if } u \neq t, \\ a_{xt} & \text{if } u = t. \end{cases}$$

Then the equation $J = TJ$ is Bellman's equation for the shortest distances $J^*(x)$ from the nodes x to node t.

A generalization is a mapping of the form

$$H(x, u, J) = p_{xt}(u) g(x, u, t) + \sum_{y=1}^{n} p_{xy}(u) \big(g(x, u, y) + J(y) \big).$$

It corresponds to a stochastic shortest path problem, which was discussed in Section 5.2 of Vol. I and will be considered again in Chapter 3. A special case is stochastic finite-horizon, finite-state DP problems.

Much of the theory of Sections 1.2 and 1.5 can be extended to the more abstract framework of this section. In particular, for a function $v : X \mapsto \Re$ with

$$v(x) > 0, \qquad \forall\, x \in X,$$

let us denote by $B(X)$ the space of real-valued functions J on X such that $J(x)/v(x)$ is bounded as x ranges over X, and as in Section 1.5, consider the weighted sup-norm

$$\|J\| = \max_{x \in X} \frac{\big|J(x)\big|}{v(x)}$$

on $B(X)$. We introduce the following assumption.

Assumption 1.6.1: (Contraction) For all $J \in B(X)$ and $\mu \in \mathcal{M}$, the functions $T_\mu J$ and TJ belong to $B(X)$. Furthermore, for some $\alpha \in (0, 1)$, we have

$$\|T_\mu J - T_\mu J'\| \leq \alpha \|J - J'\|, \qquad \forall\, J, J' \in B(X),\ \mu \in \mathcal{M}. \tag{1.57}$$

An equivalent way to state the condition (1.57) is

$$\frac{\big|H(x, u, J) - H(x, u, J')\big|}{v(x)} \leq \alpha \|J - J'\|,\ \ \forall\, x \in X,\ u \in U(x),\ J, J' \in B(X).$$

Note that Eq. (1.57) implies that

$$\|TJ - TJ'\| \leq \alpha \|J - J'\|, \qquad \forall\ J, J' \in B(X). \tag{1.58}$$

To see this we write

$$(T_\mu J)(x) \leq (T_\mu J')(x) + \alpha \|J - J'\|\, v(x), \qquad \forall\ x \in X,$$

from which, by taking infimum of both sides over $\mu \in \mathcal{M}$, we have

$$\frac{(TJ)(x) - (TJ')(x)}{v(x)} \leq \alpha \|J - J'\|, \qquad \forall\ x \in X.$$

Reversing the roles of J and J', we also have

$$\frac{(TJ')(x) - (TJ)(x)}{v(x)} \leq \alpha \|J - J'\|, \qquad \forall\ x \in X,$$

and combining the preceding two relations, and taking the supremum of the left side over $x \in X$, we obtain Eq. (1.58).

It can be seen that the Contraction Assumption 1.6.1 is satisfied for the mapping H in Examples 1.6.1-1.6.3, with v equal to the unit function e, i.e., $v(x) \equiv 1$. Generally, the assumption is not satisfied in Example 1.6.4, but we will see in Chapter 3 that it is satisfied for the special case of the stochastic shortest path problem of Section 5.2 of Vol. I. In that case, however, we cannot take $v(x) \equiv 1$, and this is one of our main motivations for considering the more general case where $v \neq e$.

The next two examples show how starting with mappings satisfying the contraction assumption, we can obtain multistep mappings with the same fixed points and a stronger contraction modulus. For any $J \in R(X)$, we denote by $T_{\mu_0} \cdots T_{\mu_k} J$ the composition of the mappings $T_{\mu_0}, \ldots, T_{\mu_k}$ applied to J, i.e,

$$T_{\mu_0} \cdots T_{\mu_k} J = \big(T_{\mu_0}\big(T_{\mu_1} \cdots (T_{\mu_{k-1}}(T_{\mu_k} J)) \cdots \big)\big).$$

Example 1.6.5 (Multistep Mappings)

Consider a set of mappings $T_\mu : \Re^n \mapsto \Re^n$, $\mu \in \mathcal{M}$, satisfying Assumption 1.6.1, let m be a positive integer, and let $\overline{\mathcal{M}}$ be the set of m-tuples $\nu = (\mu_0, \ldots, \mu_{m-1})$, where $\mu_k \in \mathcal{M}$, $k = 1, \ldots, m-1$. For each $\nu = (\mu_0, \ldots, \mu_{m-1}) \in \overline{\mathcal{M}}$, define the mapping $\overline{T}_\nu$, by

$$\overline{T}_\nu J = T_{\mu_0} \cdots T_{\mu_{m-1}} J, \qquad \forall\ J \in B(X).$$

Then we have the contraction properties

$$\|\overline{T}_\nu J - \overline{T}_\nu J'\| \leq \alpha^m \|J - J'\|, \qquad \forall\ J, J' \in B(X),$$

and

$$\|\overline{T}J - \overline{T}J'\| \leq \alpha^m \|J - J'\|, \qquad \forall\ J, J' \in B(X),$$

where $\overline{T}$ is defined by

$$(\overline{T}J)(x) = \inf_{(\mu_0,\ldots,\mu_{m-1})\in\overline{\mathcal{M}}} (T_{\mu_0}\cdots T_{\mu_{m-1}}J)(x), \qquad \forall\ J \in B(X),\ x \in X.$$

Thus the mappings $\overline{T}_\nu$, $\nu \in \overline{\mathcal{M}}$, satisfy Assumption 1.6.1, and have contraction modulus α^m.

The following example considers mappings underlying weighted Bellman equations that arise in various computational contexts in approximate DP, and will be encountered in Chapters 6 and 7.

Example 1.6.6 (Weighted Multistep Mappings)

Consider a set of mappings $L_\mu : B(X) \mapsto B(X)$, $\mu \in \mathcal{M}$, satisfying Assumption 1.6.1, i.e., for some $\alpha \in (0,1)$,

$$\|L_\mu J - L_\mu J'\| \leq \alpha \|J - J'\|, \qquad \forall\ J, J' \in B(X),\ \mu \in \mathcal{M}.$$

Consider also the mappings $T_\mu : B(X) \mapsto B(X)$ defined by

$$(T_\mu J)(x) = \sum_{\ell=1}^{\infty} w_\ell(x)(L_\mu^\ell J)(x), \qquad x \in X,\ J \in \Re^n,$$

where $w_\ell(x)$ are nonnegative scalars such that for all $x \in X$,

$$\sum_{\ell=1}^{\infty} w_\ell(x) = 1.$$

Then it follows that

$$\|T_\mu J - T_\mu J'\| \leq \sum_{\ell=1}^{\infty} w_\ell(x)\alpha^\ell \|J - J'\|,$$

showing that T_μ is a contraction with modulus

$$\bar{\alpha} = \max_{x\in X} \sum_{\ell=1}^{\infty} w_\ell(x)\,\alpha^\ell \leq \alpha.$$

Moreover L_μ and T_μ have a common fixed point for all $\mu \in \mathcal{M}$, and the same is true for the corresponding mappings L and T.

We will now consider some general questions, first under the Contraction Assumption 1.6.1, and then under an additional monotonicity assumption.

1.6.1 Basic Results Under Contraction and Monotonicity

The contraction property of T_μ and T together with the theory of Section 1.5 can be used to show the following proposition.

Proposition 1.6.1: Let Assumption 1.6.1 hold. Then:

(a) The mappings T_μ and T are contraction mappings with modulus α over $B(X)$, and have unique fixed points in $B(X)$, denoted J_μ and J^*, respectively.

(b) For any $J \in B(X)$ and $\mu \in \mathcal{M}$,

$$\lim_{k\to\infty} \|J_\mu - T_\mu^k J\| = 0, \qquad \lim_{k\to\infty} \|J^* - T^k J\| = 0.$$

(c) We have $T_\mu J^* = TJ^*$ if and only if $J_\mu = J^*$.

(d) For any $J \in B(X)$,

$$\|J^* - J\| \le \frac{1}{1-\alpha}\|TJ - J\|, \qquad \|J^* - TJ\| \le \frac{\alpha}{1-\alpha}\|TJ - J\|.$$

(e) For any $J \in B(X)$ and $\mu \in \mathcal{M}$,

$$\|J_\mu - J\| \le \frac{1}{1-\alpha}\|T_\mu J - J\|, \qquad \|J_\mu - T_\mu J\| \le \frac{\alpha}{1-\alpha}\|T_\mu J - J\|.$$

Proof: We have already shown that T_μ and T are contractions with modulus α over $B(X)$ [cf. Eqs. (1.57) and (1.58)]. Parts (a) and (b) follow from Prop. 1.5.1. To show part (c), note that if $T_\mu J^* = TJ^*$, then in view of $TJ^* = J^*$, we have $T_\mu J^* = J^*$, which implies that $J^* = J_\mu$, since J_μ is the unique fixed point of T_μ. Conversely, if $J_\mu = J^*$, we have $T_\mu J^* = T_\mu J_\mu = J_\mu = J^* = TJ^*$.

To show part (d), we use the triangle inequality to write for every k,

$$\|T^k J - J\| \le \sum_{\ell=1}^{k} \|T^\ell J - T^{\ell-1} J\| \le \sum_{\ell=1}^{k} \alpha^{\ell-1}\|TJ - J\|.$$

Taking the limit as $k \to \infty$ and using part (b), the left-hand side inequality follows. The right-hand side inequality follows from the left-hand side and the contraction property of T. The proof of part (e) is similar to part (d) [indeed it is the special case of part (d) where T is equal to T_μ, i.e., when $U(x) = \big\{\mu(x)\big\}$ for all $x \in X$]. **Q.E.D.**

Part (c) of the preceding proposition shows that there exists a $\mu \in \mathcal{M}$ such that $J_\mu = J^*$ if and only if the minimum of $H(x, u, J^*)$ over $U(x)$ is attained for all $x \in X$. Of course the minimum is attained if $U(x)$ is finite for every x, but otherwise this is not guaranteed in the absence of additional assumptions. Part (d) provides a useful error bound: we can evaluate the proximity of any function $J \in B(X)$ to the fixed point J^* by applying T to J and computing $\|TJ - J\|$. The left-hand side inequality of part (e) (with $J = J^*$) shows that for every $\epsilon > 0$, there exists a $\mu_\epsilon \in \mathcal{M}$ such that $\|J_{\mu_\epsilon} - J^*\| \leq \epsilon$, which may be obtained by letting $\mu_\epsilon(x)$ minimize $H(x, u, J^*)$ over $U(x)$ within an error of $(1-\alpha)\epsilon\, v(x)$, for all $x \in X$.

The Role of Monotonicity

Our analysis so far in this section relies only on the contraction assumption. We have made no use of the monotonicity property of the DP models of this chapter (cf. Section 1.1.2). We now introduce a generalized form of this property.

