Dynamic Programming and Optimal Control

Volume II

Dimitri P. Bertsekas

Massachusetts Institute of Technology

Athena Scientific, Belmont, Massachusetts

Athena Scientific
Post Office Box 391
Belmont, Mass. 02178-9998
U.S.A.

Email: athenasc@world.std.com

Cover Design: *Ann Gallager*

Portions of this volume are adapted and reprinted from the author's *Dynamic Programming: Deterministic and Stochastic Models*, Prentice-Hall, 1987, by permission of Prentice-Hall, Inc.

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 1995 519.703 95-075941

ISBN 1-886529-12-4 (Vol. I)
ISBN 1-886529-13-2 (Vol. II)
ISBN 1-886529-11-6 (Vol. I and II)

Contents

CONTENTS OF VOLUME I

ABOUT THE AUTHOR

Dimitri Bertsekas studied Mechanical and Electrical Engineering at the National Technical University of Athens, Greece, and obtained his Ph.D. in system science from the Massachusetts Institute of Technology. He has held faculty positions with the Engineering-Economic Systems Dept., Stanford University and the Electrical Engineering Dept. of the University of Illinois, Urbana. He is currently Professor of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology. He consults regularly with private industry and has held editorial positions in several journals. He has been elected Fellow of the IEEE.

Professor Bertsekas has done research in a broad variety of subjects from control theory, optimization theory, parallel and distributed computation, data communication networks, and systems analysis. He has written numerous papers in each of these areas. This book is his fourth on dynamic programming and optimal control.

Other books by the author:

1) *Dynamic Programming and Stochastic Control*, Academic Press, 1976.
2) *Stochastic Optimal Control: The Discrete-Time Case*, Academic Press, 1978 (with S. E. Shreve; translated in Russian).
3) *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, 1982 (translated in Russian).
4) *Dynamic Programming: Deterministic and Stochastic Models*, Prentice-Hall, 1987.
5) *Data Networks*, Prentice-Hall, 1987 (with R. G. Gallager; translated in Russian and Japanese); 2nd Edition 1992.
6) *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, 1989 (with J. N. Tsitsiklis).
7) *Linear Network Optimization: Algorithms and Codes*, M.I.T. Press 1991.

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. 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, I have added a final chapter that provides an introductory treatment of infinite 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. 1, and Chapters 1, 2, and 4 of Vol. II.

(c) 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.

(d) A one-quarter engineering course that covers the first three chapters and parts of Chapters 4 through 6 of Vol. I.

(e) 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 Athena Scientific or 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
bertsekas@lids.mit.edu

1

Infinite Horizon – Discounted Problems

Contents

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

We recall from Chapter 7 of Vol. I that there are four classes of infinite horizon problems of major interest.

(a) Discounted problems with bounded cost per stage.

(b) Stochastic shortest path problems.

(c) Discounted and undiscounted problems with unbounded cost per stage.

(d) Average cost per stage problems.

Each one of the first four chapters of the present volume considers one of the above problem classes, while the final chapter extends the analysis to continuous-time problems with a countable number of states. Throughout this volume we concentrate on the perfect information case, where each decision is made with exact knowledge of the current system state. Imperfect state information problems can be treated, as in Chapter 5 of Vol. I, by reformulation into perfect information problems involving a sufficient statistic.

1.1 MINIMIZATION OF TOTAL COST – INTRODUCTION

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

Total Cost Infinite Horizon Problem

Consider the stationary discrete-time dynamic system

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

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

Given an initial state x_0, we want to find a policy $\pi = \{\mu_0, \mu_1, \ldots\}$, where $\mu_k : S \mapsto C$, $\mu_k(x_k) \in U(x_k)$, for all $x_k \in S$, $k = 0, 1, \ldots$, 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 $g : S \times C \times D \mapsto \Re$ is given, and α is a positive scalar referred to as the *discount factor*.

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

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

A *stationary policy* is an admissible policy of the form $\pi = \{\mu, \mu, \ldots\}$, and its corresponding cost function is denoted by J_μ. For brevity, we refer to $\{\mu, \mu, \ldots\}$ 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 allow arbitrary state and control spaces, we require that the disturbance space be countable. This is necessary to avoid the mathematical complications discussed in Section 1.5 of Vol. I. The countability assumption, however, is satisfied in many problems of interest, notably for deterministic optimal control problems and problems with a finite or a countable number of states. For other problems, our main results can typically be proved (under additional technical conditions) by following the same line of argument as the one given here, but also by dealing with the mathematical complications of various measure-theoretic frameworks; see [BeS78].

The cost $J_\pi(x_0)$ given by Eq. (1.2) represents the limit of expected finite horizon costs. These costs are well defined as discussed in Section

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

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

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

The DP Algorithm for the Finite-Horizon Version of the Problem

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

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

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 of Section 1.3 of Vol. I. This algorithm expresses the optimal $(N-k)$-stage cost starting from state x, denoted by $J_k(x)$, as the minimum of the expected sum of the cost of stage $N-k$ and the optimal $(N-k-1)$-stage cost starting from the next state. It is given by

$$J_k(x) = \min_{u \in U(x)} E\big\{ \alpha^{N-k} g(x,u,w) + J_{k+1}\big(f(x,u,w)\big) \big\}, \quad k = 0, 1, \ldots, N-1, \tag{1.3}$$

with the initial condition

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

For all initial states x, the optimal N-stage cost is the function $J_0(x)$ obtained from the last step of the DP algorithm.

Let us consider for all k and x, the functions V_k given by

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

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

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

with the initial condition

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

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

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

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

Some Shorthand Notation

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

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

$$(TJ)(x) = \min_{u \in U(x)} \mathop{E}_{w} \left\{ g(x,u,w) + \alpha J\big(f(x,u,w)\big) \right\}, \qquad x \in S. \tag{1.4}$$

Since $(TJ)(\cdot)$ is itself a function defined on the state space S, we view T as a mapping that transforms the function J on S into the function TJ on S. Note that *TJ is the optimal cost function for the one-stage problem that has stage cost g and terminal cost αJ.*

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

Similarly, for any function $J : S \mapsto \Re$ and any control function $\mu : S \mapsto C$, we denote

$$(T_\mu J)(x) = \mathop{E}_{w}\left\{g\big(x,\mu(x),w\big) + \alpha J\big(f(x,\mu(x),w)\big)\right\}, \qquad x \in S. \tag{1.5}$$

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

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

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

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

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

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

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

and

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

It can be verified by induction that $(T^k J)(x)$ *is the optimal cost for the k-stage, α-discounted problem with initial state x, cost per stage g, and terminal cost function $\alpha^k J$.* Similarly, $(T_\mu^k J)(x)$ *is the cost of a policy $\{\mu,\mu,\ldots\}$ for the same problem.* To illustrate the case where $k = 2$, note that

$$(T^2 J)(x) = \min_{u\in U(x)} \mathop{E}_{w}\big\{g(x,u,w) + \alpha(TJ)\big(f(x,u,w)\big)\big\}$$

$$= \min_{u_0\in U(x)} \mathop{E}_{w_0}\left\{g(x,u_0,w_0) + \alpha \min_{u_1\in U(f(x,u_0,w_0))} \mathop{E}_{w_1}\left\{g\big(f(x,u_0,w_0),u_1,w_1\big) + \alpha J\big(f\big(f(x,u_0,w_0),u_1,w_1\big)\big)\right\}\right\}$$

$$= \min_{u_0\in U(x)} \mathop{E}_{w_0}\left\{g(x,u_0,w_0) + \min_{u_1\in U(f(x,u_0,w_0))} \mathop{E}_{w_1}\left\{\alpha g\big(f(x,u_0,w_0),u_1,w_1\big) + \alpha^2 J\big(f\big(f(x,u_0,w_0),u_1,w_1\big)\big)\right\}\right\}.$$

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

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

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

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 $\alpha^k J$.*

Some Basic Properties

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

Lemma 1.1: (Monotonicity Lemma) For any functions $J : S \mapsto \Re$ and $J' : S \mapsto \Re$, such that

$$J(x) \le J'(x), \qquad \text{for all } x \in S,$$

and for any function $\mu : S \mapsto C$ with $\mu(x) \in U(x)$, for all $x \in S$, we have

$$(T^k J)(x) \le (T^k J')(x), \qquad \text{for all } x \in S,\ k = 1,2,\ldots,$$

$$(T_\mu^k J)(x) \le (T_\mu^k J')(x), \qquad \text{for all } x \in S,\ k = 1,2,\ldots$$

Proof: The result follows by viewing $(T^k J)(x)$ and $(T_\mu^k J)(x)$ as k-stage problem costs, since as the terminal cost function increases uniformly so will the k-stage costs. (One can also prove the lemma by using a straightforward induction argument.) **Q.E.D.**

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

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

With this notation, Lemma 1.1 is stated as

$$J \le J' \quad \Rightarrow \quad T^k J \le T^k J', \qquad k = 1,2,\ldots,$$

$$J \le J' \quad \Rightarrow \quad T_\mu^k J \le T_\mu^k J', \qquad k = 1,2,\ldots$$

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

$$e(x) = 1, \qquad \text{for all } x \in S. \tag{1.6}$$

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

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

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

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

Lemma 1.2: For every k, function $J : S \mapsto \Re$, stationary policy μ, and scalar r, we have

$$\big(T^k(J + re)\big)(x) = (T^k J)(x) + \alpha^k r, \qquad \text{for all } x \in S, \tag{1.7}$$

$$\big(T_\mu^k(J + re)\big)(x) = (T_\mu^k J)(x) + \alpha^k r, \qquad \text{for all } x \in S. \tag{1.8}$$

A Preview of Infinite Horizon Results

It is worth at this point to speculate on the type of results that we will be aiming for.

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

$$J^*(x) = \lim_{k \to \infty} (T^k J_0)(x), \qquad x \in S. \tag{1.9}$$

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

$$J^*(x) = \lim_{k \to \infty} (T^k J)(x), \qquad \text{for all } x \in S \text{ and bounded functions } J. \tag{1.10}$$

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

$$(T^{k+1} J_0)(x) = \min_{u \in U(x)} E_w \big\{ g(x, u, w) + \alpha (T^k J_0)\big(f(x, u, w)\big) \big\}, \tag{1.11}$$

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 $k \to \infty$,

$$J^*(x) = \min_{u \in U(x)} E_w \big\{ g(x, u, w) + \alpha J^*\big(f(x, u, w)\big) \big\}, \qquad x \in S, \tag{1.12}$$

or, equivalently,

$$J^* = TJ^*. \tag{1.13}$$

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 2, the preceding conjectures are correct. For problems with unbounded costs per stage and for stochastic shortest path problems where our assumptions of Chapter 2 are violated, there may be counterintuitive mathematical phenomena that invalidate some of the preceding conjectures. This illustrates that infinite horizon problems should be approached carefully and with mathematical precision.

1.2 DISCOUNTED PROBLEMS WITH BOUNDED COST PER STAGE

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

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

$$\big|g(x, u, w)\big| \leq M, \qquad \text{for all } (x, u, w) \in S \times C \times D, \tag{2.1}$$

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

Boundedness of the cost per stage is not as restrictive as might appear. It holds for problems where the spaces S, C, and D are finite sets. Even if these spaces are not finite, during the computational solution of the problem they will ordinarily be approximated by finite sets. Also, it is often possible to reformulate the problem so that it is defined over bounded regions of the state and control spaces over which the cost is bounded.

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

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

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 2.1: (Convergence of the DP Algorithm) For any bounded function $J : S \mapsto \Re$, the optimal cost function satisfies

$$J^*(x) = \lim_{N\to\infty} (T^N J)(x), \qquad \text{for all } x \in S. \tag{2.2}$$

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

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

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

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

Using these relations, it follows that

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

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

$$\begin{aligned} J^*(x_0) - \frac{\alpha^K M}{1-\alpha} - \alpha^K \max_{x\in S}\big|J(x)\big| & \\ \le (T^K J)(x_0) & \\ \le J^*(x_0) + \frac{\alpha^K M}{1-\alpha} + \alpha^K \max_{x\in S}\big|J(x)\big|, & \end{aligned} \tag{2.3}$$

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

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

Given any stationary policy μ, we can consider a modified discounted problem, which is the same as the original except that the control constraint set contains only one element for each state x, the control $\mu(x)$; that is, the control constraint set is $\tilde{U}(x) = \{\mu(x)\}$ instead of $U(x)$. Proposition 2.1 applies to this modified problem and yields the following corollary:

Corollary 2.1.1: For every stationary policy μ, the associated cost function satisfies

$$J_\mu(x) = \lim_{N\to\infty} (T_\mu^N J)(x), \qquad \text{for all } x \in S. \tag{2.4}$$

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

Proposition 2.2: (Bellman's Equation) The optimal cost function J^* satisfies

$$J^*(x) = \min_{u\in U(x)} E_w\big\{g(x,u,w) + \alpha J^*\big(f(x,u,w)\big)\big\}, \qquad \text{for all } x \in S, \tag{2.5}$$

or, equivalently,

$$J^* = TJ^*. \tag{2.6}$$

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

Proof: From Eq. (2.3), we have for all $x \in S$ and N,

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

where J_0 is the zero function [$J_0(x) = 0$ for all $x \in S$]. Applying the mapping T to this relation and using the Monotonicity Lemma 1.1 as well as Lemma 1.2, we obtain for all $x \in S$ and N

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

Since $(T^{N+1}J_0)(x)$ converges to $J^*(x)$ (cf. Prop. 2.1), by taking the limit as $N \to \infty$ in the preceding relation, we obtain $J^* = TJ^*$.

To show uniqueness, observe that if J is bounded and satisfies $J = TJ$, then $J = \lim_{N\to\infty} T^N J$, so by Prop. 2.1, we have $J = J^*$. **Q.E.D.**

Based on the same reasoning we used to obtain Cor. 2.1.1 from Prop. 2.1, we have:

Corollary 2.2.1: For every stationary policy μ, the associated cost function satisfies

$$J_\mu(x) = \mathop{E}_{w}\Big\{g\big(x, \mu(x), w\big) + \alpha J_\mu\big(f(x, \mu(x), w)\big)\Big\}, \qquad \text{for all } x \in S, \tag{2.7}$$

or, equivalently,

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

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

The next proposition characterizes stationary optimal policies.

Proposition 2.3: (Necessary and Sufficient Condition for Optimality) A stationary policy μ is optimal if and only if $\mu(x)$ attains the minimum in Bellman's equation (2.5) for each $x \in S$; that is,

$$TJ^* = T_\mu J^*. \tag{2.8}$$

Proof: If $TJ^* = T_\mu J^*$, then using Bellman's equation ($J^* = TJ^*$), we have $J^* = T_\mu J^*$, so by the uniqueness part of Cor. 2.2.1, we obtain $J^* = J_\mu$;

that is, μ is optimal. Conversely, if the stationary policy μ is optimal, we have $J^* = J_\mu$, which by Cor. 2.2.1, yields $J^* = T_\mu J^*$. Combining this with Bellman's equation ($J^* = TJ^*$), we obtain $TJ^* = T_\mu J^*$. **Q.E.D.**

Note that Prop. 2.3 implies the existence of an optimal stationary policy when the minimum in the right-hand side of Bellman's equation is attained for all $x \in S$. In particular, when $U(x)$ is finite for each $x \in S$, an optimal stationary policy is guaranteed to exist.

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

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

This relation is obtained by combining Bellman's equation and the following result:

Proposition 2.4: For any two bounded functions $J : S \mapsto \Re$, $J' : S \mapsto \Re$, and for all $k = 0, 1, \ldots$, there holds

$$\max_{x\in S}\big|(T^k J)(x) - (T^k J')(x)\big| \leq \alpha^k \max_{x\in S}\big|J(x) - J'(x)\big|. \tag{2.9}$$

Proof: Denote

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

Then we have

$$J(x) - c \leq J'(x) \leq J(x) + c, \qquad x \in S.$$

Applying T^k in this relation and using the Monotonicity Lemma 1.1 as well as Lemma 1.2, we obtain

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

It follows that

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

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

As earlier, we have:

Corollary 2.4.1: For any two bounded functions $J : S \mapsto \Re$, $J' : S \mapsto \Re$, and any stationary policy μ, we have

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

Example 2.1 (Machine Replacement)

Consider an infinite horizon discounted version of a problem we formulated in Section 1.1 of Vol. 1. Here, we want to operate efficiently a machine that can be in any one of n states, denoted $1, 2, \ldots, n$. 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. 2.2) takes the form

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

By Prop. 2.3, a stationary policy is optimal if it replaces at states i where

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

and it does not replace at states i where

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

We can use the convergence of the DP algorithm (cf. Prop. 2.1) to characterize the optimal cost function using properties of the finite horizon cost functions. In particular, the DP algorithm starting from the zero function takes the form

$$J_0(i) = 0,$$

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

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

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

$$\sum_{j=1}^{n} p_{ij} J(j) \leq \sum_{j=1}^{n} p_{(i+1)j} J(j), \qquad i = 1, \ldots, n-1, \tag{2.10}$$

for all functions $J(i)$, which are monotonically nondecreasing in i. It can be shown that this assumption is satisfied if and only if, for every k, $\sum_{j=k}^{n} p_{ij}$ is monotonically nondecreasing in i (see [Ros83b], p. 252). The assumption (2.10) is satisfied in particular 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., there is greater chance of ending at a bad state j if we start at a worse state i. Since $g(i)$ is nondecreasing in i, we have that $(TJ_0)(i)$ is nondecreasing in i, and in view of the assumption (2.10) on the transition probabilities, 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

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

and let

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

Then, an optimal policy takes the form

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

as shown in Fig. 1.2.1.

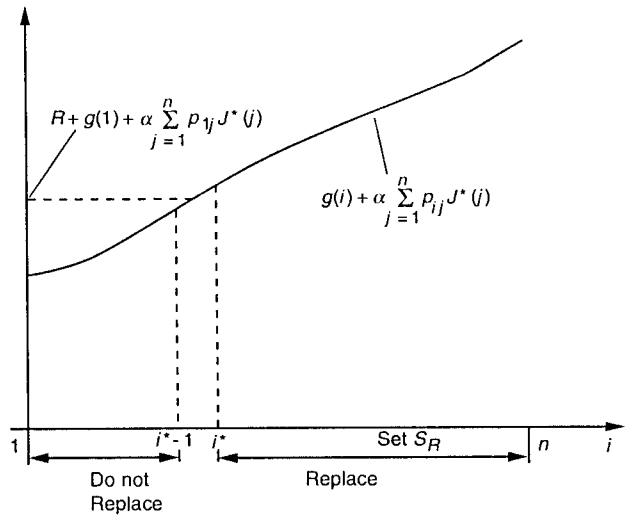


Figure 1.2.1 Determining the optimal policy in the machine replacement example.

1.3 FINITE-STATE SYSTEMS – COMPUTATIONAL METHODS

In this section we discuss several alternative approaches for numerically solving the discounted problem with bounded cost per stage. The first approach, value iteration, is essentially the DP algorithm and yields in the limit the optimal cost function and an optimal policy, as discussed in the preceding section. We will describe some variations aimed at accelerating convergence. Two other approaches, policy iteration and linear programming, terminate in a finite number of iterations (assuming the number of states and controls are finite). However, when the number of states is large, these approaches are impractical because of large overhead per iteration. Another approach, adaptive aggregation, bridges the gap between value iteration and policy iteration, and in a sense combines the best features of both methods.

In Section 2.3 we will consider some additional methods, which are well-suited for dynamic systems that are hard to model but relatively easy to simulate. In particular, we will assume in Section 2.3 that the transition probabilities of the problem are unknown, but the system's dynamics and cost structure can be observed through simulation. We will then discuss the methods of temporal differences and Q-learning, which also provide conceptual vehicles for approximate forms of value iteration and policy

iteration using, for example, neural networks.

Throughout this section we assume a discounted problem (Assumption D holds). We further assume that the state, control, and disturbance spaces underlying the problem are finite sets, so that we are dealing in effect with the control of a finite-state Markov chain.

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

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

We denote by $p_{ij}(u)$ the transition probabilities

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

These transition probabilities may be given a priori or 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

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

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

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

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

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

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

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

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

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

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

and by g_μ the cost vector

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

We can then write in vector notation

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

The cost function J_μ corresponding to a stationary policy μ is, by Cor. 2.2.1, the unique solution of the equation

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

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

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

or, equivalently,

$$J_\mu = (I - \alpha P_\mu)^{-1} g_\mu, \tag{3.1}$$

where I denotes the $n \times n$ identity matrix. The invertibility of the matrix $I - \alpha P_\mu$ is assured since we have proved that the system of equations representing $J_\mu = T_\mu J_\mu$ has a unique solution for any vector g_μ (cf. Cor. 2.2.1). For another way to see that $I - \alpha P_\mu$ is an invertible matrix, note that the eigenvalues of any transition probability matrix lie within the unit circle of the complex plane. Thus no eigenvalue of αP_μ can be equal to 1, which is the necessary and sufficient condition for $I - \alpha P_\mu$ to be invertible.

1.3.1 Value Iteration and Error Bounds

Here we start with any n-dimensional vector J and successively compute $TJ, T^2J, \ldots$. By Prop. 2.1, we have for all i

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

Furthermore, by Prop. 2.4 [using $J' = J^*$ in Eq. (2.9)], the error sequence $\big|(T^k J)(i) - J^*(i)\big|$ is bounded by a constant multiple of α^k, for all $i \in S$. This method is called *value iteration* or *successive approximation*. The method can be substantially improved thanks to certain monotonic error bounds, which are easily obtained as a byproduct of the computation.

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

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

It follows that

$$g_\mu + \left(\frac{\alpha\underline{\beta}}{1-\alpha}\right) e \le J_\mu \le g_\mu + \left(\frac{\alpha\overline{\beta}}{1-\alpha}\right) e, \tag{3.2}$$

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

$$\underline{\beta} = \min_i g\big(i,\mu(i)\big), \qquad \overline{\beta} = \max_i g\big(i,\mu(i)\big).$$

Using the definition of $\underline{\beta}$ and $\overline{\beta}$, we can strengthen the bounds (3.2) as follows:

$$\left(\frac{\underline{\beta}}{1-\alpha}\right) e \le g_\mu + \left(\frac{\alpha\underline{\beta}}{1-\alpha}\right) e \le J_\mu \le g_\mu + \left(\frac{\alpha\overline{\beta}}{1-\alpha}\right) e \le \left(\frac{\overline{\beta}}{1-\alpha}\right) e. \tag{3.3}$$

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

Suppose that we have a vector J and we compute

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

By subtracting this equation from the relation

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

we obtain

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

This equation can be viewed as a *variational* form of the equation $J_\mu = T_\mu J_\mu$, and implies that $J_\mu - J$ *is the cost vector associated with the stationary policy* μ *and a cost per stage vector equal to* $T_\mu J - J$. Therefore, the bounds (3.3) apply with J_μ replaced by $J_\mu - J$ and g_μ replaced by $T_\mu J - J$. It follows that

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

where

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

Equivalently, for every vector J, we have

$$J + \frac{\underline{c}}{\alpha} e \le T_\mu J + \underline{c} e \le J_\mu \le T_\mu J + \overline{c} e \le J + \frac{\overline{c}}{\alpha} e,$$

where

$$\underline{c} = \frac{\alpha\underline{\gamma}}{1-\alpha}, \qquad \overline{c} = \frac{\alpha\overline{\gamma}}{1-\alpha}.$$

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

Proposition 3.1: For every vector J, state i, and k, we have

$$\begin{aligned}(T^k J)(i) + \underline{c}_k &\le (T^{k+1} J)(i) + \underline{c}_{k+1} \\ &\le J^*(i) \\ &\le (T^{k+1} J)(i) + \overline{c}_{k+1} \\ &\le (T^k J)(i) + \overline{c}_k,\end{aligned} \tag{3.4}$$

where

$$\underline{c}_k = \frac{\alpha}{1-\alpha} \min_{i=1,\ldots,n} \big[(T^k J)(i) - (T^{k-1} J)(i)\big], \tag{3.5}$$

$$\overline{c}_k = \frac{\alpha}{1-\alpha} \max_{i=1,\ldots,n} \big[(T^k J)(i) - (T^{k-1} J)(i)\big]. \tag{3.6}$$

Proof: Denote

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

We have

$$J + \underline{\gamma} e \le TJ. \tag{3.7}$$

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

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

and, combining this relation with Eq. (3.7), we obtain

$$J + (1+\alpha)\underline{\gamma} e \le TJ + \alpha\underline{\gamma} e \le T^2 J. \tag{3.8}$$

This process can be repeated, first applying T to obtain

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

and then using Eq. (3.7) to write

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

After k steps, this results in the inequalities

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

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

$$J + \left(\frac{\underline{c}_1}{\alpha}\right) e \le TJ + \underline{c}_1 e \le T^2 J + \alpha \underline{c}_1 e \le J^*, \tag{3.9}$$

where $\underline{c}_1$ is defined by Eq. (3.5). Replacing J by $T^k J$ in this inequality, we have

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

which is the second inequality in Eq. (3.4).

From Eq. (3.8), we have

$$\alpha\underline{\gamma} \le \min_{i=1,\ldots,n} \big[(T^2 J)(i) - (TJ)(i)\big].$$

and consequently

$$\alpha c_1 \leq c_2.$$

Using this relation in Eq. (3.9) yields

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

and replacing J by $T^{k-1}J$, we have the first inequality in Eq. (3.1). An analogous argument shows the last two inequalities in Eq. (3.4). **Q.E.D.**

We note that the preceding proof does not rely on the finiteness of the state space, and indeed Prop. 3.1 can be proved for an infinite state space (see also Exercise 1.9). The following example demonstrates the nature of the error bounds.

Example 3.1 (Illustration of the Error Bounds)

Consider a problem where there are two states and two controls

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

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

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

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

The transition costs are

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

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

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

The scalars $\underline{c}_k$ and $\bar{c}_k$ of Eqs. (3.5) and (3.6) are given by

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

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

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

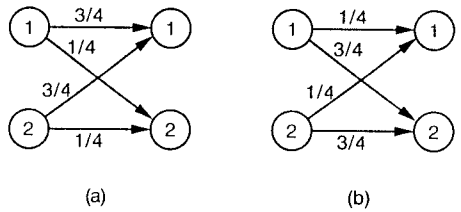


Figure 1.3.1 State transition diagram for Example 3.1: (a) $u = u^1$; (b) $u = u^2$.

k	$(T^k J_0)(1)$	$(T^k J_0)(2)$	$(T^k J_0)(1)$ $+\underline{c}_k$	$(T^k J_0)(1)$ $+\bar{c}_k$	$(T^k J_0)(2)$ $+\underline{c}_k$	$(T^k J_0)(2)$ $+\bar{c}_k$
0	0	0				
1	0.500	1.000	5.000	9.500	5.500	10.000
2	1.287	1.562	6.350	8.375	6.625	8.650
3	1.841	2.220	6.856	7.767	7.232	8.144
4	2.414	2.745	7.129	7.540	7.460	7.870
5	2.896	3.247	7.232	7.417	7.583	7.768
6	3.343	3.686	7.287	7.371	7.629	7.712
7	3.740	4.086	7.308	7.345	7.654	7.692
8	4.099	4.441	7.319	7.336	7.663	7.680
9	4.422	4.767	7.324	7.331	7.669	7.676
10	4.713	5.057	7.326	7.329	7.671	7.674
11	4.974	5.319	7.327	7.328	7.672	7.673
12	5.209	5.554	7.327	7.328	7.672	7.673
13	5.421	5.766	7.327	7.328	7.672	7.673
14	5.612	5.957	7.328	7.328	7.672	7.672
15	5.783	6.128	7.328	7.328	7.672	7.672

Figure 1.3.2 Performance of the value iteration method with and without the error bounds of Prop. 3.1 for the problem of Example 3.1.

Termination Issues – Optimality of the Obtained Policy

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

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

$$\max_i \big[J_\mu(i) - J^*(i)\big] \leq \frac{\alpha}{1-\alpha}\left(\max_i\big[(TJ)(i) - J(i)\big] - \min_i\big[(TJ)(i) - J(i)\big]\right). \tag{3.10}$$

To see this, apply Eq. (3.4) with $k = 1$ to obtain for all i

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

and also apply Eq. (3.4) with $k = 1$ and with T_μ replacing T to obtain

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

Subtracting the above two equations, we obtain the estimate (3.10).

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

$$\hat{J}_k = T^k J + \left(\frac{\overline{c}_k + \underline{c}_k}{2}\right) e \tag{3.11}$$

or the "average"

$$\hat{J}_k = T^k J + \frac{\alpha}{n(1-\alpha)} \sum_{i=1}^{n} \big((T^k J)(i) - (T^{k-1}J)(i)\big) e. \tag{3.12}$$

Both of these vectors lie in the region delineated by the error bounds. Then, the estimate (3.10) provides a bound on the suboptimality of the policy μ attaining the minimum in the calculation of $T^k J$.

The bound (3.10) can also be used to show that after a sufficiently large number of value iterations, the stationary policy μ^k that attains the minimum in the kth value iteration [i.e. $(T_{\mu^k} T^{k-1})J = T^k J$] is optimal. Indeed, since the number of stationary policies is finite, there exists an $\overline{\epsilon} > 0$ such that if a stationary policy μ satisfies

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

then μ is optimal. Now let $\overline{k}$ be such that for all $k \geq \overline{k}$ we have

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

Then from Eq. (3.10) we see that for all $k \geq \overline{k}$, the stationary policy that attains the minimum in the kth value iteration is optimal.

Rate of Convergence

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

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

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

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

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

It turns out, however, that when error bounds are used, the rate at which the iterates $\hat{J}_k$ and $\tilde{J}_k$ of Eqs. (3.11) and (3.12) approach the optimal cost vector J^* is governed by the modulus of the *subdominant* eigenvalue of the transition probability matrix P_{μ^*}, that is, the eigenvalue with second largest modulus. The proof of this is outlined in Exercise 1.8. For a sketch of the ideas involved, let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of P_{μ^*} ordered according to decreasing modulus; that is

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

with λ_1 equal to 1 and λ_2 being the subdominant eigenvalue. Assume that there is a set of linearly independent eigenvectors $c_1, c_2, \ldots, c_n$ corresponding to $\lambda_1, \lambda_2, \ldots, \lambda_n$ with $c_1 = e = (1, 1, \ldots, 1)'$. Then the initial error $J - J_{\mu^*}$ can be expressed as a linear combination of the eigenvectors

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

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

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

Thus the error after k iterations can be written as

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

Using the error bounds of Prop. 3.1 amounts to a translation of $T^k_{\mu^*} J$ along the vector e. Thus, at best, the error bounds are tight enough to eliminate the component $\alpha^k \xi_1 e$ of the error, but cannot affect the remaining term $\alpha^k \sum_{j=2}^{n} \lambda_j^k \xi_j c_j$, which diminishes like $\alpha^k |\lambda_2|^k$ with λ_2 being the subdominant eigenvalue.

Problems where Convergence is Slow

In Example 3.1, the convergence of value iteration with the error bounds is very fast. For this example, it can be verified that $\mu^*(1) = u^2$, $\mu^*(2) = u^1$, and that

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

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

As an example, consider the following three simple deterministic problems, each having a single policy and more than one recurrent class of states:

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

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

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

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

Figure 1.3.3 shows the number of iterations needed by the value iteration method with and without the error bounds of Prop. 3.1 to find J_μ within an error per coordinate of less than or equal to $10^{-6} \max_i \big|J_\mu(i)\big|$. The starting function in all cases was taken to be zero. The performance is rather unsatisfactory but, nonetheless, is typical of situations where the subdominant eigenvalue modulus of the optimal transition probability matrix is close to 1. One possible approach to improve the performance of value iteration for such problems is based on the adaptive aggregation method to be discussed in Section 1.3.3.

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

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

Elimination of Nonoptimal Actions in Value Iteration

We know from Prop. 2.3 that, if $\tilde{u} \in U(i)$ is such that

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

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

$$g(i, \tilde{u}) + \alpha \sum_{j=1}^{n} p_{ij}(\tilde{u}) \underline{J}(j) > \overline{J}(i), \tag{3.13}$$

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

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

The preceding observation is the basis for a useful application of the error bounds given earlier in Prop. 3.1. As these bounds are computed in the course of the value iteration method, the inequality (3.13) can be simultaneously checked and nonoptimal actions can be eliminated from the admissible set with attendant savings in subsequent computations. Since the upper and lower bound functions $\overline{J}$ and $\underline{J}$ converge to J^*, it can be seen [taking into account the finiteness of the constraint set $U(i)$] that eventually all nonoptimal $\tilde{u} \in U(i)$ will be eliminated, thereby reducing the set $U(i)$ after a finite number of iterations to the set of controls that are optimal at i. In this manner the computational requirements of value iteration can be substantially reduced. However, the amount of computer memory required to maintain the set of controls not as yet eliminated at each $i \in S$ may be increased.

Gauss-Seidel Version of Value Iteration

In the value iteration method described earlier, the estimate of the cost function is iterated for all states simultaneously. An alternative is to iterate one state at a time, while incorporating into the computation the interim results. This corresponds to using what is known as the *Gauss-Seidel method* for solving the nonlinear system of equations $J = TJ$ (see [BeT89a] or [OrR70]).

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

$$(FJ)(1) = \min_{u \in U(1)} \left[g(1,u) + \alpha \sum_{j=1}^{n} p_{1j}(u) J(j) \right] \tag{3.14}$$

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

$$(FJ)(i) = \min_{u \in U(i)} \left[g(i,u) + \alpha \sum_{j=1}^{i-1} p_{ij}(u)(FJ)(j) + \alpha \sum_{j=i}^{n} p_{ij}(u) J(j) \right]. \tag{3.15}$$

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

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

Proposition 3.2: Let J, J' be two n-dimensional vectors. Then for any $k = 0, 1, \ldots$,

$$\max_{i \in S} \left| (F^k J)(i) - (F^k J')(i) \right| \leq \alpha^k \max_{i \in S} \left| J(i) - J'(i) \right|. \tag{3.16}$$

Furthermore, we have

$$(FJ^*)(i) = J^*(i), \qquad i \in S, \tag{3.17}$$

$$\lim_{k \to \infty} (F^k J)(i) = J^*(i), \qquad i \in S. \tag{3.18}$$

Proof: It is sufficient to prove Eq. (3.16) for $k = 1$. We have by the definition of F and Prop. 2.4,

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

Also, using this inequality,

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

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

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

so Eq. (3.16) is proved for $k = 1$. The equation $FJ^* = J^*$ follows from the definition (3.14) and (3.15) of F, and Bellman's equation $J^* = TJ^*$. The convergence property (3.18) follows from Eqs. (3.16) and (3.17). **Q.E.D.**

Proposition 3.3: If an n-dimensional vector J satisfies

$$J(i) \leq (TJ)(i) \leq J^*(i), \qquad i = 1, \ldots, n,$$

then

$$(T^k J)(i) \leq (F^k J)(i) \leq J^*(i), \qquad i = 1, \ldots, n, \quad k = 1, 2, \ldots \tag{3.19}$$

Proof: The proof follows by using the definition (3.14) and (3.15) of F, and the monotonicity property of T (Lemma 1.1). **Q.E.D.**

The preceding proposition provides the main motivation for employing the mapping F in place of T in the value iteration method. The result indicates that the Gauss-Seidel version converges faster than the ordinary value iteration method. The faster convergence property can be substantiated by further analysis (see e.g., [BeT89a]) and has been confirmed in practice through extensive experimentation. This comparison is somewhat misleading, however, because the ordinary method will normally be used in conjunction with the error bounds of Prop. 3.1. One may also employ error bounds in the Gauss-Seidel version (see Exercise 1.9). However, there is no clear superiority of one method over the other when bounds are introduced. Furthermore, the ordinary method is better suited for parallel computation than the Gauss-Seidel version.

We note that there is a more flexible form of the Gauss-Seidel method, which selects states in arbitrary order to update their costs. This method maintains an approximation J to the optimal vector J^*, and at each iteration, it selects a state i and replaces $J(i)$ by $(TJ)(i)$. The remaining values $J(j)$, $j \neq i$, are left unchanged. The choice of the state i at each iteration is arbitrary, except for the restriction that all states are selected infinitely often. This method is an example of an *asynchronous fixed point iteration* and can be shown to converge to J^* starting from any initial J. Analyses of this type of method are given in [Ber82a], and in Chapter 6 of [BeT89a]; see also Exercise 1.15.

Generic Rank-One Corrections

We may view value iteration coupled with the error bounds of Prop. 3.1 as a method that makes a correction to the results of value iteration along the unit vector e. It is possible to generalize the idea of correction along a fixed vector so that it works for any type of convergent linear iteration.

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

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

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

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

An example is the Gauss-Seidel iteration of Section 1.3.1, and some other examples are given in Exercises 1.4, 1.5, and 1.7, and in Section 5.3. Also, the value iteration method for stochastic shortest path problems and a single stationary policy, to be discussed in Section 2.2, is of the above form.

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

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

where $\tilde{J}$ is related to J by

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

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

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

which, by denoting

$$z = Q_\mu d,$$

can be written as

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

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

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

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

$$J := MJ,$$

where

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

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

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

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

[compare with Eq. (3.12)]. Thus in this case the iteration $J := M(J)$ shifts the result $T_\mu J$ of value iteration to a vector that lies somewhere in the middle of the error bound range given by Prop. 3.1. By the result of this proposition it follows that the iteration converges to J_μ.

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

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

for some scalars $\xi_1, \ldots, \xi_n$, where $e_1, \ldots, e_n$ are eigenvectors of Q_μ, and $\lambda_1, \ldots, \lambda_n$ are corresponding eigenvalues. Suppose also that λ_1 is the

unique dominant eigenvalue, that is, $|\lambda_j| < |\lambda_1|$ for $j = 2, \ldots, n$. Then the difference $F^{k+1}J - F^kJ$ is nearly equal to $\xi_1(\lambda_1^{k+1} - \lambda_1^k)e_1$ for large k and can be used to estimate the dominant eigenvector e_1. In order to decide whether k has been chosen large enough, one can test to see if the angle between the successive differences $F^{k+1}J - F^kJ$ and $F^kJ - F^{k-1}J$ is very small; if this is so, the components of $F^{k+1}J - F^kJ$ along the eigenvectors $e_2, \ldots, e_n$ must also be very small. (For a more sophisticated version of this argument, see [Ber93], where the generic rank-one correction method is developed in more general form.)

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

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

is sufficiently close to one.

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

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

$$(FJ)(i) = \min_{u \in U(i)} \left[h_i(u) + \sum_{j=1}^{n} q_{ij}(u)J(j) \right], \qquad i = 1, \ldots, n. \tag{3.20}$$

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

(1) Iteratively compute the cost vectors of the policies generated by a policy iteration scheme of the type discussed in the next subsection.

(2) Guess at an optimal policy within the first phase, switch to the second phase, and then return to the first phase if the policy changes "substantially" during the second phase. In particular, in the first phase, the iteration $J := FJ$ is used, where F is the nonlinear mapping of Eq. (3.20). Upon switching to the second phase, the vector z is taken to be equal to $Q_{\mu^*}d$, where μ^* is the policy that attains the minimum in Eq. (3.20) at the time of the switch. The second phase consists of the iteration

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

where F is the nonlinear mapping of Eq. (3.20), and $\tilde{\gamma}$ is again given by

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

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

It should be mentioned, however, that the rank-one correction method is ineffective if there is little or no separation between the dominant and the subdominant eigenvalue moduli, both because the convergence rate of the method for obtaining d is slow, and also because the convergence rate of the modified iteration $J := MJ$ is not much faster than the one of the regular iteration $J := FJ$. For such problems, one should try corrections over subspaces of dimension larger than one (see [Ber93], and the adaptive aggregation and multiple-rank correction methods given in Section 1.3.3).

Infinite State Space – Approximate Value Iteration

The value iteration method is valid under the assumptions of Prop. 2.1, so it is guaranteed to converge to J^* for problems with infinite state and control spaces. However, for such problems, the method may be implementable only through approximations. In particular, given a function J, one may only be able to calculate a function $\tilde{J}$ such that

$$\max_{x \in S} |\tilde{J}(x) - (TJ)(x)| \leq \epsilon, \tag{3.21}$$

where ϵ is a given positive scalar. A similar situation may occur even when the state space is finite but the number of states is very large. Then instead of calculating $(TJ)(x)$ for all states x, one may do so only for some states and estimate $(TJ)(x)$ for the remaining states x by some form of interpolation, or by a least-squares error fit of $(TJ)(x)$ with a function from a suitable parametric class (compare with the discussion of Section 2.3). Then the function $\tilde{J}$ thus obtained will satisfy a relation such as (3.21).

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

$$\max_{x \in S} |J_{k+1}(x) - (TJ_k)(x)| \leq \epsilon, \qquad k = 0, 1, \ldots \tag{3.22}$$

starting from an arbitrary bounded function J_0. Generally, such a sequence "converges" to J^* to within an error of $\epsilon/(1-\alpha)$. To see this, note that Eq. (3.22) yields

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

By applying T to this relation, we obtain

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

so by using Eq. (3.22) to write

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

we have

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

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

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

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

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

It is also possible to obtain versions of the error bounds of Prop. 3.1 for the approximate value iteration method. We have from that proposition

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

By using Eq. (3.22) in the above relation, we obtain

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

or

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

These bounds hold even when the state space is infinite because the bounds of Prop. 3.1 can be shown for an infinite state space as well. However, for these bounds to be useful, one should know ϵ.

1.3.2 Policy Iteration

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

The algorithm is based on the following proposition.

Proposition 3.4: Let μ and $\overline{\mu}$ be stationary policies such that $T_{\overline{\mu}} J_\mu = TJ_\mu$, or equivalently, for $i = 1, \ldots, n$,

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

Then we have

$$J_{\overline{\mu}}(i) \leq J_\mu(i), \qquad i = 1, \ldots, n. \tag{3.23}$$

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

Proof: Since $J_\mu = T_\mu J_\mu$ (Cor. 2.2.1) and, by hypothesis, $T_{\overline{\mu}} J_\mu = TJ_\mu$, we have for every i,

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

Applying repeatedly $T_{\overline{\mu}}$ on both sides of this inequality and using the monotonicity of $T_{\overline{\mu}}$ (Lemma 1.1) and Cor. 2.1.1, we obtain

$$J_\mu \geq T_{\overline{\mu}} J_\mu \geq \cdots \geq T_{\overline{\mu}}^k J_\mu \geq \cdots \geq \lim_{N\to\infty} T_{\overline{\mu}}^N J_\mu = J_{\overline{\mu}},$$

proving Eq. (3.23).

If $J_\mu \neq J_{\overline{\mu}}$, then from the preceding relation it follows that $J_\mu \neq T_{\overline{\mu}} J_\mu$ and since by hypothesis we have $T_{\overline{\mu}} J_\mu = TJ_\mu$, we obtain $J_\mu \neq TJ_\mu$, implying that $J_\mu \neq J^*$ by Prop. 2.2. Thus μ must be optimal. It follows that if μ is not optimal, then $J_{\overline{\mu}}(i) < J_\mu(i)$ for some state i. **Q.E.D.**

Policy Iteration Algorithm

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

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

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

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

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

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

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

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

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

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

Example 3.1 (continued)

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

Initialization: We select the initial stationary policy

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

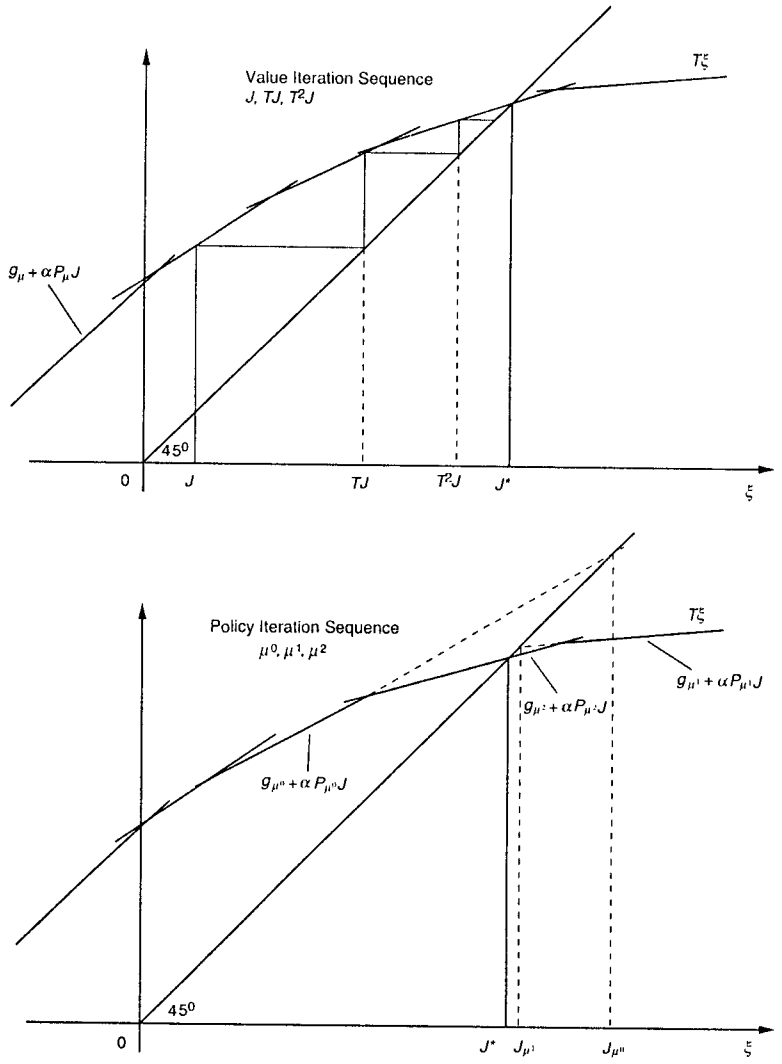


Figure 1.3.4 Geometric interpretation of policy iteration and value iteration. Each stationary policy μ defines the linear function $g_\mu + \alpha P_\mu J$ of the vector J, and TJ is the piecewise linear function $\min_\mu[g_\mu + \alpha P_\mu J]$. The optimal cost J^* satisfies $J^* = TJ^*$, so it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

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

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

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

Substituting the data of the problem, we have

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

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

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

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

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

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

$$(TJ_{\mu^0})(1) = \min \left\{ 1 + 0.9 \left(\frac{3}{4} \cdot 24.12 + \frac{1}{4} \cdot 25.96 \right), \right.$$
$$\left. 3 + 0.9 \left(\frac{1}{4} \cdot 24.12 + \frac{3}{4} \cdot 25.96 \right) \right\}$$
$$= \min\{23.12, 25.95\} = 23.12.$$

The minimizing controls are

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

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

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

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

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

$$J_{\mu^1}(1) \sim 7.33, \qquad J_{\mu^1}(2) \sim 7.67.$$

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

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

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

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

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

Modified Policy Iteration

When the number of states is large, solving the linear system $(I - \alpha P_{\mu^k}) J_{\mu^k} = g_{\mu^k}$ in the policy evaluation step by direct methods such as Gaussian elimination can be prohibitively time-consuming. One way to get around this difficulty is to solve the linear system iteratively by using value iteration. In fact, we may consider solving the system only approximately by executing a limited number of value iterations. This is called the *modified policy iteration algorithm*.

To formalize this method, let J_0 be an arbitrary n-dimensional vector. Let $m_0, m_1, \ldots$ be positive integers, and let the vectors $J_1, J_2, \ldots$ and the stationary policies $\mu_0, \mu_1, \ldots$ be defined by

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

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

Proposition 3.5: Let $\{J_k\}$ and $\{\mu_k\}$ be the sequences generated by the modified policy iteration algorithm. Then $\{J_k\}$ converges to J^*. Furthermore, there exists an integer $\overline{k}$ such that for all $k \geq \overline{k}$, μ^k is optimal.

Proof: Let r be a scalar such that the vector $\overline{J}_0$, defined by $\overline{J}_0 = J_0 + re$, satisfies $T\overline{J}_0 \leq \overline{J}_0$. [Any scalar r such that $\max_i[(TJ_0)(i) - J_0(i)] \leq (1-\alpha)r$ has this property.] Define for all k, $\overline{J}_{k+1} = T_{\mu^k}^{m_k}\overline{J}_k$. Then, it can be seen by induction that for all k and $m = 0, 1, \ldots, m_k$, the vectors $T_{\mu^k}^m J_k$ and $T_{\mu^k}^m \overline{J}_k$ differ by the multiple of the unit vector $r\alpha^{m_0+\cdots+m_{k-1}+m}e$. It follows that if J_0 is replaced by $\overline{J}_0$ as the starting vector in the algorithm, the same sequence of policies $\{\mu_k\}$ will be obtained; that is, we have for all k

$$T_{\mu^k}\overline{J}_k = T\overline{J}_k.$$

Now we will show that for all k we have $\overline{J}_k \leq T^k\overline{J}_0$. Indeed, we have $T_{\mu^0}\overline{J}_0 = T\overline{J}_0 \leq \overline{J}_0$, from which we obtain

$$T_{\mu^0}^m\overline{J}_0 \leq T_{\mu^0}^{m-1}\overline{J}_0, \qquad m = 1, 2, \ldots$$

so that

$$T_{\mu^1}\overline{J}_1 = T\overline{J}_1 \leq T_{\mu^0}\overline{J}_1 = T_{\mu^0}^{m_0+1}\overline{J}_0 \leq T_{\mu^0}^{m_0}\overline{J}_0 = \overline{J}_1 \leq T_{\mu^0}\overline{J}_0 = T\overline{J}_0.$$

This argument can be continued to show that for all k, we have $\overline{J}_k \leq T\overline{J}_{k-1}$, so that

$$\overline{J}_k \leq T^k\overline{J}_0, \qquad k = 0, 1, \ldots$$

On the other hand, since $T\overline{J}_0 \leq \overline{J}_0$, we have $J^* \leq \overline{J}_0$, and it follows that application of any number of mappings of the form T_μ to $\overline{J}_0$ produces functions that are bounded from below by J^*. Thus,

$$J^* \leq \overline{J}_k \leq T^k\overline{J}_0, \qquad k = 0, 1, \ldots.$$

By taking the limit as $k \to \infty$, we obtain $\lim_{k\to\infty}\overline{J}_k(i) = J^*(i)$ for all i, and since $\lim_{k\to\infty}(J_k - \overline{J}_k) = 0$, we obtain

$$\lim_{k\to\infty} J_k(i) = J^*(i), \qquad i = 1, \ldots, n.$$

Since the number of stationary policies is finite, there exists an $\overline{\epsilon} > 0$ such that if a stationary policy μ satisfies

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

then μ is optimal. Now let $\overline{k}$ be such that for all $k \geq \overline{k}$ we have

$$\frac{\alpha}{1-\alpha}\left(\max_i\big[(TJ_k)(i) - J_k(i)\big] - \min_i\big[(TJ_k)(i) - J_k(i)\big]\right) < \overline{\epsilon}.$$

Then from Eq. (3.10) we see that for all $k \geq \overline{k}$, the stationary policy μ^k that satisfies $T_{\mu^k}J_k = TJ_k$ is optimal. **Q.E.D.**

Note that if $m_k = 1$ for all k in the modified policy iteration algorithm, we obtain the value iteration method, while if $m_k = \infty$ we obtain the policy iteration method, where the policy evaluation step is performed iteratively by means of value iteration. Analysis and computational experience suggest that it is usually best to take m_k larger than 1 according to some heuristic scheme. A key idea here is that a value iteration involving a single policy (evaluating $T_\mu J$ for some μ and J) is much less expensive than an iteration involving all policies (evaluating TJ for some J), when the number of controls available at each state is large. Note that error bounds such as the ones of Prop. 3.1 can be used to improve the approximation process. Furthermore, Gauss-Seidel iterations can be used in place of the usual value iterations.

Infinite State Space – Approximate Policy Iteration

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

$$T_{\mu^{k+1}}J_{\mu^k} = TJ_{\mu^k}, \qquad k = 0, 1, \ldots.$$

The proof of Prop. 3.4 can then be used to show that the generated sequence of policies $\{\mu^k\}$ is improving in the sense that $J_{\mu^{k+1}} \leq J_{\mu^k}$ for all k. However, for infinite state space problems, the policy evaluation step and/or the policy improvement step of the method may be implementable only through approximations. A similar situation may occur even when the state space is finite but the number of states is very large.

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

$$\max_{x\in S}|J_k(x) - J_{\mu^k}(x)| \leq \delta, \qquad k = 0, 1, \ldots \tag{3.24}$$

and

$$\max_{x\in S}|(T_{\mu^{k+1}}J_k)(x) - (TJ_k)(x)| \leq \epsilon, \qquad k = 0, 1, \ldots. \tag{3.25}$$

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

Proposition 3.6: The sequence $\{\mu^k\}$ generated by the approximate policy iteration algorithm satisfies

$$\limsup_{k\to\infty} \max_{x\in S}\big(J_{\mu^k}(x) - J^*(x)\big) \leq \frac{\epsilon + 2\alpha\delta}{(1-\alpha)^2}. \tag{3.26}$$

Proof: From Eqs. (3.24) and (3.25), we have for all k

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

where $e = (1, 1, \ldots, 1)'$ is the unit vector, while from Eq. (3.24), we have for all k

$$TJ_k \leq TJ_{\mu^k} + \alpha\delta e.$$

By combining these two relations, we obtain for all k

$$T_{\mu^{k+1}}J_{\mu^k} \leq TJ_{\mu^k} + (\epsilon + 2\alpha\delta)e \leq T_{\mu^k}J_{\mu^k} + (\epsilon + 2\alpha\delta)e. \tag{3.27}$$

From Eq. (3.27) and the equation $T_{\mu^k}J_{\mu^k} = J_{\mu^k}$, we have

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

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

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

which can be written as

$$J_{\mu^{k+1}} - J_{\mu^k} \leq \alpha F_k + (\epsilon + 2\alpha\delta)e, \tag{3.28}$$

where F_k is the function given by

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

Let

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

Then we have $F_k(x) \leq \xi_k$ for all $x \in S$, and Eq. (3.28) yields

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

or

$$\xi_k \leq \frac{\epsilon + 2\alpha\delta}{1-\alpha}. \tag{3.29}$$

Let

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

From Eq. (3.27) and the relation

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

which follows from Prop. 2.4, we have

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

We also have

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

and by subtracting the last two relations, we obtain

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

From this relation we see that

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

By taking the limit superior as $k \to \infty$ and by using Eq. (3.29), we obtain

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

This relation simplifies to

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

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

Proposition 3.6 suggests that the approximate policy iteration method makes steady progress up to a point and then the iterates J_{μ^k} oscillate within a neighborhood of the optimum J^*. This behavior appears to be typical in practice. Note that for $\delta = 0$ and $\epsilon = 0$, Prop. 3.6 shows that the cost sequence $\{J_{\mu^k}\}$ generated by the (exact) policy iteration algorithm converges to J^*, even when the state space is infinite.

1.3.3 Adaptive Aggregation

Let us now consider an alternative to value iteration for performing approximate evaluation of a stationary policy μ, that is, for solving approximately the system

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

This alternative is recommended for problems where convergence of value iteration, even with error bounds, is very slow. The idea here is to solve instead of the system $J_\mu = T_\mu J_\mu$, another system of smaller dimension, which is obtained by lumping together the states of the original system into subsets $S_1, S_2, \ldots, S_m$ that can be viewed as *aggregate states*. These subsets are disjoint and cover the entire state space, that is,

$$S = S_1 \cup S_2 \cup \cdots \cup S_m.$$

Consider the $n \times m$ matrix W whose ith column has unit entries at coordinates corresponding to states in S_i and all other entries equal to zero. Consider also an $m \times n$ matrix Q such that the ith row of Q is a probability distribution $(q_{i1}, \ldots, q_{in})$ with $q_{is} = 0$ if $s \notin S_i$. The structure of Q implies two useful properties:

(a) $QW = I$.

(b) The matrix

$$R = QP_\mu W$$

is an $m \times m$ transition probability matrix. In particular, the ijth component of R is equal to

$$r_{ij} = \sum_{s \in S_i} q_{is} \sum_{t \in S_j} p_{st}\big(\mu(s)\big),$$

and gives the probability that the next state will belong to aggregate state S_j given that the current state is drawn from the aggregate state S_i according to the probability distribution $\{q_{is} \mid s \in S_i\}$. The transition probability matrix R defines a Markov chain, called the *aggregate Markov chain*, whose states are the m aggregate states. Figure 1.3.5 illustrates the aggregate Markov chain.

Aggregate Markov chains are most useful when their transition behavior captures the broad attributes of the behavior of the original chain. This is generally true if the states of each aggregate state are "similar" in some sense. Let us describe one such situation. In particular, suppose that we have an estimate J of J_μ and that *we postulate that over the states s of every aggregate state S_i the variation $J_\mu(s) - J(s)$ is constant*. This amounts to hypothesizing that for some m-dimensional vector y we have

$$J_\mu - J = Wy.$$

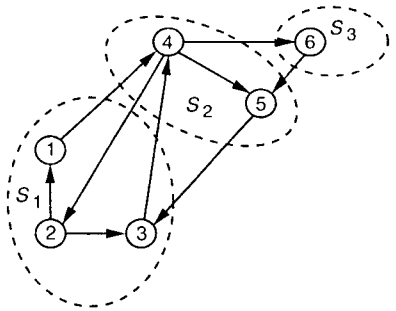


Figure 1.3.5 Illustration of the aggregate Markov chain. In this example, the aggregate states are $S_1 = \{1,2,3\}$, $S_2 = \{4,5\}$, and $S_3 = \{6\}$. The matrix W has columns $(1,1,1,0,0,0)'$, $(0,0,0,1,1,0)'$, and $(0,0,0,0,0,1)'$. In this example, the matrix Q is chosen so that each of its rows defines a uniform probability distribution over the states of the corresponding aggregate state. Thus the rows of Q are $(1/3,1/3,1/3,0,0,0)$, $(0,0,0,1/2,1/2,0)$, and $(0,0,0,0,0,1)$. The aggregate Markov chain has transition probabilities $r_{11} = \frac{1}{3}(p_{21}+p_{23})$, $r_{12} = \frac{1}{3}(p_{14}+p_{34})$, $r_{13} = 0$, $r_{21} = \frac{1}{2}(p_{42}+p_{53})$, $r_{22} = \frac{1}{2}p_{45}$, $r_{23} = \frac{1}{2}p_{46}$, $r_{31} = 0$, $r_{32} = p_{56}$, and $r_{33} = 0$.

By combining the equations $T_\mu J = g_\mu + \alpha P_\mu J$ and $g_\mu = (I - \alpha P_\mu)J_\mu$, we have

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

This is the variational form of the equation $J_\mu = T_\mu J_\mu$ discussed earlier in connection with error bounds in Section 1.3.1, and can be used equally well for evaluating J_μ. Let us multiply both sides with Q and use the equation $J_\mu - J = Wy$. We obtain

$$Q(I - \alpha P_\mu)Wy = Q\big(T_\mu J - J\big),$$

which, by using the equations $QW = I$ and $R = QP_\mu W$, is written as

$$(I - \alpha R)y = Q\big(T_\mu J - J\big).$$

This equation can be solved for y, since R is a transition probability matrix and therefore the matrix $I - \alpha R$ is invertible. Also, by applying T_μ to both sides of the equation $J_\mu = J + Wy$, we obtain

$$J_\mu = T_\mu J_\mu = T_\mu J + \alpha P_\mu W y.$$

We thus conclude that, if the variation of $J_\mu(s) - J(s)$ is roughly constant over the states s of each aggregate state, then the vector $T_\mu J + \alpha P_\mu W y$ is a good approximation for J_μ. Starting with J, this approximation is obtained as follows.

Aggregation Iteration

Step 1: Compute $T_\mu J$.

Step 2: Delineate the aggregate states (i.e., define W) and specify the matrix Q.

Step 3: Solve for y the system

$$(I - \alpha R)y = Q(T_\mu J - J), \tag{3.30}$$

where $R = QP_\mu W$, and approximate J_μ using

$$J := T_\mu J + \alpha P_\mu W y. \tag{3.31}$$

Note that the aggregation iteration (3.31) can be equivalently written as

$$J := T_\mu(J + Wy),$$

so it differs from a value iteration in that it operates with T_μ on $J + Wy$ rather than J.

Solving the system (3.30) in the aggregation iteration has an interesting interpretation. It can be seen that y is the α-discounted cost vector corresponding to the transition probability matrix R and the cost-per-stage vector $Q(T_\mu J - J)$. Thus, calculating y can be viewed as a policy evaluation step for the aggregate Markov chain when the cost per stage for each aggregate state S_i is equal to

$$\sum_{s \in S_i} q_{is}\big((T_\mu J)(s) - J(s)\big).$$

which is the average $T_\mu J - J$ over the aggregate state S_i according to the distribution $\{q_{is} \mid s \in S_i\}$. A key attractive aspect of the aggregation iteration is that the dimension of the system (3.30) is m (the number of aggregate states), which can be much smaller than n (the dimension of the system $J_\mu = T_\mu J_\mu$ arising in the policy evaluation step of policy iteration).

Delineating the Aggregate States

A key issue is how to identify the aggregate states $S_1, \ldots, S_m$ in a way that the error $J_\mu - J$ is of similar magnitude in each one. One way to do this is to view $T_\mu J$ as an approximation to J_μ and to group together states i with comparable magnitudes of $(T_\mu J)(i) - J(i)$. Thus the interval $[\underline{c}, \overline{c}]$, where

$$\underline{c} = \min_i\big[(T_\mu J)(i) - J(i)\big], \qquad \overline{c} = \max_i\big[(T_\mu J)(i) - J(i)\big],$$

is divided into m segments and membership of a state i in an aggregate state is determined by the segment within which $(T_\mu J)(i) - J(i)$ lies. By this we mean that for each state i, we set $i \in S_1$ if $(T_\mu J)(i) - J(i) = \underline{c}$, and we set

$$i \in S_k \qquad \text{if} \qquad (T_\mu J)(i) - J(i) - \underline{c} - (k-1)\delta \in (0, \delta],$$

where

$$\delta = \frac{\overline{c} - \underline{c}}{m}.$$

This choice is based on the conjecture that, at least near convergence, $(T_\mu J)(i) - J(i)$ will be of comparable magnitude for states i for which $J_\mu(i) - J(i)$ is of comparable magnitude. Analysis and experimentation given in [BeC89] has shown that the preceding scheme often works well with a small number of aggregates states m (say 3 to 6), although the properties of the method are yet fully understood.

Note that the aggregate states can change from one iteration to the next, so the aggregation scheme "adapts" to the progress of the computation. The criterion used to delineate the aggregate states does not exploit any special problem structure. In some cases, however, it is possible to take advantage of existing special structure and modify accordingly the method used to form the aggregate states.

Adaptive Aggregation Methods

It is possible to construct a number of methods that calculate J_μ by using aggregation iterations. One possibility is simply to perform a sequence of aggregation iterations using the preceding method to partition the state space into a few, say 3 to 10, aggregate states. This method can be greatly improved by interleaving each aggregation iteration with multiple value iterations (applications of the mapping T_μ on the current iterate). This is recommended based on experimentation and analysis given in [BeC89], to which we refer for further discussion. An interesting empirically observed phenomenon is that the error between the iterate and J_μ is often increased by an aggregation iteration, but then unusually large improvements are made during the next few value iterations. This suggests that the number of value iterations following an aggregation iteration should be based on algorithmic progress; that is, an aggregation iteration should be performed when the progress of the value iterations becomes relatively small. Some experimentation may be needed with a given problem to determine an appropriate criterion for switching from the value iterations to an aggregation iteration.

There is no proof of convergence of the scheme just described. On the basis of computational experimentation, it appears reliable in practice. Its convergence nonetheless can be guaranteed by introducing a feature that

enforces some irreversible progress via the value iteration method following an aggregation iteration. In particular, one may calculate the error bounds of Prop. 3.1 at the value iteration Step 1, and impose a requirement that the subsequent aggregation iteration is skipped if these error bounds do not improve by a certain factor over the bounds computed prior to the preceding aggregation iteration.

To illustrate the effectiveness of the adaptive aggregation method, consider the three deterministic problems described earlier (cf. Fig. 1.3.3), and the performance of the method with two, three, and four aggregate states, starting from the zero function. The results, given in Fig. 1.3.6, should be compared with those of Fig. 1.3.3.

It is intuitively clear that the performance of the aggregation method should improve as the number of aggregate states increases, and indeed the computational results bear this out. The two extreme cases where $m = n$ and $m = 1$ are of interest. When $m = n$, each aggregate state has a single state and we obtain the policy iteration algorithm. When $m = 1$, there is only one aggregate state, W is equal to the unit vector $e = (1, \ldots, 1)'$, and a straightforward calculation shows that for the choice $Q = (1/n, \ldots, 1/n)$, the solution of the aggregate system (3.30) is

$$y = \frac{1}{n(1-\alpha)} \sum_{i=1}^{n} \big((T_\mu J)(i) - J(i)\big).$$

From this equation (using also the fact $P_\mu e = e$), we obtain the iteration

$$J := T_\mu J + \frac{\alpha}{n(1-\alpha)} \sum_{i=1}^{n} \big((T_\mu J)(i) - J(i)\big) e, \tag{3.32}$$

which is the same as the rank-one correction formula (3.12) obtained in Section 1.3.1 and amounts to shifting the result $T_\mu J$ of value iteration within the error bound range given by Prop. 3.1. Thus we may view the aggregation scheme as a continuum of algorithms with policy iteration and value iteration (coupled with the error bounds of Prop. 3.1) included as the two extreme special cases.

Adaptive Multiple-Rank Corrections

One may observe that the aggregation iteration

$$J := T_\mu(J + Wy),$$

amounts to applying T_μ to a correction of J along the subspace spanned by the columns of W. Once the matrix W is computed based on the adaptive procedure discussed above, we may consider choosing the vector y in alternative ways. An interesting possibility, which leads to a generalization

No. of aggregate states	Pr. 1 $\alpha = .9$	Pr. 1 $\alpha = .99$	Pr. 2 $\alpha = .9$	Pr. 2 $\alpha = .99$	Pr. 3 $\alpha = .9$	Pr. 3 $\alpha = .99$
2	14	13	9	9	83	505
3	1	1	3	3	64	367
4			3	3	26	354

Figure 1.3.6 Number of iterations of adaptive aggregation methods with two, three, and four aggregate states to solve the problems of Fig. 1.3.3. Each row of Q was chosen to define a uniform probability distribution over the states of the corresponding aggregate state.

of the rank-one correction method of the preceding subsection, is to select y so that

$$\big\|J + Wy - T_\mu(J + Wy)\big\|^2 \tag{3.33}$$

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

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

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

$$J := T_\mu(J + W\tilde{y}) = T_\mu J + \alpha P_\mu W \tilde{y}.$$

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

1.3.4 Linear Programming

Since $\lim_{N\to\infty} T^N J = J^*$ for all J (cf. Prop. 2.1), we have

$$J \le TJ \qquad \Rightarrow \qquad J \le J^* = TJ^*.$$

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

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

and delineates a polyhedron in $\Re^n$. The optimal cost vector J^* is the "northeast" corner of this polyhedron, as illustrated in Fig. 1.3.7. In particular, $J^*(1), \ldots, J^*(n)$ solve the following problem (in $\lambda_1, \ldots, \lambda_n$):

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

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

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

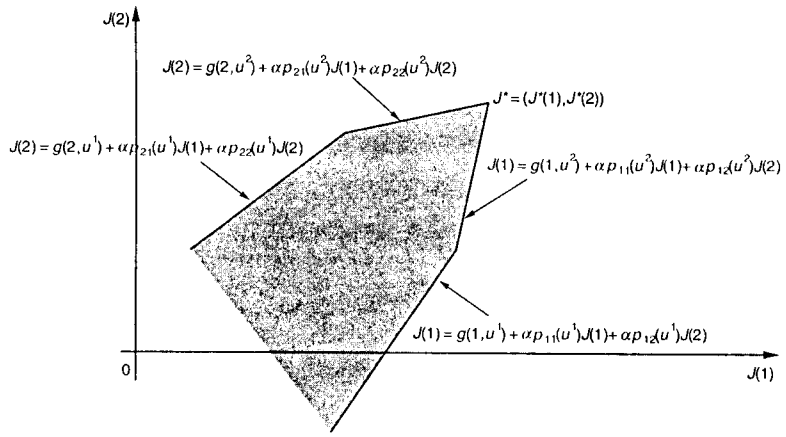


Figure 1.3.7 Linear programming problem associated with the discounted infinite horizon problem. The constraint set is shaded and the objective to maximize is $J(1) + J(2)$.

Example 3.1 (continued)

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

$$\text{maximize } \lambda_1 + \lambda_2$$

$$\text{subject to} \quad \lambda_1 \leq 2 + 0.9\left(\tfrac{3}{4}\lambda_1 + \tfrac{1}{4}\lambda_2\right), \quad \lambda_1 \leq 0.5 + 0.9\left(\tfrac{1}{4}\lambda_1 + \tfrac{3}{4}\lambda_2\right),$$
$$\lambda_2 \leq 1 + 0.9\left(\tfrac{3}{4}\lambda_1 + \tfrac{1}{4}\lambda_2\right), \quad \lambda_2 \leq 3 + 0.9\left(\tfrac{1}{4}\lambda_1 + \tfrac{3}{4}\lambda_2\right).$$

Cost Approximation Based on Linear Programming

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

$$\tilde{J}(x,r) = \sum_{k=1}^{m} r_k w_k(x), \tag{3.34}$$

where $r = (r_1, \ldots, r_m)$ is a vector of parameters, and for each state x, $w_k(x)$ are some fixed and known scalars. This amounts to approximating the cost function $J^*(x)$ by a linear combination of m given functions $w_k(x)$, where $k = 1, \ldots, m$. These functions play the role of a *basis* for the space of cost function approximations $\tilde{J}(x,r)$ that can be generated with different choices of r (see also the discussion of approximations in Section 2.3.3).

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

$$\text{maximize} \quad \sum_{x \in \tilde{S}} \tilde{J}(x,r)$$

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

where $\tilde{S}$ is either the state space S or a suitably chosen finite subset of S, and $\tilde{U}(x)$ is either $U(x)$ or a suitably chosen finite subset of $U(x)$. Because $\tilde{J}(x,r)$ is linear in the parameter vector r, the above program is linear in the parameters $r_1, \ldots, r_m$. Thus if m is small, the number of variables of the linear program is small. The number of constraints is as large as $s \cdot q$, where s is the number of elements of $\tilde{S}$ and q is the maximum number of elements of the sets $\tilde{U}(x)$. However, linear programs with a small number of variables and a large number of constraints can often be solved relatively quickly with the use of special large-scale linear programming methods known as cutting plane or column generation methods (see e.g. [Dan63], [Ber95a]). Thus, the preceding linear programming approach may be practical even for problems with a very large number of states.

Approximate Policy Evaluation Using Linear Programming

In the case of a very large or infinite state space, it is also possible to use linear programming to evaluate approximately the cost function J_μ of a stationary policy μ in the context of the approximate policy iteration scheme discussed in Section 1.3.2. Suppose that we wish to approximate J_μ by a function $\tilde{J}(\cdot, r)$ of a given form, which is parameterized by the vector $r = (r_1, \ldots, r_m)$. The bound of Prop. 3.6 suggests that we should try to determine the parameter vector r so as to minimize

$$\max_{x \in S} \left| \tilde{J}(x, r) - J_\mu(x) \right|.$$

From the error bounds given just prior to Prop. 3.1, it can also be seen that we have

$$\max_{x \in S} \left| \tilde{J}(x, r) - J_\mu(i) \right| \leq \frac{1}{1 - \alpha} \max_{x \in S} \left| \tilde{J}(x, r) - (T_\mu \tilde{J})(x, r) \right|.$$

This motivates choosing r by solving the problem

$$\min_r \max_{x \in \tilde{S}} \left| \tilde{J}(x, r) - (T_\mu \tilde{J})(x, r) \right|,$$

where $\tilde{S}$ is either the state space S or a suitably chosen finite subset of S. The preceding problem is equivalent to

$$\text{minimize} \quad z$$

$$\text{subject to} \quad \left| \tilde{J}(x, r) - g\big(x, \mu(x)\big) - \alpha \sum_{y \in S} p_{xy}\big(\mu(x)\big) \tilde{J}(y, r) \right| \leq z, \quad x \in \tilde{S}.$$

When $\tilde{J}(x, r)$ has the linear form (3.34), this is a linear program in the variables z and $r_1, \ldots, r_m$.

1.4 THE ROLE OF CONTRACTION MAPPINGS

Two key structural properties in DP models are responsible for most of the mathematical results one can prove about them. The first is the *monotonicity property* of the mappings T and T_μ (cf. Lemma 1.1 in Section 1.1). This property is fundamental for total cost infinite horizon problems. For example, it forms the basis for the results on positive and negative DP models to be shown in Chapter 3.

When the cost per stage is bounded and there is discounting, however, we have another property that strengthens the effects of monotonicity: the mappings T and T_μ are *contraction mappings*. In this section, we explain the meaning and implications of this property. The material in this section is conceptually very important, since contraction mappings are present in several additional DP models. However, the main result of this section (Prop. 4.1) will not be used explicitly in any of the proofs given later in this book.

Let $B(S)$ denote the set of all bounded real-valued functions on S. With every function $J : S \mapsto \Re$ that belongs to $B(S)$, we associate the scalar

$$\|J\| = \max_{x \in S} \left| J(x) \right|. \tag{4.1}$$

[As an aid for the advanced reader, we mention that the function $\|\cdot\|$ may be shown to be a norm on the linear space $B(S)$, and with this norm $B(S)$ becomes a complete normed linear space [Lue69].] The following definition and proposition are specializations to $B(S)$ of a more general notion and result that apply to such a space (see, e.g., references [LiS61] and [Lue69]).

Definition 4.1: A mapping $H : B(S) \mapsto B(S)$ is said to be a *contraction mapping* if there exists a scalar $\rho < 1$ such that

$$\left\| HJ - HJ' \right\| \leq \rho \|J - J'\|, \qquad \text{for all } J, J' \in B(S),$$

where $\|\cdot\|$ is the norm of Eq. (4.1). It is said to be an *m-stage contraction mapping* if there exists a positive integer m and some $\rho < 1$ such that

$$\left\| H^m J - H^m J' \right\| \leq \rho \|J - J'\|, \qquad \text{for all } J, J' \in B(S), \tag{4.2}$$

where H^m denotes the composition of H with itself m times.

The main result concerning contraction mappings is the following. For a proof, see references [LiS61] and [Lue69].

Proposition 4.1: (Contraction Mapping Fixed-Point Theorem) If $H : B(S) \mapsto B(S)$ is a contraction mapping or an m-stage contraction mapping, then there exists a unique fixed point of H; that is, there exists a unique function $J^* \in B(S)$ such that

$$HJ^* = J^*.$$

Proposition 5.1: Let $B = \max_i \max_{x^i} |R^i(x^i)|$. For fixed x, the optimal reward function $J(x, M)$ has the following properties as a function of M:

(a) $J(x, M)$ is convex and monotonically nondecreasing.

(b) $J(x, M)$ is constant for $M \leq -B/(1-\alpha)$.

(c) $J(x, M) = M$ for all $M \geq B/(1-\alpha)$.

Proof: Consider the value iteration method starting with the function

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

Successive iterates are generated by

$$J_{k+1}(x, M) = \max\left[M, \max_i L^i(x, M, J_k)\right], \qquad k = 0, 1, \ldots. \tag{5.6}$$

and we know from Prop. 2.1 of Section 1.2 that

$$\lim_{k \to \infty} J_k(x, M) = J(x, M), \qquad \text{for all } x, M. \tag{5.7}$$

We show inductively that $J_k(x, M)$ has the properties (a) to (c) stated in the proposition and, by taking the limit as $k \to \infty$, we establish the same properties for J. Clearly, $J_0(x, M)$ satisfies properties (a) to (c). Assume that $J_k(x, M)$ satisfies (a) to (c). Then from Eqs. (5.5) and (5.6) it follows that $J_{k+1}(x, M)$ is convex and monotonically nondecreasing in M, since the expectation and maximization operations preserve these properties. Verification of (b) and (c) is straightforward, and is left for the reader. **Q.E.D.**

Consider now a problem where there is only one project that can be worked on, say project i. The optimal reward function for this problem is denoted $J^i(x^i, M)$ and has the properties indicated in Prop. 5.1. A typical form for $J^i(x^i, M)$, viewed as a function of M for fixed x^i, is shown in Fig. 1.5.1. Clearly, there is a minimal value $m^i(x^i)$ of M for which $J^i(x^i, M) = M$; that is,

$$m^i(x^i) = \min\{M \mid J^i(x^i, M) = M\}, \qquad \text{for all } x^i. \tag{5.8}$$

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

Our objective is to show the optimality of the index rule (5.3) for the index function defined by Eq. (5.8).

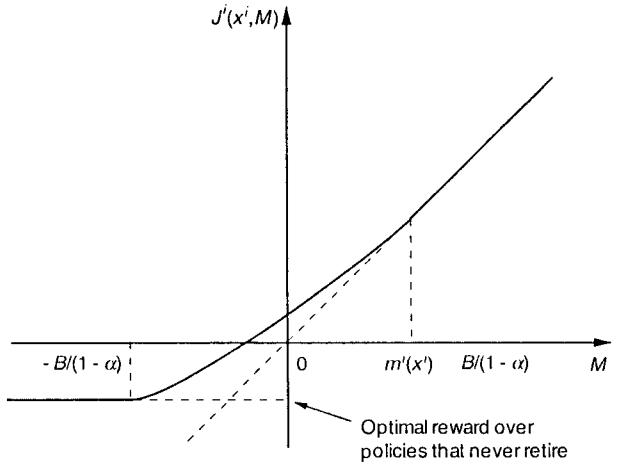


Figure 1.5.1 Form of the ith project reward function $J^i(x^i, M)$ for fixed x^i and definition of the index $m^i(x^i)$.

Project-by-Project Retirement Policies

Consider first a problem with a single project, say project i, and a fixed retirement reward M. Then by the definition (5.8) of the index, an optimal policy is to

$$\text{retire project } i \text{ if } \quad m^i(x^i) < M, \tag{5.9a}$$

$$\text{work on project } i \text{ if } \quad m^i(x^i) \geq M. \tag{5.9b}$$

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

$$S^i = \{x^i \mid m^i(x^i) < M\}. \tag{5.10}$$

At that time the project is permanently retired.

Consider now the multiproject problem for fixed retirement reward M. Suppose 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 i with state x^i such that $m^i(x^i) > M$? The answer is negative. Retiring when $m^i(x^i) > M$ cannot be optimal, since if we operate project i exclusively up to the time that its state x^i falls within the retirement set S^i of Eq. (5.10) and then retire, we will gain

Furthermore, if J is any function in $B(S)$ and H^k is the composition of H with itself k times, then

$$\lim_{k\to\infty} \left\| H^k J - J^* \right\| = 0.$$

Now consider the mappings T and T_μ defined by Eqs. (1.4) and (1.5). Proposition 2.4 and Cor. 2.4.1 show that T and T_μ are contraction mappings ($\rho = \alpha$). As a result, the convergence of the value iteration method to the unique fixed point of T follows directly from the contraction mapping theorem. Note also that, by Prop. 3.2, the mapping F corresponding to the Gauss-Seidel variant of the value iteration method is also a contraction mapping with $\rho = \alpha$, and the convergence result of Prop. 3.2 is again a special case of the contraction mapping theorem.