Assumption 1.6.2: (Monotonicity) If $J, J' \in R(X)$ and $J \leq J'$, then

$$H(x, u, J) \leq H(x, u, J'), \qquad \forall\, x \in X,\ u \in U(x).$$

Note that the assumption is equivalent to

$$J \leq J' \quad \Rightarrow \quad T_\mu J \leq T_\mu J', \qquad \forall\, \mu \in \mathcal{M},$$

and implies that

$$J \leq J' \quad \Rightarrow \quad TJ \leq TJ'.$$

An important consequence of monotonicity of H, when it holds in addition to contraction, is that it implies an optimality property of J^*.

Proposition 1.6.2: Let Assumptions 1.6.1 and 1.6.2 hold. Then

$$J^*(x) = \min_{\mu \in \mathcal{M}} J_\mu(x), \qquad \forall\, x \in X.$$

Furthermore, for every $\epsilon > 0$, there exists $\mu_\epsilon \in \mathcal{M}$ such that

$$J^*(x) \leq J_{\mu_\epsilon}(x) \leq J^*(x) + \epsilon, \qquad \forall\, x \in X. \tag{1.59}$$

Proof: We note that the right-hand side of Eq. (1.59) holds by Prop. 1.6.1(e) (see the remark following its proof). Thus $\min_{\mu \in \mathcal{M}} J_\mu(x) \leq J^*(x)$

for all $x \in X$. To show the reverse inequality as well as the left-hand side of Eq. (1.59), we note that for all $\mu \in \mathcal{M}$, we have $TJ^* \leq T_\mu J^*$, and since $J^* = TJ^*$, it follows that $J^* \leq T_\mu J^*$. By applying repeatedly T_μ to both sides of this inequality and by using the Monotonicity Assumption 1.6.2, we obtain $J^* \leq T_\mu^k J^*$ for all $k > 0$. Taking the limit as $k \to \infty$, we see that $J^* \leq J_\mu$ for all $\mu \in \mathcal{M}$. **Q.E.D.**

Note that without monotonicity, we may have $\min_{\mu \in \mathcal{M}} J_\mu(x) < J^*(x)$ for some x. This is illustrated by the following example.

Example 1.6.7 (Counterexample without Monotonicity)

Let $X = \{x_1, x_2\}$, $U = \{u_1, u_2\}$, and let

$$H(x_1, u, J) = \begin{cases} -\alpha J(x_2) & \text{if } u = u_1, \\ -1 + \alpha J(x_1) & \text{if } u = u_2, \end{cases} \qquad H(x_2, u, J) = \begin{cases} 0 & \text{if } u = u_1, \\ B & \text{if } u = u_2, \end{cases}$$

where B is a positive scalar. Then it can be seen that

$$J^*(x_1) = -\frac{1}{1-\alpha}, \qquad J^*(x_2) = 0,$$

and $J_{\mu^*} = J^*$ where $\mu^*(x_1) = u_2$ and $\mu^*(x_2) = u_1$. On the other hand, for $\mu(x_1) = u_1$ and $\mu(x_2) = u_2$, we have $J_\mu(x_1) = -\alpha B$ and $J_\mu(x_2) = B$, so $J_\mu(x_1) < J^*(x_1)$ for B sufficiently large.

Propositions 1.6.1 and 1.6.2 collectively address the problem of finding $\mu \in \mathcal{M}$ that minimizes $J_\mu(x)$ simultaneously for all $x \in X$, consistently with DP theory. The optimal value of this problem is $J^*(x)$, and μ is optimal for all x if and only if $T_\mu J^* = TJ^*$. For this we just need the contraction and monotonicity assumptions. We do not need any additional structure of H, such as for example a discrete-time dynamic system, transition probabilities, etc. While identifying the proper structure of H, and verifying its contraction and monotonicity may require some analysis that is specific to each type of problem, once this is done significant results are obtained quickly.

Nonstationary Policies

The connection with DP motivates us to consider the set Π of all sequences $\pi = \{\mu_0, \mu_1, \ldots\}$ with $\mu_k \in \mathcal{M}$ for all k (nonstationary policies in the DP context), and define

$$J_\pi(x) = \liminf_{k \to \infty} (T_{\mu_0} \cdots T_{\mu_k} J)(x), \qquad \forall\, x \in X,$$

with J being any function in $R(X)$, where $T_{\mu_0} \cdots T_{\mu_k} J$ denotes the composition of the mappings $T_{\mu_0}, \ldots, T_{\mu_k}$ applied to J, i.e,

$$T_{\mu_0} \cdots T_{\mu_k} J = \big(T_{\mu_0}\big(T_{\mu_1} \cdots (T_{\mu_{k-1}}(T_{\mu_k} J)) \cdots\big)\big).$$

Note that the choice of J in the definition of J_π does not matter since for any two $J, J' \in B(X)$, we have from the Contraction Assumption 1.6.1,

$$\|T_{\mu_0}T_{\mu_1}\cdots T_{\mu_k}J - T_{\mu_0}T_{\mu_1}\cdots T_{\mu_k}J'\| \le \alpha^{k+1}\|J - J'\|,$$

so the value of $J_\pi(x)$ is independent of J. Since by Prop. 1.6.1(b), $J_\mu(x) = \lim_{k\to\infty}(T_\mu^k J)(x)$ for all $\mu \in \mathcal{M}$, $J \in B(X)$, and $x \in X$, in the DP context we recognize J_μ as the cost function of the stationary policy $\{\mu, \mu, \ldots\}$.

We now claim that under our Assumptions 1.6.1 and 1.6.2, J^*, the fixed point of T, is equal to the optimal value of J_π, i.e.,

$$J^*(x) = \min_{\pi\in\Pi} J_\pi(x), \qquad \forall\ x \in X.$$

Indeed, since $\mathcal{M}$ defines a subset of Π, we have from Prop. 1.6.2,

$$J^*(x) = \min_{\mu\in\mathcal{M}} J_\mu(x) \ge \min_{\pi\in\Pi} J_\pi(x), \qquad \forall\ x \in X,$$

while for every $\pi \in \Pi$ and $x \in X$, we have

$$J_\pi(x) = \liminf_{k\to\infty}(T_{\mu_0}T_{\mu_1}\cdots T_{\mu_k}J)(x) \ge \lim_{k\to\infty}(T^{k+1}J)(x) = J^*(x)$$

[the Monotonicity Assumption 1.6.2 can be used to show that

$$T_{\mu_0}T_{\mu_1}\cdots T_{\mu_k}J \ge T^{k+1}J,$$

and the last equality holds by Prop. 1.6.1(b)]. Combining the preceding relations, we obtain $J^*(x) = \min_{\pi\in\Pi} J_\pi(x)$.

Thus, in DP terms, we may view J^* as an optimal cost function over all policies. At the same time, Prop. 1.6.2 states that stationary policies are sufficient in the sense that the optimal cost can be attained to within arbitrary accuracy with a stationary policy [uniformly for all $x \in X$, as Eq. (1.59) shows].

Periodic Policies

Consider the multistep mappings $\overline{T}_\nu = T_{\mu_0}\cdots T_{\mu_{m-1}}$, $\nu \in \overline{\mathcal{M}}$, defined in Example 1.6.5, where $\overline{\mathcal{M}}$ is the set of m-tuples $\nu = (\mu_0, \ldots, \mu_{m-1})$, with $\mu_k \in \mathcal{M}$, $k = 1, \ldots, m-1$, and m is a positive integer. Assuming that the mappings T_μ satisfy Assumptions 1.6.1 and 1.6.2, the same is true for the mappings $\overline{T}_\nu$ (with the contraction modulus of $\overline{T}_\nu$ being α^m). Thus the unique fixed point of $\overline{T}_\nu$ is J_π, where π is the nonstationary but periodic policy

$$\pi = \{\mu_0, \ldots, \mu_{m-1}, \mu_0, \ldots, \mu_{m-1}, \ldots\}.$$

Moreover it can be seen that the mappings $T_{\mu_0}\cdots T_{\mu_{m-1}}$, $T_{\mu_1}\cdots T_{\mu_{m-1}}T_{\mu_0}$, $\ldots$, $T_{\mu_{m-1}}T_{\mu_0}\cdots T_{\mu_{m-2}}$, have unique corresponding fixed points $J_0, J_1, \ldots$, J_{m-1}, which satisfy

$$J_0 = T_{\mu_0}J_1,\ \ J_1 = T_{\mu_1}J_2,\ \ \ldots\ \ J_{\mu_{m-2}} = T_{\mu_{m-2}}J_{\mu_{m-1}},\ \ J_{\mu_{m-1}} = T_{\mu_{m-1}}J_0.$$

To verify these equations, multiply the fixed point relation

$$J_1 = T_{\mu_1}\cdots T_{\mu_{m-1}}T_{\mu_0}J_1$$

with T_{μ_0} to show that $T_{\mu_0}J_1$ is the fixed point of $T_{\mu_0}\cdots T_{\mu_{m-1}}$, i.e., is equal to J_0, etc. Note that even though $\overline{T}_\nu$ defines the cost functions of periodic policies, $\overline{T}$ has the same fixed point as T, namely J^*. This gives rise to the computational possibility of working with $\overline{T}_\nu$ in place of T_μ in an effort to approximate J^*. We will later discuss situations where this may be advantageous.

Error Bounds Under Monotonicity

The assumptions of contraction and monotonicity together can be characterized in a form that is useful for analysis. This form is reminiscent of the Constant Shift Lemma 1.1.2, and is given in the following proposition.

Proposition 1.6.3: (Weighted Shift Property) The Contraction and Monotonicity Assumptions 1.6.1 and 1.6.2 hold if and only if for all $J, J' \in B(X)$, $\mu \in \mathcal{M}$, and scalar $c \geq 0$, we have

$$J \leq J' + c\,v \quad \Rightarrow \quad T_\mu J \leq T_\mu J' + \alpha c\,v, \tag{1.60}$$

where v is the weight function of the weighted sup-norm $\|\cdot\|$.

Proof: Let the contraction and monotonicity assumptions hold. If $J' \leq J + c\,v$, we have

$$H(x,u,J') \leq H(x,u,J+c\,v) \leq H(x,u,J) + \alpha c\,v(x), \quad \forall\ x \in X,\ u \in U(x), \tag{1.61}$$

where the left-side inequality follows from the monotonicity assumption and the right-side inequality follows from the contraction assumption, which together with $\|v\| = 1$, implies that

$$\frac{H(x,u,J+c\,v) - H(x,u,J)}{v(x)} \leq \alpha\|J + c\,v - J\| = \alpha c.$$

The condition (1.61) implies the desired condition (1.60). Conversely, condition (1.60) for $c = 0$ yields the monotonicity assumption, while for $c = \|J' - J\|$ it yields the contraction assumption. **Q.E.D.**

We can use Prop. 1.6.3 to derive some useful variants of the bounds of parts (d) and (e) of Prop. 1.6.1 (which assumes only the contraction assumption). These variants will be used in the derivation of error bounds for various computational methods in Chapter 2.

Proposition 1.6.4: (Error Bounds Under Contraction and Monotonicity) Let Assumptions 1.6.1 and 1.6.2 hold.