1.5 STOCHASTIC SCHEDULING AND THE MULTIARMED BANDIT

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

$$x_{k+1}^i = f^i(x_k^i, w_k^i), \qquad \text{if } i \text{ is worked on at time } k, \tag{5.1}$$

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

$$x_{k+1}^i = x_k^i, \qquad \text{if } i \text{ is idle at time } k. \tag{5.2}$$

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

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

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

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

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

$$\text{retire} \qquad \text{if} \qquad M > \max_j \left\{ m^j(x_k^j) \right\}, \tag{5.3a}$$

$$\text{work on project } i \qquad \text{if} \qquad m^i(x_k^i) = \max_j \left\{ m^j(x_k^j) \right\} \geq M. \tag{5.3b}$$

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

The problem of this section is known as a *multiarmed bandit problem*. An analogy here is drawn between project scheduling and selecting 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 the tradeoff here 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

$$J(x, M) = \max \left[M, \max_i L^i(x, M, J) \right], \qquad \text{for all } x, \tag{5.4}$$

where L^i is defined by

$$L^i(x, M, J) = R^i(x^i) + \alpha \mathop{E}_{w^i} \left\{ J\left(x^1, \ldots, x^{i-1}, f^i(x^i, w^i), x^{i+1}, \ldots, x^n, M\right) \right\}. \tag{5.5}$$

The next proposition gives some useful properties of J.

a higher expected reward. [This follows from the definition (5.8) of the index and the nature of the optimal policy (5.9) for the single-project problem.]

2. Does it ever make sense to work on a project i with state in the retirement set S^i of Eq. (5.10)? 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 $x = (x^1, \ldots, x^n)$,

$$\text{permanently retires project } i \quad \text{if} \quad x^i \in S^i, \tag{5.11a}$$

$$\text{works on some project} \quad \text{if} \quad x^j \notin S^j \text{ for some } j, \tag{5.11b}$$

where S^i is the ith project retirement set of Eq. (5.11). 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 5.2: There exists an optimal PPR policy.

Proof: In view of Eqs. (5.4), (5.5), and (5.11), existence of a PPR policy is equivalent to having, for all i,

$$M > L^i(x, M, J), \qquad \text{for all } x \text{ with } x^i \in S^i, \tag{5.12a}$$

$$M \leq L^i(x, M, J), \qquad \text{for all } x \text{ with } x^i \notin S^i, \tag{5.12b}$$

where L^i is given by

$$L^i(x, M, J) = R^i(x^i) + \alpha \mathop{E}_{w^i}\Big\{J\big(x^1, \ldots, x^{i-1}, f^i(x^i, w^i), x^{i+1}, \ldots, x^n, M\big)\Big\}, \tag{5.13}$$

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

The ith single-project optimal reward function J^i clearly satisfies, for all x^i,

$$J^i(x^i, M) \leq J(x^1, \ldots, x^{i-1}, x^i, x^{i+1}, \ldots, x^n, M), \tag{5.14}$$

since having the option of working at projects other than i cannot decrease the optimal reward. Furthermore, from the definition of the retirement set S^i [cf. Eq. (5.10)],

$$x^i \notin S^i, \qquad \text{if } M \leq R^i(x^i) + \alpha \mathop{E}_{w^i}\Big\{J^i\big(f^i(x^i, w^i), M\big)\Big\}. \tag{5.15}$$

Using Eqs. (5.13) to (5.15), we obtain Eq. (5.12b).

It will suffice to show Eq. (5.12a) for $i = 1$. 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 $x = (x^1, \underline{x})$:

$$\underline{J}(\underline{x}, M) \leq J(x^1, \underline{x}, M) \leq \underline{J}(\underline{x}, M) + \big(J^1(x^1, M) - M\big). \tag{5.16}$$

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. (5.12a) for $i = 1$, it will suffice to show Eq. (5.16). Indeed, when $x^1 \in S^1$, then $J^1(x^1, M) = M$, so from Eq. (5.16) we obtain $J(x^1, \underline{x}, M) = \underline{J}(\underline{x}, M)$, which is in turn equivalent to Eq. (5.12a) for $i = 1$.

We now turn to the proof of Eq. (5.16). Its left side is evident. To show the right side, we proceed by induction on the value iteration recursions

$$J_{k+1}(x^1, \underline{x}) = \max\Big[M, R^1(x^1) + \alpha E\big\{J_k\big(f^1(x^1, w^1), \underline{x}\big)\big\}, \max_{i \neq 1}\big\{R^i(x^i) + \alpha E\big\{J_k\big(x^1, F^i(\underline{x}, w^i)\big)\big\}\big\}\Big], \tag{5.17a}$$

$$\underline{J}_{k+1}(\underline{x}) = \max\Big[M, \max_{i \neq 1}\big[R^i(x^i) + \alpha E\big\{\underline{J}_k\big(F^i(\underline{x}, w^i)\big)\big\}\big]\Big], \tag{5.17b}$$

$$J^1_{k+1}(x^1) = \max\big[M, R^1(x^1) + \alpha E\big\{J^1_k\big(f^1(x^1, w^1)\big)\big\}\big], \tag{5.17c}$$

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

$$F^i(\underline{x}, w^i) = \big(x^2, \ldots, x^{i-1}, f^i(x^i, w^i), x^{i+1}, \ldots, x^n\big). \tag{5.18}$$

The initial conditions for the recursions (5.17) are

$$J_0(x^1, \underline{x}) = M, \qquad \text{for all } (x^1, \underline{x}), \tag{5.19a}$$

$$\underline{J}_0(\underline{x}) = M, \qquad \text{for all } \underline{x}, \tag{5.19b}$$

$$J^1_0(x^1) = M, \qquad \text{for all } x^1. \tag{5.19c}$$

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. (5.16) it will suffice to show that for all k and $x = (x^1, \underline{x})$ we have

$$J_k(x^1, \underline{x}) \leq \underline{J}_k(\underline{x}) + \big(J^1_k(x^1) - M\big). \tag{5.20}$$

In view of the definitions (5.19), we see that Eq. (5.20) holds for $k = 0$. Assume that it holds for some k. We will show that it holds for $k+1$. From Eq. (5.17) and the induction hypothesis (5.20), we have

$$J_{k+1}(x^1, \underline{x}) \leq \max\Big[M,\ R^1(x^1) + \alpha E\big\{\underline{J}_k(\underline{x}) + J_k^1\big(f^1(x^1, w^1)\big)\big\} - M\big\},$$
$$\max_{i \neq 1}\big[R^i(x^i) + \alpha E\big\{\underline{J}_k\big(F^i(\underline{x}, w^i)\big) + J_k^1(x^1) - M\big\}\big]\Big].$$

Using the facts $\underline{J}_k(\underline{x}) \geq M$ and $J_k^1(x^1) \geq M$ [cf. Eq. (5.17)], and the preceding equation, we see that

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

where

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

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

Using Eqs. (5.17b), (5.17c), and the preceding equations, we see that

$$J_{k+1}(x^1, \underline{x}) \leq \max\big[J_{k+1}^1(x^1) + \underline{J}_k(\underline{x}) - M,\ \underline{J}_{k+1}(\underline{x}) + J_k^1(x^1) - M\big]. \quad (5.21)$$

It can be seen from Eqs. (5.17) and (5.19) 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. (5.21) we obtain that Eq. (5.20) holds for $k+1$. The induction is complete. **Q.E.D.**

As a first step towards showing optimality of the index rule, we use the preceding proposition to derive an expression for the partial derivative of $J(x, M)$ with respect of M.

Lemma 5.1: For fixed x, let K_M denote the retirement time under an optimal policy when the retirement reward is M. Then for all M for which $\partial J(x, M)/\partial M$ exists we have

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

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

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

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

Note that the convexity of $J(x, \cdot)$ with respect to M (Prop. 4.1) implies that the derivative $\partial J(x, M)/\partial M$ exists almost everywhere with respect to Lebesgue measure [Roc70]. Furthermore, it can be shown that $\partial J(x, M)/\partial M$ exists for all M for which the optimal policy is unique.

For a given M, initial state x, and optimal PPR policy, let T_i be the retirement time of project i if it were the only project available, let T be the retirement time for the multiproject problem. Both T_i and T take values that are either nonnegative or ∞. The existence of an optimal PPR policy implies that we must have

$$T = T_1 + \cdots + T_n$$

and in addition T_i, $i = 1, \ldots, n$, are independent random variables. Therefore,

$$E\{\alpha^T\} = E\{\alpha^{T_1 + \cdots + T_n}\} = \prod_{i=1}^{n} E\{\alpha^{T_i}\}.$$

Using Lemma 5.1, we obtain

$$\frac{\partial J(x, M)}{\partial M} = \prod_{i=1}^{n} \frac{\partial J^i(x^i, M)}{\partial M}. \quad (5.22)$$

Optimality of the Index Rule

We are now ready to show our main result.

Proposition 5.3: The index rule (5.3) is an optimal stationary policy.

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

$$m(x) = \max_j \{m^j(x^j)\},$$

and let i be such that

$$m^i(x^i) = \max_j \{m^j(x^j)\}.$$

If $m(x) < M$ the optimality of the index rule (5.3a) at state x follows from the existence of an optimal PPR policy. If $m(x) \geq M$, we note that

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

and then use this relation together with Eq. (5.22) to write

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

and finally

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

where

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

(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

$$J(x, m(x)) = L^i(x, m(x), J).$$

Therefore, the convex functions $J(x, M)$ and $L^i(x, M, J)$ viewed as functions of M for fixed x are equal for $M = m(x)$ and have equal derivative for almost all $M \leq m(x)$. It follows that for all $M \leq m(x)$ we have

$$J(x, M) = L^i(x, M, J).$$

This implies that the index rule (5.3b) 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 (5.3), since optimization of a multiproject problem has been reduced to n separate single-project optimization problems. Nonetheless, solution of each of these single-project problems can be complicated. Under certain circumstances, however, the situation simplifies.

Suppose that for all i, x^i, and w^i that can occur with positive probability, we have either

$$m^i(x^i) \leq m^i(f^i(x^i, w^i)) \tag{5.23}$$

or

$$m^i(x^i) \geq m^i(f^i(x^i, w^i)). \tag{5.24}$$

Under Eq. (5.23) [or Eq. (5.24)] 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. (5.24) implies that if retirement is optimal when at state x^i then it is also optimal at each state $f^i(x^i, w^i)$. Therefore, for all x^i such that $M = m^i(x^i)$ we have, for all w^i,

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

From Bellman's equation

$$J^i(x^i, M) = \max\left[M, R^i(x^i) + \alpha E\{J^i(f^i(x^i, w^i), M)\}\right]$$

we obtain

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

or

$$m^i(x^i) = \frac{R^i(x^i)}{1 - \alpha}. \tag{5.25}$$

Thus the optimal policy in the deteriorating case is

retire if $M > \max_i \frac{R^i(x^i)}{1-\alpha}$ and otherwise engage the project i with maximal one-step reward $R^i(x^i)$.

Example 5.1 (Treasure Hunting)

Consider a search problem involving N sites. Each site i may contain a treasure with expected value v_i. A search at site i costs c_i and reveals the treasure with probability β_i (assuming a treasure is there). Let P_i be the probability that there is a treasure at site i. We take P_i as the state of the project corresponding to searching site i. Then the corresponding one-step reward is

$$R'(P_i) = \beta_i P_i v_i - c_i. \tag{5.26}$$

If a search at site i does not reveal the treasure, the probability P_i drops to

$$\overline{P}_i = \frac{P_i(1-\beta_i)}{P_i(1-\beta_i)+1-p_i},$$

as can be verified using Bayes' rule. If the search finds the treasure, the probability P_i drops to zero, since the treasure is removed from the site. Based on this and the fact that $R'(P_i)$ is increasing with P_i [cf. Eq. (5.26)], it is seen that the deteriorating condition (5.24) holds. Therefore, it is optimal to search the site i for which the expression $R'(P_i)$ of Eq. (5.26) is maximal, provided $\max_i R'(P_i) > 0$, and to retire if $R'(P_i) \leq 0$ for all i.

1.6 NOTES, SOURCES, AND EXERCISES

Many authors have contributed to the analysis of the discounted problem with bounded cost per stage, most notably Shapley [Sha53], Bellman [Bel57], and Blackwell [Bla65]. For variations and extensions of the problem involving multiple criteria, weighted criteria, and constraints, see [FeS94], [Gho90], [Ros89], and [WhK80]. The mathematical issues relating to measurability concerns are analyzed extensively in [BeS78], [DyY79], and [Her89].

The error bounds given in Section 1.3 and Exercise 1.9 are improvements on results of [McQ66] (see [Por71], [Por75], [Ber76], and [PoT78]). The corresponding convergence rate was discussed in [Mor71] and [MoW77]. The Gauss-Seidel method for discounted problems was proposed in [Kus71] (see also [Has68]). An extensive discussion of the convergence aspects of the method and related background is given in Section 2.6 of [BeT89a]. The material on the generic rank-one correction, including the convergence analysis of Exercise 1.8, is new; see [Ber93], which also describes a multiple-rank correction method where the effect of several eigenvalues is nullified. Value iteration is particularly well-suited for parallel computation; see e.g., [AMT93], [BeT89a].

Policy iteration for discounted problems was proposed in [Bel57]. The modified policy iteration algorithm was suggested and analyzed in [PuS78] and [PuS82]. The approximate policy iteration analysis and the convergence proof of policy iteration for an infinite state space (Prop. 3.6) are new and were developed in collaboration with J. Tsitsiklis. The relation between policy iteration and Newton's method (Exercise 1.10) was pointed out in [PoA69] and was further discussed in [PuB78].

The material on adaptive aggregation is due to [BeC89]. In an alternative aggregation approach [SPK89], the aggregate states are fixed. Changing adaptively the aggregate states from one iteration to the next depending on the progress of the computation has a potentially significant effect on the efficiency of the computation for difficult problems where the ordinary value iteration method is very slow.

The linear programming approach of Section 1.3.4 was proposed in [D'Ep60]. There is a relation between policy iteration and the simplex method applied to solving the linear program associated with the discounted problem. In particular, it can be shown that the simplex method for linear programming with a block pivoting rule is mathematically equivalent to the policy iteration algorithm. There are also duality connections that relate the linear programming approach with randomized policies, constraints, and multiple criteria; see e.g., [Kal83], [Put94]. Approximation methods using basis functions and linear programming were proposed in [ScS85].

A complexity analysis of finite-state infinite horizon problems is given in [PaT87]. Discretization methods that approximate infinite state space systems with finite-state Markov chains, are discussed in [Ber75], [Fox71], [HaL86], [Whi78], [Whi79], and [Whi80a]. For related multigrid approximation methods and associated complexity analysis, see [ChT89] and [ChT91]. A different approach to deal with infinite state spaces, which is based on randomization, has been introduced in [Rus94]; see also [Rus95]. Further material on computational methods may be found in [Put78].

The role of contraction mappings in discounted problems was first recognized and exploited in [Sha53], which considers two-player dynamic games. Abstract DP models and the implications of monotonicity and contraction have been explored in detail in [Den67], [Ber77], [BeS78], [VeP84], and [VeP87].

The index rule solution of the multiarmed bandit problem is due to [Git79] and [GiJ74]. Subsequent contributions include [Whi80b], [Kel81], [Whi81], and [Whi82]. The proof given here is due to [Tsi86]. Alternative proofs and analysis are given in [VWB85], [NTW89], [Tso91], [Web92], [BeN93], [Tsi93b], [BPT94a], and [BPT94b]. Much additional work on the subject is described in [Kum85] and [KuV86].

Finally, we note that even though our analysis in this chapter requires a countable disturbance space, it may still serve as the starting point of analysis of problems with uncountable disturbance space. This can be done by reducing such problems to deterministic problems with state space a set of probability measures. The basic idea of this reduction is demonstrated

in Exercise 1.13. The advanced reader may consult [BeS78] (Section 9.2), and see how such a reduction can be effected for a very broad class of finite and infinite horizon problems.

EXERCISES

1.1

Write a computer problem and compute iteratively the vector J_μ satisfying

$$J_\mu = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + \alpha \begin{pmatrix} 3/4 & 1/4 & 0 \\ 1/4 & 3/4 - \epsilon & \epsilon \\ 0 & \epsilon & 1 - \epsilon \end{pmatrix} J_\mu.$$

Do your computations for all combinations of $\alpha = 0.9$ and $\alpha = 0.999$, and $\epsilon = 0.5$ and $\epsilon = 0.001$. Try value iteration with and without error bounds, and also adaptive aggregation with two aggregate classes of states. Discuss your results.

1.2

The purpose of this problem is to show that shortest path problems with a discount factor make little sense. Suppose that we have a graph with a nonnegative length a_{ij} for each arc (i, j). The cost of a path $(i_0, i_1, \ldots, i_m)$ 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.

1.3

Consider a problem similar to that of Section 1.1 except that when we are at state x_k, there is a probability β, where $0 < \beta < 1$, that the next state x_{k+1} will be determined according to $x_{k+1} = f(x_k, u_k, w_k)$ and a probability $(1-\beta)$ that the system will move to a termination state, where it stays permanently thereafter at no cost. Show that even if $\alpha = 1$, the problem can be put into the discounted cost framework.

1.4

Consider a problem similar to that of Section 1.2 except that the discount factor α depends on the current state x_k, the control u_k, and the disturbance w_k; that is, 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

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

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

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

Argue that the results and algorithms of Sections 1.2 and 1.3 have direct counterparts for such problems.

1.5 (Column Reduction [Por75])

The purpose of this problem is to provide a transformation of a certain type of discounted problem into another discounted problem with smaller discount factor. Consider the n-state discounted problem under the assumptions of Section 1.3. The cost per stage is $g(i, u)$, the discount factor is α, and the transition probabilities are $p_{ij}(u)$. For each $j = 1, \ldots, n$, 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 $\alpha\big(1 - \sum_{j=1}^n m_j\big)$, and transition probabilities $\tilde{p}_{ij}(u)$. Show that this problem has the same optimal policies as the original, and that its optimal cost vector J' satisfies

$$J^* = J' + \frac{\alpha \sum_{j=1}^n m_j J'(j)}{1 - \alpha} e,$$

where J^* is the optimal cost vector of the original problem and e is the unit vector.

1.6

Let $\overline{J} : S \mapsto \Re$ be any bounded function on S and consider the value iteration method of Section 1.3 with a starting function $J : S \mapsto \Re$ of the form

$$J(x) = \overline{J}(x) + r, \qquad x \in S,$$

where r is some scalar. Show that the bounds $(T^k J)(x)+\underline{c}_k$ and $(T^k J)(x)+\overline{c}_k$ of Prop. 3.1 are independent of the scalar r for all $x \in S$. Show also that if S consists of a single state $\tilde{x}$ (i.e., $S = \{\tilde{x}\}$), then

$$(TJ)(\tilde{x}) + \underline{c}_1 = (TJ)(\tilde{x}) + \overline{c}_1 = J^*(\tilde{x}).$$

1.7 (Jacobi Version of Value Iteration)

Consider the problem of Section 1.3 and the version of the value iteration method that starts with an arbitrary function $J : S \mapsto \Re$ and generates recursively $FJ, F^2J, \ldots$, where F is the mapping given by

$$(FJ)(i) = \min_{u \in U(i)} \frac{g(i,u) + \alpha \sum_{j \neq i} p_{ij}(u) J(j)}{1 - \alpha p_{ii}(u)}.$$

Show that $(F^k J)(i) \to J^*(i)$ as $k \to \infty$ and provide a rate of convergence estimate that is at least as favorable as the one for the ordinary method (cf. Prop. 2.3).

1.8 (Convergence Properties of Rank-One Correction [Ber93])

Consider the solution of the system $J = FJ$, where $F : \Re^n \mapsto \Re^n$ is the mapping

$$FJ = h + QJ,$$

h is a given vector in $\Re^n$, and Q is an $n \times n$ matrix. Consider the generic rank-one correction iteration $J := MJ$, where $M : \Re^n \mapsto \Re^n$ is the mapping

$$MJ = FJ + \gamma z,$$

and

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

(a) Show that any solution J^* of the system $J = FJ$ satisfies $J^* = MJ^*$.

(b) Verify that the value iteration method that uses the error bounds in the manner of Eq. (3.12) is a special case of the iteration $J := MJ$ with d equal to the unit vector.

(c) Assume that d is an eigenvector of Q, let λ be the corresponding eigenvalue, and let $\lambda_1, \ldots, \lambda_{n-1}$ be the remaining eigenvalues. Show that MJ can be written as

$$MJ = h + RJ,$$

where h is some vector in $\Re^n$ and

$$R = Q - \frac{\lambda}{(1-\lambda)\|d\|^2} dd'(I - Q).$$

Show also that $Rd = 0$ and that for all k and J,

$$R^k = RQ^{k-1}, \qquad M^k J = M(F^{k-1}J).$$

Furthermore, the eigenvalues of R are $0, \lambda_1, \ldots, \lambda_{n-1}$. (This last statement requires a somewhat complicated proof; see [Ber93].)

(d) Let d be as in part (c), and suppose that $e_1, \ldots, e_{n-1}$ are eigenvectors corresponding to $\lambda_1, \ldots, \lambda_{n-1}$. Suppose that a vector J can be written as

$$J = J^* + \xi e + \sum_{i=1}^{n-1} \xi_i e_i,$$

where J^* is a solution of the system. Show that, for all $k > 1$,

$$M^k J = J^* + \sum_{i=1}^{n-1} \xi_i \lambda_i^{k-1} R e_i,$$

so that if λ is a dominant eigenvalue and $\lambda_1, \ldots, \lambda_{n-1}$ lie within the unit circle, $M^k J$ converges to J^* at a rate governed by the subdominant eigenvalue. *Note*: This result can be generalized for the case where Q does not have a full set of linearly independent eigenvectors, and for the case where F is modified through multiple-rank corrections [Ber93].

1.9 (Generalized Error Bounds [Ber76])

Let S be a set and $B(S)$ be the set of all bounded real-valued functions on S. Let $T : B(S) \mapsto B(S)$ be a mapping with the following two properties:

(1) $TJ \leq TJ'$ for all $J, J' \in B(S)$ with $J \leq J'$.

(2) For every scalar $r \neq 0$ and all $x \in S$,

$$\alpha_1 \leq \frac{\big(T(J + re)\big)(x) - (TJ)(x)}{r} \leq \alpha_2,$$

where α_1, α_2 are two scalars with $0 \leq \alpha_1 \leq \alpha_2 < 1$.

(a) Show that T is a contraction mapping on $B(S)$, and hence for every $J \in B(S)$ we have

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

where J^* is the unique fixed point of T in $B(S)$.

(b) Show that for all $J \in B(S)$, $x \in S$, and $k = 1, 2, \ldots$,

$$(T^k J)(x) + \underline{c}_k \le (T^{k+1} J)(x) + \underline{c}_{k+1} \le J^*(x) \le (T^{k+1} J)(x) + \bar{c}_{k+1} \le (T^k J)(x) + \bar{c}_k,$$

where for all k

$$\underline{c}_k = \min\left\{ \frac{\alpha_1}{1-\alpha_1} \min_{x\in S}\left[(T^k J)(x) - (T^{k-1} J)(x)\right], \frac{\alpha_2}{1-\alpha_2} \min_{x\in S}\left[(T^k J)(x) - (T^{k-1} J)(x)\right] \right\}, \tag{6.1}$$

$$\bar{c}_k = \max\left\{ \frac{\alpha_1}{1-\alpha_1} \max_{x\in S}\left[(T^k J)(x) - (T^{k-1} J)(x)\right], \frac{\alpha_2}{1-\alpha_2} \max_{x\in S}\left[(T^k J)(x) - (T^{k-1} J)(x)\right] \right\}. \tag{6.2}$$

A geometric interpretation of these relations for the case where S consists of a single element is provided in Fig. 1.6.1.

(c) Consider the following algorithm:

$$J_k(x) = (TJ_{k-1})(x) + \gamma_k, \qquad x \in S,$$

where J_0 is any function in $B(S)$, γ_k is any scalar in the range $[\underline{c}_k, \bar{c}_k]$, and $\underline{c}_k$ and $\bar{c}_k$ are given by Eqs. (6.1) and (6.2) with $(T^k J)(x) - (T^{k-1} J)(x)$ replaced by $(TJ_{k-1})(x) - J_{k-1}(x)$. Show that for all k,

$$\max_{i\in S}\left|J_k(x) - J^*(x)\right| \le \alpha_2^k \max_{x\in S}\left|J_0(x) - J^*(x)\right|.$$

(d) Let $J \in \Re^n$ and consider the equation $J = TJ$, where

$$TJ = h + MJ$$

and the vector $h \in \Re^n$ and the matrix M are given. Let s_i be the ith row sum of M, that is,

$$s_i = \sum_{j=1}^{n} m_{ij},$$

and let $\alpha_1 = \min_i s_i$, $\alpha_2 = \max_i s_i$. Show that if the elements m_{ij} of M are all nonnegative and $\alpha_2 < 1$, then the conclusions of parts (a) and (b) hold.

(e) [Por75] Consider the Gauss-Seidel method for solving the system $J = g + \alpha PJ$, where $0 < \alpha < 1$ and P is a transition probability matrix. Use part (d) to obtain suitable error bounds.

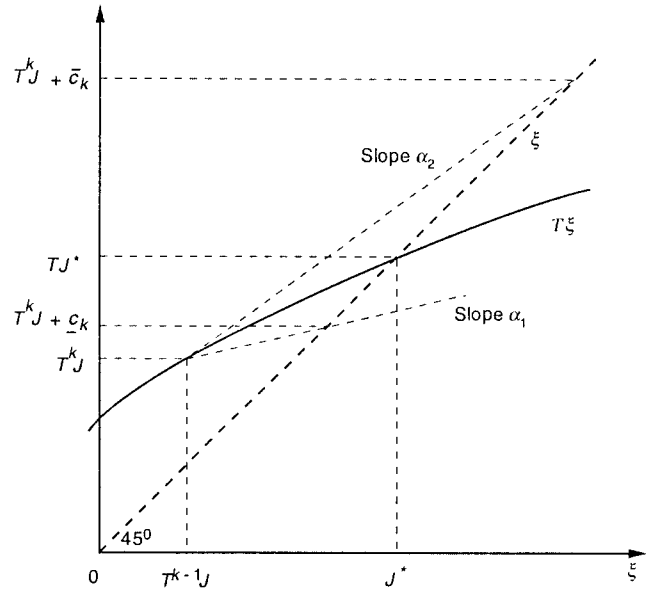


Figure 1.6.1 Graphical interpretation of the error bounds of Exercise 1.9.

1.10 (Policy Iteration and Newton's Method)

The purpose of this problem is to demonstrate a relation between policy iteration and Newton's method for solving nonlinear equations. Consider an equation of the form $F(J) = 0$, where $F : \Re^n \mapsto \Re^n$. Given a vector $J_k \in \Re^n$, Newton's method determines J_{k+1} by solving the linear system of equations

$$F(J_k) + \frac{\partial F(J_k)}{\partial J}(J_{k+1} - J_k) = 0,$$

where $\partial F(J_k)/\partial J$ is the Jacobian matrix of F evaluated at J_k.

(a) Consider the discounted finite-state problem of Section 1.3 and define

$$F(J) = TJ - J.$$

Show that if there is a unique μ such that

$$T_\mu J = TJ,$$

then the Jacobian matrix of F at J is

$$\frac{\partial F(J)}{\partial J} = \alpha P_\mu - I,$$

where I is the $n \times n$ identity.

(b) Show that the policy iteration algorithm can be identified with Newton's method for solving $F(J) = 0$ (assuming it gives a unique policy at each step).

1.11 (Minimax Problems)

Provide analogs of the results and algorithms of Sections 1.2 and 1.3 for the minimax problem where the cost is

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

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 D for each $(x, u) \in S \times C$. (Compare with Exercise 1.5 in Chapter 1 of Vol. I.)

1.12 (Data Transformations [Sch72])

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 $J(i)$:

$$J(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^{n} m_{ij}(u) J(j) \right], \qquad i = 1, \ldots, n, \tag{6.3}$$

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,

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

and that a solution $\{J(i) \mid i = 1, \ldots, n\}$ of Eq. (6.3) is also a solution of the equations

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

1.13 (Stochastic to Deterministic Problem Transformation)

Under the assumptions and notation of Section 1.3, consider the controlled system

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

where p_k is a probability distribution over S viewed as a row vector, and P_{μ_k} is the transition probability matrix corresponding to the control function μ_k. The state is p_k and the control is μ_k. Consider also the cost function

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

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 for the discounted cost problem of Section 1.3.

1.14 (Threshold Policies and Policy Iteration)

(a) Consider the machine replacement example of Section 1.2, and assume that the condition (2.10) holds. Let us define a *threshold* policy to be a stationary policy that replaces if and only if the state is greater than or equal to some fixed state i. Suppose that we start the policy iteration algorithm using a threshold policy. Show that all the subsequently generated policies will be threshold policies, so that the algorithm will terminate after at most n iterations.

(b) Prove the result of part (a) for the asset selling example of Section 1.2, assuming that there is a finite number of values that the offer x_k can take. Here, a threshold policy is a stationary policy that sells the asset if the offer is higher than a certain fixed number.

1.15 (Distributed Asynchronous DP [Ber82a], [BeT89a])

The value iteration method is well suited for distributed (or parallel) computation since the iteration

$$J(i) := (TJ)(i)$$

can be executed in parallel for all states i. Consider the finite-state discounted problem of Section 1.3, and assume that the above iteration is executed *asynchronously* at a different processor i for each state i. By this we mean that the ith processor holds a vector J^i and updates the ith component of that vector at *arbitrary* times with an iteration of the form

$$J^i(i) := (TJ^i)(i),$$

and at *arbitrary* times transmits the results of the latest computation to other processors m who then update $J^m(i)$ according to

$$J^m(i) := J^i(i).$$

Assume that all processors never stop computing and transmitting the results of their computation to the other processors. Show that the estimates J_t^i of the optimal cost function available at each processor i at time t converge to the optimal solution function J^* as $t \to \infty$. *Hint*: Let $\overline{J}$ and $\underline{J}$ be two functions such that $\underline{J} \leq T\underline{J}$ and $T\overline{J} \leq \overline{J}$, and suppose that for all initial estimates J_0^i of the processors, we have $\underline{J} \leq J_0^i \leq \overline{J}$. Show that the estimates J_t^i of the processors at time t satisfy $\underline{J} \leq J_t^i \leq \overline{J}$ for all $t \geq 0$, and $T\underline{J} \leq J_t^i \leq T\overline{J}$ for t sufficiently large.

1.16

Assume that we have two gold mines, Anaconda and Bonanza, and a gold-mining machine. Let x_A and x_B be the current amounts of gold in Anaconda and Bonanza, respectively. When the machine is used in Anaconda (or Bonanza), there is a probability p_A (or p_B, respectively) that $r_A x_A$ (or $r_B x_B$, respectively) of the gold will be mined without damaging the machine, and a probability $1 - p_A$ (or $1 - p_B$, respectively) that the machine will be damaged beyond repair and no gold will be mined. We assume that $0 < r_A < 1$ and $0 < r_B < 1$.

(a) Assume that $p_A = p_B = p$, where $0 < p < 1$. Find the mine selection policy that maximizes the expected amount of gold mined before the machine breaks down. *Hint*: This problem can be viewed as a discounted multiarmed bandit problem with a discount factor p.

(b) Assume that $p_A < 1$ and $p_B = 1$. Argue that the optimal expected amount of gold mined has the form $J^*(x_A, x_B) = \tilde{J}_A(x_A) + x_B$, where $\tilde{J}_A(x_A)$ is the optimal expected amount of gold mined if mining is restricted just to Anaconda. Show that there is no policy that attains the optimal amount $J^*(x_A, x_B)$.

1.17 (The Tax Problem [VWB85])

This problem is similar to the multiarmed bandit problem. The only difference is that, if we engage project i at period k, we pay a tax $\alpha^k C^j(x^j)$ for every other project j [for a total of $\alpha^k \sum_{j \neq i} C^j(x^j)$], instead of earning a reward $\alpha^k R^i(x^i)$. The objective is to find a project selection policy that minimizes the total tax paid. Show that the problem can be converted into a bandit problem with reward function for project i equal to

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

1.18 (The Restart Problem [KaV87])

The purpose of this problem 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

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

2

Stochastic Shortest Path Problems

Contents

In this chapter we consider a stochastic version of the shortest path problem discussed in Chapter 2 of Vol. I. An introductory analysis of this problem was given in Section 7.2 of Vol. I. The analysis of this chapter is more sophisticated and uses weaker assumptions. In particular, we make assumptions that generalize those made for deterministic shortest path problems in Chapter 2 of Vol. I.

In this chapter we also discuss another major topic of this book. In particular, in Section 2.3 we develop simulation-based methods, possibly involving approximations, which are suitable for complex problems that involve a large number of states and/or a lack of an explicit mathematical model. These methods are most economically developed in the context of stochastic shortest path problems. They can then be extended to discounted problems, and this is done in Section 2.3.4. Further extensions to average cost per stage problems are discussed in Section 4.3.4.

2.1 MAIN RESULTS

Suppose that we have a graph with nodes $1, 2, \ldots, n, t$, where t is a special state called the *destination* or the *termination state*. We can view the deterministic shortest path problem of Chapter 2 of Vol. I as follows: we want to choose for each node $i \neq t$, a successor node $\mu(i)$ so that $\big(i, \mu(i)\big)$ is an arc, and the path formed by a sequence of successor nodes starting at any node j terminates at t and has the minimum sum of arc lengths over all paths that start at j and terminate at t.

The stochastic shortest path problem is a generalization whereby at each node i, we must select a probability distribution over all possible successor nodes j out of a given set of probability distributions $p_{ij}(u)$ parameterized by a control $u \in U(i)$. For a given selection of distributions and for a given origin node, the path traversed as well as its length are now random, but we wish that the path leads to the destination t with probability one and has minimum expected length. Note that if every feasible probability distribution assigns a probability of 1 to a single successor node, we obtain the deterministic shortest path problem.

We formulate this problem as the special case of the total cost infinite horizon problem where:

(a) There is no discounting ($\alpha = 1$).

(b) The state space is $S = \{1, 2, \ldots, n, t\}$ with transition probabilities denoted by

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

Furthermore, the destination t is absorbing, that is, for all $u \in U(t)$,

$$p_{tt}(u) = 1.$$

(c) The control constraint set $U(i)$ is a finite set for all i.

(d) A cost $g(i, u)$ is incurred when control $u \in U(i)$ is selected. Furthermore, the destination is *cost-free*; that is, $g(t, u) = 0$ for all $u \in U(t)$.

Note that as in Section 1.3, we assume that the cost per stage does not depend on w. This amounts to using expected cost per stage in all calculations. In particular, if the cost of using u at state i and moving to state j is $\tilde{g}(i, u, j)$, we use as cost per stage the expected cost

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

We are interested in problems where either reaching the destination is inevitable or else there is an incentive to reach the destination in a finite expected number of stages, so that the essence of the problem is to reach the destination with minimum expected cost. We will be more specific about this shortly.

Note that since the destination is a cost-free and absorbing state, the cost starting from t is zero for every policy. Accordingly, for all cost functions, we ignore the component that corresponds to t and define the mappings T and T_μ on functions J with components $J(1), \ldots, J(n)$. We will also view the functions J as n-dimensional vectors. Thus

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

$$(T_\mu J)(i) = g\big(i, \mu(i)\big) + \sum_{j=1}^{n} p_{ij}\big(\mu(i)\big)J(j), \qquad i = 1, \ldots, n.$$

As in Section 1.3, for any stationary policy μ, we use the compact notation

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

and

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

We can then write in vector notation

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

In terms of this notation, the cost function of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ can be written as

$$J_\pi = \limsup_{N\to\infty} T_{\mu_0}\cdots T_{\mu_{N-1}}J_0 = \limsup_{N\to\infty}\left(g_{\mu_0} + \sum_{k=1}^{N-1} P_{\mu_0}\cdots P_{\mu_{k-1}}g_{\mu_k}\right),$$

where J_0 denotes the zero vector. The cost function of a stationary policy μ can be written as

$$J_\mu = \limsup_{N\to\infty} T_\mu^{N-1}J_0 = \limsup_{N\to\infty}\sum_{k=0}^{N-1} P_\mu^k g_\mu.$$

The stochastic shortest path problem was discussed in Section 7.2 of Vol. I, under the assumption that all policies lead to the destination with probability 1, regardless of the initial state. In order to analyze the problem under weaker conditions, we introduce the notion of a proper policy.

Definition 1.1: A stationary policy μ is said to be *proper* if, when using this policy, there is positive probability that the destination will be reached after at most n stages, regardless of the initial state; that is, if

$$\rho_\mu = \max_{i=1,\ldots,n} P\{x_n \neq t \mid x_0 = i, \mu\} < 1. \tag{1.1}$$

A stationary policy that is not proper is said to be *improper*.

With a little thought, it can be seen that μ is proper if and only if in the Markov chain corresponding to μ, each state i is connected to the destination with a path of positive probability transitions. Note from the definition (1.1) that

$$\begin{aligned} P\{x_{2n} \neq t \mid x_0 = i, \mu\} &= P\{x_{2n} \neq t \mid x_n \neq t, x_0 = i, \mu\} \\ &\quad \times P\{x_n \neq t \mid x_0 = i, \mu\} \\ &\leq \rho_\mu^2. \end{aligned}$$

More generally, for a proper policy μ, the probability of not reaching the destination after k stages diminishes as $\rho_\mu^{\lfloor k/n \rfloor}$ regardless of the initial state; that is,

$$P\{x_k \neq t \mid x_0 = i, \mu\} \leq \rho_\mu^{\lfloor k/n \rfloor}, \qquad i = 1, \ldots, n. \tag{1.2}$$

Thus the destination will eventually be reached with probability one under a proper policy. Furthermore, the limit defining the associated total cost vector J_μ will exist and be finite, since the expected cost incurred in the kth period is bounded in absolute value by

$$\rho_\mu^{\lfloor k/n \rfloor} \max_{i=1,\ldots,n} \left|g\big(i, \mu(i)\big)\right|. \tag{1.3}$$

Note that under a proper policy, the cost structure is similar to the one for discounted problems, the main difference being that the effective discount factor depends on the current state and stage, but builds up to at least ρ_μ per n stages.

Throughout this section, we assume the following:

Assumption 1.1: There exists at least one proper policy.

Assumption 1.2: For every improper policy μ, the corresponding cost $J_\mu(i)$ is ∞ for at least one state i; that is, some component of the sum $\sum_{k=0}^{N-1} P_\mu^k g_\mu$ diverges to ∞ as $N \to \infty$.

In the case of a deterministic shortest path problem, Assumption 1.1 is satisfied if and only if every node is connected to the destination with a path, while Assumption 1.2 is satisfied if and only if each cycle that does not contain the destination has positive length. A simple condition that implies Assumption 1.2 is that the cost $g(i, u)$ is strictly positive for all $i \neq t$ and $u \in U(i)$. Another important case where Assumptions 1.1 and 1.2 are satisfied is when *all* policies are proper, that is, when termination is inevitable under all stationary policies (this was assumed in Section 7.2 of Vol. I). Actually, for this case, it is possible to show that mappings T and T_μ are contraction mappings with respect to some norm [not necessarily the maximum norm of Eq. (4.1) in Chapter 1]; see Section 4.3 of [BeT89a], or [Tse90]. As a result of this contraction property, the results shown for discounted problems can also be shown for stochastic shortest path problems where termination is inevitable under all stationary policies. It turns out, however, that similar results can be shown even when some improper policies exist: the results that we prove under Assumptions 1.1 and 1.2 are almost as strong as those for discounted problems with bounded cost per stage. In particular, we show that:

(a) The optimal cost vector is the unique solution of Bellman's equation $J^* = TJ^*$.