(a) For any $J \in B(X)$ and $c \geq 0$, we have

$$TJ \leq J + c\,v \quad \Rightarrow \quad J^* \leq J + \frac{c}{1-\alpha}v,$$

$$J \leq TJ + c\,v \quad \Rightarrow \quad J \leq J^* + \frac{c}{1-\alpha}v.$$

(b) For any $J \in B(X)$, $\mu \in \mathcal{M}$, and $c \geq 0$, we have

$$T_\mu J \leq J + c\,v \quad \Rightarrow \quad J_\mu \leq J + \frac{c}{1-\alpha}v,$$

$$J \leq T_\mu J + c\,v \quad \Rightarrow \quad J \leq J_\mu + \frac{c}{1-\alpha}v.$$

(c) For all $J \in B(X)$, $c \geq 0$, and $k = 0, 1, \ldots$, we have

$$TJ \leq J + c\,v \qquad \Rightarrow \qquad J^* \leq T^k J + \frac{\alpha^k c}{1-\alpha}v,$$

$$J \leq TJ + c\,v \qquad \Rightarrow \qquad T^k J \leq J^* + \frac{\alpha^k c}{1-\alpha}v.$$

Proof: (a) We show the first relation. Applying Eq. (1.60) with J' and J replaced by J and TJ, respectively, and taking minimum over $u \in U(x)$ for all $x \in X$, we see that if $TJ \leq J + c\,v$, then $T^2 J \leq TJ + \alpha c\,v$. Proceeding similarly, it follows that

$$T^\ell J \leq T^{\ell-1} J + \alpha^{\ell-1} c\,v.$$

We now write for every k,

$$T^k J - J = \sum_{\ell=1}^{k} (T^\ell J - T^{\ell-1} J) \leq \sum_{\ell=1}^{k} \alpha^{\ell-1} c\,v,$$

from which, by taking the limit as $k \to \infty$, we obtain $J^* \leq J + \big(c/(1-\alpha)\big)v$. The second relation follows similarly.

(b) This part is the special case of part (a) where T is equal to T_μ.

(c) We show the first relation. From part (a), the inequality $TJ \leq J + c\,v$ implies that

$$J^* \leq J + \frac{c}{1-\alpha} v.$$

Applying T^k to both sides of this inequality, and using the monotonicity and fixed point property of T^k, we have

$$J^* \leq T^k \left(J + \frac{c}{1-\alpha} v \right).$$

Using Eq. (1.60) with T_μ and α replaced by T^k and α^k, respectively, we obtain

$$T^k \left(J + \frac{c}{1-\alpha} v \right) \leq T^k J + \frac{\alpha^k c}{1-\alpha} v,$$

and the first relation to be shown follows from the preceding two relations. The second relation follows similarly. **Q.E.D.**

1.6.2 Discounted Dynamic Games

We will now discuss an application of the preceding framework to zero sum games. In the simplest such game there are two players that choose actions just once: the first (called the *minimizer*) may choose a move i out of n moves and the second (called the *maximizer*) may choose a move j out of m moves. Then the minimizer gives a specified amount a_{ij} to the maximizer, called a *payoff*. The minimizer wishes to minimize a_{ij}, and the maximizer wishes to maximize a_{ij}.

The players use mixed strategies, whereby the minimizer selects a probability distribution $u = (u_1, \ldots, u_n)$ over his n possible moves and the maximizer selects a probability distribution $v = (v_1, \ldots, v_m)$ over his m possible moves. Since the probability of selecting i and j is $u_i v_j$, the expected payoff is $\sum_{i,j} a_{ij} u_i v_j$ or $u'Av$, where A is the $n \times m$ matrix with components a_{ij}. If each player adopts a worst case viewpoint, whereby he optimizes his choice against the worst possible selection by the other player, the minimizer must minimize $\max_{v \in V} u'Av$ and the maximizer must maximize $\min_{u \in U} u'Av$, where U and V are the sets of probability distributions over $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$, respectively. A fundamental result (which will not be proved here) is that these two optimal values are equal,

$$\min_{u \in U} \max_{v \in V} u'Av = \max_{v \in V} \min_{u \in U} u'Av, \tag{1.62}$$

implying that there is an amount that can be meaningfully viewed as the value of the game for its participants.

We will now consider a dynamic zero-sum game, where a separate game of the type just described is played at each stage. The game played at a given stage is represented by a "state" x that takes values in a finite set X. The state evolves according to transition probabilities $q_{xy}(i,j)$ where i and j are the moves selected by the minimizer and the maximizer, respectively (here y represents the next game to be played after moves i and j are chosen at the game represented by x). When the state is x, under $u \in U$ and $v \in V$, the one-stage expected payoff is $u'A(x)v$, where $A(x)$ is the $n \times m$ payoff matrix, and the state transition probabilities are

$$p_{xy}(u,v) = \sum_{i=1}^{n} \sum_{j=1}^{m} u_i v_j q_{xy}(i,j) = u' Q_{xy} v,$$

where Q_{xy} is the $n \times m$ matrix that has components $q_{xy}(i,j)$. Payoffs are discounted by $\alpha \in (0,1)$, and the objectives of the minimizer and maximizer, roughly speaking, are to minimize and to maximize the total discounted expected payoff.

We introduce the mappings G and H given by

$$\begin{aligned} G(x,u,v,J) &= u'A(x)v + \alpha \sum_{y \in X} p_{xy}(u,v) J(y) \\ &= u' \left(A(x) + \alpha \sum_{y \in X} Q_{xy} J(y) \right) v, \end{aligned} \tag{1.63}$$

$$H(x,u,J) = \max_{v \in V} G(x,u,v,J).$$

It can be verified that H satisfies the Contraction Assumption 1.6.1 (with $v(x) \equiv 1$) and the Monotonicity Assumption 1.6.2, so Props. 1.6.1 and 1.6.2 apply. Thus the corresponding mapping T is an unweighted sup-norm contraction, and its unique fixed point J^* satisfies

$$J^*(x) = \min_{u \in U} \max_{v \in V} G(x,u,v,J^*), \qquad \forall\ x \in X.$$

We now note that since

$$A(x) + \alpha \sum_{y \in X} Q_{xy} J(y)$$

[cf. Eq. (1.63)] is a matrix that is independent of u and v, we may view $J^*(x)$ as the value of a static game the depends on the state x. In particular, from the fundamental minimax equality (1.62), we have

$$\min_{u \in U} \max_{v \in V} G(x,u,v,J^*) = \max_{v \in V} \min_{u \in U} G(x,u,v,J^*), \qquad \forall\ x \in X.$$

This implies that J^* is also the unique fixed point of the mapping

$$(\overline{T}J)(x) = \max_{v \in V} \overline{H}(x,v,J),$$

where

$$\overline{H}(x, v, J) = \min_{u \in U} G(x, u, v, J),$$

i.e., J^* is the fixed point regardless of the order in which minimizer and maximizer select mixed strategies at each stage.

There is another interpretation of $J^*(x)$ as the value of a game where the players choose policies μ and ν rather than (single-stage) moves u and v. This interpretation requires additional analysis and will be described only briefly. For given x, we may view $J^*(x)$ as the best possible payoff that the minimizer (or maximizer) can achieve starting from x and using a policy $\mu : X \mapsto U$ (or $\nu : X \mapsto V$, respectively) against the worst possible policy choice of the maximizer (or minimizer, respectively). More specifically, fix a policy μ of the minimizer, and consider the discounted DP problem of maximizing the expected payoff of the maximizer by optimal choice of a policy ν. Then it can be shown that J_μ is the maximal value function of this DP problem, and $J^* = \min_{\mu \in \mathcal{M}} J_\mu$. Similarly, by reordering minimization and maximization, $J^* = \max_{\nu \in \mathcal{N}} \overline{J}_\nu$, where $\mathcal{N}$ is the set of policies of the maximizer, and for fixed ν, $\overline{J}_\nu$ is the optimal cost function of the discounted DP problem of minimizing the expected payoff by optimal choice of $\mu \in \mathcal{M}$.

1.7 NOTES, SOURCES, AND EXERCISES

Sections 1.1-1.2: Many authors have contributed to the analysis of the discounted problem with bounded cost per stage, most notably Shapley [Sha53], Bellman [Bel57], and Blackwell [Bla65a]. 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].

In this book, for mathematical rigor, we have assumed a countable disturbance space. However, our analysis may still serve as the starting point of the mathematical treatment 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 illustrated in Exercise 1.5. This line of analysis was adopted in the book [BeS78] (Chapter 9) for the resolution of measurability questions in infinite horizon stochastic control problems.

Section 1.3: 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], who simplified the earlier proof by Whittle [Whi80b]. Another simple proof for the case of a finite state space was given by Tsitsiklis [Tsi94a], following an earlier proof by Weiss [Wei88]. For additional analysis of the multiarmed bandit problem, see Kumar [Kum85], Varaiya, Walrand, and Buyukkoc [VWB85], Kumar and Varaiya [KuV86], Nain, Tsoucas, and Walrand [NTW89], Weber [Web92], Bertsimas and Nino-Mora [BeN96], and Bertsimas, Paschalidis, and Tsitsiklis [BPT94a], [BPT94b].

Section 1.4: 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]. Semi-Markov decision models were introduced by Jewell [Jew63] and are also discussed by Ross [Ros70].

Section 1.5: The role of contraction mappings in discounted problems was first recognized and exploited by Shapley [Sha53], who considered two-player dynamic games. Countable-state discounted problems with unbounded cost per stage (cf. Section 1.5.2) were discussed by Harrison [Har72], Lippman [Lip73], [Lip75a], van Nunen [Van76], Wessels [Wes77], van Nunen and Wessels [VaW78], and Cavazos-Cadena [Cav86].

Section 1.6: Abstract DP models under unweighted sup-norm contraction assumptions were introduced by Denardo [Den67]. Our treatment here, extends the theory to weighted sup-norm contractions, and was given in the author's survey paper [Ber12]. Abstract DP models were investigated by the author [Ber77] in the absence of contraction properties, relying only on the type of monotonicity properties that are common in DP, and they were used extensively in the book [BeS78], Chapters 2-6. Abstract models also served as the basis for the development of an associated asynchronous distributed DP framework developed in the author's paper [Ber82a], which we will use on several occasions in this book, starting with Section 2.6 in the next chapter. For related work, see Verd'u and Poor [VeP84], [VeP87]. The author's monograph [Ber18] provides an extensive treatment of abstract models and includes several advanced applications to stochastic optimal control, beyond the ones discussed in Chapters 3 and 4.

EXERCISES

1.1

The purpose of this exercise is to show that shortest path problems with a discount factor may 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 $(i_0, i_1, \ldots, i_m)$ 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.

1.2

Consider a problem similar to that of Section 1.1 except that when we are at state x_k, there is a probability $\beta \in (0, 1)$ 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. What happens if β is replaced by a state-dependent probability $\beta_x \in (0, 1)$?