(b) The value iteration method converges to the optimal cost vector J^* for an arbitrary starting vector.

(c) A stationary policy μ is optimal if and only if $T_\mu J^* = TJ^*$.

(d) The policy iteration algorithm yields an optimal proper policy starting from an arbitrary proper policy.

The following proposition provides some basic preliminary results:

Proposition 1.1:

(a) For a proper policy μ, the associated cost vector J_μ satisfies

$$\lim_{k\to\infty} (T_\mu^k J)(i) = J_\mu(i), \qquad i = 1, \ldots, n, \tag{1.4}$$

for every vector J. Furthermore,

$$J_\mu = T_\mu J_\mu,$$

and J_μ is the unique solution of this equation.

(b) A stationary policy μ satisfying for some vector J,

$$J(i) \geq (T_\mu J)(i), \qquad i = 1, \ldots, n,$$

is proper.

Proof: (a) Using an induction argument, we have for all $J \in \Re^n$ and $k \geq 1$

$$T_\mu^k J = P_\mu^k J + \sum_{m=0}^{k-1} P_\mu^m g_\mu. \tag{1.5}$$

Equation (1.2) implies that for all $J \in \Re^n$, we have

$$\lim_{k\to\infty} P_\mu^k J = 0,$$

so that

$$\lim_{k\to\infty} T_\mu^k J = \lim_{k\to\infty} \sum_{m=0}^{k-1} P_\mu^m g_\mu = J_\mu,$$

where the limit above can be shown to exist using Eq. (1.2).

Also we have by definition

$$T_\mu^{k+1} J = g_\mu + P_\mu T_\mu^k J,$$

and by taking the limit as $k \to \infty$, we obtain

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

which is equivalent to $J_\mu = T_\mu J_\mu$.

Finally, to show uniqueness, note that if $J = T_\mu J$, then we have $J = T_\mu^k J$ for all k, so that $J = \lim_{k\to\infty} T_\mu^k J = J_\mu$.

(b) The hypothesis $J \geq T_\mu J$, the monotonicity of T_μ, and Eq. (1.5) imply that

$$J \geq T_\mu^k J = P_\mu^k J + \sum_{m=0}^{k-1} P_\mu^m g_\mu, \qquad k = 1, 2, \ldots$$

If μ were not proper, by Assumption 1.2, some component of the sum in the right-hand side of the above relation would diverge to ∞ as $k \to \infty$, which is a contradiction. **Q.E.D.**

The following proposition is the main result of this section, and provides analogs to the main results for discounted cost problems (Props. 2.1-2.3 in Section 1.2).

Proposition 1.2:

(a) The optimal cost vector J^* satisfies Bellman's equation

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

Furthermore, J^* is the unique solution of this equation.

(b) We have

$$\lim_{k\to\infty} (T^k J)(i) = J^*(i), \qquad i = 1, \ldots, n,$$

for every vector J.

(c) A stationary policy μ is optimal if and only if

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

Proof: (a), (b) We first show that T has at most one fixed point. Indeed, if J and J' are two fixed points, then we select μ and μ' such that $J = TJ = T_\mu J$ and $J' = TJ' = T_{\mu'} J'$; this is possible because the control constraint set is finite. By Prop. 1.1(b), we have that μ and μ' are proper, and Prop. 1.1(a) implies that $J = J_\mu$ and $J' = J_{\mu'}$. We have $J = T^k J \leq T_{\mu'}^k J$ for all $k \geq 1$, and by Prop. 1.1(a), we obtain $J \leq \lim_{k\to\infty} T_{\mu'}^k J = J_{\mu'} = J'$. Similarly, $J' \leq J$, showing that $J = J'$ and that T has at most one fixed point.

We next show that T has at least one fixed point. Let μ be a proper policy (there exists one by Assumption 1.1). Choose μ' such that

$$T_{\mu'}J_\mu = TJ_\mu.$$

Then we have $J_\mu = T_\mu J_\mu \geq T_{\mu'}J_\mu$. By Prop. 1.1(b), μ' is proper, and using the monotonicity of $T_{\mu'}$ and Prop. 1.1(a), we obtain

$$J_\mu \geq \lim_{k\to\infty} T_{\mu'}^k J_\mu = J_{\mu'}. \tag{1.6}$$

Continuing in the same manner, we construct a sequence $\{\mu^k\}$ such that each μ^k is proper and

$$J_{\mu^k} \geq TJ_{\mu^k} \geq J_{\mu^{k+1}}, \qquad k = 0, 1, \ldots \tag{1.7}$$

Since the set of proper policies is finite, some policy μ must be repeated within the sequence $\{\mu^k\}$, and by Eq. (1.7), we have

$$J_\mu = TJ_\mu.$$

Thus J_μ is a fixed point of T, and in view of the uniqueness property shown earlier, J_μ is the unique fixed point of T.

Next we show that the unique fixed point of T is equal to the optimal cost vector J^*, and that $T^kJ \to J^*$ for all J. The construction of the preceding paragraph provides a proper μ such that $TJ_\mu = J_\mu$. We will show that $T^kJ \to J_\mu$ for all J and that $J_\mu = J^*$. Let $e = (1, 1, \ldots, 1)$, let $\delta > 0$ be some scalar, and let $\hat{J}$ be the vector satisfying

$$T_\mu \hat{J} = \hat{J} - \delta e.$$

There is a unique such vector because the equation $\hat{J} = T_\mu \hat{J} + \delta e$ can be written as $\hat{J} = g_\mu + \delta e + P_\mu \hat{J}$, so $\hat{J}$ is the cost vector corresponding to μ for g_μ replaced by $g_\mu + \delta e$. Since μ is proper, by Prop. 1.1(a), $\hat{J}$ is unique. Furthermore, we have $J_\mu \leq \hat{J}$, which implies that

$$J_\mu = TJ_\mu \leq T\hat{J} \leq T_\mu \hat{J} = \hat{J} - \delta e \leq \hat{J}.$$

Using the monotonicity of T and the preceding relation, we obtain

$$J_\mu = T^kJ_\mu \leq T^k\hat{J} \leq T^{k-1}\hat{J} \leq \hat{J}, \qquad k \geq 1.$$

Hence, $T^k\hat{J}$ converges to some vector $\tilde{J}$, and we have

$$T\tilde{J} = T\Big(\lim_{k\to\infty} T^k\hat{J}\Big).$$

The mapping T can be seen to be continuous, so we can interchange T with the limit in the preceding relation, thereby obtaining $\tilde{J} = T\tilde{J}$. By the uniqueness of the fixed point of T shown earlier, we must have $\tilde{J} = J_\mu$. It is also seen that

$$J_\mu - \delta e = TJ_\mu - \delta e \leq T(J_\mu - \delta e) \leq TJ_\mu = J_\mu.$$

Thus, $T^k(J_\mu - \delta e)$ is monotonically increasing and bounded above. As earlier, it follows that $\lim_{k\to\infty} T^k(J_\mu - \delta e) = J_\mu$. For any J, we can find $\delta > 0$ such that

$$J_\mu - \delta e \leq J \leq \hat{J}.$$

By the monotonicity of T, we then have

$$T^k(J_\mu - \delta e) \leq T^kJ \leq T^k\hat{J}, \qquad k \geq 1,$$

and since $\lim_{k\to\infty} T^k(J_\mu - \delta e) = \lim_{k\to\infty} T^k\hat{J} = J_\mu$, it follows that

$$\lim_{k\to\infty} T^kJ = J_\mu.$$

To show that $J_\mu = J^*$, take any policy $\pi = \{\mu_0, \mu_1, \ldots\}$. We have

$$T_{\mu_0} \cdots T_{\mu_{k-1}} J_0 \geq T^kJ_0,$$

where J_0 is the zero vector. Taking the limsup of both sides as $k \to \infty$ in the preceding inequality, we obtain

$$J_\pi \geq J_\mu,$$

so μ is an optimal stationary policy and $J_\mu = J^*$.

(c) If μ is optimal, then $J_\mu = J^*$ and, by Assumptions 1.1 and 1.2, μ is proper, so by Prop. 1.1(a), $T_\mu J^* = T_\mu J_\mu = J_\mu = J^* = TJ^*$. Conversely, if $J^* = TJ^* = T_\mu J^*$, it follows from Prop. 1.1(b) that μ is proper, and by using Prop. 1.1(a), we obtain $J^* = J_\mu$. Therefore, μ is optimal. **Q.E.D.**

The results of Prop. 1.2 can also be proved (with minor changes) assuming, in place of Assumption 1.2, that $g(i, u) \geq 0$ for all i and $u \in U(i)$, and that there exists an optimal proper policy; see Exercise 2.12.

Compact Control Constraint Sets

It turns out that the finiteness assumption on the control constraint $U(i)$ can be weakened. It is sufficient that, for each i, $U(i)$ be a compact subset of a Euclidean space, and that $p_{ij}(u)$ and $g(i, u)$ be continuous in u over $U(i)$, for all i and j. Under these compactness and continuity assumptions, and also Assumptions 1.1 and 1.2, Prop. 1.2 holds as stated. The proof is similar to the one given above, but is technically much more complex. It can be found in [BeT91b].

Underlying Contractions

We mentioned in Section 1.4 that the strong results we derived for discounted problems in Chapter 1 owe their validity to the contraction property of the mapping T. Despite the similarity of Prop. 1.2 with the corresponding discounted cost results of Section 1.2, under Assumptions 1.1 and 1.2, the mapping T of this section need not be a contraction mapping with respect to any norm; see Exercise 2.13 for a counterexample. On the other hand there is an important special case where T is a contraction mapping with respect to a *weighted sup norm*. In particular, it can be shown that *if all stationary policies are proper*, then there exist positive constants $v_1, \ldots, v_n$ and some γ with $0 \leq \gamma < 1$, such that we have for all vectors J_1 and J_2,

$$\max_{i=1,\ldots,n} \frac{1}{v_i} \big|(TJ_1)(i) - (TJ_2)(i)\big| \leq \gamma \max_{i=1,\ldots,n} \frac{1}{v_i} \big|J_1(i) - J_2(i)\big|.$$

A proof of this fact is outlined in Exercise 2.14.

Pathologies of Stochastic Shortest Path Problems

We now give two examples that illustrate the sensitivity of our results to seemingly minor changes in our assumptions.

Example 1.1 (The Blackmailer's Dilemma [Whi82])

This example shows that the assumption of a finite or compact control constraint set cannot be easily relaxed. Here, there are two states, state 1 and the destination state t. At state 1, we can choose a control u with $0 < u \leq 1$; we then move to state t at no cost with probability u^2, and stay in state 1 at a cost $-u$ with probability $1 - u^2$. Note that every stationary policy is proper in the sense that it leads to the destination with probability one.

We may regard u as a demand made by a blackmailer, and state 1 as the situation where the victim complies. State t is the situation where the victim refuses to yield to the blackmailer's demand. The problem then can be seen as one whereby the blackmailer tries to maximize his total gain by balancing his desire for increased demands with keeping his victim compliant.

If controls were chosen from a *finite* subset of the interval $(0, 1]$, the problem would come under the framework of this section. The optimal cost would then be finite, and there would exist an optimal stationary policy. It turns out, however, that *without the finiteness restriction the optimal cost starting at state 1 is $-\infty$ and there exists no optimal stationary policy*. Indeed, for any stationary policy μ with $\mu(1) = u$, we have

$$J_\mu(1) = -u + (1 - u^2)J_\mu(1)$$

from which

$$J_\mu(1) = -\frac{1}{u}.$$

Therefore, $\min_\mu J_\mu(1) = -\infty$ and $J^*(1) = -\infty$, but there is no stationary policy that achieves the optimal cost. Note also that this situation would not change if the constraint set were $u \in [0, 1]$ (i.e., $u = 0$ were an allowable control), although in this case the stationary policy that applies $\mu(1) = 0$ is improper and its corresponding cost vector is zero, thus violating Assumption 1.2.

An interesting fact about this problem is that there is an optimal *nonstationary* policy π for which $J_\pi(1) = -\infty$. This is the policy $\pi = \{\mu_0, \mu_1, \ldots\}$ that applies $\mu_k(1) = \gamma/(k+1)$ at time k and state 1, where γ is a scalar in the interval $(0, 1/2)$. We leave the verification of this fact to the reader. What happens with the policy π is that the blackmailer requests diminishing amounts over time, which nonetheless add to ∞. However, the probability of the victim's refusal diminishes at a much faster rate over time, and as a result, the probability of the victim remaining compliant forever is strictly positive, leading to an infinite total expected payoff to the blackmailer.

Example 1.2 (Pure Stopping Problems)

This example illustrates why we need to assume that all improper policies have infinite cost for at least some initial state (Assumption 1.2). Consider an optimal stopping problem where a state-dependent cost is incurred only when invoking a stopping action that drives the system to the destination; all costs are zero prior to stopping. Eventual stopping is a requirement here, so to properly formulate such a stopping problem as a total cost infinite horizon problem, it is essential to make the stopping costs negative (by adding a negative constant to all stopping costs if necessary), providing an incentive to stop. We then come under the framework of this section but with Assumption 1.2 violated because the improper policy that never stops does not yield infinite cost for any starting state. Unfortunately, this seemingly small relaxation of our assumptions invalidates our results as shown by the example of Fig. 2.1.1. This example is in effect a deterministic shortest path problem involving a cycle with zero length. In particular, in the example there is a (nonoptimal) improper policy that yields finite cost for all initial states (rather than infinite cost for some initial state), and T has multiple fixed points.

2.2 COMPUTATIONAL METHODS

All the methods developed in connection with the discounted cost problem in Section 1.3, have stochastic shortest path analogs. For example, value iteration works as shown by Prop. 1.2(b). Furthermore, the (exact and approximate) linear programming approach also has a straightforward extension (cf. Section 1.3.4), since J^* is the largest solution of the system of inequalities $J \leq TJ$. In this section, we will discuss in more detail some

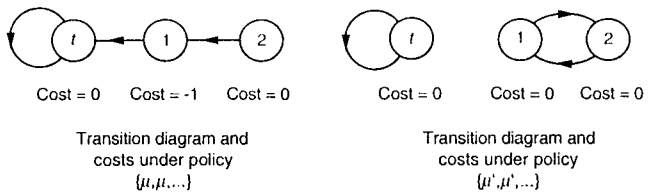


Figure 2.1.1 Example where Prop. 1.2 fails to hold when Assumption 1.2 is violated. There are two stationary policies, μ and μ', with transition probabilities and costs as shown. The equation $J = TJ$ is given by

$$J(1) = \min\{-1, J(2)\},$$

$$J(2) = J(1),$$

and is satisfied by any J of the form

$$J(1) = \delta, \qquad J(2) = \delta,$$

with $\delta \leq -1$. Here the proper policy μ is optimal and the corresponding optimal cost vector is

$$J(1) = -1, \qquad J(2) = -1.$$

The difficulty is that the improper policy μ' has finite (zero) cost for all initial states.

of the major methods, and we will also focus on some stochastic shortest path problems with special structure. It turns out that by exploiting this special structure, we can improve the convergence properties of some of the methods. For example, in deterministic shortest path problems, value iteration terminates finitely (Section 2.1 of Vol. I), whereas this does not happen for any significant class of discounted cost problems.

2.2.1 Value Iteration

As shown by Prop. 1.2(b), value iteration works for stochastic shortest path problems. Furthermore, several of the enhancements and variations of value iteration for discounted problems have stochastic shortest path analogs. In particular, there are error bounds similar to the ones of Prop. 3.1 in Section 1.3 (although not quite as powerful; see Section 7.2 of Vol. I). It can also be shown that the Gauss-Seidel version of the method works and that its rate of convergence is typically faster than that of the ordinary method (Exercise 2.6). Furthermore, the rank-one correction method described in Section 1.3.1 is straightforward and effective, as long as there is some separation between the dominant and the subdominant eigenvalue moduli.

Finite Termination of Value Iteration

Generally, the value iteration method requires an infinite number of iterations in stochastic shortest path problems. However, under special circumstances, the method can terminate finitely. A prominent example is the case of a deterministic shortest path problem, but there are other more general circumstances where termination occurs. In particular, let us *assume that the transition probability graph corresponding to some optimal stationary policy μ^* is acyclic*. By this we mean that there are no cycles in the graph that has as nodes the states $1, \ldots, n, t$, and has an arc (i, j) for each pair of states i and j such that $p_{ij}(\mu^*(i)) > 0$. We assume in particular that there are no positive self-transition probabilities $p_{ii}(\mu^*(i))$ for $i \neq t$, but it turns out that under Assumptions 1.1 and 1.2, a stochastic shortest path problem with such self-transitions can be converted into another stochastic shortest path problem where $p_{ii}(u) = 0$ for all $i \neq t$ and $u \in U(i)$. In particular, it can be shown (Exercise 2.8) that the modified stochastic shortest path problem that has costs

$$\tilde{g}(i, u) = g(i, u) + \frac{g(i, u)p_{ii}(u)}{1 - p_{ii}(u)}, \qquad i = 1, \ldots, n,$$

in place of $g(i, u)$, and transition probabilities

$$\tilde{p}_{ij}(u) = \begin{cases} 0 & \text{if } j = i, \\ \frac{p_{ij}(u)}{1 - p_{ii}(u)} & \text{if } j \neq i, \end{cases} \qquad i = 1, \ldots, n,$$

instead of $p_{ij}(u)$ is equivalent to the original in the sense that it has the same optimal costs and policies.

We claim that, under the preceding acyclicity assumption, *the value iteration method will yield J^* after at most n iterations when started from the vector J given by*

$$J(i) = \infty, \qquad i = 1, \ldots, n. \tag{2.1}$$

To show this, consider the sets of states $S_0, S_1, \ldots$, defined by

$$S_0 = \{t\}, \tag{2.2}$$

$$S_{k+1} = \left\{ i \mid p_{ij}(\mu^*(i)) = 0 \text{ for all } j \notin \cup_{m=0}^{k} S_m \right\}, \qquad k = 0, 1, \ldots, \tag{2.3}$$

and let $S_{\overline{k}}$ be the last of these sets that is nonempty. Then in view of our acyclicity assumption, we have

$$\cup_{m=0}^{\overline{k}} S_m = \{1, \ldots, n, t\}. \tag{2.4}$$

Let us show by induction that, starting from the vector J of Eq. (2.1), the value iteration method will yield for $k = 0, 1, \ldots, \overline{k}$,

$$(T^k J)(i) = J^*(i), \qquad \text{for all } i \in \cup_{m=0}^{k} S_m,\ i \neq t.$$

Indeed, this is so for $k = 0$. Assume that $(T^k J)(i) = J^*(i)$ if $i \in \cup_{m=0}^{k} S_m$. Then, by the monotonicity of T, we have for all i,

$$J^*(i) \leq (T^{k+1}J)(i),$$

while we have by the induction hypothesis, the definition of the sets S_k, and the optimality of μ^*,

$$\begin{aligned}(T^{k+1}J)(i) &\leq g\big(i,\mu^*(i)\big) + \sum_{j \in \cup_{m=0}^{k} S_m} p_{ij}\big(\mu^*(i)\big) J^*(j) \\ &= J^*(i), \qquad \text{for all } i \in \cup_{m=0}^{k+1} S_m,\ i \neq t.\end{aligned}$$

The last two relations complete the induction.

Thus, we have shown that under the acyclicity assumption, at the kth iteration, the value iteration method will set to the optimal values the costs of states in the set S_k. In particular, all optimal costs will be found after k iterations.

Consistently Improving Policies

The properties of value iteration can be further improved if there is an optimal policy μ^* under which from a given state, we can only go to a state of lower cost; that is, for all i, we have

$$p_{ij}\big(\mu^*(i)\big) > 0 \qquad \Rightarrow \qquad J^*(i) > J^*(j).$$

We call such a policy *consistently improving*.

A case where a consistently improving policy exists arises in deterministic shortest path problems when all the arc lengths are positive. Another important case arises in continuous-space shortest path problems; see [Tsi93a] and Exercise 2.10.

The transition probability graph corresponding to a consistently improving policy is seen to be acyclic, so when such a policy exists, by the preceding discussion, the value iteration method terminates finitely. However, a stronger property can be proved. As discussed in Chapter 2 of Vol. I, for shortest path problems with positive arc lengths, one can use Dijkstra's algorithm. This is the label correcting method, which removes from the OPEN list a node with minimum label at each iteration and requires just one iteration per node. A similar property holds for stochastic shortest path problems if there is a consistently improving policy: if one removes from the OPEN list a state j with minimum cost estimate $J(j)$, the Gauss-Seidel version of the value iteration method requires just one iteration per state; see Exercise 2.11.

For problems where a consistently improving policy exists, it is also appropriate to use straightforward adaptations of the label correcting shortest path methods discussed in Section 2.3.1 of Vol. I. In particular, one may approximate the policy of removing from the OPEN list a minimum cost state by using the SLF and LLL strategies (see [PBT95]).

2.2.2 Policy Iteration

The policy iteration algorithm is based on the construction used in the proof of Prop. 1.2 to show that T has a fixed point. In the typical iteration, given a proper policy μ and the corresponding cost vector J_μ, one obtains a new proper policy $\overline{\mu}$ satisfying $T_{\overline{\mu}} J_\mu = T J_\mu$. It was shown in Eq. (1.6) that $J_{\overline{\mu}} \leq J_\mu$. It can be seen also that strict inequality $J_{\overline{\mu}}(i) < J_\mu(i)$ holds for at least one state i, if μ is nonoptimal; otherwise we would have $J_\mu = T J_\mu$ and by Prop. 1.2(c), μ would be optimal. Therefore, the new policy is strictly better if the current policy is nonoptimal. Since the number of proper policies is finite, the policy iteration algorithm terminates after a finite number of iterations with an optimal proper policy.

It is possible to execute approximately the policy evaluation step of policy iteration, using a finite number of value iterations, as in the discounted case. Here we start with some vector J_0. For all k, a stationary policy μ^k is defined from J_k according to $T_{\mu^k} J_k = T J_k$, the cost J_{μ^k} is approximately evaluated by $m_k - 1$ additional value iterations, yielding the vector J_{k+1}, which is used in turn to define μ^{k+1}. The proof of Prop. 3.5 in Section 1.3 can be essentially repeated to show that $J_k \to J^*$, assuming that the initial vector J_0 satisfies $T J_0 \leq J_0$. Unfortunately, the requirement $T J_0 \leq J_0$ is essential for the convergence proof, unless all stationary policies are proper, in which case T is a contraction mapping (cf. Exercise 2.14).

As in Section 1.3.3, it is possible to use adaptive aggregation in conjunction with approximate policy evaluation. However, it is important that the destination t forms by itself an aggregate state, which will play the role of the destination in the aggregate Markov chain.

Approximate Policy Iteration

Let us consider an approximate policy iteration algorithm that generates a sequence of stationary policies $\{\mu^k\}$ and a corresponding sequence of approximate cost vectors $\{J_k\}$ satisfying

$$\max_{i=1,\ldots,n} |J_k(i) - J_{\mu^k}(i)| \leq \delta, \qquad k = 0, 1, \ldots \tag{2.5}$$

and

$$\max_{i=1,\ldots,n} |(T_{\mu^{k+1}} J_k)(i) - (T J_k)(i)| \leq \epsilon, \qquad k = 0, 1, \ldots \tag{2.6}$$

where δ and ϵ are some positive scalars, and μ^0 is some proper policy. One difficulty with such an algorithm is that, even if the current policy μ^k is proper, the next policy μ^{k+1} may not be proper. In this case, we have $J_{\mu^{k+1}}(i) = \infty$ for some i, and the method breaks down. Note, however, that for a sufficiently small ϵ, Eq. (2.6) implies that $T_{\mu^{k+1}} J_k = T J_k$, so by Prop. 1.1(b), μ^{k+1} will be proper. In any case, we will analyze the method

under the assumption that all generated policies are proper. The following proposition parallels Prop. 3.6 in Section 1.3. It provides an estimate of the difference $J_{\mu^k} - J^*$ in terms of the scalar

$$\rho = \max_{\substack{i=1,\ldots,n \\ \mu: \text{proper}}} P\{x_n \neq t \mid x_0 = i, \mu\}.$$

Note that for every proper policy μ and state i, we have $P\{x_n \neq t \mid x_0 = i, \mu\} < 1$ by the definition of a proper policy, and since the number of proper policies is finite, we have $\rho < 1$.

Proposition 2.1: Assume that the stationary policies μ^k generated by the approximate policy iteration algorithm are all proper. Then

$$\limsup_{k\to\infty} \max_{i=1,\ldots,n} \big(J_{\mu^k}(i) - J^*(i)\big) \leq \frac{n(1-\rho+n)(\epsilon+2\delta)}{(1-\rho)^2}. \tag{2.7}$$

Proof: The proof is similar to the one of Prop. 3.6 in Section 1.3. We modify the arguments in order to use the relations $T_\mu(J+re) \leq T_\mu J + re$ and $P_\mu^n e \leq \rho e$, which hold for all proper policies μ and positive scalars r. We use Eqs. (2.5) and (2.6) to obtain for all k

$$T_{\mu^{k+1}} J_{\mu^k} \leq T J_{\mu^k} + (\epsilon+2\delta)e \leq T_{\mu^k} J_{\mu^k} + (\epsilon+2\delta)e. \tag{2.8}$$

From Eq. (2.8) and the equation $T_{\mu^k} J_{\mu^k} = J_{\mu^k}$, we have

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

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

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

This relation can be written as

$$J_{\mu^{k+1}} - J_{\mu^k} \leq P_{\mu^{k+1}}(J_{\mu^{k+1}} - J_{\mu^k}) + (\epsilon+2\delta)e, \tag{2.9}$$

where $P_{\mu^{k+1}}$ is the transition probability matrix corresponding to μ^{k+1}. Let

$$\xi_k = \max_{i=1,\ldots,n} \big(J_{\mu^{k+1}}(i) - J_{\mu^k}(i)\big).$$

Then Eq. (2.9) yields

$$\xi_k e \leq \xi_k P_{\mu^{k+1}} e + (\epsilon+2\delta)e.$$

By multiplying this relation with $P_{\mu^{k+1}}$ and by adding $(\epsilon+2\delta)e$, we obtain

$$\xi_k e \leq \xi_k P_{\mu^{k+1}} e + (\epsilon+2\delta)e \leq \xi_k P_{\mu^{k+1}}^2 e + 2(\epsilon+2\delta)e.$$

By repeating this process for a total of $n-1$ times, we have

$$\xi_k e \leq \xi_k P_{\mu^{k+1}}^n e + n(\epsilon+2\delta)e \leq \rho \xi_k e + n(\epsilon+2\delta)e.$$

Thus,

$$\xi_k \leq \frac{n(\epsilon+2\delta)}{1-\rho}. \tag{2.10}$$

Let μ^* be an optimal stationary policy. From Eq. (2.8), we have

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

We also have

$$T_{\mu^{k+1}} J_{\mu^k} = J_{\mu^{k+1}} + T_{\mu^{k+1}} J_{\mu^k} - T_{\mu^{k+1}} J_{\mu^{k+1}} = J_{\mu^{k+1}} + P_{\mu^{k+1}}(J_{\mu^k} - J_{\mu^{k+1}}).$$

By subtracting the last two relations, and by using the definition of ξ_k and Eq. (2.10), we obtain

$$\begin{aligned} J_{\mu^{k+1}} - J^* &\leq P_{\mu^*}(J_{\mu^k} - J^*) + P_{\mu^{k+1}}(J_{\mu^{k+1}} - J_{\mu^k}) + (\epsilon+2\delta)e \\ &\leq P_{\mu^*}(J_{\mu^k} - J^*) + \xi_k P_{\mu^{k+1}} e + (\epsilon+2\delta)e \\ &\leq P_{\mu^*}(J_{\mu^k} - J^*) + \xi_k e + (\epsilon+2\delta)e \\ &\leq P_{\mu^*}(J_{\mu^k} - J^*) + \frac{(1-\rho+n)(\epsilon+2\delta)}{1-\rho} e. \end{aligned} \tag{2.11}$$

Let

$$\zeta_k = \max_{i=1,\ldots,n} \big(J_{\mu^k}(i) - J^*(i)\big).$$

Then Eq. (2.11) yields, for all k,

$$\zeta_{k+1} e \leq \zeta_k P_{\mu^*} e + \frac{(1-\rho+n)(\epsilon+2\delta)}{1-\rho} e.$$

By multiplying this relation with P_{μ^*} and by adding $(1-\rho+n)(\epsilon+2\delta)e/(1-\rho)$, we obtain

$$\zeta_{k+2} e \leq \zeta_{k+1} P_{\mu^*} e + \frac{(1-\rho+n)(\epsilon+2\delta)}{1-\rho} e \leq \zeta_k P_{\mu^*}^2 e + \frac{2(1-\rho+n)(\epsilon+2\delta)}{1-\rho} e.$$

By repeating this process for a total of $n-1$ times, we have

$$\zeta_{k+n}c \leq \zeta_k P_\mu^n c + \frac{n(1-\rho+n)(\epsilon+2\delta)}{1-\rho}c \leq \rho\zeta_k c + \frac{n(1-\rho+n)(\epsilon+2\delta)}{1-\rho}c.$$

By taking the limit superior as $k \to \infty$, we obtain

$$(1-\rho)\limsup_{k\to\infty}\zeta_k \leq \frac{n(1-\rho+n)(\epsilon+2\delta)}{1-\rho},$$

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

The error bound (2.7) uses the worst-case estimate of the number of stages required to reach t with positive probability, which is n. We can strengthen the error bound if we have a better estimate. In particular, for all $m \geq 1$, let

$$\rho_m = \max_{\substack{i=1,\ldots,n \\ \mu:\text{proper}}} P\{x_m \neq t \mid x_0 = i, \mu\},$$

and let $\overline{m}$ be the minimal m for which $\rho_m < 1$. Then the proof of Prop. 2.1 can be adapted to show that

$$\limsup_{k\to\infty}\max_{i=1,\ldots,n}\left(J_{\mu^k}(i) - J^*(i)\right) \leq \frac{\overline{m}(1-\rho_{\overline{m}}+\overline{m})(\epsilon+2\delta)}{(1-\rho_{\overline{m}})^2}.$$

2.3 SIMULATION-BASED METHODS

The computational methods described so far apply when there is a mathematical model of the cost structure and the transition probabilities of the system. In many problems, however, such a model is not available, but instead, the system and cost structure can be simulated. By this we mean that the state space and the control space are known, and there is a computer program that simulates, for a given control u, the probabilistic transitions from any given state i to a successor state j according to the transition probabilities $p_{ij}(u)$, and also generates a corresponding transition cost $g(i,u,j)$. It is then of course possible to use repeated simulation to calculate (at least approximately) the transition probabilities of the system and the expected stage costs by averaging, and then to apply the methods discussed earlier.

The methodology discussed in this section, however, is geared towards an alternative possibility, which is much more attractive when one is faced with a large and complex system, and one contemplates approximations: rather than estimate explicitly the transition probabilities and costs, we

estimate the cost function of a given policy by generating a number of simulated system trajectories and associated costs, and by using some form of "least-squares fit."

Within this context, there are a number of possible approximation techniques, which for the most part are patterned after the value and the policy iteration methods. We focus first on exact methods where estimates of various cost functions are maintained in a "look-up table" that contains one entry per state. We later develop approximate methods where cost functions are maintained in a "compact" form; that is, they are represented by a function chosen from a parametric class, perhaps involving a feature extraction mapping or a neural network. We first consider these methods for the stochastic shortest path problem, and we later adapt them for the discounted cost problem in Section 2.3.4.

To make the notation better suited for the simulation context, we make a slight change in the problem definition. In particular, instead of considering the expected cost $g(i,u)$ at state i under control u, we allow the cost g to depend on the next state j. Thus our notation for the cost per stage is now $g(i,u,j)$. All the results and the entire analysis of the preceding sections can be rewritten in terms of the new notation by replacing $g(i,u)$ with $\sum_{j=1}^n p_{ij}(u)g(i,u,j)$.

2.3.1 Policy Evaluation by Monte-Carlo Simulation

Consider the stochastic shortest path problem of Section 2.1. Suppose that we are given a *fixed* stationary policy μ and we want to calculate by simulation the corresponding cost vector J_μ. One possibility is of course to generate, starting from each i, many sample state trajectories and average the corresponding costs to obtain an approximation to $J_\mu(i)$. We can do this separately for each possible initial state, but a more efficient method is to use each trajectory to obtain a cost sample for many states by considering the costs of the trajectory portions that start at these states. If a state is encountered multiple times within the same trajectory, the corresponding cost samples can be treated as multiple independent samples for that state.†

To simplify notation, in what follows we do not show the dependence of various quantities on the given policy. In particular, the transition probability from i to j, and the corresponding stage cost are denoted by p_{ij} and $g(i,j)$, in place of $p_{ij}\big(\mu(i)\big)$ and $g\big(i,\mu(i),j\big)$, respectively.

To formalize the process, suppose that we perform an infinite number of simulation runs, each ending at the termination state t. Assume also that within the total number of runs, each state is encountered an infinite

† The validity of doing so is not quite obvious because in the case of multiple visits to the same state within the same trajectory, the corresponding multiple cost samples are correlated, since portions of the corresponding cost sequence are shared by these cost samples. For a justifying analysis, see Exercise 2.15.

number of times. Consider the mth time a given state i is encountered, and let $(i, i_1, i_2, \ldots, i_N, t)$ be the remainder of the corresponding trajectory. Let $c(i,m)$ be the corresponding cost of reaching state t,

$$c(i,m) = g(i,i_1) + g(i_1,i_2) + \cdots + g(i_N,t).$$

We assume that the simulations correctly average the desired quantities; that is, for all states i, we have

$$J_\mu(i) = \lim_{M\to\infty} \frac{1}{M} \sum_{m=1}^{M} c(i,m). \tag{3.1}$$

We can iteratively calculate the sums appearing in the above equations by using the update formulas

$$J_\mu(i) := J_\mu(i) + \gamma_m\big(c(i,m) - J_\mu(i)\big), \qquad m = 1,2,\ldots,$$

where

$$\gamma_m = \frac{1}{m}, \qquad m = 1,2,\ldots,$$

and the initial conditions are, for all i,

$$J_\mu(i) = 0.$$

The normal way to implement the preceding algorithm is to update the costs $J_\mu(i)$ at the end of each simulation run that generates the state trajectory $(i_1, i_2, \ldots, i_N, t)$, by using for each $k = 1, \ldots, N$, the formula

$$J_\mu(i_k) := J_\mu(i_k) + \gamma_{m_k}\big(g(i_k,i_{k+1}) + g(i_{k+1},i_{k+2}) + \cdots + g(i_N,t) - J_\mu(i_k)\big), \tag{3.2}$$

where m_k is the number of visits thus far to state i_k and $\gamma_{m_k} = 1/m_k$. There are also forms of the law of large numbers, which allow the use of a different stepsize γ_{m_k} in the above equation. It can be shown that for convergence of iteration (3.2) to the correct cost value $J_\mu(i_k)$, it is sufficient that γ_{m_k} be diminishing at the rate of one over the number of visits to state i_k.

Monte-Carlo Simulation Using Temporal Differences

An alternative (and essentially equivalent) method to implement the Monte-Carlo simulation update (3.2), is to update $J_\mu(i_1)$ immediately after $g(i_1,i_2)$ and i_2 are generated, then update $J_\mu(i_1)$ and $J_\mu(i_2)$ immediately after $g(i_2,i_3)$ and i_3 are generated, and so on. The method uses the quantities

$$d_k = g(i_k,i_{k+1}) + J_\mu(i_{k+1}) - J_\mu(i_k), \qquad k = 1,\ldots,N, \tag{3.3}$$

with $i_{N+1} = t$, which are called *temporal differences*. They represent the difference between the current estimate $J_\mu(i_k)$ of *expected cost* to go to the termination state and the *predicted cost* to go to the termination state,

$$g(i_k,i_{k+1}) + J_\mu(i_{k+1}),$$

based on the simulated outcome of the current stage. Given a sample state trajectory $(i_1, i_2, \ldots, i_N, t)$, the cost update formula (3.2) can be rewritten in terms of the temporal differences d_k as follows [to see this, just add the formulas below and use the fact $J_\mu(i_{N+1}) = J_\mu(t) = 0$]:

Following the state transition (i_1, i_2), set

$$J_\mu(i_1) := J_\mu(i_1) + \gamma_{m_1} d_1 = J_\mu(i_1) + \gamma_{m_1}\big(g(i_1,i_2) + J_\mu(i_2) - J_\mu(i_1)\big).$$

Following the state transition (i_2, i_3), set

$$J_\mu(i_1) := J_\mu(i_1) + \gamma_{m_1} d_2 = J_\mu(i_1) + \gamma_{m_1}\big(g(i_2,i_3) + J_\mu(i_3) - J_\mu(i_2)\big),$$

$$J_\mu(i_2) := J_\mu(i_2) + \gamma_{m_2} d_2 = J_\mu(i_2) + \gamma_{m_2}\big(g(i_2,i_3) + J_\mu(i_3) - J_\mu(i_2)\big),$$

$$\ldots \qquad \ldots$$

Following the state transition (i_N, t), set

$$J_\mu(i_1) := J_\mu(i_1) + \gamma_{m_1} d_N = J_\mu(i_1) + \gamma_{m_1}\big(g(i_N,t) - J_\mu(i_N)\big),$$

$$J_\mu(i_2) := J_\mu(i_2) + \gamma_{m_2} d_N = J_\mu(i_2) + \gamma_{m_2}\big(g(i_N,t) - J_\mu(i_N)\big),$$

$$\ldots \qquad \ldots$$

$$J_\mu(i_N) := J_\mu(i_N) + \gamma_{m_N} d_N = J_\mu(i_N) + \gamma_{m_N}\big(g(i_N,t) - J_\mu(i_N)\big).$$

The stepsizes γ_{m_k}, $k = 1,\ldots,N$, are given by $\gamma_{m_k} = 1/m_k$, where m_k is the number of visits already made to state i_k. In the case where the sample trajectory involves at most one visit to each state, the preceding updates are equivalent to the update (3.2). If there are multiple visits to some state during a sample trajectory, there is a difference between the preceding updates and the update (3.2), because the updates corresponding to each visit to the given state affect the updates corresponding to subsequent visits to the same state. However, this is an effect which is of second order in the stepsize γ, so once γ becomes small, the difference is negligible.

TD(λ)

The preceding implementation of the Monte-Carlo simulation method for evaluating the cost of a policy μ is known as TD(1) (here TD stands for Temporal Differences). A generalization of TD(1) is TD(λ), where λ is a parameter with

$$0 \leq \lambda \leq 1.$$

Given a sample trajectory $(i_1, \ldots, i_N, t)$ with a corresponding cost sequence $g(i_1, i_2), \ldots, g(i_N, t)$, TD($\lambda$) updates the cost estimates $J_\mu(i_1), \ldots, J_\mu(i_N)$ using the temporal differences

$$d_k = g(i_k, i_{k+1}) + J_\mu(i_{k+1}) - J_\mu(i_k), \qquad k = 1, \ldots, N,$$

and the equations

$$J_\mu(i_1) := J_\mu(i_1) + \gamma_{m_1} d_1, \qquad \text{following the transition } (i_1, i_2).$$

$$\begin{cases} J_\mu(i_1) := J_\mu(i_1) + \gamma_{m_1} \lambda d_2, \\ J_\mu(i_2) := J_\mu(i_2) + \gamma_{m_2} d_2, \end{cases} \qquad \text{following the transition } (i_2, i_3),$$

and more generally for $k = 1, \ldots, N$,

$$\begin{cases} J_\mu(i_1) := J_\mu(i_1) + \gamma_{m_1} \lambda^{k-1} d_k, \\ J_\mu(i_2) := J_\mu(i_2) + \gamma_{m_2} \lambda^{k-2} d_k, \\ \cdots \qquad \cdots \\ J_\mu(i_k) := J_\mu(i_k) + \gamma_{m_k} d_k, \end{cases} \qquad \text{following the transition } (i_k, i_{k+1}).$$

The use of a value of λ less than 1 tends to discount the effect of the temporal differences of state transitions far into the future on the cost estimate of the current state. In the case where $\lambda = 0$, we obtain the TD(0) algorithm, which following the transition (i_k, i_{k+1}) updates $J_\mu(i_k)$ by

$$J_\mu(i_k) := J_\mu(i_k) + \gamma_{m_k} \big(g(i_k, i_{k+1}) + J_\mu(i_{k+1}) - J_\mu(i_k) \big). \tag{3.4}$$

This algorithm is a special case of an important stochastic iterative algorithm known as the *stochastic approximation* (or *Robbins-Monro*) method (see e.g., [BMP90], [BeT89a], [LjS83]) for solving Bellman's equations

$$E\big\{ g(i_k, i_{k+1}) + J_\mu(i_{k+1}) \big\} - J_\mu(i_k) = 0.$$

In this algorithm, the expected value above is approximated by using a single sample at each iteration [cf. Eq. (3.4)].

The stepsizes γ_{m_k} need not be equal to $1/m_k$, where m_k is the number of visits thus far to state i_k, but they should diminish to zero with the number of visits to each state. For example one may use the same stepsize $\gamma_m = 1/m$ for all states within the mth simulation trajectory. With such

a stepsize and under some technical conditions, chief of which is that each state $i = 1, \ldots, n$ is visited infinitely often in the course of the simulation, it can be shown that for all $\lambda \in [0, 1]$, the cost estimates $J_\mu(i)$ generated by TD(λ) converge to the correct values with probability 1.

While TD(λ) yields the correct values of $J_\mu(i)$ in the limit regardless of the value of λ, the choice of λ may have a substantial effect on the rate of convergence. Some experience suggests that using $\lambda < 1$ (rather than $\lambda = 1$), often reduces the number of sample trajectories needed to attain the same variance of error between $J_\mu(i)$ and its estimate. However, at present there is no analysis relating to this phenomenon.

Simulation-Based Policy Iteration

The policy evaluation procedures discussed above can be embedded within a simulation-based policy iteration approach. Let us introduce the notion of the *Q-factor* of a state-control pair (i, u) and a stationary policy μ, defined as

$$Q_\mu(i, u) = \sum_{j=1}^{n} p_{ij}(u) \big(g(i, u, j) + J_\mu(j) \big). \tag{3.5}$$

It is the expected cost corresponding to starting at state i, using control u at the first stage, and using the stationary policy μ at the second and subsequent stages.

The Q-factors can be evaluated by first evaluating J_μ as above, and then using further simulation and averaging (if necessary) to compute the right-hand side of Eq. (3.5) for all pairs (i, u). Once this is done, one can execute a policy improvement step using the equation

$$\overline{\mu}(i) = \arg \min_{u \in U(i)} Q_\mu(i, u), \qquad i = 1, \ldots, n. \tag{3.6}$$

We thus obtain a version of the policy iteration algorithm that combines policy evaluation using simulation, and policy improvement using Eq. (3.6) and further simulation, if necessary. In particular, given a policy μ and its associated cost vector J_μ, the cost of the improved policy $J_{\overline{\mu}}$ is computed by simulation, with $\overline{\mu}(i)$ determined using Eq. (3.6) on-line.

2.3.2 Q-Learning

We now introduce an alternative method for cases where there is no explicit model of the system and the cost structure. This method is analogous to value iteration and has the advantage that it can be used directly in the case of multiple policies. Instead of approximating the cost function of a particular policy, it updates directly the Q-factors associated with an *optimal* policy, thereby avoiding the multiple policy evaluation

steps of the policy iteration method. These Q-factors are defined, for all pairs (i, u) by

$$Q(i,u) = \sum_{j=1}^{n} p_{ij}(u)\big(g(i,u,j) + J^*(j)\big).$$

From this definition and Bellman's equation, we see that the Q-factors satisfy for all pairs (i, u),

$$Q(i,u) = \sum_{j=1}^{n} p_{ij}(u)\left(g(i,u,j) + \min_{u' \in U(j)} Q(j,u')\right), \tag{3.7}$$

and it can be shown that the Q-factors are the unique solution of the above system of equations. The proof is essentially the same as the proof of existence and uniqueness of solution of Bellman's equation; see Prop. 1.2 of Section 2.1. In fact, by introducing a system whose states are the original states $1, \ldots, n, t$ together with all the pairs (i, u), the above system of equations can be seen to be a special case of Bellman's equation (see Exercise 2.17). Furthermore, the Q-factors can be obtained by the iteration

$$Q(i,u) := \sum_{j=1}^{n} p_{ij}(u)\left(g(i,u,j) + \min_{u' \in U(j)} Q(j,u')\right), \qquad \text{for all } (i,u),$$

which is analogous to value iteration. A more general version of this is

$$Q(i,u) := (1-\gamma)Q(i,u) + \gamma \sum_{j=1}^{n} p_{ij}(u)\left(g(i,u,j) + \min_{u' \in U(j)} Q(j,u')\right), \tag{3.8}$$

where γ is a stepsize parameter with $\gamma \in (0, 1]$, that may change from one iteration to the next. The Q-learning method is an approximate version of this iteration, whereby the expected value is replaced by a single sample, i.e.,

$$Q(i,u) := Q(i,u) + \gamma\left(g(i,u,j) + \min_{u' \in U(j)} Q(j,u') - Q(i,u)\right).$$

Here j and $g(i, u, j)$ are generated from the pair (i, u) by simulation, that is, according to the transition probabilities $p_{ij}(u)$. Thus Q-learning can be viewed as a combination of value iteration and simulation.

Because Q-learning works using a single sample per iteration, it is well suited for a simulation context. By contrast, there is no single sample version of the value iteration method, except in special cases [see Exercise 2.9(d)]. The reason is that, while it is possible to use a single-sample approximation of a term of the form $E\{\min[\cdot]\}$, such as the one appearing in the Q-factor equation (3.8), it is not possible to do so for a term of the form $\min[E\{\cdot\}]$, such as the one appearing in Bellman's equation.

To guarantee the convergence of the Q-learning algorithm to the optimal Q-factors, all state-control pairs (i, u) must be visited infinitely often, and the stepsize γ should be chosen in some special way. In particular, if the iteration corresponds to the mth visit of the pair (i, u), one may use in the Q-learning iteration the stepsize $\gamma_k = c/m$, where c is a positive constant. We refer to [Tsi94] for a proof of convergence of Q-learning under very general conditions.

2.3.3 Approximations

We now consider approximation/suboptimal control schemes that are suitable for problems with a large number of states. The discounted versions of these schemes, which are discussed in Section 2.3.4, can be adapted for the case of an infinite state space. Generally there are two types of approximations to consider:

(a) Approximation of the optimal cost function J^*. This is done by using a function that, given a state i, produces an approximation $\tilde{J}(i, r)$ of $J^*(i)$ where r is a parameter/weight vector that is typically determined by some form of optimization; for example, by using some type of least squares framework. Once $\tilde{J}(i, r)$ is known, it can be used in real-time to generate a suboptimal control at any state i according to

$$\tilde{\mu}(i) = \arg\min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u)\big(g(i,u,j) + \tilde{J}(j,r)\big).$$

An alternative possibility, which does not require the real-time calculation of the expected value in the above formula, is to obtain approximations $\tilde{Q}(i, u, r)$ of the Q-factors $Q(i, u)$, and then to generate a suboptimal control at any state i according to

$$\tilde{\mu}(i) = \arg\min_{u \in U(i)} \tilde{Q}(i,u,r).$$

It is also possible to use approximations $\tilde{J}_\mu(i, r)$ of the cost functions J_μ of policies μ in an approximate policy iteration scheme. Note that the cost approximation approach can be enhanced if we have additional information on the true functions $J^*(i)$, $Q(i, u)$, or $J_\mu(i)$. For example, if we know that $J^*(i) \geq 0$ for all i, we may first compute the approximation $\tilde{J}(i, r)$ by using some method, and then replace this approximation by $\max\{0, \tilde{J}(i, r)\}$. This idea applies to all the approximation procedures of this section.

(b) Approximation of a policy μ, or the optimal policy μ^*. Again this approximation will be done by some function parameterized by a

parameter/weight vector r, which given a state i, produces an approximation $\tilde{\mu}(i,r)$ of $\mu(i)$ or an approximation $\tilde{\mu}^*(i,r)$ of $\mu^*(i)$. The parameter/weight vector r can be determined by some type of least squares optimization framework.

In this section we discuss several possibilities, emphasizing primarily the case of cost approximation. The choice of the structure of the approximating functions is very significant for the success of the approximation approach. One possibility is to use the *linear* form

$$\tilde{J}(i,r) = \sum_{k=1}^{m} r_k w_k(i), \tag{3.9}$$

where $r = (r_1, \ldots, r_m)$ is the parameter vector and $w_k(i)$ are some fixed and known scalars. This amounts to approximating the cost function J^* by a linear combination of m given basis functions $\big(w_k(1), \ldots, w_k(n)\big)$, where $k = 1, \ldots, m$.

Example 3.1: (Polynomial Approximations)

An important example of linear cost approximation is based on polynomial basis functions. Suppose that the state consists of q integer components $x_1, \ldots, x_q$, each taking values within some limited range of the nonnegative integers. For example, in a queueing system, x_k may represent the number of customers in the kth queue, where $k = 1, \ldots, q$. Suppose that we want to use an approximating function that is quadratic in the components x_k. Then we can define a total of $1 + q + q^2$ basis functions that depend on the state $x = (x_1, \ldots, x_q)$ via

$$w_0(x) = 1, \qquad w_k(x) = x_k, \qquad w_{ks}(x) = x_k x_s, \quad k, s = 1, \ldots, q.$$

An approximating function that is a linear combination of these functions is given by

$$\tilde{J}(x,r) = r_0 + \sum_{k=1}^{q} r_k x_k + \sum_{k=1}^{q} \sum_{s=1}^{q} r_{ks} x_k x_s,$$

where the parameter vector r has components r_0, r_k, and r_{ks}, with $k, s = 1, \ldots, q$. In fact, any kind of approximating function that is polynomial in the components $x_1, \ldots, x_q$ can be constructed in this way.

Example 3.2: (Feature Extraction)

Suppose that through intuition or analysis we can identify a number of characteristics of the state that affect the optimal cost function in a substantial way. We assume that these characteristics can be numerically quantified, and that they form a q-dimensional vector $f(i) = \big(f_1(i), \ldots, f_q(i)\big)$, called the *feature vector* of state i. For example, in computer chess (Section 6.3.2 of Vol. 1) where the state is the current board position, appropriate features are material balance, piece mobility, king safety, and other positional factors. Features, when well-chosen, can capture the dominant nonlinearities of the optimal cost function J^*, and can be used to approximate J^* through the linear combination

$$\tilde{J}(i,r) = r_0 + \sum_{k=1}^{q} r_k f_k(i),$$

where $r_0, r_1, \ldots, r_q$ are appropriately chosen weights.

It is also possible to combine feature extraction with more general polynomial approximations of the type discussed in Example 3.1. For example, a feature extraction mapping f followed by a quadratic polynomial mapping, yields an approximating function of the form

$$\tilde{J}(i,r) = r_0 + \sum_{k=1}^{q} r_k f_k(i) + \sum_{k=1}^{q} \sum_{s=1}^{q} r_{ks} f_k(i) f_s(i),$$

where the parameter vector r has components r_0, r_k, and r_{ks}, with $k, s = 1, \ldots, q$. This function can be viewed as a linear cost approximation that uses the basis functions

$$w_0(i) = 1, \qquad w_k(i) = f_k(i), \qquad w_{ks}(i) = f_k(i) f_s(i), \quad k, s = 1, \ldots, q.$$

Note that more that one state may map into the same feature vector, so that each distinct value of feature vector corresponds to a subset of states. This subset may be viewed as an "aggregate state." The optimal cost function J^* is approximated by a function that is constant over each aggregate state. We will discuss this viewpoint shortly.

It can be seen from the preceding examples that linear approximating functions of the form (3.9) are well suited for a broad variety of situations. There are also interesting nonlinear approximating functions $\tilde{J}(i,r)$, including those defined by neural networks, perhaps in combination with feature extraction mappings. In our discussion, we will not address the choice of the structure of $\tilde{J}(i,r)$, but rather focus on various methods for obtaining a suitable parameter vector r. We will primarily discuss three approaches:

(a) *Feature-based aggregation*, where r is determined as the cost vector of an "aggregate stochastic shortest path problem."

(b) *Minimizing the Bellman equation error*, where r is determined so that the approximate cost function $\tilde{J}(i,r)$ nearly satisfies Bellman's equation.

(c) *Approximate policy iteration*, where the cost functions J_μ of the generated policies μ are approximated by $\tilde{J}_\mu(i,r)$, with r chosen according to a least-squares error criterion.

We note, however, that the methods described in this subsection are not fully understood. We have chosen to present them because of their potential to deal with problems that are too complex to be handled in any other way.

Feature-Based Aggregation

We mentioned earlier in Example 3.2 that a feature extraction mapping divides the state space into subsets. The states of each subset are mapped into the same feature vector, and are "similar" in that they "share the same features." With this context in mind, let the set of states $\{1,\ldots,n\}$ of the given stochastic shortest path problem be partitioned in m disjoint subsets S_k, $k=1,\ldots,m$. We approximate the optimal cost $J^*(i)$ by a function that is constant over each set S_k, that is,

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

where $r=(r_1,\ldots,r_m)'$ is a vector of parameters and

$$w_k(i)=\begin{cases}1 & \text{if } i\in S_k,\\ 0 & \text{if } i\in S_k.\end{cases}$$

Equivalently, the approximate cost function $\big(\tilde{J}(1,r),\ldots,\tilde{J}(n,r)\big)'$ is represented as Wr, where W is the $n\times m$ matrix whose entry in the ith row and kth column is $w_k(i)$. The ith row of W may be viewed as the feature vector corresponding to state i (cf. Example 3.2).

In the aggregation approach, the parameters r_k are obtained as the optimal costs of an "aggregate stochastic shortest path problem" whose states are the subsets S_k. Thus r_k is chosen to be the optimal cost of the aggregate state S_k in an aggregate problem, which is formulated similar to the aggregation method of Section 1.3.3. In particular, let Q be an $m\times n$ matrix such that the kth row of Q is a probability distribution $(q_{k1},\ldots,q_{kn})$ with $q_{ki}=0$ if $i\notin S_k$. As in Section 1.3.3, the structure of Q implies that for each stationary policy μ, the matrix

$$R_\mu=QP_\mu W$$

is an $m\times m$ transition probability matrix. The states of the aggregate stochastic shortest path problem are the sets $S_1,\ldots,S_m$ together with the termination state t; the stationary policies select at aggregate state S_k a control $u\in U(i)$ for each $i\in S_k$ and thus can be identified with stationary policies of the original stochastic shortest path problem; finally the transition probability matrix corresponding to μ in the aggregate stochastic shortest path problem is R_μ. Given a stationary policy μ, the state transition mechanism in the aggregate stochastic shortest path problem can be described as follows: at aggregate state S_k, we move to state i with probability q_{ki}, then we move to state j with probability $p_{ij}\big(\mu(i)\big)$, and finally, if j is not the termination state t, we move to the aggregate state S_l corresponding to j ($j\in S_l$).

Suppose now that $r=(r_1,\ldots,r_m)'$ is the optimal cost function of the aggregate stochastic shortest path problem. Then r solves the corresponding Bellman equation, which has the form

$$r_k=\sum_{i=1}^{n} q_{ki}\min_{u\in U(i)}\sum_{j=1}^{n} p_{ij}(u)\left(g(i,u,j)+\sum_{s=1}^{m} r_s w_s(j)\right),\qquad k=1,\ldots,m.$$

One way to obtain r is policy iteration based on Monte-Carlo simulation, as described in Section 2.3.1. An alternative, due to [TsV94], is to use a simulation-based form of value iteration for the aggregate problem. Here, at each iteration we choose a subset S_k, we randomly select a state $i\in S_k$ according to the probabilities q_{ki}, and we update r_k according to

$$r_k:=(1-\gamma)r_k+\gamma\min_{u\in U(i)}\sum_{j=1}^{n} p_{ij}(u)\left(g(i,u,j)+\sum_{s=1}^{m} r_s w_s(j)\right), \tag{3.10}$$

where γ is a positive stepsize that diminishes to zero as the algorithm progresses. The following example illustrates the method. We refer to [TsV94] for experimental results relating to this example as well for convergence analysis of the method.

Example 3.3: (Tetris [TsV94])

Tetris is a popular video game played on a two-dimensional grid. Each square in the grid can be full or empty, making up a "wall of bricks" with "holes" and a "jagged top". The squares fill up as blocks of different shapes fall at a constant rate from the top of the grid and are added to the top of the wall. As a given block falls, the player can move horizontally and rotate the block in all possible ways, subject to the constraints imposed by the sides of the grid and the top of the wall. There is a finite set of standard shapes for the falling blocks. The game starts with an empty grid and ends when a square in the top row becomes full and the top of the wall reaches the top of the grid. However, when a row of full squares is created, this row is removed, the bricks lying above this row move one row downward, and the player scores a point. The player's objective is to maximize the score attained (total number of rows removed) up to termination of the game.

Assuming that, for every policy, the game terminates with probability one (something that is not really known at present), we can model the problem of finding an optimal tetris playing strategy as a stochastic shortest path problem. The control, denoted by u, is the horizontal positioning and rotation applied to the falling block. The state consists of two components:

(1) The board position, that is, a binary description of the full/empty status of each square, denoted by x.

(2) The shape of the current falling block, denoted by y.

The component y is generated according to a probability distribution $p(y)$, independently of the control. Exercise 2.9 shows that under these circumstances, it is possible to derive a reduced form of Bellman's equation

involving a cost function $\hat{J}$ that depends only on the component x of the state (see also Exercise 1.22 of Vol. 1). This equation has the intuitive form

$$\hat{J}(x) = \sum_y p(y) \max_u \big[g(x,y,u) + \hat{J}\big(f(x,y,u)\big)\big], \qquad \text{for all } x,$$

where $g(x,y,u)$ and $f(x,y,u)$ are the number of points scored (rows removed), and the board position when the state is (x,y) and control u is applied, respectively.

Unfortunately, the number of states is extremely large. It is equal to $m2^{hw}$, where m is the number of different shapes of falling blocks, and h and w are the height and width of the grid, respectively. In particular, for the reasonable numbers $m = 7$, $h = 20$, and $w = 7$ we have over 10^{12} states. Thus it is essential to use approximations.

An approximating function that involves feature extraction is particularly attractive here, since the quality of a given position can be described quite well by a few features that are easily recognizable by experienced players. These features include the current height of the wall, and the presence of "holes" and "glitches" (severe irregularities) in the first few rows. Suppose that, based on human experience and trial and error, we obtain a method to map each board position x into a vector of features. Suppose that there is a finite number of possible feature vectors, say m, and define

$$w_k(x) = \begin{cases} 1 & \text{if board position } x \text{ maps into the } k\text{th feature vector,} \\ 0 & \text{otherwise.} \end{cases}$$

The approximating function $\tilde{J}(x,r)$ is given by

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

where $r = (r_1, \ldots, r_m)$ is the parameter vector. The simulation-based value iteration (3.10) takes the form

$$r_k := (1-\gamma) r_k + \gamma \max_u \big[g(x,y,u) + \tilde{J}\big(f(x,y,u),r\big)\big],$$

where the positive stepsize γ diminishes with the number of visits to position x.

One way to implement the method is as follows: The game is simulated many times from start to finish, starting from a variety of "representative" board positions. At each iteration, we have the current board position x and we determine the feature vector k to which x maps. Then we randomly generate a falling block y according to a known and fixed probabilistic mechanism, and we update r_k using the above iteration. Let u^* be the choice of u that attains the maximum in the iteration,

$$u^* = \arg\max_u \big[g(x,y,u) + \tilde{J}\big(f(x,y,u),r\big)\big].$$

Then the board position subsequent to x in the simulation is $f(x,y,u^*)$, and this position is used as the current state for the next iteration.

In the aggregate stochastic shortest path problem formulated above, policies consist of a different control choice for each state. A somewhat different aggregate stochastic shortest path problem is obtained by requiring that, for each k, the same control is used at all states of S_k. This control must be chosen from a suitable set $\overline{U}(k)$ of admissible controls for the states in S_k. The optimal cost function $r = (r_1, \ldots, r_m)'$ corresponding to this aggregate stochastic shortest path problem solves the following Bellman equation

$$r_k = \min_{u \in \overline{U}(k)} \sum_{i=1}^{n} q_{ki} \sum_{j=1}^{n} p_{ij}(u) \left(g(i,u,j) + \sum_{s=1}^{m} r_s w_s(j) \right), \qquad k = 1, \ldots, m.$$

This equation can be solved by Q-learning, particularly when m is relatively small and the number of controls in the sets $\overline{U}(k)$ is also small.

Note also that the aggregate problem need not be solved exactly, but can itself be solved approximately by any of the simulation-based methods to be discussed subsequently in this section. In this context, aggregation is used as a feature extraction mapping that maps each state i to the corresponding feature vector $w(i) = \big(w_1(i), \ldots, w_m(i)\big)$. This feature vector becomes the input to some other approximating function (see Fig. 2.3.1).

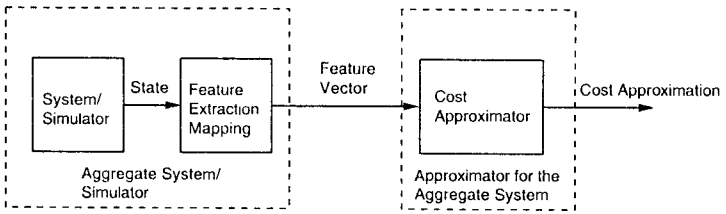


Figure 2.3.1 View of a cost function approximation scheme that consists of a feature extraction mapping followed by an approximator. The scheme conceptually separates into an aggregate system and a cost approximator for the aggregate system.

We finally mention an extension of the aggregation approach whereby we represent the approximate cost function $\big(\tilde{J}(1,r), \ldots, \tilde{J}(n,r)\big)'$ as Wr, where each row of the $n \times m$ matrix W is a probability distribution. Thus

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

where $w_k(i)$ is the (i,k)th entry of the matrix W, and we have

$$\sum_{k=1}^{n} w_k(i) = 1, \qquad w_k(i) \geq 0, \quad i = 1, \ldots, n, \quad k = 1, \ldots, m.$$

The transition probability matrix of the aggregate stochastic shortest path problem corresponding to μ is still $R_\mu = QP_\mu W$, and we may use as parameter vector r the optimal cost vector of this aggregate problem.

Approximation Based on Bellman's Equation

Another possibility for approximation of the optimal cost by a function $\tilde{J}(i,r)$, where r is a vector of unknown parameters, is based on minimizing the error in Bellman's equation; for example by solving the problem

$$\min_r \sum_{i \in \tilde{S}} \Big| \tilde{J}(i,r) - \min_{u \in U(i)} \sum_j p_{ij}(u)\big(g(i,u,j) + \tilde{J}(j,r)\big) \Big|^2, \tag{3.11}$$

where $\tilde{S}$ is a suitably chosen subset of "representative" states. This minimization may be attempted by using some type of gradient or Gauss-Newton method.

A gradient-like method that can be used to solve this problem is obtained by making a correction to r that is proportional to the gradient of the squared error term in Eq. (3.11). This method is given by

$$\begin{aligned} r &:= r - \gamma D(i,r) \nabla D(i,r) \\ &= r - \gamma D(i,r) \Big(\sum_j p_{ij}(\overline{u}) \nabla \tilde{J}(j,r) - \nabla \tilde{J}(i,r) \Big), \end{aligned} \tag{3.12}$$

where ∇ denotes the gradient with respect to r, $D(i,r)$ is the error in Bellman's equation, given by

$$D(i,r) = \min_{u \in U(i)} \sum_j p_{ij}(u)\big(g(i,u,j) + \tilde{J}(j,r)\big) - \tilde{J}(i,r),$$

$\overline{u}$ is given by

$$\overline{u} = \arg \min_{u \in U(i)} \sum_j p_{ij}(u)\big(g(i,u,j) + \tilde{J}(j,r)\big),$$

and γ is a stepsize, which may change from one iteration to the next. The method should perform many such iterations at each of the representative states. Typically one should cycle through the set of representative states $\tilde{S}$ in some order, which may change (perhaps randomly) from one cycle to the next.

Note that in iteration (3.12) we approximate the gradient of the term

$$\min_{u \in U(i)} \sum_j p_{ij}(u)\big(g(i,u,j) + \tilde{J}(j,r)\big) \tag{3.13}$$

by

$$\sum_j p_{ij}(\overline{u}) \nabla \tilde{J}(j,r),$$

which can be shown to be correct only when the above minimum is attained at a unique $\overline{u} \in U(i)$ [otherwise the function (3.13) is nondifferentiable with respect to r]. Thus the convergence of iteration (3.12) should be analyzed using the theory of nondifferentiable optimization. One possibility to avoid this complication is to replace the nondifferentiable term (3.13) by a smooth approximation, which can be arbitrarily accurate (see [Ber82b], Ch. 3).

An interesting special case arises when we want to approximate the cost function of a given policy μ by a function $\tilde{J}_\mu(i,r)$, where r is a parameter vector. The iteration (3.12) then takes the form

$$\begin{aligned} r &:= r - \gamma E_j\big\{ d_\mu(i,j,r) \mid i,\mu \big\} E_j\big\{ \nabla d_\mu(i,j,r) \mid i,\mu \big\} \\ &= r - \gamma E_j\big\{ d_\mu(i,j,r) \mid i,\mu \big\} \big(E_j\big\{ \nabla \tilde{J}_\mu(j,r) \mid i,\mu \big\} - \nabla \tilde{J}_\mu(i,r) \big), \end{aligned} \tag{3.14}$$

where

$$d_\mu(i,j,r) = g\big(i,\mu(i),j\big) + \tilde{J}_\mu(j,r) - \tilde{J}_\mu(i,r).$$

and $E_j\{\cdot \mid i,\mu\}$ denotes expected value over j using the transition probabilities $p_{ij}\big(\mu(i)\big)$. There is a simpler version of iteration (3.14) that does not require averaging over the successor states j. In this version, the two expected values in iteration (3.14) are replaced by two independent single sample values. In particular, r is updated by

$$r := r - \gamma d_\mu(i,j,r)\big(\nabla \tilde{J}_\mu(\overline{j},r) - \nabla \tilde{J}_\mu(i,r)\big), \tag{3.15}$$

where j and $\overline{j}$ correspond to two independent transitions starting from i. It is necessary to use two independently generated states j and $\overline{j}$ in order that the expected value (over j and $\overline{j}$) of the product

$$d_\mu(i,j,r)\big(\nabla \tilde{J}_\mu(\overline{j},r) - \nabla \tilde{J}_\mu(i,r)\big),$$

given i, is equal to the term

$$E_j\big\{ d_\mu(i,j,r) \mid i,\mu \big\} \big(E_j\big\{ \nabla \tilde{J}_\mu(j,r) \mid i,\mu \big\} - \nabla \tilde{J}_\mu(i,r) \big)$$

appearing in the right-hand side of Eq. (3.14).

There are also versions of the above iterations that update Q-factor approximations rather than cost approximations. In particular, let us introduce an approximation $\tilde{Q}(i,u,r)$ to the Q-factor $Q(i,u)$, where r is an

unknown parameter vector. Bellman's equation for the Q-factors is given by [cf. Eq. (3.7)]

$$Q(i,u) = \sum_j p_{ij}(u)\Big(g(i,u,j) + \min_{u' \in U(j)} Q(j,u')\Big),$$

so in analogy with problem (3.11), we determine the parameter vector r by solving the least squares problem

$$\min_r \sum_{(i,u)\in\tilde{V}} \Big|\tilde{Q}(i,u,r) - \sum_j p_{ij}(u)\Big(g(i,u,j) + \min_{u'\in U(i)} \tilde{Q}(j,u',r)\Big)\Big|^2, \quad (3.16)$$

where $\tilde{V}$ is a suitably chosen subset of "representative" state-control pairs. The analog of the gradient-like methods (3.12) and (3.14) is given by

$$\begin{aligned} r :&= r - \gamma E\big\{d_u(i,j,r) \mid i,u\big\} E\big\{\nabla d_u(i,j,r) \mid i,u\big\} \\ &= r - \gamma E\big\{d_u(i,j,r) \mid i,u\big\}\Big(\sum_j p_{ij}(u)\nabla\tilde{Q}(j,\overline{u},r) - \nabla\tilde{Q}(i,u,r)\Big), \end{aligned}$$

where $d_u(i,j,r)$ is given by

$$d_u(i,j,r) = g(i,u,j) + \min_{u'\in U(i)} \tilde{Q}(j,u',r) - \tilde{Q}(i,u,r),$$

$\overline{u}$ is obtained by

$$\overline{u} = \arg\min_{u'\in U(i)} \tilde{Q}(j,u',r),$$

and γ is a stepsize parameter. In analogy with Eq. (3.15), the two-sample version of this iteration is given by

$$r := r - \gamma d_u(i,j,r)\big(\nabla\tilde{Q}(\overline{j},\overline{u},r) - \nabla\tilde{Q}(i,u,r)\big),$$

where j and $\overline{j}$ are two states independently generated from i according to the transition probabilities corresponding to u, and

$$\overline{u} = \arg\min_{u'\in U(i)} \tilde{Q}(\overline{j},u',r).$$

Note that there is no two-sample version of iteration (3.12), which is based on optimal cost approximation. This is the advantage of using Q-factor approximations rather than optimal cost approximations. The point is that it is possible to use single-sample or two-sample approximations in gradient-like methods for terms of the form $E\{\min[\cdot]\}$, such as the one appearing in Eq. (3.16), but not for terms of the form $\min[E\{\cdot\}]$, such as the one appearing in Eq. (3.11). The following example illustrates the use of the two-sample approximation idea.

Example 3.4: (Tetris Continued)

Consider the game of tetris described in Example 3.3, and suppose that an approximation of a given form $\tilde{J}(x,r)$ is desired, where the parameter vector r is obtained by solving the problem

$$\min_r \sum_{x\in\tilde{S}} \Big|\tilde{J}(x,r) - \sum_y p(y)\max_u\big[g(x,y,u) + \tilde{J}\big(f(x,y,u),r\big)\big]\Big|^2,$$

where $\tilde{S}$ is a suitably chosen set of "representative" states. Because this problem involves a term of the form $E\{\max[\cdot]\}$, a two-sample gradient-like method is possible. It has the form

$$r := r - \gamma d(x,y,r)\big(\nabla\tilde{J}\big(f(x,\overline{y},\overline{u}),r\big) - \nabla\tilde{J}(x,r)\big),$$

where y and $\overline{y}$ are two falling blocks that are randomly and independently generated,

$$d(x,y,r) = \max_u\big[g(x,y,u) + \tilde{J}\big(f(x,y,u),r\big)\big] - \tilde{J}(x,r),$$

and

$$\overline{u} = \arg\max_{u'}\big[g(x,\overline{y},u') + \tilde{J}\big(f(x,\overline{y},u',r)\big)\big].$$

Similar to Example 3.3, consider a feature-based approximating function $\tilde{J}(x,r)$ given by

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

where $r = (r_1,\ldots,r_m)$ is the parameter vector and

$$w_k(x) = \begin{cases} 1 & \text{if board position } x \text{ maps into the } k\text{th feature vector,} \\ 0 & \text{otherwise.} \end{cases}$$

For this approximating function, the preceding two-sample gradient iteration takes the relatively simple form

$$r_k := r_k - \gamma d(x,y,r)\big(w_k\big(f(x,\overline{y},\overline{u})\big) - w_k(x)\big). \qquad k = 1,\ldots,m.$$

Note that this iteration updates at most two parameters [the ones corresponding to the feature vectors to which the board positions x and $f(x,\overline{y},\overline{u})$ map, assuming that these feature vectors are different]. To implement the method, a set $\tilde{S}$ containing a large number of states x is selected and at each $x \in \tilde{S}$, two falling blocks y and $\overline{y}$ are independently generated. The controls u and $\overline{u}$ that are optimal for (x,y) and $(x,\overline{y})$, based on the current parameter vector r, are calculated, and the parameters of the feature vectors associated with x and $f(x,\overline{y},\overline{u})$ are adjusted according to the preceding formula.

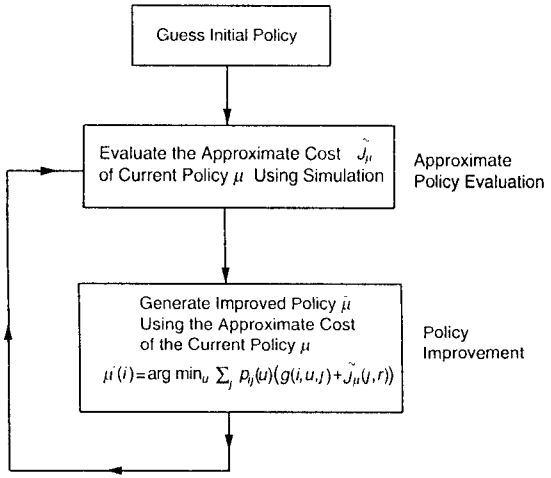


Figure 2.3.2 Block diagram of approximate policy iteration.

Approximate Policy Iteration Using Monte-Carlo Simulation

We now discuss an approximate form of the policy iteration method, where we use approximations $\tilde{J}_\mu(i,r)$ to the cost J_μ of stationary policies μ, and/or approximations $\tilde{Q}_\mu(i,u,r)$ to the corresponding Q-factors. The theoretical basis of the method was discussed in Section 2.2.2 (cf. Prop. 2.1).

Similar to our earlier discussion on simulation, suppose that for a fixed stationary policy μ, we have a subset of "representative" states $\tilde{S}$ (perhaps chosen in the course of the simulation), and that for each $i \in \tilde{S}$, we have $M(i)$ samples of the cost $J_\mu(i)$. The mth such sample is denoted by $c(i,m)$. Then, we can introduce approximate costs $\tilde{J}_\mu(i,r)$, where r is a parameter/weight vector obtained by solving the following least-squares optimization problem

$$\min_r \sum_{i\in\tilde{S}} \sum_{m=1}^{M(i)} \left|\tilde{J}_\mu(i,r) - c(i,m)\right|^2.$$

Once the optimal value of r has been determined, we can approximate the costs $J_\mu(i)$ of the policy μ by $\tilde{J}_\mu(i,r)$. Then, we can evaluate approximate Q-factors using the formula

$$\tilde{Q}_\mu(i,u,r) = \sum_j p_{ij}(u)\big(g(i,u,j) + \tilde{J}_\mu(j,r)\big). \tag{3.17}$$

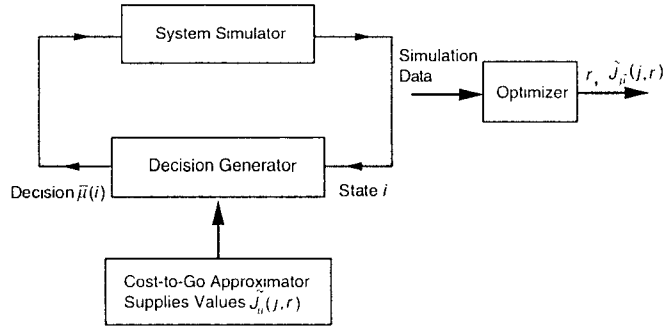


Figure 2.3.3 Structure of approximate policy iteration algorithm.

and we can obtain an improved policy $\bar{\mu}$ using the formula

$$\begin{aligned}\bar{\mu}(i) &= \arg\min_{u\in U(i)} \tilde{Q}_\mu(i,u,r) \\ &= \arg\min_{u\in U(i)} \sum_j p_{ij}(u)\big(g(i,u,j) + \tilde{J}_\mu(j,r)\big), \qquad \text{for all } i.\end{aligned} \tag{3.18}$$

We thus obtain an algorithm that alternates between approximate policy evaluation steps and policy improvement steps, as illustrated in Fig. 2.3.2. The algorithm requires a single approximation per policy iteration, namely the approximation $\tilde{J}_\mu(i,r)$ associated with the current policy μ. The parameter vector r determines the Q-factors via Eq. (3.17) and the next policy $\bar{\mu}$ via Eq. (3.18).

For another view of the approximate policy iteration algorithm, note that it consists of four modules (see Fig. 2.3.3):

(a) The *simulator*, which given a state-decision pair (i,u), generates the next state j according to the correct transition probabilities.

(b) The *decision generator*, which generates the decision $\bar{\mu}(i)$ of the improved policy at the current state i [cf. Eq. (3.18)] for use in the simulator.

(c) The *cost-to-go approximator*, which is the function $\tilde{J}_\mu(j,r)$ that is consulted by the decision generator for approximate cost-to-go values to use in the minimization of Eq. (3.18).

(d) The *optimizer*, which accepts as input the sample trajectories produced by the simulator and solves the problem

$$\min_{\bar{r}} \sum_{i\in\tilde{S}} \sum_{m=1}^{M(i)} \left|\tilde{J}_{\bar{\mu}}(i,\bar{r}) - c(i,m)\right|^2 \tag{3.19}$$

to obtain the approximation $\tilde{J}_{\overline{\mu}}(i,\overline{r})$ of the cost of $\overline{\mu}$.

Note that in very large problems, the policy $\overline{\mu}$ cannot be evaluated and stored in explicit form, and thus the optimization in Eq. (3.18) must be evaluated "on the fly" during the simulation. When this is the case, the parameter vector r associated with μ remains unchanged as we evaluate the cost of the improved policy $\overline{\mu}$ by generating the simulation data and by solving the least squares problem (3.19).

One way to solve this latter problem is to use gradient-like methods. Given a sample state trajectory $(i_1, i_2, \ldots, i_N, t)$ generated using the policy $\overline{\mu}$, which is defined by Eq. (3.18), the parameter vector $\overline{r}$ associated with $\overline{\mu}$ is updated by

$$\overline{r} := \overline{r} - \gamma \sum_{k=1}^{N} \nabla \tilde{J}_{\mu}(i_k,\overline{r}) \left(\tilde{J}_{\overline{\mu}}(i_k,\overline{r}) - \sum_{m=k}^{N} g\big(i_m,\overline{\mu}(i_m),i_{m+1}\big) \right), \quad (3.20)$$

where γ is a stepsize. The summation in the right-hand side above is a sample gradient corresponding to a term in the least squares summation of problem (3.19).

We finally mention two variants of the approximate policy iteration algorithm, both of which require additional approximations per policy iteration. In the first variant, instead of calculating the approximate Q-factors via Eq. (3.17), we form an approximation $\tilde{Q}_{\mu}(i,u,y)$, where the parameter vector y is determined by solving the least squares problem

$$\min_{y} \sum_{(i,u)\in\tilde{V}} \big|\tilde{Q}_{\mu}(i,u,y) - \tilde{Q}_{\mu}(i,u,r)\big|^2, \quad (3.21)$$

where $\tilde{V}$ is a "representative" set of state-control pairs (i,u), and $\tilde{Q}_{\mu}(i,u,r)$ is evaluated using Eq. (3.17) and either exact calculation or simulation. This variant is useful if it speeds up the calculations of the policy improvement step [cf. Eq. (3.18)].