1.3 (Column Reduction [Por75])

The purpose of this exercise 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 where $U(i)$ is a finite set for all states i. The cost per stage is $g(i, u)$, the discount factor is α, and the transition probabilities are $p_{ij}(u)$. For each $j = 1, \ldots, n$, let

$$m_j = \min_{i=1,\ldots,n} \min_{u \in U(i)} p_{ij}(u).$$

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.

1.4 (Data Transformations [Sch72])

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 $J(i)$:

$$J(i) = \min_{u \in U(i)} \left[g(i,u) + \sum_{j=1}^{n} m_{ij}(u) J(j) \right], \qquad i = 1, \ldots, n, \tag{1.64}$$

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

$$\alpha = \max_{i=1,\ldots n} \left\{ \frac{M_i(u) - m_{ii}(u)}{1 - m_{ii}(u)} \right\},$$

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases}$$

and define, for all i and j,

$$\overline{g}(i,u) = \frac{g(i,u)(1-\alpha)}{1 - M_i(u)},$$

$$\overline{m}_{ij}(u) = \delta_{ij} + \frac{(1-\alpha)(m_{ij}(u) - \delta_{ij})}{1 - M_i(u)}.$$

Show that for all i and j,

$$\sum_{j=1}^{n} \overline{m}_{ij}(u) = \alpha < 1, \qquad \overline{m}_{ij}(u) \geq 0,$$

and that a solution $\{J(i) \mid i = 1, \ldots, n\}$ of Eq. (1.64) is also a solution of the equations

$$J(i) = \min_{u \in U(i)} \left[\overline{g}(i,u) + \alpha \sum_{j=1}^{n} \overline{p}_{ij}(u) J(j) \right], \qquad i = 1, \ldots, n,$$

where $\overline{p}_{ij}(u)$ are the transitions probabilities defined by

$$\overline{p}_{ij}(u) = \frac{\overline{m}_{ij}(u)}{\alpha}.$$

1.5 (Stochastic to Deterministic Problem Transformation)

Consider the controlled system

$$p_{k+1} = p_k P_{\mu_k}, \qquad k = 0, 1, \ldots,$$

where p_k is a probability distribution over X 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

$$\lim_{N\to\infty} \sum_{k=0}^{N-1} \alpha^k p_k g_{\mu_k}.$$

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 of a corresponding discounted cost problem.

1.6

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 $J^*(x_A, x_B)$.

1.7 (The Tax Problem [VWB85])

This problem is similar to the multiarmed bandit problem. The only difference is that if we engage project ℓ at period k, we pay a tax $\alpha^k C^{\bar{\ell}}(x^{\bar{\ell}})$ for every other project $\bar{\ell}$ [for a total of $\alpha^k \sum_{\bar{\ell}\neq\ell} C^{\bar{\ell}}(x^{\bar{\ell}})$], instead of earning a reward $\alpha^k R^\ell(x^\ell)$. 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 ℓ equal to

$$R^\ell(x^\ell) = C^\ell(x^\ell) - \alpha E\big\{C^\ell\big(f^\ell(x^\ell, w^\ell)\big)\big\}.$$

1.8 (The Restart Problem [KaV87])

The purpose of this exercise is to show that the index of a project in the multiarmed 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

$$J(x) = \max\big[R(x_0) + \alpha E\big\{J\big(f(x_0, w)\big)\big\},\ R(x) + \alpha E\big\{J\big(f(x, w)\big)\big\}\big].$$

1.9 (Multiarmed Bandit Problems and Separable Approximations)

Consider the multiarmed bandit problem of Section 1.3, but with two differences:

(1) When a project ℓ is not worked on, its state changes according to

$$x^\ell_{k+1} = \bar f^\ell(x^\ell_k, \overline{w}^\ell_k),$$

where $\bar f^\ell$ is a given function and $\overline{w}^\ell_k$ is a random disturbance with distribution depending on x^ℓ_k but not on prior disturbances. Furthermore, a reward $\overline{R}^\ell(x^\ell_k)$ is earned, where $\bar R^\ell$ 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.)

Suppose that the optimal reward function $J^*(x^1, \ldots, x^n)$ is approximated by a separable function of the form $\sum_{\ell=1}^n \tilde J^\ell(x^\ell)$, where each $\tilde J^\ell$ is a function corresponding to the contribution of the ℓth project to the total reward. The corresponding one-step lookahead policy selects the project ℓ that maximizes

$$R^\ell(x^\ell) + \sum_{j\neq\ell} \bar R^j(x^j) + \alpha E\big\{\tilde J^\ell\big(f^\ell(x^\ell, w^\ell)\big)\big\} + \alpha \sum_{j\neq\ell} E\left\{\tilde J^j\left(\bar f^j(x^j, \overline{w}^j)\right)\right\}.$$

Show that this policy takes the form

$$\text{work on project } \ell \qquad \text{if} \qquad \tilde m^\ell(x^\ell) = \max_j \big\{\tilde m^j(x^j)\big\},$$

where for all ℓ,

$$\tilde m^\ell(x^\ell) = R^\ell(x^\ell) - \bar R^\ell(x^\ell) + \alpha E\big\{\tilde J^\ell\big(f^\ell(x^\ell, w^\ell)\big) - \tilde J^\ell\big(\bar f^\ell(x^\ell, \overline{w}^\ell)\big)\big\}.$$

Thus, we may view $\tilde m^\ell(x^\ell)$ as an approximate index for project ℓ, induced by the separable reward function approximation $\sum_{\ell=1}^n \tilde J^\ell(x^\ell)$.

1.10 (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 $i = 1, \ldots, n$. 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 components

$$p_i(t) = P\big\{i(t) = i \mid x_0\big\}, \qquad i = 1, \ldots, n.$$

We have

$$dp(t)/dt = p(t)A,$$

where $p(0)$ is the row vector with ith component 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 $\nu \geq \nu_i$, $i = 1, \ldots, n$. 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 $i = 1, \ldots, n$ we have

$$p_i(t) = \text{Prob}\big\{i(t) = i \text{ in the uniform Markov chain}\big\}.$$

1.11

A person has an asset to sell for which she receives offers that can take one of n given values. The offer values and the times between successive offers are random, independent, and identically distributed with given distributions. 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.

1.12

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

$$J_\pi(x_0) = \lim_{N\to\infty} \mathop{E}_{\substack{w_k\\ k=0,1,\ldots}} \left\{ \sum_{k=0}^{N-1} \alpha_{\pi,k}\, g\big(x_k, \mu_k(x_k), w_k\big) \right\},$$

where

$$\alpha_{\pi,k} = \alpha\big(x_0, \mu_0(x_0), w_0\big)\alpha\big(x_1, \mu_1(x_1), w_1\big)\cdots\alpha\big(x_k, \mu_k(x_k), w_k\big),$$

and $\alpha(x,u,w)$ is a given function satisfying

$$\begin{aligned} 0 &\le \min\big\{\alpha(x,u,w) \mid x\in X, u\in U, w\in W\big\}\\ &\le \max\big\{\alpha(x,u,w) \mid x\in X, u\in U, w\in W\big\}\\ &< 1. \end{aligned}$$

Use the analysis of Section 1.6 to provide counterparts of the results of Section 1.2.

1.13 (Minimax Problems)

Use the analysis of Section 1.6 to provide counterparts of the results of Section 1.2 for the minimax problem where the cost is

$$J_\pi(x_0) = \lim_{N\to\infty} \max_{\substack{w_k\in W(x_k,\mu_k(x_k))\\ k=0,1\ldots}} \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big),$$

g is bounded, x_k is generated by $x_{k+1} = f\big(x_k, \mu_k(x_k), w_k\big)$, and $W(x,u)$ is a given nonempty subset of W for each $(x,u) \in S\times U$. (Compare with Section 1.6 of Vol. I, and see the monograph [Ber18].)

1.14 (Infinite Horizon Formulation of Finite Horizon Problems)

Consider the N-stage basic problem of Chapter 1 of Vol. I, and the following special case of the abstract DP problem of Section 1.6. Let $X = X_0 \cup \cdots \cup X_N$, where $X_0, \ldots, X_N$ are the state spaces of the N-stage problem, and for $k \leq N-1$ and $x \in X_k$, let $U(x) = U_k(x)$. For $J : X \mapsto \Re$, denote by J_k the restriction of J on X_k, i.e., $J_k(x) = J(x)$ for $x \in X_k$, and $J = (J_0, \ldots, J_N)$. Define

$$H(x, u, J) = E\Big\{g_k(x, u, w_k) + J_{k+1}\big(f_k(x, u, w_k)\big)\Big\}$$

if $k \leq N-1$, $x \in X_k$, $u \in U_k(x)$, and

$$H(x, u, J) = g_N(x)$$

if $x \in X_N$. Let T be the mapping defined by

$$(TJ)(x) = \min_{u \in U(x)} H(x, u, J), \qquad x \in X.$$

(a) Show that the fixed point of T is $J^* = (J_0^*, \ldots, J_N^*)$, where J_k^* is the optimal cost-to-go function at stage k of the N-stage problem.

(b) Show that the finite horizon DP algorithm is equivalent to $N+1$ applications of T starting with any $J = (J_0, \ldots, J_N)$, or equivalently N applications of T starting with any $J = (J_0, \ldots, J_N)$ with $J_N = g_N$.

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.

[ABF93] 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.

[AMS07] Antos, A., Munos, R., and Szepesvari, C., 2007. “Fitted Q- Iteration in Continuous Action-Space MDPs,” Proc. of NIPS, pp. 9-16.

[ABJ06] Ahamed, T. P. I., Borkar, V. S., and Juneja, S., 2006. “Adaptive Importance Sampling Technique for Markov Chains Using Stochastic Approximation,” Operations Research, Vol. 54, pp. 489-504.

[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.

[ASM08] Antos, A., Szepesvari, C., and Munos, R., 2008. “Learning Near-Optimal Policies with Bellman-Residual Minimization Based Fitted Policy Iteration and a Single Sample Path,” Machine Learning, Vol. 71, pp. 89-129.

[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.

[AsG10] Asmussen, S., and Glynn, P. W., 2010. Stochastic Simulation: Algorithms and Analysis, Springer, N. Y.

[Ash70] Ash, R. B., 1970. Basic Probability Theory, Wiley, N. Y.

[Att03] Attias, H. 2003. "Planning by Probabilistic Inference," in C. M. Bishop and B. J. Frey, (Eds.), Proc. of the 9th Int. Workshop on Artificial Intelligence and Statistics.

[AyR91] Ayoun, S., and Rosberg, Z., 1991. "Optimal Routing to Two Parallel Heterogeneous Servers with Resequencing," IEEE Trans. on Aut. Control, Vol. 36, pp. 1436-1449.

[BBB08] Basu, A., Bhattacharyya, and Borkar, V., 2008. "A Learning Algorithm for Risk-Sensitive Cost," Math. of Operations Research, Vol. 33, pp. 880-898.

[BBD10] Busoniu, L., Babuska, R., De Schutter, B., and Ernst, D., 2010. Reinforcement Learning and Dynamic Programming Using Function Approximators, CRC Press, N. Y.