In the second variant of the algorithm, we first perform the approximate policy evaluation step to obtain $\tilde{J}_{\mu}(i,r)$. Then we compute the improved policy $\overline{\mu}(i)$ by the formula (3.18) only for states i in a "representative" subset $\tilde{S}$. We then obtain an "improved" policy $\tilde{\mu}(i,v)$, which is defined over all states, by introducing a parameter vector v and by solving the least squares problem

$$\min_{v} \sum_{i\in\tilde{S}} \big\|\overline{\mu}(i) - \tilde{\mu}(i,v)\big\|^2. \quad (3.22)$$

Here, we assume that the controls are elements of some Euclidean space and $\|\cdot\|$ denotes the norm on that space. This approach accelerates the policy improvement step [cf. Eq. (3.18)] at the expense of solving an additional least squares problem per policy iteration.

Approximate Policy Iteration Using TD(1)

Just as there is a temporal differences implementation of Monte-Carlo simulation, there is also a temporal differences implementation of the gradient iteration (3.20). The temporal differences d_k are given by

$$d_k = g\big(i_k,\overline{\mu}(i_k),i_{k+1}\big) + \tilde{J}_{\overline{\mu}}(i_{k+1},\overline{r}) - \tilde{J}_{\overline{\mu}}(i_k,\overline{r}), \qquad k = 1,\ldots,N, \quad (3.23)$$

and the iteration (3.20) can be alternatively written as follows [just add the equations below using the temporal difference expression (3.23) to obtain the iteration (3.20)]:

Following the state transition (i_1,i_2), set

$$\overline{r} := \overline{r} + \gamma d_1 \nabla \tilde{J}_{\overline{\mu}}(i_1,\overline{r}). \quad (3.24)$$

Following the state transition (i_2,i_3), set

$$\overline{r} := \overline{r} + \gamma d_2 \big(\nabla \tilde{J}_{\overline{\mu}}(i_1,\overline{r}) + \nabla \tilde{J}_{\overline{\mu}}(i_2,\overline{r})\big). \quad (3.25)$$

$\ldots \qquad \ldots$

Following the state transition (i_N,t), set

$$\overline{r} := \overline{r} + \gamma d_N \big(\nabla \tilde{J}_{\overline{\mu}}(i_1,\overline{r}) + \nabla \tilde{J}_{\overline{\mu}}(i_2,\overline{r}) + \cdots + \nabla \tilde{J}_{\overline{\mu}}(i_N,\overline{r})\big). \quad (3.26)$$

The vector $\overline{r}$ may be updated at each transition, although the gradients $\nabla \tilde{J}_{\overline{\mu}}(i_k,\overline{r})$ are evaluated for the value of $\overline{r}$ that prevails at the time i_k is generated. Also, for convergence, the stepsize γ should diminish over time. A popular choice is to use during the mth trajectory $\gamma = c/m$, where c is a constant.

A variant of this method that has been proposed under the name TD(λ) uses a parameter $\lambda \in [0,1]$ in the formulas (3.23)-(3.26). It has the following form:

For $k = 1,\ldots,N$, following the state transition (i_k,i_{k+1}), set

$$\overline{r} := \overline{r} + \gamma d_k \sum_{m=1}^{k} \lambda^{k-m} \nabla \tilde{J}_{\overline{\mu}}(i_m,\overline{r}).$$

While this method has received wide attention, its validity has been questioned. Examples have been constructed [Ber95b] where the approximating function $\tilde{J}_{\overline{\mu}}(i,\overline{r})$ obtained in the limit by TD(λ) is an increasingly poor approximation to $J_{\mu}(i)$ as λ decreases towards 0, and the approximation obtained by TD(0) is very poor. It is possible, however, to use the two-sample gradient iteration (3.15) for a simulation-based, approximate evaluation of the cost functions of various policies in an approximate policy iteration scheme. This iteration resembles the TD(0) formula but aims at minimizing the error in Bellman's equation.

Optimistic Policy Iteration

In the approximate policy iteration approach discussed so far, the least squares problem that evaluates the cost of the improved policy $\overline{\mu}$ must be solved completely for the vector $\overline{r}$. An alternative is to solve this problem approximately and replace the policy μ with the policy $\overline{\mu}$ after a single or a few simulation runs. An extreme possibility is to replace μ with $\overline{\mu}$ at the end of each state transition, as in the next algorithm:

Following the state transition (i_k, i_{k+1}), set

$$\overline{r} := \overline{r} + \gamma d_k \sum_{m=1}^{k} \nabla \tilde{J}_{\overline{\mu}}(i_m, \overline{r}),$$

and generate the next transition (i_{k+1}, i_{k+2}) by simulation using the control

$$\overline{\mu}(i_{k+1}) = \arg \min_{u \in U(i)} \sum_j p_{i_{k+1}j}(u)\big(g(i_{k+1}, u, j) + \tilde{J}_{\overline{\mu}}(j, \overline{r})\big).$$

The theoretical convergence properties of this method have not been investigated so far, although its TD(λ) version has been used with success in solving some challenging problems [Tes92].

Variations Involving Multistage Lookahead

To reduce the effect of the approximation error

$$J_\mu(i) - \tilde{J}_\mu(i, r)$$

between the true and approximate costs of a policy μ, one can consider a lookahead of several stages in computing the improved policy $\overline{\mu}$. The method adopted earlier for generating the decisions $\overline{\mu}(i)$ of the improved policy,

$$\overline{\mu}(i) = \arg \min_{u \in U(i)} \sum_j p_{ij}(u)\big(g(i, u, j) + \tilde{J}_\mu(j, r)\big), \qquad \text{for all } i,$$

corresponds to a single stage lookahead. At a given state i, it finds the optimal decision for a one-stage problem with stage cost $g(i, u, j)$ and terminal cost (after the first stage) $\tilde{J}_\mu(j, r)$.

An m-stage lookahead version finds the optimal policy for an m-stage problem, whereby we start at the current state i, make the m subsequent decisions with perfect state information, incur the corresponding costs of the m stages, and pay a terminal cost $\tilde{J}_\mu(j, r)$, where j is the state after m stages. This is a finite horizon stochastic optimal control problem that may be tractable, depending on the horizon m and the number of possible

successor states from each state. If $u_1(i)$ is the first decision of the m-stage lookahead optimal policy starting at state i, the improved policy is defined by

$$\overline{\mu}(i) = u_1(i).$$

Note that if $\tilde{J}_\mu(j, r)$ is equal to the exact cost $J_\mu(i)$ for all states i, that is, there is no approximation, the multistage version of policy iteration can be shown to terminate with an optimal policy under the same conditions as ordinary policy iteration (see Exercise 2.16).

Multistage lookahead can also be used in the real-time calculation of a suboptimal control policy, once an approximation $\tilde{J}(i, r)$ of the optimal cost has been obtained by any one of the methods of this subsection. An example is the computer chess programs discussed in Section 6.3 of Vol. I. In that case, the approximation of the cost-to-go function (the position evaluator discussed in Section 6.3 of Vol. I) is relatively primitive. It is derived from the features of the position (material balance, piece mobility, king safety, etc.), appropriately weighted with factors that are either heuristically determined by trial and error, or (in the case of a champion program, IBM's Deep Thought) by training on examples from grandmaster play. It is well-known that the quality of play of computer chess programs crucially depends on the size of the lookahead. This indicates that in many types of problems, the multistage lookahead versions of the methods of this subsection should be much more effective than their single stage lookahead counterparts. This improvement in performance must of course be weighted against the considerable increase in computation required to optimally solve the associated multistage problems.

Approximation in Policy Space

We finally mention a conceptually different approximation possibility that aims at direct optimization over policies of a given type. Here we hypothesize a stationary policy of a certain structural form, say $\tilde{\mu}(i, r)$, where r is a vector of unknown parameters/weights that is subject to optimization. We also assume that for a fixed r, the cost of starting at i and using the stationary policy $\tilde{\mu}(\cdot, r)$, call it $\tilde{J}(i, r)$, can be evaluated by simulation. We may then minimize over r

$$E_i\big\{\tilde{J}(i, r)\big\}, \tag{3.27}$$

where the expectation is taken with respect to some probability distribution over the set of initial states. This minimization will typically be carried out by some method that does not require the use of gradients if the gradient of $\tilde{J}(i, r)$ with respect to r cannot be easily calculated. If the simulation can produce the value of the gradient $\nabla \tilde{J}(i, r)$ together with $\tilde{J}(i, r)$, then a gradient-based method can be used. Generally, the minimization of the cost

function (3.27) tends to be quite difficult if the dimension of the parameter vector r is large (say over 10). As a result, the method is most likely effective only when adequate optimal policy approximation is possible with very few parameters.

2.3.4 Extension to Discounted Problems

We now discuss adaptations of the simulation-based methods for the case of a discounted problem. Consider first the evaluation of policies by simulation. One difficulty here is that trajectories do not terminate, so we cannot obtain sample costs corresponding to different states. One way to get around this difficulty is to approximate a discounted cost by a finite horizon cost of sufficiently large horizon. Another possibility is to convert the α-discounted problem to an equivalent stochastic shortest path problem by introducing an artificial termination state t and a transition probability $1-\alpha$ from each state $i \neq t$ to the termination state t. The remaining transition probabilities are scaled by multiplication with α (see Vol. I, Section 7.3). Bellman's equation for this stochastic shortest path problem is identical with Bellman's equation for the original α-discounted problem, so the optimal cost functions and the optimal policies of the two problems are identical.

The preceding approaches may lead to long simulation runs, involving many transitions. An alternative possibility that is useful in some cases is based on identifying a special state, called the *reference state*, that is assumed to be reachable from all other states under the given policy. Suppose that such a state can be identified and for concreteness assume that it is state 1. Thus, we assume that the Markov chain corresponding to the given policy has a single recurrent class and state 1 belongs to that class (see Appendix D of Vol. I). If there are multiple recurrent classes, the procedure described in what follows can be modified so that there is a reference state for each class.

To simplify notation, we do not show the dependence of various quantities on the given policy. In particular, the transition probability from i to j and the corresponding stage cost are denoted by p_{ij} and $g(i,j)$, in place of $p_{ij}\big(\mu(i)\big)$ and $g\big(i,\mu(i),j\big)$, respectively. For each initial state i, let $C(i)$ denote the average discounted cost incurred up to reaching the reference state 1. Let also m_i denote the *first passage time* from state i to state 1, that is, the number of transitions required to reach state 1 starting from state i. Note that m_i is a random variable. We denote

$$D(i) = E\{\alpha^{m_i}\}.$$

By dividing the cost $J_\mu(i)$ into the portion up to reaching state 1 and the remaining portion starting from state 1, we have

$$\begin{aligned} J_\mu(i) &= E\left\{\sum_{k=0}^{m_i-1} \alpha^k g(x_k,x_{k+1}) \,|\, x_0 = i\right\} \\ &\quad + E\left\{\sum_{k=m_i}^{\infty} \alpha^k g(x_k,x_{k+1}) \,|\, x_{m_i} = 1\right\} \\ &= C(i) + D(i)J_\mu(1). \end{aligned} \tag{3.28}$$

Applying this equation for $i=1$, we have $J_\mu(1) = C(1) + D(1)J_\mu(1)$, so that

$$J_\mu(1) = \frac{C(1)}{1-D(1)}. \tag{3.29}$$

Combining Eqs. (3.28) and (3.29), we obtain

$$J_\mu(i) = C(i) + \frac{D(i)C(1)}{1-D(1)}, \qquad i = 1,\ldots,n.$$

Therefore, to calculate the cost vector J_μ, it is sufficient to calculate the costs $C(i)$, and in addition to calculate the expected discount terms $D(i)$. Both of these can be computed, similar to the stochastic shortest path problem, by generating many sample system trajectories, and averaging the corresponding sample costs and discount terms up to reaching the reference state 1.

Note here that because $C(1)$ and $D(1)$ crucially affect the calculated values $J_\mu(i)$, it may be worth doing some extra simulations starting from the reference state 1 to ensure that $C(1)$ and $D(1)$ are accurately calculated.

Once a simulation method is available to evaluate (perhaps approximately) the cost of various policies, it can be embedded within a (perhaps approximate) policy iteration algorithm along the lines discussed for the stochastic shortest path problem.

We note also that there is a straightforward extension of the Q-learning algorithm to discounted problems. The optimal Q-factors are the unique solution of the equation

$$\begin{aligned} Q(i,u) &= \sum_{j=1}^{n} p_{ij}(u)\big(g(i,u,j) + \alpha J^*(j)\big) \\ &= \sum_{j=1}^{n} p_{ij}(u)\left(g(i,u,j) + \alpha \min_{u' \in U(j)} Q(j,u')\right). \end{aligned}$$

This is again proved by introducing a system whose states are pairs (i,u), so that the above system of equations becomes a special case of Bellman's

equation. With similar observations, it follows that the vector of Q-factors can be obtained by the value iteration

$$Q(i,u) := \sum_{j=1}^{n} p_{ij}(u)\left(g(i,u,j) + \alpha \min_{u' \in U(j)} Q(j,u')\right).$$

The Q-learning method is an approximate version of this iteration, whereby the expected value is replaced by a single sample, i.e.,

$$Q(i,u) := Q(i,u) + \gamma\left(g(i,u,j) + \alpha \min_{u' \in U(j)} Q(j,u') - Q(i,u)\right).$$

Here j and $g(i,u,j)$ are generated from the pair (i,u) by simulation, that is, according to the transition probabilities $p_{ij}(u)$.

We finally note that approximation based on minimization of the error in Bellman's equation can also be used in the case of a discounted cost. One simply needs to introduce the discount factor at the appropriate places in the various iterations given above. For example, the variant of iteration (3.15) for evaluating the discounted cost of a policy μ is

$$r := r - \gamma d_\mu(i,j,r)\big(\alpha \nabla \tilde{J}_\mu(j,r) - \nabla \tilde{J}_\mu(i,r)\big),$$

where α is the discount factor and

$$d_\mu(i,j,r) = g\big(i,\mu(i),j\big) + \alpha \tilde{J}_\mu(j,r) - \tilde{J}_\mu(i,r).$$

2.3.5 The Role of Parallel Computation

It is well-known that Monte-Carlo simulation is very well-suited for parallelization; one can simply carry out multiple simulation runs in parallel and occasionally merge the results. Also several DP-related methods are well-suited for parallelization; for example, each value iteration can be parallelized by executing the cost updates of different states in different parallel processors (see e.g., [AMT93]). In fact the parallel updates can be asynchronous. By this we mean that different processors may execute cost updates as fast as they can, without waiting to acquire the most recent updates from other processors; these latter updates may be late in coming because some of the other processors may be slow or because some of the communication channels connecting the processors may be slow. Asynchronous parallel value iteration can be shown to have the same convergence properties as its synchronous counterpart, and is often substantially faster. We refer to [Ber82a] and [BeT89a] for an extensive discussion.

There are similar parallelization possibilities in approximate DP. Indeed, approximate policy iteration may be viewed as a combination of two operations:

(a) *Simulation*, which produces many pairs $\big(i, c(i)\big)$ of states i and sample costs $c(i)$ associated with the improved policy $\overline{\mu}$.

(b) *Training*, which obtains the state-sample cost pairs produced by the simulator and uses them in the least-squares optimization of the parameter vector $\overline{r}$ of the approximate cost function $\tilde{J}_{\overline{\mu}}(\cdot,\overline{r})$.

The simulation operation can be parallelized in the usual way by executing multiple independent simulations in multiple processors. The training operation can also be parallelized to a great extent. For example, one may parallelize the gradient iteration

$$\overline{r} := \overline{r} - \gamma \sum_{k=1}^{N} \nabla \tilde{J}_{\overline{\mu}}(i_k,\overline{r})\big(\tilde{J}_{\overline{\mu}}(i_k,\overline{r}) - c(i_k)\big),$$

that is used for training [cf. Eq. (3.20)]. There are two possibilities here:

(1) To assign different components of $\overline{r}$ to different processors and to execute the component updates in parallel.

(2) To parallelize the computation of the sample gradient

$$\sum_{k=1}^{N} \nabla \tilde{J}_{\overline{\mu}}(i_k,\overline{r})\big(\tilde{J}_{\overline{\mu}}(i_k,\overline{r}) - c(i_k)\big)$$

in the gradient iteration, by assigning different blocks of state-sample cost pairs to different processors.

There are several straightforward versions of these parallelization methods, and it is also valid to use asynchronous versions of them ([BeT89a], Ch. 7).

There is still another parallelization approach for the training process. It is possible to divide the state space S into several subsets S_m, $m = 1,\ldots,M$, and to calculate a different approximation $\tilde{J}_{\overline{\mu}}(i,\overline{r}_m)$ for each subset S_m. In other words, the parameter vector $\overline{r}_m$ that is used to calculate the approximate cost $\tilde{J}_{\overline{\mu}}(i,\overline{r}_m)$ depends on the subset S_m to which state i belongs. The parameters $\overline{r}_m$ can be obtained by a parallel training process using the applicable simulation data, that is, the state-sample cost pairs $\big(i,c(i)\big)$ with $i \in S_m$. Note that the extreme case where each set S_m corresponds to a single state, corresponds to the case where there is no approximation.

2.4 NOTES, SOURCES, AND EXERCISES

The analysis of the stochastic shortest path problems of Section 2.1 is taken from [BeT89a] and [BeT91b]. The latter reference proves the results

shown here under a more general compactness assumption on $U(i)$ and continuity assumption on $g(i,u)$ and $p_{ij}(u)$. Stochastic shortest path problems were first formulated and analyzed in [EaZ62] under the assumption $q(i,u) > 0$ for all $i = 1,\ldots,n$ and $u \in U(i)$. Finitely terminating value iteration algorithms have been developed for several types of stochastic shortest path problems (see [NgP88], [PoT92], [PsT93], [Tsi93a]). The use of a Dijkstra-like algorithm for continuous space shortest path problems involving a consistently improving policy was proposed in [Tsi93a] (see Exercise 2.10). A Dijkstra-like algorithm was also proposed for another class of problems involving a consistently improving policy in [NgP88]. The algorithm of Exercise 2.11 is new in the general form given here. The error bound on the performance of approximate policy iteration (Prop. 2.1), which was developed in collaboration with J. Tsitsiklis, is also new. Two-player dynamic game versions of the stochastic shortest path problem have been discussed in [PoA69] (see also the survey [RaF91]).

Several approximation methods that are not based on simulation were given in [ScS85]. The interest in simulation-based methods is relatively recent. In the artificial intelligence community, these methods are collectively referred to as *reinforcement learning*. In the engineering community, these methods are also referred to as *neuro-dynamic programming*. The method of temporal differences was proposed in an influential paper by Sutton [Sut88]. Q-learning was proposed by Watkins [Wat89]. A convergence proof of Q-learning under fairly weak assumptions was given in [Tsi94]; see also [JJS93], which discusses the convergence of TD(λ). For a nice survey of related methods, which also includes historical references, see [BBS93]. A variant of Q-learning is the method of advantage updating developed in [Bai93], [Bai94], [Bai95], and [HBK94] (see Exercise 2.18). The material on feature-based aggregation has been adapted from [TsV94]. The two-sample simulation-based gradient method for minimizing the error in Bellman's equation was proposed in [Ber95b]; see also [HBK94]. The optimistic policy iteration method was used in an application to backgammon described in [Tes92].

EXERCISES

2.1

Suppose that you want to travel from a start point S to a destination point D in minimum average time. There are two options:

(1) Use a direct route that requires a time units.

(2) Take a potential shortcut that requires b time units to go to an intermediate point I. From I you can either go to the destination D in c time units or return to the start (this will take an additional b time units). You will find out the value of c once you reach the intermediate point I. What you know a priori is that c has one of the m values $c_1,\ldots,c_m$ with corresponding probabilities $p_1,\ldots,p_m$. Consider two cases: (i) The value of c is constant over time, and (ii) The value of c changes each time you return to the start independently of the value at the previous time periods.

(a) Formulate the problem as a stochastic shortest path problem. Write Bellman's equation and characterize the optimal stationary policies as best as you can in terms of the given problem data. Solve the problem for the case $a = 2$, $b = 1$, $c_1 = 0$, $c_2 = 5$, $p_1 = 0.5$, $p_2 = 0.5$.

(b) Formulate as a stochastic shortest path problem the variation where once you reach the intermediate point I, you can wait there. Each d time units the value of c changes to one of the values $c_1,\ldots,c_m$ with probabilities $p_1,\ldots,p_m$, independently of its earlier values. Each time the value of c changes, you have the option of waiting for an extra d units, returning to the start, or going to the destination. Characterize the optimal stationary policies as best as you can.

2.2

A gambler engages in a game of successive coin flipping over an infinite horizon. He wins one dollar each time heads comes up, and loses $m > 0$ dollars each time two successive tails come up (so the sequence TTTT loses $3m$ dollars). The gambler at each time period either flips a fair coin or else cheats by flipping a two-headed coin. In the latter case, however, he gets caught with probability $p > 0$ before he flips the coin, the game terminates, and the gambler keeps his earnings thus far. The gambler wishes to maximize his expected earnings.

(a) View this as a stochastic shortest path problem and identify all proper and all improper policies.

(b) Identify a critical value $\overline{m}$ such that if $m > \overline{m}$, then all improper policies give an infinite cost for some initial state.

(c) Assume that $m > \overline{m}$, and show that it is then optimal to try to cheat if the last flip was tails and to play fair otherwise.

(d) Show that if $m < \overline{m}$ it is optimal to always play fair.

2.3

Consider a stochastic shortest path problem where all stationary policies are proper. Show that for every policy π there exists an $m > 0$ such that

$$P(x_m = t \mid x_0 = i, \pi) > 0$$

for all $i = 1, \ldots, n$. *Abbreviated Proof*: Assume the contrary; that is, there exists a nonstationary $\pi = \{\mu_0, \mu_1, \ldots\}$ and an initial state i such that $P(x_m = t \mid x_0 = i, \pi) = 0$ for all m. For each state j, let $m(j)$ be the minimum integer m such that state j is reachable from i with positive probability under policy π; that is,

$$m(j) = \min\bigl\{m \mid P(x_m = j \mid x_0 = i, \pi) > 0\bigr\},$$

where we adopt the convention that $m(j) = \infty$ if j is not reachable from i under π, i.e., $P(x_m = j \mid x_0 = i, \pi) = 0$ for all m. In particular, we have $m(i) = 0$ and $m(t) = \infty$. Consider any stationary policy μ such that $\mu(j) = \mu_{m(j)}(j)$ for all j with $m(j) < \infty$. Argue that for any two states j and j' with $m(j) < \infty$ and $m(j') = \infty$, we have $p_{jj'}\bigl(\mu(j)\bigr) = 0$. Thus, states j' with $m(j') = \infty$ (including t) are not reachable under the stationary policy μ from states j with $m(j) < \infty$ (including i), thereby contradicting the hypothesis.

2.4

Consider the stochastic shortest path problem, and assume that $g(i, u) \leq 0$ for all i and $u \in U(i)$. Show that either the optimal cost is $-\infty$ for some initial state, or else, under every policy, the system eventually enters with probability one a set of cost-free states and never leaves that set thereafter.

2.5

Consider the stochastic shortest path problem, and assume that there exists at least one proper policy. Proposition 1.2 implies that if, for each improper policy μ, we have $J_\mu(i) = \infty$ for at least one state i, then there is no improper policy μ' such that $J_{\mu'}(j) = -\infty$ for at least one state j. Give an alternative proof of this fact that does not use Prop. 1.2. *Hint*: Suppose that there exists an improper policy μ' such that $J_{\mu'}(j) = -\infty$ for at least one state j. Combine this policy with a proper policy to produce another improper policy μ'' for which $J_{\mu''}(i) < \infty$ for all i.

2.6 (Gauss-Seidel Method for Stochastic Shortest Paths)

Show that the Gauss-Seidel version of the value iteration method for stochastic shortest paths converges under the same assumptions as the ordinary method (Assumptions 1.1 and 1.2). *Hint*: Consider two functions $\underline{J}$ and $\overline{J}$ that differ by a constant from J^* at all states except the destination, and are such that $\underline{J} \leq T\underline{J}$ and $T\overline{J} \leq \overline{J}$.

2.7 (Sequential Space Decomposition)

Consider the stochastic shortest path problem, and suppose that there is a finite sequence of subsets of states $S_1, S_2, \ldots, S_M$ such that each of the states $i = 1, \ldots, n$ belongs to one and only one of these subsets, and the following property holds:

> For all $m = 1, \ldots, M$ and states $i \in S_m$, the successor state j is either the termination state t or else belongs to one of the subsets $S_m, S_{m-1}, \ldots, S_1$ for all choices of the control $u \in U(i)$.

(a) Show that the solution of this problem decomposes into the solution of M stochastic shortest path problems, each involving the states in a subset S_m plus a termination state.

(b) Show also that a finite horizon problem with N stages can be viewed as a stochastic shortest path problem with the property given above.

2.8

Consider a stochastic shortest path problem under Assumptions 1.1 and 1.2. Assuming $p_{it}(u) < 1$ for all $i \neq t$ and $u \in U(i)$, consider another stochastic shortest path problem that has transition probabilities $p_{ij}(u)/\bigl(1 - p_{it}(u)\bigr)$ for all $i \neq t$ and $j \neq i$, and costs

$$\tilde{g}(i, u) = g(i, u) + \frac{g(i, u)p_{it}(u)}{1 - p_{it}(u)}.$$

(a) Show that the two problems are equivalent in that they have the same optimal costs and policies. How would you deal with the case where $p_{it}(u) = 1$ for some $i \neq t$ and $u \in U(i)$?

(b) Interpret $\tilde{g}(i, u)$ as an average cost incurred between arrival to state i and transition to a state $j \neq i$.

2.9 (Simplifications for Uncontrollable State Components)

Consider a stochastic shortest path problem under Assumptions 1.1 and 1.2, where the state is a composite (i, y) of two components i and y, and the evolution of the main component i can be directly affected by the control u, but the evolution of the other component y cannot (cf. Section 1.4 and Exercise 1.22 of Vol. I). In particular, we assume that given the state (i, y) and the control u, the next state (j, z) is determined as follows: first j is generated according to transition probabilities $p_{ij}(u, y)$, and then z is generated according to conditional probabilities $p(z \mid j)$ that depend on the main component j of the new state. We also assume that the cost per stage is $g(i, y, u, j)$ and does not depend on the second component z of the next state (j, z). For functions $\hat{J}(i)$, $i = 1, \ldots, n$, consider the mapping

$$(\hat{T}\hat{J})(i) = \sum_y p(y \mid i) \left(\min_{u \in U(i,y)} \sum_{j=1}^{n} p_{ij}(u, y)\bigl(g(i, y, u, j) + \hat{J}(j)\bigr) \right),$$

and the corresponding mapping of a stationary policy μ,

$$(\hat{T}_\mu \hat{J})(i) = \sum_y p(y \mid i) \sum_{j=1}^{n} p_{ij}\big(\mu(i,y), y\big)\Big(g\big(i, y, \mu(i,y), j\big) + \hat{J}(j)\Big).$$

(a) Show that $\hat{J} = \hat{T}\hat{J}$ is a form of Bellman's equation and can be used to characterize the optimal stationary policies. *Hint.* Given $J(i,y)$, define $\hat{J}(i) = \sum_y p(y \mid i) J(i,y)$.

(b) Show the validity of a modified value iteration algorithm that starts with an arbitrary function $\hat{J}$ and sequentially produces $\hat{T}\hat{J}$, $\hat{T}^2\hat{J}$, ...

(c) Show the validity of a modified policy iteration algorithm whose typical iteration, given the current policy $\mu^k(i,y)$, consists of two steps: (1) The policy evaluation step, which computes the unique function $\hat{J}_{\mu^k}$ that solves the linear system of equations $\hat{J}_{\mu^k} = \hat{T}_{\mu^k}\hat{J}_{\mu^k}$. (2) The policy improvement step, which computes the improved policy $\mu^{k+1}(i,y)$ from the equation $\hat{T}_{\mu^{k+1}}\hat{J}_{\mu^k} = \hat{T}\hat{J}_{\mu^k}$.

(d) Suppose that y is the only source of randomness in the problem; that is, for each (i,y,u), there is a state j such that $p_{ij}(u,y) = 1$. Justify the use of the following single sample version of value iteration (cf. the Q-learning algorithm of Section 7.6.2)

$$\hat{J}(i) := \hat{J}(i) + \gamma\left(\min_{u \in U(i,y)}\big[g(i,y,u,j) + \hat{J}(j)\big] - \hat{J}(i)\right).$$

Here, given i, we generate y according to the probability distribution $p(y \mid i)$, and j is the unique state corresponding to (i,y,u).

2.10 (Discretized Shortest Path Problems [Tsi93a])

Suppose that the states are the grid points of a grid on the plane. The set of neighbors of each grid point x is denoted $U(x)$ and includes between two and four grid points. At each grid point x, we have two options:

(1) Choose two neighbors $x^+, x^- \in U(x)$ and a probability $p \in [0,1]$, pay a cost $g(x)\sqrt{p^2 + (1-p)^2}$, and move to x^+ or to x^- with probability p or $1-p$, respectively. Here g is a function such that $g(x) > 0$ for all x.

(2) Stop and pay a cost $t(x)$.

Show that there exists a consistently improving optimal policy for this problem. *Note*: This problem can be used to model discretized versions of deterministic continuous space 2-dimensional shortest path problems. (Compare also with Exercise 6.11 in Chapter 6 of Vol. I.)

2.11 (Dijkstra's Algorithm and Consistently Improving Policies)

Consider the stochastic shortest path problem under Assumptions 1.1 and 1.2, and assume that there exists a consistently improving optimal stationary policy.

(a) Show that the transition probability graph of this policy is acyclic.

(b) Consider the following algorithm, which maintains two subsets of states P and L, and a function J defined on the state space. (To relate the algorithm with Dijkstra's method of Section 2.3.1 of Vol. I, associate J with the node labels, L with the OPEN list, and P with the subset of nodes that have already exited the OPEN list.) Initially, $P = \emptyset$, $L = \{t\}$, and

$$J(i) = \begin{cases} \infty & \text{if } i = 1, \ldots, n, \\ 0 & \text{if } i = t. \end{cases}$$

At the typical iteration, select a state j^* from L such that

$$j^* = \arg\min_{j \in L} J(j).$$

(If L is empty the algorithm terminates.) Remove j^* from L and place it in P. In addition, for all $i \notin P$ such that there exists a $u \in U(i)$ with $p_{ij^*}(u) > 0$, and

$$p_{ij}(u) = 0 \qquad \text{for all } j \notin P,$$

define

$$\tilde{U}(i) = \big\{u \in U(i) \mid p_{ij^*}(u) > 0 \text{ and } p_{ij}(u) = 0 \text{ for all } j \notin P\big\},$$

set

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

and place i in L if it is not already there. Show that the algorithm is well defined in the sense that $\tilde{U}(i)$ is nonempty and the set L does not become empty until all states are in P. Furthermore, each state j is removed from L once, and at the time it is removed, we have $J(j) = J^*(j)$.

2.12 (Alternative Assumptions for Prop. 1.2)

Consider a variation of Assumption 1.2, whereby we assume that $g(i,u) \geq 0$ for all i and $u \in U(i)$, and that there exists an optimal proper policy. Prove the assertions of Prop. 1.2, except that, in part (a), uniqueness of the solution of Bellman's equation should be shown within the set $\Re^+ = \{J \mid J \geq 0\}$ (rather than within $\Re^n$), and the vector J in part (b) must belong to $\Re^+$.

Hint: Proposition 1.1 is not valid, so a somewhat different proof is needed. Complete the details of the following argument. The assumptions guarantee that J^* is finite and $J^* \in \Re^+$. [We have $J^* \geq 0$ because $g(i,u) \geq 0$, and $J^*(i) < \infty$ because a proper policy exists.] The idea now is to show that $J^* \geq TJ^*$, and then to choose μ such that $T_\mu J^* = TJ^*$ and show that μ is optimal and proper. Let $\pi = \{\mu_0, \mu_1, \ldots\}$ be a policy. We have for all i,

$$J_\pi(i) = g\big(i, \mu_0(i)\big) + \sum_{j=1}^{n} p_{ij}\big(\mu_0(i)\big) J_{\pi_1}(j)$$

where π_1 is the policy $\{\mu_1, \mu_2, \ldots\}$. Since $J_{\pi_1} \geq J^*$, we obtain

$$J_\pi(i) \geq g\big(i, \mu_0(i)\big) + \sum_{j=1}^{n} p_{ij}\big(\mu_0(i)\big) J^*(j) = (T_{\mu_0} J^*)(i) \geq (TJ^*)(i).$$

Taking the infimum over π in the preceding equation, we obtain

$$J^* \geq TJ^*. \tag{4.1}$$

Let μ be such that $T_\mu J^* = TJ^*$. From Eq. (4.1), we have $J^* \geq T_\mu J^*$, and using the monotonicity of T_μ, we obtain

$$J^* \geq T_\mu^N J^* = P_\mu^N J^* + \sum_{k=0}^{N-1} P_\mu^k g_\mu \geq \sum_{k=0}^{N-1} P_\mu^k g_\mu, \qquad N \geq 1. \tag{4.2}$$

By taking limit superior as $N \to \infty$, we obtain $J^* \geq J_\mu$. Therefore, μ is an optimal proper policy, and $J^* = J_\mu$. Since μ was selected so that $T_\mu J^* = TJ^*$, we obtain, using $J^* = J_\mu$ and $J_\mu = T_\mu J_\mu$, that $J^* = TJ^*$. For the rest of the proof, use the vector δe similar to the proof of Prop. 1.2.

2.13 (A Contraction Counterexample)

Consider a stochastic shortest path problem with a single state 1, in addition to the termination state t. At state 1 there are two controls u and u'. Under u the cost is 1 and the system remains in state 1 for one more stage; under u' the cost is 2 and the system moves to t. Show that Assumptions 1.1 and 1.2 are satisfied, but T is not a contraction mapping with respect to any norm.

2.14 (Contraction Property – All Stationary Policies are Proper)

Assume that all stationary policies are proper. Show that the mappings T and T_μ are contraction mappings with respect to some weighted sup norm

$$\|J\|_v = \max_{i=1,\ldots,n} \frac{1}{v_i} |J(i)|,$$

where v is a vector whose components $v_1, \ldots, v_n$ are positive.

Abbreviated proof (from [BeT89a], p. 325; see also [Tse90]): Partition the state space as follows. Let $S_1 = \{1\}$ and for $k = 2, 3, \ldots$, define sequentially

$$S_k = \left\{ i \;\middle|\; i \notin S_1 \cup \cdots \cup S_{k-1} \text{ and } \min_{u \in U(i)} \max_{j \in S_1 \cup \cdots \cup S_{k-1}} p_{ij}(u) > 0 \right\}.$$

Let S_m be the last of these sets that is nonempty. We claim that the sets S_k cover the entire state space, that is, $\cup_{k=1}^m S_k = S$. To see this, suppose that the set $S_\infty = \left\{ i \mid i \notin \cup_{k=1}^m S_k \right\}$ is nonempty. Then for each $i \in S_\infty$, there exists some $u_i \in U(i)$ such that $p_{ij}(u_i) = 0$ for all $j \notin S_\infty$. Take any μ such that $\mu(i) = u_i$ for all $i \in S_\infty$. The stationary policy μ satisfies $[P_\mu^N]_{ij} = 0$ for all $i \in S_\infty$, $j \notin S_\infty$, and N, and therefore cannot be proper. This contradicts the hypothesis.

We will choose a vector $v > 0$ so that T is a contraction mapping with respect to $\|\cdot\|_v$. We will take the ith component v_i to be the same for states i in the same set S_k. In particular, we will choose the components v_i of the vector v by

$$v_i = y_k \qquad \text{if} \qquad i \in S_k,$$

where $y_1, \ldots, y_m$ are appropriately chosen scalars satisfying

$$1 = y_1 < y_2 < \cdots < y_m. \tag{4.3}$$

Let

$$c = \min_{k=2,\ldots,m} \min_{\mu \in M} \min_{i \in S_k} \sum_{j \in S_1 \cup \cdots \cup S_{k-1}} [P_\mu]_{ij}, \tag{4.4}$$

and note that $0 < c \leq 1$. We will show that it is sufficient to choose $y_2, \ldots, y_m$ so that for some $\gamma < 1$, we have

$$\frac{y_m}{y_k}(1-c) + \frac{y_{k-1}}{y_k} c \leq \gamma < 1, \qquad k = 2, \ldots, m, \tag{4.5}$$

and then show that such a choice of $y_2, \ldots, y_m$ exists.

Indeed, for vectors J and J' in $\Re^n$, let μ be such that $T_\mu J = TJ$. Then we have for all i,

$$\begin{aligned} (TJ')(i) - (TJ)(i) &= (TJ')(i) - (T_\mu J)(i) \\ &\leq (T_\mu J')(i) - (T_\mu J)(i) \\ &= \sum_{j=1}^{n} p_{ij}\big(\mu(i)\big)\big(J'(j) - J(j)\big). \end{aligned} \tag{4.6}$$

Let $k(j)$ be such that j belongs to the set $S_{k(j)}$. Then we have for any constant c,

$$\|J' - J\|_v \leq c \qquad \Rightarrow \qquad J'(j) - J(j) \leq c y_{k(j)}, \qquad j = 2, \ldots, n,$$

and Eq. (4.6) implies that for all i,

$$\begin{aligned}\frac{(TJ')(i)-(TJ)(i)}{cy_{k(i)}} &\le \frac{1}{y_{k(i)}}\sum_{j=1}^{n} p_{ij}\big(\mu(i)\big)y_{k(j)}\\ &\le \frac{y_{k(i)-1}}{y_{k(i)}}\sum_{j\in S_1\cup\cdots\cup S_{k(i)-1}} p_{ij}\big(\mu(i)\big)\\ &\quad+\frac{y_m}{y_{k(i)}}\sum_{j\in S_{k(i)}\cup\cdots\cup S_m} p_{ij}\big(\mu(i)\big)\\ &=\left(\frac{y_{k(i)-1}}{y_{k(i)}}-\frac{y_m}{y_{k(i)}}\right)\sum_{j\in S_1\cup\cdots\cup S_{k(i)-1}} p_{ij}\big(\mu(i)\big)+\frac{y_m}{y_{k(i)}}\\ &\le \left(\frac{y_{k(i)-1}}{y_{k(i)}}-\frac{y_m}{y_{k(i)}}\right)c+\frac{y_m}{y_{k(i)}}\le\gamma,\end{aligned}$$

where the second inequality follows from Eq. (4.3), the third inequality uses Eq. (4.4) and the fact $y_{k(i)-1}-y_m\le 0$, and the last inequality follows from Eq. (4.5). Thus, we have

$$\frac{(TJ')(i)-(TJ)(i)}{v_i}\le c\gamma,\qquad i=1,\ldots,n,$$

and we obtain

$$\max_i \frac{(TJ')(i)-(TJ)(i)}{v_i}\le c\gamma,$$

or

$$\|TJ-TJ'\|_v\le c\gamma,\qquad \text{for all } J,J'\in\Re^n \text{ with } \|J-J'\|_v\le c.$$

It follows that T is a contraction mapping with respect to $\|\cdot\|_v$.