[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, and D. Wunsch, (Eds.), IEEE Press, N. Y.

[BBP11] Bhatnagar, S., Borkar, V. S., and Prashanth, L. A., 2011. "Adaptive Feature Pursuit: Online Adaptation of Features in Reinforcement Learning," appears in *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, by F. Lewis and D. Liu (eds.), IEEE Press, Computational Intelligence Series, 2012.

[BBS87] Bean, J. C., Birge, J. R., and Smith, R. L., 1987. "Aggregation in Dynamic Programming," Operations Research, Vol. 35, pp. 215-220.

[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 μc-Rule is Often Optimal," Report SRR 83-1, Department of Electrical Engineering, University of Maryland.

[BED09] Busoniu, L., Ernst, D., De Schutter, B., and Babuska, R., 2009. "Online Least-Squares Policy Iteration for Reinforcement Learning Control," unpublished report, Delft Univ. of Technology, Delft, NL.

[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.

[BKM05] de Boer, P. T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. 2005. "A Tutorial on the Cross-Entropy Method," Annals of Operations Research, Vol. 134, pp. 19-67.

[BMP90] Benveniste, A., Metivier, M., and Priouret, P., 1990. Adaptive Algorithms and Stochastic Approximations, Springer-Verlag, N. Y.

[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 Aut. Control, Vol. 40, 1995, pp. 2063-2075.

[BPW12] Browne, C., Powley, E., Whitehouse, D., Lucas, L., Cowling, P. I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S., 2012. "A Survey of Monte Carlo Tree Search Methods," IEEE Trans. on Computational Intellegence and AI in Games, Vol. 4, pp. 1-43.

[BSA83] Barto, A. G., Sutton, R. S., and Anderson, C. W., 1983. "Neuronlike Elements that Can Solve Difficult Learning Control Problems," IEEE Trans. on Systems, Man, and Cybernetics, Vol. 13, pp. 835-846.

[BaB01] Baxter, J., and Bartlett, P. L., 2001. "Infinite-Horizon Policy-Gradient Estimation," J. Artificial Intelligence Research, Vol. 15, pp. 319–350.

[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 Aut. Control, Vol. AC-34, pp. 589-598.

[BeG82] Bertsekas, D. P., and Gafni, E. M., 1982. "Projection Methods for Variational Inequalities with Application to the Traffic Assignment Problem," Mathematical Programming Study, Vol. 17, pp. 139-159.

[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.

[BeN96] Bertsimas, D., and Nino-Mora, J., 1996. "Conservation Laws, Extended Polymatroids, and the Multiarmed Bandit Problem: A Unified Polyhedral Approach," Mathematics of Operations Research, Vol. 21, pp. 257-306.

[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," Aut. a, 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.

[BeT08] Bertsekas, D. P., and Tsitsiklis, J. N., 2008. Introduction to Probability, (2nd Edition), Athena Scientific, Belmont, MA.

[BeY07] Bertsekas, D. P., and Yu, H., 2007. "Solution of Large Systems of Equations Using Approximate Dynamic Programming Methods," Lab. for Information and Decision Systems Report LIDS-P-2754, MIT.

[BeY09] Bertsekas, D. P., and Yu, H., 2009. "Projected Equation Methods for Approximate Solution of Large Linear Systems," J. of Computational and Applied Mathematics, Vol. 227, pp. 27-50.

[BeY10a] Bertsekas, D. P., and Yu, H., 2010. "Q-Learning and Enhanced Policy Iteration in Discounted Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-P-2831, MIT; Math. of Operations Research, Vol. 37, 2012, pp. 66-94.

[BeY10b] Bertsekas, D. P., and Yu, H., 2010. “Asynchronous Distributed Policy Iteration in Dynamic Programming,” Proc. of Allerton Conf. on Communication, Control and Computing, Allerton Park, Ill, pp. 1368-1374.

[BeY16] Bertsekas, D. P., and Yu, H., 2016. “Stochastic Shortest Path Problems Under Weak Conditions,” Lab. for Information and Decision Systems Report LIDS-2909, January 2016.

[Ber10c] Bertsekas, D. P., 2010. “Williams-Baird Counterexample for Q-factor Asynchronous Policy Iteration,” from the author’s website,
http://www.mit.edu/ dimitrib/publ.html.

[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. Aut. Control, Vol. AC-17, pp. 604-613.

[Ber73a] Bertsekas, D. P., 1973. “Stochastic Optimization Problems with Nondifferentiable 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. Aut. Control, Vol. AC-18, pp. 314-315.

[Ber75a] Bertsekas, D. P., 1975. “Convergence of Discretization Procedures in Dynamic Programming,” IEEE Trans. Aut. Control, Vol. AC-20, pp. 415-419.

[Ber75b] Bertsekas, D. P., 1975. “Monotone Mappings in Dynamic Programming,” 1975 IEEE Conference on Decision and Control, pp. 20-25.

[Ber76] Bertsekas, D. P., 1976. “On Error Bounds for Successive Approximation Methods,” IEEE Trans. Aut. 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. Aut. 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.

[Ber96] Bertsekas, D. P., 1996. Lecture at NSF Workshop on Reinforcement Learning, Hilltop House, Harper’s Ferry, N. Y.

[Ber97] Bertsekas, D. P., 1997. “Differential Training of Rollout Policies,” Proc. of the 35th Allerton Conference on Communication, Control, and Computing, Allerton Park, Ill.

[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.

[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 LIDS-P-2646, MIT.

[Ber09a] Bertsekas, D. P., 2009. Convex Optimization Theory, Athena Scientific, Belmont, MA.

[Ber09b] Bertsekas, D. P., 2009. “Projected Equations, Variational Inequalities, and Temporal Difference Methods,” Lab. for Information and Decision Systems Report LIDS-P-2808, MIT.

[Ber10a] Bertsekas, D. P., 2010. “Approximate Policy Iteration: A Survey and Some New Methods,” Lab. for Information and Decision Systems Report LIDS-P-2833, MIT; J. of Control Theory and Applications, Vol. 9, 2011, pp. 310-335.

[Ber10b] Bertsekas, D. P., 2010. “Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey,” Lab. for Information and Decision Systems Report LIDS-P-2848, MIT; also in “Optimization for Machine Learning,” by S. Sra, S. Nowozin, and S. J. Wright, MIT Press, Cambridge, MA, 2012, pp. 85-119.

[Ber10c] Bertsekas, D. P., 2010. “Williams-Baird Counterexample for Q-

Factor Asynchronous Policy Iteration,"
http://web.mit.edu/dimitrib/www/Williams-Baird Counterexample.pdf.

[Ber11a] Bertsekas, D. P., 2011. "Temporal Difference Methods for General Projected Equations," IEEE Trans. on Aut. Control, Vol. 56, pp. 2128-2139.

[Ber11b] Bertsekas, D. P., 2011. "λ-Policy Iteration: A Review and a New Implementation," Lab. for Information and Decision Systems Report LIDS-P-2874, MIT; in *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, by F. Lewis and D. Liu (eds.), IEEE Press, Computational Intelligence Series, 2012.

[Ber12] Bertsekas, D. P., 2012. "Weighted Sup-Norm Contractions in Dynamic Programming: A Review and Some New Applications," Lab. for Information and Decision Systems Report LIDS-P-2884, MIT.

[Ber14] Bertsekas, D. P., 2014. "Robust Shortest Path Planning and Semicontractive Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-P-2915, MIT, Feb. 2014 (revised Jan. 2015 and June 2016); arXiv preprint arXiv:1608.01670; to appear in Naval Research Logistics.

[Ber16a] Bertsekas, D. P., 2016. "Affine Monotonic and Risk-Sensitive Models in Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-3204, MIT, June 2016; arXiv preprint arXiv:1608.01393.

[Ber16b] Bertsekas, D. P., 2016. "Proximal Algorithms and Temporal Differences for Large Linear Systems: Extrapolation, Approximation, and Simulation," Report LIDS-P-3205, MIT; arXiv preprint
arXiv:1610.1610.05427.

[Ber16c] Bertsekas, D. P., 2016. Nonlinear Programming, (3rd Edition), Athena Scientific, Belmont, MA.

[Ber17a] Bertsekas, D. P., 2017. Dynamic Programming and Optimal Control, Vol. I, 4th Edition, Athena Scientific, Belmont, MA.

[Ber17b] Bertsekas, D. P., 2017. "Value and Policy Iteration in Deterministic Optimal Control and Adaptive Dynamic Programming," IEEE Transactions on Neural Networks and Learning Systems, Vol. 28, pp. 500-509.

[Ber17c] Bertsekas, D. P., 2017. "Stable Optimal Control and Semicontractive Dynamic Programming," Report LIDS-P-3506, MIT, May 2017; to appear in SIAM J. on Control and Optimization.

[Ber17d] Bertsekas, D. P., 2017. "Proper Policies in Infinite-State Stochastic Shortest Path Problems", Report LIDS-P-3507, MIT, May 2017.

[Ber17e] Bertsekas, D. P., 2017. "Proximal Algorithms and Temporal Differences for Solving Fixed Point Problems," to appear in Computational Optimization and Applications J.

[Ber18] Bertsekas, D. P., 2018. Abstract Dynamic Programming, 2nd Edition, Athena Scientific, Belmont, MA.

[BhE91] Bhattacharya, P. P., and Ephremides, A., 1991. "Optimal Allocations of a Server Between Two Queues with Due Times," IEEE Trans. on Aut. Control, Vol. 36, pp. 1417-1423.

[Bla62] Blackwell, D., 1962. "Discrete Dynamic Programming," Ann. Math. Statist., Vol. 33, pp. 719-726.

[Bla65a] Blackwell, D., 1965. "Discounted Dynamic Programming," Ann. Math. Statist., Vol. 36, pp. 226-235.

[Bla65b] Blackwell, D., 1965. "Positive Dynamic Programming," Proc. Fifth Berkeley Symposium Math. Statistics and Probability, pp. 415-418.

[Bla70] Blackwell, D., 1970. "On Stationary Policies," J. Roy. Statist. Soc. Ser. A, Vol. 133, pp. 33-38.

[BoM00] Borkar, V. S., and Meyn, S. P., 2000. "The O.D.E. Method for Convergence of Stochastic Approximation and Reinforcement Learning, SIAM J. Control and Optimization, Vol. 38, pp. 447-469.

[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.

[Bor08] Borkar, V. S., 2008. Stochastic Approximation: A Dynamical Systems Viewpoint, Cambridge Univ. Press, N. Y.

[Bor09] Borkar, V. S., 2009. "Reinforcement Learning: A Bridge Between Numerical Methods and Monte Carlo," in World Scientific Review, Vol. 9, Ch. 4.

[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.

[Bur97] Burgiel, H., 1997. “How to Lose at Tetris,” The Mathematical Gazette, Vol. 81, pp. 194-200.

[CFH07] Chang, H. S., Fu, M. C., Hu, J., Marcus, S. I., 2007. Simulation-Based Algorithms for Markov Decision Processes, Springer, N. Y.