We now show how to choose the scalars $y_1,y_2,\ldots,y_m$ so that Eqs. (4.3) and (4.5) hold. Let $y_0=0$, $y_1=1$, and suppose that $y_1,y_2,\ldots,y_k$ have been chosen. If $c=1$, we choose $y_{k+1}=y_k+1$. If $c<1$, we choose y_{k+1} to be

$$y_{k+1}=\frac{1}{2}(y_k+M_k),$$

where

$$M_k=\min_{1\le i\le k}\left[y_i+\frac{c}{1-c}(y_i-y_{i-1})\right].$$

Using the fact

$$M_{k+1}=\min\left\{M_k,\ y_{k+1}+\frac{c}{1-c}(y_{k+1}-y_k)\right\},$$

it is seen by induction that for all k,

$$y_k<y_{k+1}<M_{k+1}.$$

In particular, we have

$$y_m<M_m=\min_{1\le i\le m}\left[y_i+\frac{c}{1-c}(y_i-y_{i-1})\right],$$

which implies Eq. (4.5).

2.15 (Multiple State Visits in Monte Carlo Simulation)

Argue that the Monte-Carlo simulation formula

$$J_\mu(i)=\lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M} c(i,m)$$

[cf. Eq. (3.1)] is valid even if a state may be revisited within the same sample trajectory. *Hint*: Suppose the M cost samples are generated from N trajectories, and that the kth trajectory involves n_k visits to state i and generates n_k corresponding cost samples. Denote $m_k=n_1+\cdots+n_k$. Write

$$\begin{aligned}\lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M} c(i,m)&=\lim_{N\to\infty}\frac{\frac{1}{N}\sum_{k=1}^{N}\sum_{m=m_{k-1}+1}^{m_k} c(i,m)}{\frac{1}{N}(n_1+\cdots+n_N)}\\ &=\frac{E\left\{\sum_{m=m_{k-1}+1}^{m_k} c(i,m)\right\}}{E\{n_k\}},\end{aligned}$$

and argue that

$$E\left\{\sum_{m=m_{k-1}+1}^{m_k} c(i,m)\right\}=E\{n_k\}J_\mu(i),$$

(or see [Ros83b], Cor. 7.2.3 for a closely related result).

2.16 (Multistage Lookahead Policy Iteration)

(a) Consider the stochastic shortest path problem under Assumptions 1.1 and 1.2. Let μ be a stationary policy, let J be a function such that $TJ\le J\le J_\mu$ ($J=J_\mu$ is one possibility), and let $\{\overline{\mu}_0,\overline{\mu}_1,\ldots,\overline{\mu}_{N-1}\}$ be an optimal policy for the N-stage problem with terminal cost function J, i.e.

$$T_{\overline{\mu}_k}T^{N-k-1}J=T^{N-k}J,\qquad k=0,1,\ldots,N-1.$$

(a) Show that

$$J_{\overline{\mu}_k}\le J_\mu,\qquad \text{for all } k=0,1,\ldots,N-1.$$

Hint: First show that $T^{k+1}J\le T^kJ\le J$ for all k, and then show that the hypothesis $T_{\overline{\mu}_k}T^{N-k-1}J=T^{N-k}J$ implies that $J_{\overline{\mu}_k}\le T^{N-k-1}J$.

(b) Use part (a) to show the validity of the multistage policy iteration algorithm discussed in Section 2.3.3.

2.17 (Viewing Q-Factors as Optimal Costs)

Consider the stochastic shortest path problem under Assumptions 1.1 and 1.2. Show that the Q-factors $Q(i,u)$ can be viewed as state costs associated with a modified stochastic shortest path problem. Use this fact to show that the Q-factors $Q(i,u)$ are the unique solution of the system of equations

$$Q(i,u) = \sum_{j=1}^{n} p_{ij}(u) \left(g(i,u,j) + \min_{u' \in U(j)} Q(j,u') \right).$$

Hint: Introduce a new state for each pair (i,u), with transition probabilities $p_{ij}(u)$ to the states $j = 1, \ldots, n, t$.

2.18 (Advantage Updating)

Consider the optimal Q-factors $Q^*(i,u)$ of the stochastic shortest path problem under Assumptions 1.1 and 1.2. Define the *advantage function* by

$$A^*(i,u) = \min_{u' \in U(i)} Q^*(i,u') - Q^*(i,u).$$

(a) Show that $A^*(i,u)$ together with the optimal costs $J^*(i)$ solve uniquely the system of equations

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

$$\max_{u \in U(i)} A^*(i,u) = 0, \qquad i = 1, \ldots, n.$$

(b) Introduce approximating functions $\tilde{A}(i,u,r)$ and $\tilde{J}(i,r)$, and derive a gradient method aimed at minimizing the sum of the squared errors of the Bellman-like equations of part (a) (cf. Section 2.3.3).

3

Undiscounted Problems

Contents

In this chapter we consider total cost infinite horizon problems where we allow costs per stage that are unbounded above or below. Also, the discount factor α does not have to be less than one. The complications resulting are substantial, and the analysis required is considerably more sophisticated than the one given thus far. We also consider applications of the theory to important classes of problems. The problem section touches on several related topics.

3.1 UNBOUNDED COSTS PER STATE

In this section we consider the total cost infinite horizon problem of Section 1.1 under one of the following two assumptions.

Assumption P: (Positivity) The cost per stage g satisfies

$$0 \leq g(x,u,w), \qquad \text{for all } (x,u,w) \in S \times C \times D. \tag{1.1}$$

Assumption N: (Negativity) The cost per stage g satisfies

$$g(x,u,w) \leq 0, \qquad \text{for all } (x,u,w) \in S \times C \times D. \tag{1.2}$$

Problems corresponding to Assumption P are sometimes referred to in the research literature as *negative DP problems*. This name was used in the original reference [Str66], where the problem of maximizing the infinite sum of negative rewards per stage was considered. Similarly, problems corresponding to Assumption N are sometimes referred to as *positive DP problems* [Bla65], [Str66]. Assumption N arises in problems where there is a nonnegative reward per stage and the total expected reward is to be *maximized*.

Note that when $\alpha < 1$ and g is either bounded above or below, we may add a suitable scalar to g in order to satisfy Eq. (1.1) or Eq. (1.2), respectively. An optimal policy will not be affected by this change since, in view of the presence of the discount factor, the addition of a constant r to g merely adds $(1-\alpha)^{-1}r$ to the cost associated with every policy.

One complication arising from unbounded costs per stage is that, for some initial states x_0 and some genuinely interesting admissible policies

$\pi = \{\mu_0, \mu_1, \ldots\}$, the cost $J_\pi(x_0)$ may be ∞ (in the case of Assumption P) or $-\infty$ (in the case of Assumption N). Here is an example:

Example 1.1

Consider the scalar system

$$x_{k+1} = \beta x_k + u_k, \qquad k = 0, 1, \ldots,$$

where $x_k \in \Re$ and $u_k \in \Re$, for all k, and β is a positive scalar. The control constraint is $|u_k| \leq 1$, and the cost is

$$J_\pi(x_0) = \lim_{N\to\infty} \sum_{k=0}^{N-1} \alpha^k |x_k|.$$

Consider the policy $\tilde{\pi} = \{\tilde{\mu}, \tilde{\mu}, \ldots\}$, where $\tilde{\mu}(x) = 0$ for all $x \in \Re$. Then

$$J_{\tilde{\pi}}(x_0) = \lim_{N\to\infty} \sum_{k=0}^{N-1} \alpha^k \beta^k |x_0|,$$

and hence

$$J_{\tilde{\pi}}(x_0) = \begin{cases} 0 & \text{if } x_0 = 0 \\ \infty & \text{if } x_0 \neq 0 \end{cases} \qquad \text{if} \qquad \alpha\beta \geq 1,$$

while

$$J_{\tilde{\pi}}(x_0) = \frac{|x_0|}{1-\alpha\beta} \qquad \text{if} \qquad \alpha\beta < 1.$$

Note a peculiarity here: if $\beta > 1$ the state x_k diverges to ∞ or to $-\infty$, but if the discount factor is sufficiently small ($\alpha < 1/\beta$), the cost $J_{\tilde{\pi}}(x_0)$ is finite.

It is also possible to verify that when $\beta > 1$ and $\alpha\beta \geq 1$ the optimal cost $J^*(x_0)$ is equal to ∞ for $|x_0| \geq 1/(\beta-1)$ and is finite for $|x_0| < 1/(\beta-1)$. The problem here is that when $\beta > 1$ the system is unstable, and in view of the restriction $|u_k| \leq 1$ on the control, it may not be possible to force the state near zero once it has reached sufficiently large magnitude.

The preceding example shows that there is not much that can be done about the possibility of the cost function being infinite for some policies. To cope with this situation, we conduct our analysis with the notational understanding that the costs $J_\pi(x_0)$ and $J^*(x_0)$ may be ∞ (or $-\infty$) under Assumption P (or N, respectively) for some initial states x_0 and policies π. In other words, we consider $J_\pi(\cdot)$ and $J^*(\cdot)$ to be extended real-valued functions. In fact, the entire subsequent analysis is valid even if the cost $g(x,u,w)$ is ∞ or $-\infty$ for some (x,u,w), as long as Assumption P or Assumption N holds.

The line of analysis of this section is fundamentally different from the one of the discounted problem of Section 1.2. For the latter problem, the analysis was based on ignoring the "tails" of the cost sequences. In

this section, the tails of the cost sequences may not be small, and for this reason, the control is much more focused on affecting the long-term behavior of the state. For example, let $\alpha = 1$, and assume that the stage cost at all states is nonzero except for a cost-free and absorbing termination state. Then, a primary task of control under Assumption P (or Assumption N) is roughly to bring the state of the system to the termination state or to a region where the cost per stage is nearly zero as *quickly* as possible (as *late* as possible, respectively). Note the difference in control objective between Assumptions P and N. It accounts for some strikingly different results under the two assumptions.

Main Results -- Bellman's Equation

We now present results that characterize the optimal cost function J^*, as well as optimal stationary policies. We also give conditions under which value iteration converges to the optimal cost function J^*. In the proofs we will often need to interchange expectation and limit in various relations. This interchange is valid under the assumptions of the following theorem.

Monotone Convergence Theorem: Let $P = (p_1, p_2, \ldots)$ be a probability distribution over $S = \{1, 2, \ldots\}$. Let $\{h_N\}$ be a sequence of extended real-valued functions on S such that for all $i \in S$ and $N = 1, 2, \ldots,$

$$0 \leq h_N(i) \leq h_{N+1}(i).$$

Let $h : S \mapsto [0, \infty]$ be the limit function

$$h(i) = \lim_{N \to \infty} h_N(i).$$

Then

$$\lim_{N \to \infty} \sum_{i=1}^{\infty} p_i h_N(i) = \sum_{i=1}^{\infty} p_i \lim_{N \to \infty} h_N(i) = \sum_{i=1}^{\infty} p_i h(i).$$

Proof: We have

$$\sum_{i=1}^{\infty} p_i h_N(i) \leq \sum_{i=1}^{\infty} p_i h(i).$$

By taking the limit, we obtain

$$\lim_{N \to \infty} \sum_{i=1}^{\infty} p_i h_N(i) \leq \sum_{i=1}^{\infty} p_i h(i),$$

so there remains to prove the reverse inequality. For every integer $M \geq 1$, we have

$$\lim_{N \to \infty} \sum_{i=1}^{\infty} p_i h_N(i) \geq \lim_{N \to \infty} \sum_{i=1}^{M} p_i h_N(i) = \sum_{i=1}^{M} p_i h(i),$$

and by taking the limit as $M \to \infty$ the reverse inequality follows. **Q.E.D.**

Similar to all the infinite horizon problems considered so far, the optimal cost function satisfies Bellman's equation.

Proposition 1.1: (Bellman's Equation) Under either Assumption P or N the optimal cost function J^* satisfies

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

or, equivalently,

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

Proof: For any admissible policy $\pi = \{\mu_0, \mu_1, \ldots\}$, consider the cost $J_\pi(x)$ corresponding to π when the initial state is x. We have

$$J_\pi(x) = E_w \{ g\big(x, \mu_0(x), w\big) + V_\pi\big(f(x, \mu_0(x), w)\big) \}, \tag{1.3}$$

where, for all $x_1 \in S$,

$$V_\pi(x_1) = \lim_{N \to \infty} E_{\substack{w_k \\ k=1,2,\ldots}} \left\{ \sum_{k=1}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\}.$$

Thus, $V_\pi(x_1)$ is the cost from stage 1 to infinity using π when the initial state is x_1. We have clearly

$$V_\pi(x_1) \geq \alpha J^*(x_1), \qquad \text{for all } x_1 \in S.$$

Hence, from Eq. (1.3),

$$\begin{aligned} J_\pi(x) &\geq E_w \{ g\big(x, \mu_0(x), w\big) + \alpha J^*\big(f(x, \mu_0(x), w)\big) \} \\ &\geq \min_{u \in U(x)} E_w \{ g(x, u, w) + \alpha J^*\big(f(x, u, w)\big) \}. \end{aligned}$$

Taking the minimum over all admissible policies, we have

$$\begin{aligned} \min_\pi J_\pi(x) &= J^*(x) \\ &\geq \min_{u \in U(x)} E_w \{ g(x, u, w) + \alpha J^*\big(f(x, u, w)\big) \} \qquad (1.4) \\ &= (TJ^*)(x). \end{aligned}$$

Thus there remains to prove that the reverse inequality also holds. We prove this separately for Assumption N and for Assumption P.

Assume P. The following proof of $J^* \leq TJ^*$ under this assumption would be considerably simplified if we knew that there exists a μ such that $T_\mu J^* = TJ^*$. Since in general such a μ need not exist, we introduce a positive sequence $\{\epsilon_k\}$, and we choose an admissible policy $\pi = \{\mu_0, \mu_1, \ldots\}$ such that

$$(T_{\mu_k} J^*)(x) \leq (TJ^*)(x) + \epsilon_k, \qquad x \in S, \quad k = 0, 1, \ldots$$

Such a choice is possible because we know that, under P, we have $-\infty < J^*(x)$ for all x. By using the inequality $TJ^* \leq J^*$ shown earlier, we obtain

$$(T_{\mu_k} J^*)(x) \leq J^*(x) + \epsilon_k, \qquad x \in S, \quad k = 0, 1, \ldots$$

Applying $T_{\mu_{k-1}}$ to both sides of this relation, we have

$$\begin{aligned}(T_{\mu_{k-1}} T_{\mu_k} J^*)(x) &\leq (T_{\mu_{k-1}} J^*)(x) + \alpha\epsilon_k \\ &\leq (TJ^*)(x) + \epsilon_{k-1} + \alpha\epsilon_k \\ &\leq J^*(x) + \epsilon_{k-1} + \alpha\epsilon_k.\end{aligned}$$

Continuing this process, we obtain

$$(T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} J^*)(x) \leq (TJ^*)(x) + \sum_{i=0}^{k} \alpha^i \epsilon_i.$$

By taking the limit as $k \to \infty$ and noting that

$$J^*(x) \leq J_\pi(x) = \lim_{k\to\infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} J_0)(x) \leq \lim_{k\to\infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} J^*)(x),$$

where J_0 is the zero function, it follows that

$$J^*(x) \leq J_\pi(x) \leq (TJ^*)(x) + \sum_{i=0}^{\infty} \alpha^i \epsilon_i, \qquad x \in S.$$

Since the sequence $\{\epsilon_k\}$ is arbitrary, we can take $\sum_{i=0}^{\infty} \alpha^i \epsilon_i$ as small as desired, and we obtain $J^*(x) \leq (TJ^*)(x)$ for all $x \in S$. Combining this with the inequality $J^*(x) \geq (TJ^*)(x)$ shown earlier, the result follows (under Assumption P).

Assume N and let J_N be the optimal cost function for the corresponding N-stage problem

$$J_N(x_0) = \min_\pi E\left\{\sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\}. \tag{1.5}$$

We first show that

$$J^*(x) = \lim_{N\to\infty} J_N(x), \qquad x \in S. \tag{1.6}$$

Indeed, in view of Assumption N, we have $J^* \leq J_N$ for all N, so

$$J^*(x) \leq \lim_{N\to\infty} J_N(x), \qquad x \in S. \tag{1.7}$$

Also, for all $\pi = \{\mu_0, \mu_1, \ldots\}$, we have

$$E\left\{\sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \geq J_N(x_0),$$

and by taking the limit as $N \to \infty$,

$$J_\pi(x) \geq \lim_{N\to\infty} J_N(x), \qquad x \in S.$$

Taking the minimum over π, we obtain $J^*(x) \geq \lim_{N\to\infty} J_N(x)$, and combining this relation with Eq. (1.7), we obtain Eq. (1.6).

For every admissible μ, we have

$$T_\mu J_N \geq J_{N+1},$$

and by taking the limit as $N \to \infty$, and using the monotone convergence theorem and Eq. (1.6), we obtain

$$T_\mu J^* \geq J^*.$$

Taking the minimum over μ, we obtain $TJ^* \geq J^*$, which combined with the inequality $J^* \geq TJ^*$ shown earlier, proves the result under Assumption N. **Q.E.D.**

Similar to Cor. 2.2.1 in Section 1.2, we have:

Corollary 1.1.1: Let μ be a stationary policy. Then under Assumption P or N, we have

$$J_\mu(x) = \mathop{E}_w\big\{g\big(x, \mu(x), w\big) + \alpha J_\mu\big(f(x, \mu(x), w)\big)\big\}, \qquad x \in S$$

or, equivalently,

$$J_\mu = T_\mu J_\mu. \tag{1.8}$$

Contrary to discounted problems with bounded cost per stage, the optimal cost function J^* under Assumption P or N need not be the unique solution of Bellman's equation. Consider the following example.

Example 1.2

Let $S = [0, \infty)$ (or $S = (-\infty, 0]$) and

$$g(x, u, w) = 0, \qquad f(x, u, w) = \frac{x}{\alpha}.$$

Then for every β, the function J given by

$$J(x) = \beta x, \qquad x \in S,$$

is a solution of Bellman's equation, so T has an infinite number of fixed points. Note, however, that there is a unique fixed point within the class of bounded functions, the zero function $J_0(x) \equiv 0$, which is the optimal cost function for this problem. More generally, it can be shown by using the following Prop. 1.2 that if $\alpha < 1$ and there exists a bounded function that is a fixed point of T, then that function must be equal to the optimal cost function J^* (see Exercise 3.5). When $\alpha = 1$, Bellman's equation may have an infinity of solutions even within the class of bounded functions. This is because if $\alpha = 1$ and $J(\cdot)$ is any solution, then for any scalar r, $J(\cdot) + r$ is also a solution.

The optimal cost function J^*, however, has the property that it is the smallest (under Assumption P) or largest (under Assumption N) fixed point of T in the sense described in the following proposition.

Proposition 1.2:

(a) Under Assumption P, if $\tilde{J} : S \mapsto (-\infty, \infty]$ satisfies $\tilde{J} \geq T\tilde{J}$ and either $\tilde{J}$ is bounded below and $\alpha < 1$, or $\tilde{J} \geq 0$, then $\tilde{J} \geq J^*$.

(b) Under Assumption N, if $\tilde{J} : S \mapsto [-\infty, \infty)$ satisfies $\tilde{J} \leq T\tilde{J}$ and either $\tilde{J}$ is bounded above and $\alpha < 1$, or $\tilde{J} \leq 0$, then $\tilde{J} \leq J^*$.

Proof: (a) Under Assumption P, let r be a scalar such that $\tilde{J}(x) + r \geq 0$ for all $x \in S$ and if $\alpha \geq 1$ let $r = 0$. For any sequence $\{\epsilon_k\}$ with $\epsilon_k > 0$, let $\tilde{\pi} = \{\tilde{\mu}_0, \tilde{\mu}_1, \ldots\}$ be an admissible policy such that, for every $x \in S$ and k,

$$E_w\big\{g\big(x, \mu_k(x), w\big) + \alpha \tilde{J}\big(f\big(x, \mu_k(x), w\big)\big)\big\} \leq (T\tilde{J})(x) + \epsilon_k. \tag{1.9}$$

Such a policy exists since $(T\tilde{J})(x) > -\infty$ for all $x \in S$. We have for any initial state $x_0 \in S$,

$$\begin{aligned} J^*(x_0) &= \min_{\pi} \lim_{N\to\infty} E\left\{\sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\ &\leq \min_{\pi} \liminf_{N\to\infty} E\left\{\alpha^N \big(\tilde{J}(x_N) + r\big) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\ &\leq \liminf_{N\to\infty} E\left\{\alpha^N \big(\tilde{J}(x_N) + r\big) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \tilde{\mu}_k(x_k), w_k\big)\right\}. \end{aligned}$$

Using Eq. (1.9) and the assumption $\tilde{J} \geq T\tilde{J}$, we obtain

$$\begin{aligned} &E\left\{\alpha^N \tilde{J}(x_N) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \tilde{\mu}(x_k), w_k\big)\right\} \\ &= E\left\{\alpha^N \tilde{J}\big(f\big(x_{N-1}, \tilde{\mu}_{N-1}(x_{N-1}), w_{N-1}\big)\big) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \tilde{\mu}_k(x_k), w_k\big)\right\} \\ &\leq E\left\{\alpha^{N-1} \tilde{J}(x_{N-1}) + \sum_{k=0}^{N-2} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} + \alpha^{N-1}\epsilon_{N-1} \\ &\leq E\left\{\alpha^{N-2} \tilde{J}(x_{N-2}) + \sum_{k=0}^{N-3} \alpha^k g\big(x_k, \tilde{\mu}_k(x_k), w_k\big)\right\} + \alpha^{N-2}\epsilon_{N-2} \\ &\qquad + \alpha^{N-1}\epsilon_{N-1} \\ &\vdots \\ &\leq \tilde{J}(x_0) + \sum_{k=0}^{N-1} \alpha^k \epsilon_k. \end{aligned}$$

Combining these inequalities, we obtain

$$J^*(x_0) \leq \tilde{J}(x_0) + \lim_{N\to\infty} \left(\alpha^N r + \sum_{k=0}^{N-1} \alpha^k \epsilon_k\right).$$

Since the sequence $\{\epsilon_k\}$ is arbitrary (except for $\epsilon_k > 0$), we may select $\{\epsilon_k\}$ so that $\lim_{N\to\infty} \sum_{k=0}^{N-1} \alpha^k \epsilon_k$ is arbitrarily close to zero, and the result follows.

(b) Under Assumption N, let r be a scalar such that $\tilde{J}(x) + r \leq 0$ for all $x \in S$, and if $\alpha \geq 1$, let $r = 0$. We have for every initial state $x_0 \in S$.

$$\begin{aligned} J^*(x_0) &= \min_{\pi} \lim_{N\to\infty} E\left\{\sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\ &\geq \min_{\pi} \limsup_{N\to\infty} E\left\{\alpha^N \big(\tilde{J}(x_N) + r\big) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\ &\geq \limsup_{N\to\infty} \min_{\pi} E\left\{\alpha^N \big(\tilde{J}(x_N) + r\big) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\}. \end{aligned} \tag{1.10}$$

where the last inequality follows from the fact that for any sequence $\{h_N(\xi)\}$ of functions of a parameter ξ we have

$$\min_{\xi} \limsup_{N\to\infty} h_N(\xi) \geq \limsup_{N\to\infty} \min_{\xi} h_N(\xi).$$

This inequality follows by writing

$$h_N(\xi) \geq \min_{\xi} h_N(\xi)$$

and by subsequently taking the limsup of both sides and the minimum over ξ of the left-hand side.

Now we have, by using the assumption $\tilde{J} \leq T\tilde{J}$,

$$\begin{aligned}
&\min_{\pi} E\left\{\alpha^N \tilde{J}(x_N) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\
&\quad = \min_{\pi} E\left\{\alpha^{N-1} \min_{u_{N-1} \in U(x_{N-1})} E_{w_{N-1}}\big\{g(x_{N-1}, u_{N-1}, w_{N-1}) \right. \\
&\qquad \left. + \alpha \tilde{J}\big(f(x_{N-1}, u_{N-1}, w_{N-1})\big)\big\} + \sum_{k=0}^{N-2} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\
&\quad \geq \min_{\pi} E\left\{\alpha^{N-1} \tilde{J}(x_{N-1}) + \sum_{k=0}^{N-2} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\
&\quad \vdots \\
&\quad \geq \tilde{J}(x_0).
\end{aligned}$$

Using this relation in Eq. (1.10), we obtain

$$J^*(x_0) \geq \tilde{J}(x_0) + \lim_{N \to \infty} \alpha^N r = \tilde{J}(x_0).$$

Q.E.D.

As before, we have the following corollary:

Corollary 1.2.1: Let μ be an admissible stationary policy.

(a) Under Assumption P, if $\tilde{J} : S \mapsto (-\infty, \infty]$ satisfies $\tilde{J} \geq T_\mu \tilde{J}$ and either $\tilde{J}$ is bounded below and $\alpha < 1$, or $\tilde{J} \geq 0$, then $\tilde{J} \geq J_\mu$.

(b) Under Assumption N, if $\tilde{J} : S \mapsto [-\infty, \infty)$ satisfies $\tilde{J} \leq T_\mu \tilde{J}$ and either $\tilde{J}$ is bounded above and $\alpha < 1$, or $\tilde{J} \leq 0$, then $\tilde{J} \leq J_\mu$.

Conditions for Optimality of a Stationary Policy

Under Assumption P, we have the same optimality condition as for discounted problems with bounded cost per stage.

Proposition 1.3: (Necessary and Sufficient Condition for Optimality under P) Let Assumption P hold. A stationary policy μ is optimal if and only if

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

Proof: If $TJ^* = T_\mu J^*$, Bellman's equation ($J^* = TJ^*$) implies that $J^* = T_\mu J^*$. From Cor. 1.2.1(a) we then obtain $J^* \geq J_\mu$, showing that μ is optimal. Conversely, if $J^* = J_\mu$, we have using Cor. 1.1.1, $TJ^* = J^* = J_\mu = T_\mu J_\mu = T_\mu J^*$. **Q.E.D.**

Unfortunately, the sufficiency part of the above proposition need not be true under Assumption N; that is, we may have $TJ^* = T_\mu J^*$ while μ is not optimal. This is illustrated in the following example.

Example 1.3

Let $S = C = (-\infty, 0]$, $U(x) = C$ for all $x \in S$, and

$$g(x, u, w) = f(x, u, w) = u,$$

for all $(x, u, w) \in S \times C \times D$. Then $J^*(x) = -\infty$ for all $x \in S$, and every stationary policy μ satisfies the condition of the preceding proposition. On the other hand, when $\mu(x) = 0$ for all $x \in S$, we have $J_\mu(x) = 0$ for all $x \in S$, and hence μ is not optimal.

It is worth noting that Prop. 1.3 implies the existence of an optimal stationary policy under Assumption P when $U(x)$ is a finite set for every $x \in S$. This need not be true under Assumption N (see Example 4.4 in Section 3.4).

Under Assumption N, we have a different characterization of an optimal stationary policy.

Proposition 1.4: (Necessary and Sufficient Condition for Optimality under N) Let Assumption N hold. A stationary policy μ is optimal if and only if

$$TJ_\mu = T_\mu J_\mu. \tag{1.11}$$

Proof: If $TJ_\mu = T_\mu J_\mu$, then from Cor. 1.1.1 we have $J_\mu = T_\mu J_\mu$, so that J_μ is a fixed point of T. Then by Prop. 1.2, we have $J_\mu \leq J^*$, which implies that μ is optimal. Conversely, if $J_\mu = J^*$, then $T_\mu J_\mu = J_\mu = J^* = TJ^* = TJ_\mu$. **Q.E.D.**

The interpretation of the preceding optimality condition is that persistently using μ is optimal if and only if this performs at least as well as using any $\overline{\mu}$ at the first stage and using μ thereafter. Under Assumption P this condition is not sufficient to guarantee optimality of the stationary policy μ, as the following example shows.

Example 1.4

Let $S = (-\infty, \infty)$, $U(x) = (0, 1]$ for all $x \in S$,

$$g(x, u, w) = |x|, \qquad f(x, u, w) = \alpha^{-1} u x,$$

for all $(x, u, w) \in S \times C \times D$. Let $\mu(x) = 1$ for all $x \in S$. Then $J_\mu(x) = \infty$ if $x \neq 0$ and $J_\mu(0) = 0$. Furthermore, we have $J_\mu = T_\mu J_\mu = T J_\mu$, as the reader can easily verify. It can also be verified that $J^*(x) = |x|$, and hence the stationary policy μ is not optimal.

The Value Iteration Method

We now turn to the question whether the DP algorithm converges to the optimal cost function J^*. Let J_0 be the zero function on S,

$$J_0(x) = 0, \qquad x \in S.$$

Then under Assumption P, we have

$$J_0 \leq T J_0 \leq T^2 J_0 \leq \cdots \leq T^k J_0 \leq \cdots,$$

while under Assumption N, we have

$$J_0 \geq T J_0 \geq T^2 J_0 \geq \cdots \geq T^k J_0 \geq \cdots$$

In either case the limit function

$$J_\infty(x) = \lim_{k \to \infty} (T^k J_0)(x), \qquad x \in S, \tag{1.12}$$

is well defined, provided we allow the possibility that J_∞ can take the value ∞ (under Assumption P) or $-\infty$ (under Assumption N). The question is whether the value iteration method is valid in the sense

$$J_\infty = J^*. \tag{1.13}$$

This question is, of course, of computational interest, but it is also of analytical interest since, if one knows that $J^* = \lim_{k \to \infty} T^k J_0$, one can infer properties of the unknown function J^* from properties of the k-stage optimal cost functions $T^k J_0$, which are defined in a concrete algorithmic manner.

We will show that $J_\infty = J^*$ under Assumption N. It turns out, however, that under Assumption P, we may have $J_\infty \neq J^*$ (see Exercise 3.1). We will later provide easily verifiable conditions that guarantee that $J_\infty = J^*$ under Assumption P. We have the following proposition.

Proposition 1.5:

(a) Let Assumption P hold and assume that

$$J_\infty(x) = (T J_\infty)(x), \qquad x \in S.$$

Then if $J : S \mapsto \Re$ is any bounded function and $\alpha < 1$, or otherwise if $J_0 \leq J \leq J^*$, we have

$$\lim_{k \to \infty} (T^k J)(x) = J^*(x), \qquad x \in S. \tag{1.14}$$

(b) Let Assumption N hold. Then if $J : S \mapsto \Re$ is any bounded function and $\alpha < 1$, or otherwise if $J^* \leq J \leq J_0$, we have

$$\lim_{k \to \infty} (T^k J)(x) = J^*(x), \qquad x \in S. \tag{1.15}$$

Proof: (a) Since under Assumption P, we have

$$J_0 \leq T J_0 \leq \cdots \leq T^k J_0 \leq \cdots \leq J^*,$$

it follows that $\lim_{k \to \infty} T^k J_0 = J_\infty \leq J^*$. Since J_∞ is also a fixed point of T by assumption, we obtain from Prop. 1.2(a) that $J^* \leq J_\infty$. It follows that

$$J_\infty = J^*, \tag{1.16}$$

and hence Eq. (1.14) is proved for the case $J = J_0$.

For the case where $\alpha < 1$ and J is bounded, let r be a scalar such that

$$J_0 - re \leq J \leq J_0 + re. \tag{1.17}$$

Applying T^k to this relation, we obtain

$$T^k J_0 - \alpha^k re \leq T^k J \leq T^k J_0 + \alpha^k re. \tag{1.18}$$

Since $T^k J_0$ converges to J^*, as shown earlier, this relation implies that $T^k J$ converges also to J^*.

In the case where $J_0 \leq J \leq J^*$, we have by applying T^k

$$T^k J_0 \leq T^k J \leq J^*, \qquad k = 0, 1, \ldots \tag{1.19}$$

Since $T^k J_0$ converges to J^*, so does $T^k J$.

(b) It was shown earlier [cf. Eq. (1.6)] that under Assumption N, we have

$$J_\infty(x) = \lim_{k \to \infty} (T^k J_0)(x) = J^*(x). \tag{1.20}$$

The proof from this point is identical to that for part (a). **Q.E.D.**

We now derive conditions guaranteeing that $J_\infty = TJ_\infty$ holds under Assumption P, which by Prop. 1.5 implies that $J_\infty = J^*$. We prove two propositions. The first admits an easy proof but requires a finiteness assumption on the control constraint set. The second is harder to prove but requires a weaker compactness assumption.

Proposition 1.6: Let Assumption P hold and assume that the control constraint set is finite for every $x \in S$. Then

$$J_\infty = TJ_\infty = J^*. \tag{1.21}$$

Proof: As shown in the proof of Prop. 1.5(a), we have for all k, $T^k J_0 \leq J_\infty \leq J^*$. Applying T to this relation, we obtain

$$\begin{aligned}(T^{k+1} J_0)(x) &= \min_{u \in U(x)} E_w\big\{g(x,u,w) + \alpha (T^k J_0)\big(f(x,u,w)\big)\big\} \\ &\leq (TJ_\infty)(x),\end{aligned} \tag{1.22}$$

and by taking the limit as $k \to \infty$, it follows that

$$J_\infty \leq TJ_\infty.$$

Suppose that there existed a state $\tilde{x} \in S$ such that

$$J_\infty(\tilde{x}) < (TJ_\infty)(\tilde{x}). \tag{1.23}$$

Let u_k minimize in Eq. (1.22) when $x = \tilde{x}$. Since $U(\tilde{x})$ is finite, there must exist some $\tilde{u} \in U(\tilde{x})$ such that $u_k = \tilde{u}$ for all k in some infinite subset K of the positive integers. By Eq. (1.22) we have for all $k \in K$

$$\begin{aligned}(T^{k+1} J_0)(\tilde{x}) &= E_w\big\{g(\tilde{x},\tilde{u},w) + \alpha (T^k J_0)\big(f(\tilde{x},\tilde{u},w)\big)\big\} \\ &\leq (TJ_\infty)(\tilde{x}).\end{aligned}$$

Taking the limit as $k \to \infty$, $k \in K$, we obtain

$$\begin{aligned}J_\infty(\tilde{x}) &= E_w\big\{g(\tilde{x},\tilde{u},w) + \alpha J_\infty\big(f(\tilde{x},\tilde{u},w)\big)\big\} \\ &\geq (TJ_\infty)(\tilde{x}) \\ &= \min_{u \in U(\tilde{x})} E_w\big\{g(\tilde{x},u,w) + \alpha J_\infty\big(f(\tilde{x},u,w)\big)\big\}.\end{aligned}$$

This contradicts Eq. (1.23), so we have $J_\infty(\tilde{x}) = (TJ_\infty)(\tilde{x})$. **Q.E.D.**

The following proposition strengthens Prop. 1.6 in that it requires a compactness rather than a finiteness assumption. We recall (see Appendix A of Vol. I) that a subset X of the n-dimensional Euclidean space $\Re^n$ is said to be *compact* if every sequence $\{x_k\}$ with $x_k \in X$ contains a subsequence $\{x_k\}_{k \in K}$ that converges to a point $x \in X$. Equivalently, X is compact if and only if it is closed and bounded. The empty set is (trivially) considered compact. Given any collection of compact sets, their intersection is a compact set (possibly empty). Given a sequence of nonempty compact sets $X_1, X_2 \ldots, X_k, \ldots$ such that

$$X_1 \supset X_2 \supset \cdots \supset X_k \supset X_{k+1} \supset \cdots \tag{1.24}$$

their intersection $\cap_{k=1}^\infty X_k$ is both nonempty and compact. In view of this fact, it follows that if $f : \Re^n \mapsto [-\infty, \infty]$ is a function such that the set

$$F_\lambda = \big\{x \in \Re^n \mid f(x) \leq \lambda\big\} \tag{1.25}$$

is compact for every $\lambda \in R$, then there exists a vector x^* minimizing f; that is, there exists an $x^* \in \Re^n$ such that

$$f(x^*) = \min_{x \in \Re^n} f(x). \tag{1.26}$$

To see this, take a sequence $\{\lambda_k\}$ such that $\lambda_k \to \min_{x \in \Re^n} f(x)$ and $\lambda_k \geq \lambda_{k+1}$ for all k. If $\min_{x \in \Re^n} f(x) < \infty$, such a sequence exists and the sets

$$F_{\lambda_k} = \{x \in \Re^n \mid f(x) \leq \lambda_k\} \tag{1.27}$$

are nonempty and compact. Furthermore, $F_{\lambda_k} \supset F_{\lambda_{k+1}}$ for all k, and hence the intersection $\cap_{k=1}^\infty F_{\lambda_k}$ is also nonempty and compact. Let x^* be any vector in $\cap_{k=1}^\infty F_{\lambda_k}$. Then

$$f(x^*) \leq \lambda_k, \qquad k = 1, 2, \ldots, \tag{1.28}$$

and taking the limit as $k \to \infty$, we obtain $f(x^*) \leq \min_{x \in \Re^n} f(x)$, proving that x^* minimizes $f(x)$. The most common case where we can guarantee

that the set F_λ of Eq. (1.25) is compact for all λ is when f is continuous and $f(x) \to \infty$ as $\|x\| \to \infty$.

Proposition 1.7: Let Assumption P hold, and assume that the sets

$$U_k(x,\lambda) = \left\{ u \in U(x) \,\middle|\, \mathop{E}_{w}\{g(x,u,w) + \alpha(T^k J_0)\big(f(x,u,w)\big)\} \le \lambda \right\} \tag{1.29}$$

are compact subsets of a Euclidean space for every $x \in S$, $\lambda \in \Re$, and for all k greater than some integer $\overline{k}$. Then

$$J_\infty = TJ_\infty = J^*. \tag{1.30}$$

Furthermore, there exists a stationary optimal policy.

Proof: As in Prop. 1.6, we have $J_\infty \le TJ_\infty$. Suppose that there existed a state $\tilde{x} \in S$ such that

$$J_\infty(\tilde{x}) < (TJ_\infty)(\tilde{x}). \tag{1.31}$$

Clearly, we must have $J_\infty(\tilde{x}) < \infty$. For every $k \ge \overline{k}$, consider the sets

$$U_k\big(\tilde{x}, J_\infty(\tilde{x})\big) = \left\{ u \in U(\tilde{x}) \,\middle|\, \mathop{E}_{w}\big\{g(\tilde{x},u,w) + \alpha(T^k J_0)\big(f(\tilde{x},u,w)\big)\big\} \le J_\infty(\tilde{x}) \right\}.$$

Let also u_k be a point attaining the minimum in

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

that is, u_k is such that

$$(T^{k+1}J_0)(\tilde{x}) = \mathop{E}_{w}\big\{g(\tilde{x},u_k,w) + \alpha(T^k J_0)\big(f(\tilde{x},u_k,w)\big)\big\}.$$

Such minimizing points u_k exist by our compactness assumption. For every $k \ge \overline{k}$, consider the sequence $\{u_i\}_{i=k}^\infty$. Since $T^k J_0 \le T^{k+1}J_0 \le \cdots \le J_\infty$, it follows that

$$\begin{aligned} &\mathop{E}_{w}\big\{g(\tilde{x},u_i,w) + \alpha(T^k J_0)\big(f(\tilde{x},u_i,w)\big)\big\} \\ &\qquad \le \mathop{E}_{w}\big\{g(\tilde{x},u_i,w) + \alpha(T^i J_0)\big(f(\tilde{x},u_i,w)\big)\big\} \\ &\qquad \le J_\infty(\tilde{x}), \qquad i \ge k. \end{aligned}$$

Therefore $\{u_i\}_{i=k}^\infty \subset U_k\big(\tilde{x}, J_\infty(\tilde{x})\big)$, and since $U_k\big(\tilde{x}, J_\infty(\tilde{x})\big)$ is compact, all the limit points of $\{u_i\}_{i=k}^\infty$ belong to $U_k\big(\tilde{x}, J_\infty(\tilde{x})\big)$ and at least one such limit point exists. Hence the same is true of the limit points of the whole sequence $\{u_i\}_{i=\overline{k}}^\infty$. It follows that if $\tilde{u}$ is a limit point of $\{u_i\}_{i=k}^\infty$ then

$$\tilde{u} \in \cap_{k=\overline{k}}^\infty U_k\big(\tilde{x}, J_\infty(\tilde{x})\big).$$

This implies by Eq. (1.29) that for all $k \ge \overline{k}$

$$J_\infty(\tilde{x}) \ge \mathop{E}_{w}\big\{g(\tilde{x},\tilde{u},w) + \alpha(T^k J_0)\big(f(\tilde{x},\tilde{u},w)\big)\big\} \ge (T^{k+1}J_0)(\tilde{x}).$$

Taking the limit as $k \to \infty$, we obtain

$$J_\infty(\tilde{x}) = \mathop{E}_{w}\big\{g(\tilde{x},\tilde{u},w) + \alpha J_\infty\big(f(\tilde{x},\tilde{u},w)\big)\big\}.$$

Since the right-hand side is greater than or equal to $(TJ_\infty)(\tilde{x})$, Eq. (1.31) is contradicted. Hence $J_\infty = TJ_\infty$ and Eq. (1.30) is proved in view of Prop. 1.5(a).