[CaC97] Cao, X. R., and Chen, H. F., 1997. “Perturbation Realization Potentials and Sensitivity Analysis of Markov Processes,” IEEE Trans. on Aut. Control, Vol. 32, pp. 1382-1393.

[CaR13] Canbolat, P. G., and Rothblum, U. G., 2013. “(Approximate) Iterated Successive Approximations Algorithm for Sequential Decision Processes,” Annals of Operations Research, Vol. 208, pp. 309-320.

[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 Trans. 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, and D. Wunsch, (Eds.), IEEE Press, N. Y.

[Cao05] Cao, X. R., 2005. “A Basic Formula for Online Policy Gradient Algorithms,” IEEE Trans. on Aut. Control, Vol. 50, pp. 696-699.

[Cao07] Cao, X. R., 2007. Stochastic Learning and Optimization: A Sensitivity-Based Approach, Springer, N. Y.

[Car96] Carriere, J., 1996. “Valuation of Early-Exercise Price of Options Using Simulations and Nonparametric Regression,” Insurance: Mathematics and Economics, Vol. 19, pp. 19-30.

[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.

[ChM82] Chatelin, F., and Miranker, W. L., 1982. "Acceleration by Aggregation of Successive Approximation Methods," Linear Algebra and its Applications, Vol. 43, pp. 17-47.

[ChT89] Chow, C.-S., and Tsitsiklis, J. N., 1989. "The Complexity of Dynamic Programming," J. of Complexity, Vol. 5, pp. 466-488.

[ChT91] Chow, C.-S., and Tsitsiklis, J. N., 1991. "An Optimal One–Way Multigrid Algorithm for Discrete–Time Stochastic Control," IEEE Trans. on Aut. 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 Aut. 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. Aut. Control, Vol. AC-29, pp. 1005-1008.

[Cou06] Coulom, R., 2006. "Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search," International Conference on Computers and Games, Springer, pp. 72-83.

[CrC91] Cruz, R. L., and Chuah, M. C., 1991. "A Minimax Approach to a Simple Routing Problem," IEEE Trans. on Aut. Control, Vol. 36, pp. 1424-1435.

[Cur54] Curtiss, J. H., 1954. "A Theoretical Comparison of the Efficiencies of Two Classical Methods and a Monte Carlo Method for Computing One Component of the Solution of a Set of Linear Algebraic Equations," Proc. Symposium on Monte Carlo Methods, pp. 191-233.

[Cur57] Curtiss, J. H., 1957. "A Monte Carlo Methods for the Iteration of Linear Operators," Uspekhi Mat. Nauk, Vol. 12, pp. 149-174.

[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).

[DFM12] Desai, V. V., Farias, V. F., and Moallemi, C. C., 2012. "Aproximate Dynamic Programming via a Smoothed Approximate Linear Program," Operations Research, Vol. 60, pp. 655-674.

[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, pp. 589-608.

[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.

[DFV04] 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.

[DKM06a] Drineas, P., Kannan, R., and Mahoney, M. W., 2006. "Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplication," SIAM J. Computing, Vol. 35, pp. 132-157.

[DKM06b] Drineas, P., Kannan, R., and Mahoney, M. W., 2006. "Fast Monte Calo Algorithms for Matrices II: Computing a Low-Rank Approximation to a Matrix," SIAM J. Computing, Vol. 36, pp. 158-183.

[DMM06] Drineas, P., Mahoney, M. W., and Muthukrishnan, S., 2006. "Sampling Algorithms for L2 Regression and Applications," Proc. 17th Annual SODA, pp. 1127-1136.

[DMM08] Drineas, P., Mahoney, M. W., and Muthukrishnan, S., 2008. "Relative-Error CUR Matrix Decompositions," SIAM J. Matrix Anal. Appl., Vol. 30, pp. 844-881.

[DMM11] Drineas, P., Mahoney, M. W., Muthukrishnan, S., and Sarlos, T., 2011. "Faster Least Squares Approximation," Numerische Mathematik, Vol. 117, pp. 219-249.

[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, and D. Wunsch, (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.

[DeR79] Denardo, E. V., and Rothblum, U. G., 1979. "Optimal Stopping, Exponential Utility, and Linear Programming," Math. Programming, Vol. 16, pp. 228-244.

[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.

[DiM10] Di Castro, D., and Mannor, S., 2010. “Adaptive Bases for Reinforcement Learning,” Machine Learning and Knowledge Discovery in Databases, Vol. 6321, pp. 312-327.

[DoD93] Douglas, C. C., and Douglas, J., 1993. “A Unified Convergence Theory for Abstract Multigrid or Multilevel Algorithms, Serial and Parallel,” SIAM J. Num. Anal., Vol. 30, pp. 136-158.

[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.

[EGW06] Ernst, D., Geurts, P., and Wehenkel, L., 2006. “Tree-Based Batch Mode Reinforcement Learning,” J. of Machine Learning Research, Vol. 6, pp. 503-556.

[ELP12] Estanjini, R. M., Li, K., and Paschalidis, I. C., 2012. “A Least Squares Temporal Difference Actor-Critic Algorithm with Applications to Warehouse Management,” Naval Research Logistics, Vol. 59, pp. 197-211.

[EVW80] Ephremides, A., Varaiya, P. P., and Walrand, J. C., 1980. “A Simple Dynamic Routing Problem,” IEEE Trans. Aut. 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. Aut. 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.

[FHS14] Feinberg, E. A., Huang, J., and Scherrer, B., 2014. "Modified Policy Iteration Algorithms are not Strongly Polynomial for Discounted Dynamic Programming," Operations Research Letters, Vol. 42, pp. 429-431.

[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.

[FRF11] Foderaro, G., Raju, V., and Ferrari, S., 2011. "A Model-Based Approximate λ-Policy Iteration Approach to Online Evasive Path Planning and the Video Game Ms. Pac-Man," J. of Control Theory and Applications, Vol. 9, pp. 391-399.

[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.

[FYG06] Fern, A., Yoon, S., and Givan, R., 2006. "Approximate Policy Iteration with a Policy Language Bias: Solving Relational Markov Decision Processes," J. of Artificial Intelligence Research, Vol. 25, pp. 75-118.

[FaV06] Farias, V. F., and Van Roy, B., 2006. "Tetris: A Study of Randomized Constraint Sampling," in Probabilistic and Randomized Methods for Design Under Uncertainty, Part II, G. Calafiore, and F. Dabbene (eds.), Springer-Verlag, pp. 189-201.

[FeL07] Feinberg, E. A., and Lewis, M. E., 2007. "Optimality Inequalities for Average Cost Markov Decision Processes and the Stochastic Cash Balance Problem," Mathematics of Operations Research, Vol. 32, pp. 769-785.

[FeS94] Feinberg, E. A., and Shwartz, A., 1994. "Markov Decision Models with Weighted Discounted Criteria," Mathematics of Operations Research, Vol. 19, pp. 1-17.

[FeS96] Feinberg, E. A., and Shwartz, A., 1996. "Constrained Discounted Dynamic Programming," Mathematics of Operations Research, Vol. 21, pp. 922-945.

[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

Designs," in Learning and Approximate Dynamic Programming, by J. Si, A. Barto, W. Powell, and D. Wunsch, (Eds.), IEEE Press, N. Y.

[Fea10] Fearnley, J., 2010. "Exponential Lower Bounds For Policy Iteration," Department of Computer Science Report, University of Warwick, UK.

[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.

[Fle84] Fletcher, C. A. J., 1984. Computational Galerkin Methods, Springer, N. Y.

[FoL50] Forsythe, G. E., and Leibler, R. A., 1950. "Matrix Inversion by a Monte Carlo Method," Mathematical Tables and Other Aids to Computation, Vol. 4, pp. 127-129.

[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, J.-Q., 1994. "Smoothed Perurbation Analysis Derivative Estimation for Markov Chains," Oper. Res. Letters, Vol. 41, pp. 241-251.

[GBC16] Goodfellow, I., Bengio, J., and Courville, A., Deep Learning, MIT Press, Cambridge, MA.

[GBL12] Grondman, I., Busoniu, L., Lopes, G. A. D., and Babuska, R., 2012. "A Survey of Actor-Critic Reinforcement Learning: Standard and Natural Policy Gradients," IEEE Trans. on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol. 42, pp. 1291-1307.

[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.

[GlI89] Glynn, P. W., and Iglehart, D. L., 1989. “Importance Sampling for Stochastic Simulations,” Management Science, Vol. 35, pp. 1367-1392.

[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.

[Gor95] Gordon, G. J., 1995. “Stable Function Approximation in Dynamic Programming,” in Machine Learning: Proceedings of the Twelfth International Conference, Morgan Kaufmann, San Francisco, CA.

[Gos03] Gosavi, A., 2003. Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning, Springer, N. Y.; a 2nd edition appeared in 2015.

[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, and D. Wunsch, (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. 730-756.

[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.

[HGD14] Hollanders, R., Gerencser, B., Delvenne, J.-C., and Jungers, R. M., 2014. “About Upper Bounds on the Complexity of Policy Iteration,” arXiv preprint arXiv:1410.7583.

[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.

[HMZ13] Hansen, T. D., Miltersen, P. B., and Zwick, U. “Strategy Iteration is Strongly Polynomial for 2-Player Turn-Based Stochastic Games with a Constant Discount Factor,” Journal of the ACM, Vol. 60.

[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. Aut. Control, Vol. AC-31, pp. 227-235.

[Haj84] Hajek, B., 1984. “Optimal Control of Two Interacting Service Stations,” IEEE Trans. Aut. Control, Vol. AC-29, pp. 491-499.

[Hal70] Halton, J. H., 1970. “A Retrospective and Prospective Survey of the Monte Carlo Method,” SIAM Review, Vol. 12, pp. 1-63.

[Han08] Hansen, E. A., 2008. “Sparse Stochastic Finite-State Controllers for POMDPs,” Proc. UAI, pp. 256-263.

[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.

[Hau00] Hauskrecht, M., 2000. “Value-Function Approximations for Partially Observable Markov Decision Processes,” J. of Artificial Intelligence Research, Vol. 13, pp. 33-95.

[Hay08] Haykin, S., 2008. Neural Networks and Learning Machines, (3rd Edition), Prentice-Hall, Englewood-Cliffs, N. J.

[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.

[JuP07] Jung, T., and Polani, D., 2007. "Kernelizing LSPE(λ)," Proc. 2007 IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning, Honolulu, Ha., pp. 338-345.

[KLM96] Kaelbling, L. P., Littman, M. L., and Moore, A. W., 1996. "Reinforcement Learning: A Survey," J. of Artificial Intelligence Res., Vol. 4, pp. 237-285.

[KMP06] Keller, P. W., Mannor, S., and Precup, D., 2006. “Aut. Basis Function Construction for Approximate Dynamic Programming and Reinforcement Learning,” Proc. of the 23rd ICML, Pittsburgh, Penn.

[KVZ72] Krasnoselskii, M. A., Vainikko, G. M., Zabreyko, R. P., and Ruticki, Ya. B., 1972. Approximate Solution of Operator Equations, Translated by D. Louvish, Wolters-Noordhoff Pub., Groningen.

[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. Aut. Control, Vol. AC-13, pp. 114-115.

[Kir11] Kirsch, A., 2011. An Introduction to the Mathematical Theory of Inverse Problems, (2nd Edition), Springer, N. Y.

[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.

[KuY03] Kushner, H. J., and Yin, G., 2003. Stochastic Approximation and Recursive Algorithms and Applications, (2nd Ed.), 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.

[LDK95] Littman, M. L., Dean, T. L., and Kaelbling, L. P., 1995. "On the Complexity of Solving Markov Decision Problems," Proc. of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence, pp. 394-402.

[LGM10] Lazaric, A., Ghavamzadeh, M., and Munos, R., 2010. "Analysis of a Classification-Based Policy Iteration Algorithm," Proceedings of ICML, pp. 607-614.

[LLL08] Lewis, F. L., Liu, D., and Lendaris, G. G., 2008. Special Issue on Adaptive Dynamic Programming and Reinforcement Learning in Feedback Control, IEEE Trans. on Systems, Man, and Cybernetics, Part B, Vol. 38, No. 4.

[LSS09] Li, Y., Szepesvari, C., and Schuurmans, D., 2009. "Learning Exercise Policies for American Options," Proc. of the Twelfth International Conference on Artificial Intelligence and Statistics, Clearwater Beach, Fla.

[L'Ec91] L'Ecuyer, P., 1991. "An Overview of Derivative Estimation," Proceedings of the 1991 Winter Simulation Conference, pp. 207-217.

[LaP03a] Lagoudakis, M. G., and Parr, R., 2003. "Least-Squares Policy Iteration," J. of Machine Learning Research, Vol. 4, pp. 1107-1149.

[LaP03b] Lagoudakis, M. G., and Parr, R., 2003. "Reinforcement Learning as Classification: Leveraging Modern Classifiers," Proc. of ICML, pp. 424-431.

[LaT85] Lancaster, P., and Tismenetsky, M., 1985. The Theory of Matrices, Academic Press, N. Y.

[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.

[LeL12] Lewis, F. L., and Liu, D., 2012. Reinforcement Learning and Approximate Dynamic Programming for Feedback Control, IEEE Press Computational Intelligence Series, N. Y.

[LeV09] Lewis, F. L., and Vrabie, D., 2009. “Reinforcement Learning and Adaptive Dynamic Programming for Feedback Control,” IEEE Circuits and Systems Magazine, 3rd Q. Issue.

[LiK84] Lin, W., and Kumar, P. R., 1984. “Optimal Control of a Queueing System with Two Heterogeneous Servers,” IEEE Trans. Aut. 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.

[Liu01] Liu, J. S., 2001. Monte Carlo Strategies in Scientific Computing, Springer, N. Y.

[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.

[LuT89] Luo, Z. Q., and Tseng P., 1989. “On the Convergence of a Matrix Splitting Algorithm for the Symmetric Monotone Linear Complementarity Problem,” SIAM J. Control and Optimization, Vol. 29, pp. 1037-1060.

[MMS06] Menache, I., Mannor, S., and Shimkin, N., 2005. “Basis Function Adaptation in Temporal Difference Reinforcement Learning,” Ann. Oper. Res., Vol. 134, pp. 215-238.

[MaT01] Marbach, P., and Tsitsiklis, J. N., 2001. “Simulation-Based Optimization of Markov Reward Processes,” IEEE Trans. on Aut. 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.

[Mah11] Mahoney, M. W., 2011. "Randomized Algorithms for Matrices and Data," Foundations and Trends in Machine Learning, Vol. 3, pp. 123-224.

[Man60] Manne A., 1960. "Linear Programming and Sequential Decisions," Management 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 Aut. 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.

[Mey07] Meyn, S., 2007. Control Techniques for Complex Networks, Cambridge Univ. Press, N. Y.

[MiV69] Miller, B. L., and Veinott, A. F., Jr., 1969. "Dynamic Programming with a Small Interest Rate," Annals of Mathematical Statistics, Vol. 40, pp. 366-370.

[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.

[MuS08] Munos, R., and Szepesvari, C, 2008. "Finite-Time Bounds for Fitted Value Iteration," J. of Machine Learning Research, Vol. 1, pp. 815-857.

[Mun03] Munos, R., 2003. "Error Bounds for Approximate Policy Iteration," Proc. 20th International Conference on Machine Learning, pp. 560-567.

[NJL09] Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A., 2009. "Robust Stochastic Approximation Approach to Stochastic Programming," SIAM J. on Optimization, Vol. 19, pp. 1574-1609.

[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.

[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 Nonlinear Equations in Several Variables, Academic Press, N. Y.

[OrS02] Ormoneit, D., and Sen, S., 2002. "Kernel-Based Reinforcement Learning," Machine Learning, Vol. 49, pp. 161-178.

[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 Aut. 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.

[PSD01] Precup, D., Sutton, R. S., and Dasgupta, S., 2001. "Off-Policy Temporal-Difference Learning with Function Approximation," Proc. 18th Int. Conf. Machine Learning, pp. 417424.

[PaB99] Patek, S. D., and Bertsekas, D. P., 1999. "Stochastic Shortest Path Games," SIAM J. on Control and Optimization, Vol. 36, pp. 804-824.

[PaF03] Pang, J. S., and Facchinei, F., 2003. Finite-Dimensional Variational Inequalities and Complementarity Problems, Springer-Verlag, N. Y.

[PaK81] Pattipati, K. R., and Kleinman, D. L., 1981. "Priority Assignment Using Dynamic Programming for a Class of Queueing Systems," IEEE Trans. on Aut. 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 Trans. on Networking, Vol. 8, pp. 171-184.

[Pal67] Pallu de la Barriere, R., 1967. Optimal Control Theory, Saunders, Phila; republished by Dover, N. Y., 1980.

[Pat01] Patek, S. D., 2001. "On Terminating Markov Decision Processes with a Risk Averse Objective Function," Automatica, Vol. 37, pp. 1379-1386.

[Pat04] Patek, S. D., 2004. "Policy Iteration Type Algorithms for Recurrent State Markov Decision Processes," Computers and Operations Research, Vol. 31, pp. 2333-2347.

[Pat07] Patek, S. D., 2007. "Partially Observed Stochastic Shortest Path Problems with Approximate Solution by Neuro-Dynamic Programming," IEEE Trans. on Systems, Man, and Cybernetics Part A, Vol. 37, pp. 710-720.

[Pin97] Pineda, F., 1997. "Mean-Field Analysis for Batched TD(λ)," Neural Computation, Vol. 9, 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.

[PoB04] Poupart, P., and Boutilier, C., 2004. "Bounded Finite State Controllers," Advances in Neural Information Processing Systems, Proc. of 2003 NIPS Conference, MIT Press, Cambridge, MA.

[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, and D. Wunsch, (Eds.), IEEE Press, N. Y.

[PoY15] Post, I., and Ye, Y. 2015. "The Simplex Method is Strongly Polynomial for Deterministic Markov Decision Processes," Mathematics of Operations Research, published on-line.

[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.

[Pow11] Powell, W. B., 2011. Approximate Dynamic Programming: Solving the Curses of Dimensionality, 2nd Edition, J. Wiley and Sons, Hoboken, N. J.

[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.

[Put94] Puterman, M. L., 1994. Markovian Decision Problems, J. Wiley, N. Y.

[RGT05] Roy, N., Gordon, G., and Thrun, S., 2005. "Finding Approximate POMDP Solutions Through Belief Compression," J. of Artificial Intelligence Research, Vol. 23, pp. 1-40.

[RPW91] Rogers, D. F., Plante, R. D., Wong, R. T., and Evans, J. R., 1991. "Aggregation and Disaggregation Techniques and Methodology in Optimization," Operations Research, Vol. 39, pp. 553-582.

[RVW82] Rosberg, Z., Varaiya, P. P., and Walrand, J. C., 1982. "Optimal Control of Service in Tandem Queues," IEEE Trans. Aut. 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 Sennot, L. I., 1992. "Optimal Stationary Policies in General State Markov Decision Chains with Finite Action Set," Math. Operations Research, Vol. 17, pp. 901-909.

[RoC10] Robert, C. P., and Casella, G., 2010. Monte Carlo Statistical Methods, Springer, N. Y.

[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.

[Rot79] Rothblum, U. G., 1979. "Iterated Successive Approximation for Sequential Decision Processes," in Stochastic Control and Optimization, by J. W. B. van Overhagen and H. C. Tijms (eds), Vrije University, Amsterdam.

[Roy88] Royden, H. L., 1988. Principles of Mathematical Analysis, (3rd Ed.), McGraw-Hill, N. Y.

[RuK04] Rubinstein, R. Y., and Kroese, D. P., 2004. The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Springer, N. Y.

[RuK08] Rubinstein, R. Y., and Kroese, D. P., 2008. Simulation and the Monte Carlo Method, (2nd Edition), J. Wiley, 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.

[SDR09] Shapiro, A., Dentcheva, D., and Ruszczynski, A., 2009. Lectures on Stochastic Programming: Modeling and Theory, SIAM, Phila., PA.

[SHS17] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T. and Lillicrap, T., 2017. “Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm,” arXiv preprint arXiv:1712.01815.

[SJJ94] 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.

[SJJ95] Singh, S. P., Jaakkola, T., and Jordan, M. I., 1995. “Reinforcement Learning with Soft State Aggregation,” in Advances in Neural Information Processing Systems 7, MIT Press, Cambridge, MA.

[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, and D. Wunsch, (Eds.), IEEE Press, N. Y.

[Saa03] Saad, Y., 2003. Iterative Methods for Sparse Linear Systems, SIAM, Phila., Pa.

[Sam59] Samuel, A. L., 1959. “Some Studies in Machine Learning Using the Game of Checkers,” IBM J. of Research and Development, pp. 210-229.

[Sam67] Samuel, A. L., 1967. “Some Studies in Machine Learning Using the Game of Checkers. II – Recent Progress,” IBM J. of Research and Development, pp. 601-617.

[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 Approximations 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.

[Sch10] Scherrer, B., 2010. "Should One Compute the Temporal Difference Fix Point or Minimize the Bellman Residual? The Unified Oblique Projection View," in ICML'10: Proc. of the 27th Annual International Conf. on Machine Learning.

[Sch11] Scherrer, B., 2011. "Performance Bounds for Lambda Policy Iteration and Application to the Game of Tetris," Report RR-6348, INRIA, France; J. of Machine Learning Research, Vol. 14, 2013, pp. 1181-1227.

[Sch12] Scherrer, B., 2012. "On the Use of Non-Stationary Policies for Infinite-Horizon Discounted Markov Decision Processes," INRIA Lorraine Report, France.

[Sch13] Scherrer, B., 2013. "Improved and Generalized Upper Bounds on the Complexity of Policy Iteration," C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, (Eds.), Advances in Neural Information Processing Systems, Vol. 26, pp. 386-394.

[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.

[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 Univ. 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.

[SpV05] Spaan, M. J. T., and Vlassis, N., 2005. "Perseus: Randomized Point-Based Value Iteration for POMDPs," J. of Artificial Intelligence Research, Vol. 24, pp. 195-220.

[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.

[StW93] Stidham, S., and Weber, R., 1993. "A Survey of Markov Decision Models for Control of Networks of Queues," Queueing Systems, Vol. 13, pp. 291-314.

[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. Aut. 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. (A draft 2nd edition is available on-line.)

[SuC91] Suk, J.-B., and Cassandras, C. G., 1991. "Optimal Scheduling of Two Competing Queues with Blocking," IEEE Trans. on Aut. 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.

[SzL06] Szita, I., and Lorinz, A., 2006. "Learning Tetris Using the Noisy Cross-Entropy Method," Neural Computation, Vol. 18, pp. 2936-2941.

[Sze10] Szepesvari, C., 2010. Algorithms for Reinforcement Learning, Morgan and Claypool Publishers, San Franscisco, CA.

[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 Aut. Control, Vol. 37, pp. 1446-1451.

[Tes92] Tesauro, G., 1992. "Practical Issues in Temporal Difference Learning," Machine Learning, Vol. 8, pp. 257-277.

[ThS09] Thiery, C., and Scherrer, B., 2009. "Improvements on Learning Tetris with Cross-Entropy," International Computer Games Association J., Vol. 32, pp. 23-33.

[ThS10a] Thiery, C., and Scherrer, B., 2010. "Least-Squares λ-Policy Iteration: Bias-Variance Trade-off in Control Problems," in ICML'10: Proc. of the 27th Annual International Conf. on Machine Learning.

[ThS10b] Thiery, C., and Scherrer, B., 2010. "Performance Bound for Approximate Optimistic Policy Iteration," Technical Report, INRIA, France.

[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.

[ToS06] Toussaint, M., and Storkey, A. J., 2006. "Probabilistic Inference for Solving Discrete and Continuous State Markov Decision Processes," Proc. of the 23nd ICML, pp. 945-952.

[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 Trans. on Aut. Control, Vol. 42, pp. 674-690.

[TsV99a] Tsitsiklis, J. N., and Van Roy, B., 1999. "Average Cost Temporal-Difference Learning," Aut. a, 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 Trans. on Aut. Control, Vol. 44, pp. 1840-1851.

[TsV01] Tsitsiklis, J. N., and Van Roy, B., 2001. "Regression Methods for

Pricing Complex American-Style Options," IEEE Trans. on Neural Networks, Vol. 12, pp. 694-703.

[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.

[Tsi84] Tsitsiklis, J. N., 1984. "Convexity and Characterization of Optimal Policies in a Dynamic Routing Problem," J. Optimization Theory Appl., Vol. 44, pp. 105-136.

[Tsi86] Tsitsiklis, J. N., 1986. "A Lemma on the Multiarmed Bandit Problem," IEEE Trans. Aut. 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. Aut. Control, Vol. AC-40, pp. 1528-1538.

[Tsi07] Tsitsiklis, J. N., 2007. "NP-Hardness of Checking the Unichain Condition in Average Cost MDPs," Operations Research Letters, Vol. 35, pp. 319-323.

[VBL07] Van Roy, B., Bertsekas, D. P., Lee, Y., and Tsitsiklis, J. N., 1997. "A Neuro-Dynamic Programming Approach to Retailer Inventory Management," Proc. of the IEEE Conference on Decision and Control, pp. 4052-4057.

[VWB85] Varaiya, P. P., Walrand, J. C., and Buyukkoc, C., 1985. "Extensions of the Multiarmed Bandit Problem: The Discounted Case," IEEE Trans. Aut. 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.

[Van95] Van Roy, B., 1995. "Feature-Based Methods for Large Scale Dynamic Programming," Lab. for Info. and Decision Systems Report LIDS-TH-2289, Massachusetts Institute of Technology, Cambridge, MA.

[Van98] Van Roy, B., 1998. Learning and Value Function Approximation in Complex Decision Processes, Ph.D. Thesis, Dept. of EECS, MIT, Cambridge, MA.

[Van06] Van Roy, B., 2006. "Performance Loss Bounds for Approximate Value Iteration with State Aggregation," Mathematics of Operations Research, Vol. 31, pp. 234-244.

[Van10] Van Roy, B., 2010. "On Regression-Based Stopping Times," Discrete Event Dynamic Systems, Vol. 20, pp. 307-324.

[Var78] Varaiya, P. P., 1978. "Optimal and Suboptimal Stationary Controls of Markov Chains," IEEE Trans. Aut. 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.

[VeR06] Verma, R., and Rao, R. P. N., 2006. "Planning and Acting in Uncertain Environments Using Probabilistic Inference," Proc. of IEEE/RSJ Intern. Conf. on Intelligent Robots and Systems.

[Vei66] Veinott, A. F., Jr., 1966. "On Finding Optimal Policies in Discrete Dynamic Programming with no Discounting," Ann. Math. Statist., Vol. 37, pp. 1284-1294.

[Vei69] 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 Aut. Control, Vol. 33, pp. 104-109.

[VlT09] Vlassis, N., and Toussaint, M., 2009. "Model-Free Reinforcement Learning as Mixture Learning," Proc. of the 26th International Conference on Machine Learning, Montreal, Canada.

[WPB09] Wang, M., Polydorides, N., and Bertsekas, D. P., 2009. "Approximate Simulation-Based Solution of Large-Scale Least Squares Problems," Lab. for Information and Decision Systems Report LIDS-P-2819, MIT.

[WaB13a] Wang, M., and Bertsekas, D. P., 2013. "Stabilization of Stochastic Iterative Methods for Singular and Nearly Singular Linear Systems," Mathematics of Operations Research, Vol. 39, pp. 1-30.

[WaB13b] Wang, M., and Bertsekas, D. P., 2013. "Convergence of Iterative Simulation-Based Methods for Singular Linear Systems," Stochastic Systems, Vol. 3, pp. 39-96.

[WaD92] Watkins, C. J. C. H., and Dayan, P., 1992. "Q-Learning," Machine Learning, Vol. 8, pp. 279-292.

[Wal88] Walrand, J., 1988. An Introduction to Queueing Networks, Prentice Hall, Englewood Cliffs, N. J.

[Wan17] Wang, M., 2017. "Primal-Dual π Learning: Sample Complexity and Sublinear Run Time for Ergodic Markov Decision Problems," arXiv preprint arXiv:1710.06100.

[Was52] Wasow, W. R., 1952. "A Note on Inversion of Matrices by Random Walks," Mathematical Tables and Other Aids to Computation, Vol. 6, pp. 78-81.

[Wat89] Watkins, C. J. C. H., Learning from Delayed Rewards, Ph.D. Thesis, Cambridge Univ., England.

[WeB99] Weaver, L., and Baxter, J., 1999. "Reinforcement Learning From State and Temporal Differences," Tech. Report, Department of Computer Science, Australian National University.

[Web92] Weber, R., 1992. "On the Gittins Index for Multiarmed Bandits," Annals of Applied Probability, Vol. 2, pp. 1024-1033.

[Wei88] Weiss, G., 1988. "Branching Bandit Processes," Probab. Eng. Inform. Sci., Vol. 2, pp. 269-278.

[Wer09] Werbos, P. J., 2009. "Intelligence in the Brain: A Theory of How it Works and How to Build it," Neural Networks, Vol. 22, pp. 200-212.

[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.

[WhS92] White, D., and Sofge, D., (Eds.), 1992. Handbook of Intelligent Control, Van Nostrand, N. Y.

[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.

[YaL08] Yao, H., and Liu, Z.-Q., 2008. "Preconditioned Temporal Difference Learning," Proc. of the 25th ICML, Helsinki, Finland.

[Ye05] Ye, Y., 2005. "A New Complexity Result on Solving the Markov Decision Problem," Mathematics of Operations Research, Vol. 30, pp. 733-749.

[Ye11] Ye, Y., 2011. "The Simplex and Policy-Iteration Methods are Strongly Polynomial for the Markov Decision Problem with a Fixed Discount Rate," Mathematics of Operations Research, Vol. 36, pp. 593-603.

[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 LIDS-P-2689, MIT; Mathematics of Operations Research, Vol. 33, 2008, pp. 1-11.

[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 LIDS-P-2697, MIT; IEEE Trans. on Aut. Control, Vol. 54, 2009, pp. 1515-1531.

[YuB07] Yu, H., and Bertsekas, D. P., 2007. "A Least Squares Q-Learning Algorithm for Optimal Stopping Problems," Proc. European Control Conference 2007, Kos, Greece, pp. 23682375; an extended version appears in

Lab. for Information and Decision Systems Report LIDS-P-2731, MIT.

[YuB08] Yu, H., and Bertsekas, D. P., 2008. "Error Bounds for Approximations from Projected Linear Equations," Lab. for Information and Decision Systems Report LIDS-P-2797, MIT, July 2008; Mathematics of Operations Research, Vol. 35, 2010, pp. 306-329.

[YuB09] Yu, H., and Bertsekas, D. P., 2009. "Basis Function Adaptation Methods for Cost Approximation in MDP," Proceedings of 2009 IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning (ADPRL 2009), Nashville, Tenn.

[YuB12] Yu, H., and Bertsekas, D. P., 2012. "Weighted Bellman Equations and their Applications in Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-P-2876, MIT.

[YuB13a] Yu, H., and Bertsekas, D. P., 2013. "Q-Learning and Policy Iteration Algorithms for Stochastic Shortest Path Problems," Annals of Operations Research, Vol. 208, pp. 95-132.

[YuB13b] Yu, H., and Bertsekas, D. P., 2013. "On Boundedness of Q-Learning Iterates for Stochastic Shortest Path Problems," Math. of OR, Vol. 38, pp. 209-227.

[YuB15] Yu, H., and Bertsekas, D. P., 2015. "A Mixed Value and Policy Iteration Method for Stochastic Control with Universally Measurable Policies," Math. of OR, Vol. 40, pp. 926-968.

[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.

[Yu11] Yu, H., 2011. "Stochastic Shortest Path Games and Q-Learning," Lab. for Information and Decision Systems Report LIDS-P-2875, MIT.

[Yu12] Yu, H., 2012. "Least Squares Temporal Difference Methods: An Analysis Under General Conditions," SIAM J. on Control and Optimization, Vol. 50, pp. 3310-3343.

[ZhH01] Zhou, R., and Hansen, E. A., 2001. "An Improved Grid-Based Approximation Algorithm for POMDPs," In Int. J. Conf. Artificial Intelligence, Seattle, WA.

[ZhL97] Zhang, N. L., and Liu, W., 1997. "A Model Approximation Scheme for Planning in Partially Observable Stochastic Domains," J. Artificial Intelligence Research, Vol. 7, pp. 199-230.