To show that there exists an optimal stationary policy, observe that Eq. (1.30) and the last relation imply that $\tilde{u}$ attains the minimum in

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

for a state $\tilde{x} \in S$ with $J^*(\tilde{x}) < \infty$. For states $\tilde{x} \in S$ such that $J^*(\tilde{x}) = \infty$, every $u \in U(\tilde{x})$ attains the preceding minimum. Hence by Prop. 1.3(a) an optimal stationary policy exists. **Q.E.D.**

The reader may verify by inspection of the preceding proof that if $\mu_k(\tilde{x})$, $k = 0, 1, \ldots$, attains the minimum in the relation

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

then if $\mu^*(\tilde{x})$ is a limit point of $\{\mu_k(\tilde{x})\}$, for every $\tilde{x} \in S$, the stationary policy μ^* is optimal. Furthermore, $\{\mu_k(\tilde{x})\}$ has at least one limit point for every $\tilde{x} \in S$ for which $J^*(\tilde{x}) < \infty$. Thus *the value iteration method under the assumptions of either Prop. 1.6 or Prop. 1.7 yields in the limit not only the optimal cost function J^* but also an optimal stationary policy.*

Other Computational Methods

Unfortunately, policy iteration is not a valid procedure under either P or N in the absence of further conditions. If μ and $\overline{\mu}$ are stationary policies such that $T_{\overline{\mu}}J_\mu = TJ_\mu$, then it can be shown that under Assumption P we have

$$J_{\overline{\mu}}(x) \le J_\mu(x), \qquad x \in S. \tag{1.32}$$

To see this, note that $T_{\overline{\mu}}J_\mu = TJ_\mu \le T_\mu J_\mu = J_\mu$ from which we obtain $\lim_{N\to\infty} T^N_{\overline{\mu}} J_\mu \le J_\mu$. Since $J_{\overline{\mu}} = \lim_{N\to\infty} T^N_{\overline{\mu}} J_0$ and $J_0 \le J_\mu$, we obtain $J_{\overline{\mu}} \le J_\mu$. However, $J_\mu \le J_\mu$ by itself is not sufficient to guarantee the validity of policy iteration. For example, it is not clear that strict inequality holds in Eq. (1.32) for at least one state $x \in S$ when μ is not optimal. The difficulty here is that the equality $J_\mu = TJ_\mu$ does not imply that μ is optimal, and additional conditions are needed to guarantee the validity of policy iteration. However, for special cases such conditions can be verified (see for example Section 3.2 and Exercise 3.16).

It is possible to devise a computational method based on mathematical programming when S, C, and D are finite sets by making use of Prop. 1.2. Under N and $\alpha = 1$, the corresponding (linear) program is (compare with Section 1.3.4)

$$\text{maximize} \quad \sum_{i=1}^{n} \lambda_i$$

$$\text{subject to} \quad \lambda_i \le g(i,u) + \sum_{j=1}^{n} p_{ij}(u)\lambda_j, \qquad i = 1,2,\ldots,n, \quad u \in U(i).$$

When $\alpha = 1$ and Assumption P holds, the corresponding program takes the form

$$\text{minimize} \quad \sum_{i=1}^{n} \lambda_i$$

$$\text{subject to} \quad \lambda_i \ge \min_{u\in U(i)} \left[g(i,u) + \sum_{j=1}^{n} p_{ij}(u)\lambda_j \right], \qquad i = 1,\ldots,n,$$

but unfortunately this program is not linear or even convex.

3.2 LINEAR SYSTEMS AND QUADRATIC COST

Consider the case of the linear system

$$x_{k+1} = Ax_k + Bu_k + w_k, \qquad k = 0,1,\ldots,$$

where $x_k \in \Re^n$, $u_k \in \Re^m$ for all k, and the matrices A, B are known. As in Sections 4.1 and 5.2 of Vol. I, we assume that the random disturbances

w_k are independent with zero mean and finite second moments. The cost function is quadratic and has the form

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

where Q is a positive semidefinite symmetric $n \times n$ matrix and R is a positive definite symmetric $m \times m$ matrix. Clearly, Assumption P of Section 3.1 holds.

Our approach will be to use the DP algorithm to obtain the functions TJ_0, $T^2J_0, \ldots$, as well as the pointwise limit function $J_\infty = \lim_{k\to\infty} T^k J_0$. Subsequently, we show that J_∞ satisfies $J_\infty = TJ_\infty$ and hence, by Prop. 1.5(a) of Section 3.1, $J_\infty = J^*$. The optimal policy is then obtained from the optimal cost function J^* by minimizing in Bellman's equation (cf. Prop. 1.3 of Section 3.1).

As in Section 4.1 of Vol. I, we have

$$J_0(x) = 0, \qquad x \in \Re^n,$$

$$(TJ_0)(x) = \min_u [x'Qx + u'Ru] = x'Qx, \qquad x \in \Re^n,$$

$$\begin{aligned}(T^2J_0)(x) &= \min_u E\{x'Qx + u'Ru + \alpha(Ax + Bu + w)'Q(Ax + Bu + w)\} \\ &= x'K_1x + \alpha E\{w'Qw\}, \qquad x \in \Re^n.\end{aligned}$$

$$(T^{k+1}J_0)(x) = x'K_kx + \sum_{m=0}^{k-1} \alpha^{k-m} E\{w'K_m w\}, \qquad x \in \Re^n, \quad k = 1,2,\ldots,$$

where the matrices $K_0, K_1, K_2, \ldots$ are given recursively by

$$K_0 = Q,$$

$$K_{k+1} = A'\big(\alpha K_k - \alpha^2 K_k B(\alpha B'K_kB + R)^{-1}B'K_k\big)A + Q, \qquad k = 0,1,\ldots$$

By defining $\tilde{R} = R/\alpha$ and $\tilde{A} = \sqrt{\alpha}A$, the preceding equation may be written as

$$K_{k+1} = \tilde{A}'\big(K_k - K_kB(B'K_kB + \tilde{R})^{-1}B'K_k\big)\tilde{A} + Q,$$

and is of the form considered in Section 4.1 of Vol. I. By using the result shown there, we have that the generated matrix sequence $\{K_k\}$ converges to a positive definite symmetric matrix K,

$$K_k \to K,$$

provided the pairs $(\tilde{A}, B)$ and $(\tilde{A}, C)$, where $Q = C'C$, are controllable and observable, respectively. Since $\tilde{A} = \sqrt{\alpha}A$, controllability and observability

of (A, B) or (A, C) are clearly equivalent to controllability and observability of $(\hat{A}, B)$ or $(\hat{A}, C)$, respectively. The matrix K is the unique solution of the equation

$$K = A'\big(\alpha K - \alpha^2 KB(\alpha B'KB + R)^{-1}B'K\big)A + Q. \tag{2.1}$$

Because $K_k \to K$, it can also be seen that the limit

$$c = \lim_{k\to\infty} \sum_{m=0}^{k-1} \alpha^{k-m} E\{w'K_m w\}$$

is well defined, and in fact

$$c = \frac{\alpha}{1-\alpha} E\{w'Kw\}. \tag{2.2}$$

Thus, in conclusion, if the pairs (A, B) and (A, C) are controllable and observable, respectively, the limit of the functions $T^k J_0$ is given by

$$J_\infty(x) = \lim_{k\to\infty} (T^k J_0)(x) = x'Kx + c. \tag{2.3}$$

Using Eqs. (2.1) to (2.3), it can be verified by straightforward calculation that for all $x \in S$

$$J_\infty(x) = (TJ_\infty)(x) = \min_u \big[x'Qx + u'Ru + \alpha E\big\{J_\infty(Ax + Bu + w)\big\}\big] \tag{2.4}$$

and hence, by Prop. 1.5(a) of Section 3.1, $J_\infty = J^*$. Another method for proving that $J_\infty = TJ_\infty$ is to show that the assumption of Prop. 1.7 of Section 3.1, is satisfied; that is, the sets

$$U_k(x, \lambda) = \big\{u \mid E\big\{x'Qx + u'Ru + \alpha(T^k J_0)(Ax + Bu + w)\big\} \le \lambda\big\}$$

are compact for all k and scalars λ. This can be verified using the fact that $T^k J_0$ is a positive semidefinite quadratic function and R is positive definite. The optimal stationary policy μ^*, obtained by minimization in Eq. (2.4), has the form

$$\mu^*(x) = -\alpha(\alpha B'KB + R)^{-1}B'KAx, \qquad x \in \Re^n.$$

This policy is attractive for practical implementation since it is linear and stationary. A number of generalized versions of the problem of this section, including the case of imperfect state information, are treated in the exercises. Interestingly, the problem can be solved by policy iteration (see Exercise 3.16), even though, as discussed in Section 3.1, policy iteration is not valid in general under Assumption P.

3.3 INVENTORY CONTROL

Let us consider a discounted, infinite horizon version of the inventory control problem of Section 4.2 in Vol. I. Inventory stock evolves according to the equation

$$x_{k+1} = x_k + u_k - w_k, \qquad k = 0, 1, \ldots \tag{3.1}$$

We assume that the successive demands w_k are independent and bounded, and have identical probability distributions. We also assume for simplicity that there is no fixed cost. The case of a nonzero fixed cost can be treated similarly. The cost function is

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

where

$$H(y) = p\max(0, -y) + h\max(0, y).$$

The DP algorithm is given by

$$J_0(x) = 0,$$

$$(T^{k+1}J_0)(x) = \min_{0\le u} E\big\{cu + H(x + u - w) + \alpha(T^k J_0)(x + u - w)\big\}. \tag{3.2}$$

We first show that the optimal cost is finite for all initial states, that is,

$$J^*(x_0) = \min_\pi J_\pi(x_0) < \infty, \qquad \text{for all } x_0 \in S. \tag{3.3}$$

Indeed, consider the policy $\tilde{\pi} = \{\tilde{\mu}, \tilde{\mu}, \ldots\}$, where $\tilde{\mu}$ is defined by

$$\tilde{\mu}(x) = \begin{cases} 0 & \text{if } x \ge 0, \\ -x & \text{if } x < 0. \end{cases}$$

Since w_k is nonnegative and bounded, it follows that the inventory stock x_k when the policy $\tilde{\pi}$ is used satisfies

$$-w_{k-1} \le x_k \le \max(0, x_0), \qquad k = 1, 2, \ldots,$$

and is bounded. Hence $\tilde{\mu}(x_k)$ is also bounded. It follows that the cost per stage incurred when $\tilde{\pi}$ is used is bounded, and in view of the presence of the discount factor we have

$$J_{\tilde{\pi}}(x_0) < \infty, \qquad x_0 \in S.$$

Since $J^* \le J_{\tilde{\pi}}$, the finiteness of the optimal cost follows.

Next we observe that, under the assumption $c < p$, the functions $T^k J_0$ are real-valued and convex. Indeed, we have

$$J_0 \leq TJ_0 \leq \cdots \leq T^k J_0 \leq \cdots \leq J^*,$$

which implies that $T^k J_0$ is real-valued. Convexity follows by induction as shown in Section 4.2 of Vol. I.

Consider now the sets

$$U_k(x,\lambda) = \big\{u \geq 0 \mid E\big\{cu + H(x+u-w) + \alpha(T^k J_0)(x_u - w)\big\} \leq \lambda\big\}. \quad (3.4)$$

These sets are bounded since the expected value within the braces above tends to ∞ as $u \to \infty$. Also, the sets $U_k(x,\lambda)$ are closed since the expected value in Eq. (3.4) is a continuous function of u [recall that $T^k J_0$ is a real-valued convex and hence continuous function]. Thus we may invoke Prop. 1.7 of Section 3.1 and assert that

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

It follows from the convexity of the functions $T^k J_0$ that the limit function J^* is a real-valued convex function. Furthermore, an optimal stationary policy μ^* can be obtained by minimizing in the right-hand side of Bellman's equation

$$J^*(x) = \min_{u\geq 0} E\big\{cu + H(x+u-w) + \alpha J^*(x+u-w)\big\}.$$

We have

$$\mu^*(x) = \begin{cases} S^* - x & \text{if } x \leq S^*, \\ 0 & \text{otherwise,} \end{cases}$$

where S^* is a minimizing point of

$$G^*(y) = cy + L(y) + \alpha E\big\{J^*(y-w)\big\},$$

with

$$L(y) = E\big\{H(y-w)\big\}.$$

It can be seen that if $p > c$, we have $\lim_{|y|\to\infty} G^*(y) = \infty$, so that such a minimizing point exists. Furthermore, by using the observation made near the end of Section 3.1, it follows that a minimizing point S^* of $G^*(y)$ may be obtained as a limit point of a sequence $\{S_k\}$, where for each k the scalar S_k minimizes

$$G_k(y) = cy + L(y) + \alpha E\big\{(T^k J_0)(y-w)\big\}$$

and is obtained by means of the value iteration method.

It turns out that the critical level S^* has a simple characterization. It can be shown that S^* minimizes over y the expression $(1-\alpha)cy + L(y)$, and it can be essentially obtained in closed form (see Exercise 3.18, and [HeS84], Ch. 2).

In the case where there is a positive fixed cost ($K > 0$), the same line of argument may be used. Similarly, we prove that J^* is a real-valued K-convex function. A separate argument is necessary to prove that J^* is also continuous (this is intuitively clear and is left for the reader). Once K-convexity and continuity of J^* are established, the optimality of a stationary (s^*, S^*) policy follows from the equation

$$J^*(x) = \min_{u\geq 0} E\big\{C(u) + H(x+u-w) + \alpha J^*(x+u-w)\big\},$$

where $C(u) = K + cu$ if $u > 0$ and $C(0) = 0$.

3.4 OPTIMAL STOPPING

Consider an infinite horizon version of the stopping problems of Section 4.4 of Vol. I. At each state x, we must choose between two actions: pay a stopping cost $s(x)$ and *stop*, or pay a cost $c(x)$ and *continue* the process according to the system equation

$$x_{k+1} = f_c(x_k, w_k), \qquad k = 0, 1, \ldots \quad (4.1)$$

The objective is to find the optimal stopping policy that minimizes the total expected cost over an infinite number of stages. It is assumed that the input disturbances w_k have the same probability distribution for all k, which depends only on the current state x_k.

This problem may be viewed as a special case of the stochastic shortest path problem of Section 2.1, but here we will not assume that the state space is finite and that only proper policies can be optimal, as we did in Section 2.1. Instead we will rely on the general theory of unbounded cost problems developed in Section 3.1.

To put the problem within the framework of the total cost infinite horizon problem, we introduce an additional state t (termination state) and we complete the system equation (4.1) as in Section 4.4 of Vol. I by letting

$$x_{k+1} = t, \qquad \text{if } u_k = \text{stop} \quad \text{or} \quad x_k = t.$$

Once the system reaches the termination state, it remains there permanently at no cost.

We first assume that

$$s(x) \geq 0, \qquad c(x) \geq 0, \qquad \text{for all } x \in S, \quad (4.2)$$

thus coming under the framework of Assumption P of Section 3.1. The case corresponding to Assumption N, where $s(x) \leq 0$ and $c(x) \leq 0$ for all $x \in S$ will be considered later. Actually, whenever there exists an $\epsilon > 0$ such that $c(x) \geq \epsilon$ for all $x \in S$, the results to be obtained under the assumption (4.2) apply also to the case where $s(x)$ is bounded below by some scalar rather than bounded by zero. The reason is that, if $c(x)$ is assumed to be greater than $\epsilon > 0$ for all $x \in S$, any policy that will not stop within a finite expected number of stages results in infinite cost and can be excluded from consideration. As a result, if we reformulate the problem and add a constant r to $s(x)$ so that $s(x) + r \geq 0$ for all $x \in S$, the optimal cost $J^*(x)$ will merely be increased by r, while optimal policies will remain unaffected.

The mapping T that defines the DP algorithm takes the form

$$(TJ)(x) = \begin{cases} \min\big[s(x),\, c(x) + E\big\{J\big(f_c(x,w)\big)\big\}\big] & \text{if } x \neq t, \\ 0 & \text{if } x = t, \end{cases} \tag{4.3}$$

where $s(x)$ is the cost of the stopping action, and $c(x) + E\big\{J\big(f_c(x,w)\big)\big\}$ is the cost of the continuation action. Since the control space has only two elements, by Prop. 1.6 of Section 3.1, we have

$$\lim_{k \to \infty} (T^k J_0)(x) = J^*(x), \qquad x \in S, \tag{4.4}$$

where J_0 is the zero function [$J_0(x) = 0$, for all $x \in S$]. By Prop. 1.3 of Section 3.1, there exists a stationary optimal policy given by

$$\text{stop} \qquad \text{if } s(x) < c(x) + E\big\{J^*\big(f_c(x,w)\big)\big\},$$

$$\text{continue} \quad \text{if } s(x) \geq c(x) + E\big\{J^*\big(f_c(x,w)\big)\big\}.$$

Let us denote by S^* the optimal stopping set (which may be empty)

$$S^* = \big\{x \in S \mid s(x) < c(x) + E\big\{J^*\big(f_c(x,w)\big)\big\}\big\}.$$

Consider also the sets

$$S_k = \big\{x \in S \mid s(x) < c(x) + E\big\{(T^k J_0)\big(f_c(x,w)\big)\big\}\big\}$$

that determine the optimal policy for finite horizon versions of the stopping problem. Since we have

$$J_0 \leq TJ_0 \leq \cdots \leq T^k J_0 \leq \cdots \leq J^*,$$

it follows that

$$S_1 \subset S_2 \subset \cdots \subset S_k \subset \cdots \subset S^*$$

and therefore $\cup_{k=1}^{\infty} S_k \subset S^*$. Also, if $\tilde{x} \notin \cup_{k=1}^{\infty} S_k$, then we have

$$s(\tilde{x}) \geq c(\tilde{x}) + E\big\{(T^k J_0)\big(f_c(\tilde{x},w)\big)\big\}, \qquad k = 0, 1, \ldots,$$

and by taking the limit and using the monotone convergence theorem and the fact $T^k J_0 \to J^*$, we obtain

$$s(\tilde{x}) \geq c(\tilde{x}) + E\big\{J^*\big(f_c(\tilde{x},w)\big)\big\},$$

from which $\tilde{x} \notin S^*$. Hence

$$S^* = \cup_{k=1}^{\infty} S_k. \tag{4.5}$$

In other words, the *optimal stopping set S^* for the infinite horizon problem is equal to the union of all the finite horizon stopping sets S_k.*

Consider now, as in Section 4.4 of Vol. I, the one-step-to-go stopping set

$$\tilde{S}_1 = \big\{x \in S \mid s(x) \leq c(x) + E\big\{J\big(f_c(x,w)\big)\big\}\big\} \tag{4.6}$$

and assume that $\tilde{S}_1$ is *absorbing* in the sense

$$f_c(x,w) \in \tilde{S}_1, \qquad \text{for all } x \in \tilde{S}_1, \quad w \in D. \tag{4.7}$$

Then, as in Section 4.4 of Vol. I, it follows that the one-step lookahead policy

$$\text{stop if and only if } x \in \tilde{S}_1$$

is optimal. We now provide some examples.

Example 4.1 (Asset Selling)

Consider the version of the asset selling example of Sections 4.4 and 7.3 of Vol. I, where the rate of interest r is zero and there is instead a maintenance cost $c > 0$ per period for which the house remains unsold. Furthermore, past offers can be accepted at any future time. We have the following optimality equation:

$$J^*(x) = \max\big[x,\, -c + E\big\{J^*\big(\max(x,w)\big)\big\}\big].$$

In this case we consider maximization of total expected reward, the continuation cost is strictly negative, and the stopping reward x is positive. Hence the assumption (4.2) is not satisfied. If, however, we assume that x takes values in a bounded interval $[0, M]$, where M is an upper bound on the possible values of offers, our analysis is still applicable [cf. the discussion following Eq. (4.2)]. Consider the one-step-to-go stopping set given by

$$\tilde{S}_1 = \big\{x \mid x \geq -c + E\big\{\max(x,w)\big\}\big\}.$$

After a calculation similar to the one given in Section 4.4 of Vol. I, we see that

$$\tilde{S}_1 = \{x \mid x \geq \overline{\alpha}\},$$

where $\overline{\alpha}$ is the scalar satisfying

$$\overline{\alpha} = P(\overline{\alpha})\overline{\alpha} + \int_{\overline{\alpha}}^{\infty} w \, dP(w) - c.$$

Clearly, $\tilde{S}_1$ is absorbing in the sense of Eq. (4.7) and therefore the one-step lookahead policy that accepts the first offer greater than or equal to $\overline{\alpha}$ is optimal.

Example 4.2 (Sequential Hypothesis Testing)

Consider the hypothesis testing problem of Section 5.5 of Vol. I for the case where the number of possible observations is unlimited. Here the states are x^0 and x^1 (true distribution of the observations is f_0 and f_1, respectively). The set S is the interval $[0,1]$ and corresponds to the sufficient statistic

$$p_k = P(x_k = x^0 \mid z_0, z_1, \ldots, z_k).$$

To each $p \in [0,1]$ we may assign the stopping cost

$$s(p) = \min\big[(1-p)L_0, pL_1\big],$$

that is, the cost associated with optimal choice between the distributions f_0 and f_1. The mapping T of Eq. (4.3) takes the form

$$(TJ)(p) = \min\left[(1-p)L_0, pL_1, c + E_z\left\{J\left(\frac{pf_0(z)}{pf_0(z)+(1-p)f_1(z)}\right)\right\}\right]$$

for all $p \in [0,1]$, where the expectation over z is taken with respect to the probability distribution

$$P(z) = pf_0(z) + (1-p)f_1(z), \qquad z \in Z.$$

The optimal cost function J^* satisfies Bellman's equation

$$J^*(p) = \min\left[(1-p)L_0, pL_1, c + E_z\left\{J^*\left(\frac{pf_0(z)}{pf_0(z)+(1-p)f_1(z)}\right)\right\}\right]$$

and is obtained in the limit through the equation

$$J^*(p) = \lim_{k\to\infty}(T^k J_0)(p), \qquad p \in [0,1],$$

where J_0 is the zero function on $[0,1]$.

Now consider the functions $T^k J_0$, $k = 0, 1, \ldots$ It is clear that

$$J_0 \le TJ_0 \le \cdots \le T^k J_0 \le \cdots \le \min\big[(1-p)L_0, pL_1\big].$$

Furthermore, in view of the analysis of Section 5.5 of Vol. I, we have that the function $T^k J_0$ is concave on $[0,1]$ for all k. Hence the pointwise limit function J^* is also concave on $[0,1]$. In addition, Bellman's equation implies that

$$J^*(0) = J^*(1) = 0,$$

$$J^*(p) \le \min\big[(1-p)L_0, pL_1\big].$$

Using the reasoning illustrated in Fig. 3.4.1 it follows that [provided $c < L_0L_1/(L_0+L_1)$] there exist two scalars $\overline{\alpha}$, $\overline{\beta}$ with $0 < \overline{\beta} \le \overline{\alpha} < 1$, that determine an optimal stationary policy of the form

$$\begin{aligned} &\text{accept } f_0 && \text{if } p \le \overline{\alpha},\\ &\text{accept } f_1 && \text{if } p \le \overline{\beta},\\ &\text{continue the observations} && \text{if } \overline{\beta} < p < \overline{\alpha}. \end{aligned}$$

In view of the optimality of the preceding stationary policy, the sequential probability ratio test described in Section 5.5 of Vol. I is justified when the number of possible observations is infinite.

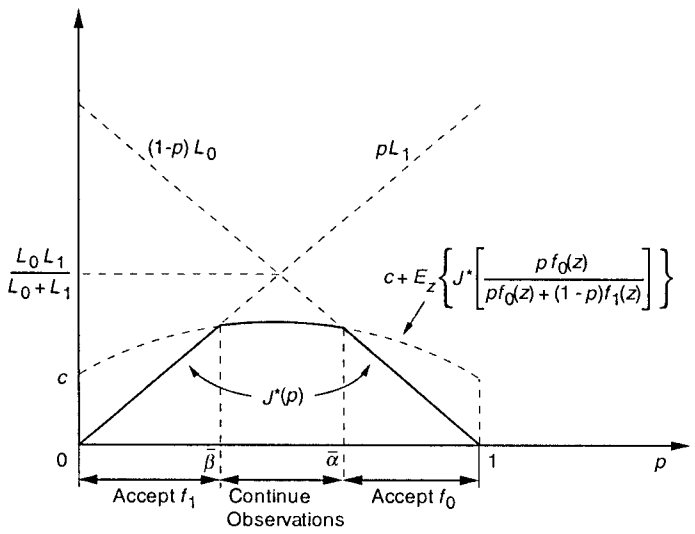


Figure 3.4.1 Derivation of the sequential probability ratio test.

The Case of Negative Transition Costs

We now consider the stopping problem under Assumption N, that is,

$$s(x) \le 0, \qquad c(x) \le 0, \qquad \text{for all } x \in S.$$

Under these circumstances there is no penalty for continuing operation of the system (although by not stopping at a given state, a favorable opportunity may be missed). The mapping T is given by

$$(TJ)(x) = \min\big[s(x), c(x) + E\big\{J\big(f_c(x,w)\big)\big\}\big].$$

The optimal cost function J^* satisfies $J^*(x) \le s(x)$ for all $x \in S$, and by using Props. 1.1 and 1.5(b) of Section 3.1, we have

$$J^* = TJ^*, \qquad J^* = \lim_{k\to\infty} T^k J_0 = \lim_{k\to\infty} T^k s,$$

where J_0 is the zero function. It can also be seen that if the one-step-to-go stopping set $\tilde{S}_1$ is *absorbing* [cf. Eq. (4.7)], a one-step lookahead policy is optimal.

Example 4.3 (The Rational Burglar)

This example was considered at the end of Section 4.4 of Vol. I where it was shown that a one-step lookahead policy is optimal for any finite horizon length. The optimality equation is

$$J^*(x) = \max\left[x, (1-p)E\left\{J^*(x+w)\right\}\right].$$

The problem is equivalent to a minimization problem where

$$s(x) = -x, \qquad c(x) = 0,$$

so Assumption N holds. From the preceding analysis, we have that $T^k s \to J^*$ and that a one-step lookahead policy is optimal if the one-step stopping set is absorbing [cf. Eqs. (4.6) and (4.7)]. It can be shown (see the analysis of Section 4.4 of Vol. I) that this condition holds, so the finite horizon optimal policy whereby the burglar retires when his accumulated earnings reach or exceed $(1-p)\overline{w}/p$ is optimal for an infinite horizon as well.

Example 4.4 (A Problem with no Optimal Policy)

This is a deterministic stopping problem where Assumption N holds, and an optimal policy does not exist, even though only two controls are available at each state (stop and continue). The states are the positive integers, and continuation from state i leads to state $i+1$ with certainty and no cost, that is, $S = \{1, 2, \ldots\}$, $c(i) = 0$, and $f_c(i, w) = i + 1$ for all $i \in S$ and $w \in D$. The stopping cost is $s(i) = -1 + (1/i)$ for all $i \in S$, so that there is an incentive to delay stopping at every state. We have $J^*(i) = -1$ for all i, and the optimal cost -1 can be approached arbitrarily closely by postponing the stopping action for a sufficiently long time. However, there does not exist an optimal policy that attains the optimal cost.

3.5 OPTIMAL GAMBLING STRATEGIES

A gambler enters a certain game played as follows. The gambler may stake at any time k any amount $u_k \geq 0$ that does not exceed his current fortune x_k (defined to be his initial capital plus his gain or minus his loss thus far). He wins his stake back and as much more with probability p and he loses his stake with probability $(1-p)$. Thus the gambler's fortune evolves according to the equation

$$x_{k+1} = x_k + w_k u_k, \qquad k = 0, 1, \ldots, \tag{5.1}$$

where $w_k = 1$ with probability p and $w_k = -1$ with probability $(1-p)$. Several games, such as playing red and black in roulette, fit this description.

The gambler enters the game with an initial capital x_0, and his goal is to increase his fortune up to a level X. He continues gambling until he either reaches his goal or loses his entire initial capital, at which point he leaves the game. The problem is to determine the optimal gambling strategy for maximizing the probability of reaching his goal. By a gambling strategy, we mean a rule that specifies what the stake should be at time k when the gambler's fortune is x_k, for every x_k with $0 < x_k < X$.

The problem may be cast within the total cost, infinite horizon framework, where we consider maximization in place of minimization. Let us assume for convenience that fortunes are normalized so that $X = 1$. The state space is the set $[0, 1] \cup \{t\}$, where t is a termination state to which the system moves with certainty from both states 0 and 1 with corresponding rewards 0 and 1. When $x_k \neq 0$, $x_k \neq 1$, the system evolves according to Eq. (5.1). The control constraint set is specified by

$$0 \leq u_k \leq x_k, \qquad 0 \leq u_k \leq 1 - x_k.$$

The reward per stage when $x_k \neq 0$ and $x_k \neq 1$ is zero. Under these circumstances the probability of reaching the goal is equal to the total expected reward. Assumption N holds since our problem is equivalent to a problem of minimizing expected total cost with nonpositive costs per stage.

The mapping T defining the DP algorithm takes the form

$$(TJ)(x) = \begin{cases} \max\limits_{\substack{0 \leq u \leq x \\ 0 \leq u \leq 1-x}} \left[pJ(x+u) + (1-p)J(x-u)\right] & \text{if } x \in (0, 1), \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x = 1, \end{cases} \tag{5.2}$$

for any function $J : [0, 1] \mapsto [0, \infty]$.

Consider now the case where

$$0 < p < \frac{1}{2},$$

that is, the game is unfair to the gambler. A discretized version of the case where $1/2 \leq p < 1$ is considered in Exercise 3.21. When $0 < p < 1/2$, it is intuitively clear that if the gambler follows a very conservative strategy and stakes a very small amount at each time, he is all but certain to lose his capital. For example, if the gambler adopts a strategy of betting $1/n$ at each time, then it may be shown (see Exercise 3.21 or [Ash70], p. 182) that his probability of attaining the target fortune of 1 starting with an initial capital i/n, $0 < i < n$, is given by

$$\left(\left(\frac{1-p}{p}\right)^i - 1\right)\left(\left(\frac{1-p}{p}\right)^n - 1\right)^{-1}.$$

If $0 < p < 1/2$, n tends to infinity, and i/n tends to a constant, the above probability tends to zero, thus indicating that placing consistently small bets is a bad strategy.

We are thus led to a policy that places large bets and, in particular, the *bold strategy* whereby the gambler stakes at each time k his entire fortune x_k or just enough to reach his goal, whichever is least. In other words, the bold strategy is the stationary policy μ^* given by

$$\mu^*(x) = \begin{cases} x & \text{if } 0 < x \leq 1/2, \\ 1 - x & \text{if } 1/2 \leq x < 1. \end{cases}$$

We will prove that the bold strategy is indeed an optimal policy. To this end it is sufficient to show that for every initial fortune $x \in [0, 1]$ the value of the reward function $J_{\mu^*}(x)$ corresponding to the bold strategy μ^* satisfies the sufficiency condition (cf. Prop. 1.4, Section 3.1)

$$TJ_{\mu^*} = J_{\mu^*},$$

or equivalently

$$J_{\mu^*}(0) = 0, \qquad J_{\mu^*}(1) = 1,$$

$$J_{\mu^*}(x) \geq pJ_{\mu^*}(x + u) + (1 - p)J_{\mu^*}(x - u), \tag{5.3}$$

for all $x \in (0, 1)$ and $u \in [0, x] \cap [0, 1 - x]$.

By using the definition of the bold strategy, Bellman's equation

$$J_{\mu^*} = T_{\mu^*}J_{\mu^*},$$

is written as

$$J_{\mu^*}(0) = 0, \qquad J_{\mu^*}(1) = 1, \tag{5.4}$$

$$J_{\mu^*}(x) = \begin{cases} pJ_{\mu^*}(2x) & \text{if } 0 < x \leq 1/2, \\ p + (1 - p)J_{\mu^*}(2x - 1) & \text{if } 1/2 \leq x < 1. \end{cases} \tag{5.5}$$

The following lemma shows that J_{μ^*} is uniquely defined from these relations.

Lemma 5.1: For every p, with $0 < p \leq 1/2$, there is only one bounded function on $[0, 1]$ satisfying Eqs. (5.4) and (5.5), the function J_{μ^*}. Furthermore, J_{μ^*} is continuous and strictly increasing on $[0, 1]$.

Proof: Suppose that there existed two bounded functions $J_1 : [0, 1] \mapsto \Re$ and $J_2 : [0, 1] \mapsto \Re$ such that $J_i(0) = 0$, $J_i(1) = 1$, $i = 1, 2$, and

$$J_i(x) = \begin{cases} pJ_i(2x) & \text{if } 0 < x \leq 1/2, \\ p + (1 - p)J_i(2x - 1) & \text{if } 1/2 \leq x < 1, \end{cases} \qquad i = 1, 2.$$

Then we have

$$J_1(2x) - J_2(2x) = \frac{J_1(x) - J_2(x)}{p}, \qquad \text{if } 0 \leq x \leq 1/2, \tag{5.6}$$

$$J_1(2x - 1) - J_2(2x - 1) = \frac{J_1(x) - J_2(x)}{1 - p}, \qquad \text{if } 1/2 \leq x < 1. \tag{5.7}$$

Let z be any real number with $0 \leq z \leq 1$. Define

$$z_1 = \begin{cases} 2z & \text{if } 0 \leq z \leq 1/2, \\ 2z - 1 & \text{if } 1/2 < z \leq 1, \end{cases}$$

$$\vdots$$

$$z_k = \begin{cases} 2z_{k-1} & \text{if } 0 \leq z_{k-1} \leq 1/2, \\ 2z_{k-1} - 1 & \text{if } 1/2 < z_{k-1} \leq 1, \end{cases}$$

for $k = 1, 2, \ldots$. Then from Eqs. (5.6) and (5.7) it follows (using $p \leq 1/2$) that

$$\left|J_1(z_k) - J_1(z_k)\right| \geq \frac{\left|J_1(z) - J_2(z)\right|}{(1 - p)^k}, \qquad k = 1, 2, \ldots$$

Since $J_1(z_k) - J_2(z_k)$ is bounded, it follows that $J_1(z) - J_2(z) = 0$, for otherwise the right side of the inequality would tend to ∞. Since $z \in [0, 1]$ is arbitrary, we obtain $J_1 = J_2$. Hence J_{μ^*} is the unique bounded function on $[0, 1]$ satisfying Eqs. (5.4) and (5.5).

To show that J_{μ^*} is strictly increasing and continuous, we consider the mapping T_{μ^*}, which operates on functions $J : [0, 1] \mapsto [0, 1]$ and is defined by

$$(T_{\mu^*}J)(x) = \begin{cases} pJ(2x) + (1 - p)J(0) & \text{if } 0 < x \leq 1/2, \\ pJ(1) + (1 - p)J(2x - 1) & \text{if } 1/2 \leq x < 1, \end{cases}$$

$$(T_{\mu^*}J)(0) = 0, \qquad (T_{\mu^*}J)(1) = 1. \tag{5.8}$$

Consider the functions J_0, $T_{\mu^*}J_0, \cdots, T^k_{\mu^*}J_0, \ldots$, where J_0 is the zero function [$J_0(x) = 0$ for all $x \in [0, 1]$]. We have

$$J_{\mu^*}(x) = \lim_{k \to \infty} (T^k_{\mu^*}J_0)(x), \qquad x \in [0, 1]. \tag{5.9}$$

Furthermore, the functions $T^k_{\mu^*}J_0$ can be shown to be monotonically nondecreasing in the interval $[0, 1]$. Hence, by Eq. (5.9), J_{μ^*} is also monotonically nondecreasing.

Consider now for $n = 0, 1, \ldots$ the sets

$$S_n = \left\{x \in [0, 1] \mid x = k2^{-n}, k = \text{nonnegative integer}\right\}.$$

It is straightforward to verify that

$$(T_{\mu^*}^m J_0)(x) = (T_{\mu^*}^n J_0)(x), \qquad x \in S_{n-1}, \quad m \geq n \geq 1.$$

As a result of this equality and Eq. (5.9),

$$J_{\mu^*}(x) = (T_{\mu^*}^n J_0)(x), \qquad x \in S_{n-1}, \quad n \geq 1. \tag{5.10}$$

A further fact that may be verified by using induction and Eqs. (5.8) and (5.10) is that for any nonnegative integers k, n for which $0 \leq k2^{-n} < (k+1)2^{-n} \leq 1$, we have

$$p^n \leq J_{\mu^*}\big((k+1)2^{-n}\big) - J_{\mu^*}\big(k2^{-n}\big) \leq (1-p)^n. \tag{5.11}$$

Since any number in $[0,1]$ can be approximated arbitrarily closely from above and below by numbers of the form $k2^{-n}$, and since J_{μ^*} has been shown to be monotonically nondecreasing, it follows from Eq. (5.11) that J_{μ^*} is continuous and strictly increasing. **Q.E.D.**

We are now in a position to prove the following proposition.

Proposition 5.1: The bold strategy is an optimal stationary gambling policy.

Proof: We will prove the sufficiency condition

$$J_{\mu^*}(x) \geq pJ_{\mu^*}(x+u) + (1-p)J_{\mu^*}(x-u), \qquad x \in [0,1], \quad u \in [0,1] \cap [0,1-x]. \tag{5.12}$$

In view of the continuity of J_{μ^*} established in the previous lemma, it it sufficient to establish Eq. (5.12) for all $x \in [0,1]$ and $u \in [0,x] \cap [0,1-x]$ that belong to the union $\cup_{n=0}^{\infty} S_n$ of the sets S_n defined by

$$S_n = \big\{ z \in [0,1] \mid z = k2^{-n},\ k = \text{nonnegative integer} \big\}.$$

We will use induction. By using the fact that $J_{\mu^*}(0) = 0$, $J_{\mu^*}(1/2) = p$, and $J_{\mu^*}(1) = 1$, we can show that Eq. (5.12) holds for all x and u in S_0 and S_1. Assume that Eq. (5.12) holds for all $x, u \in S_n$. We will show that it holds for all $x, u \in S_{n+1}$.

For any $x, u \in S_{n+1}$ with $u \in [0,x] \cap [0,1-x]$, there are four possibilities:

1. $x + u \leq 1/2$,
2. $x - u \geq 1/2$,
3. $x - u \leq x \leq 1/2 \leq x + u$,
4. $x - u \leq 1/2 \leq x \leq x + u$,

We will prove Eq. (5.12) for each of these cases.

Case 1. If $x, u \in S_{n+1}$, then $2x \in S_n$, and $2u \in S_n$, and by the induction hypothesis

$$J_{\mu^*}(2x) - pJ_{\mu^*}(2x+2u) - (1-p)J_{\mu^*}(2x-2u) \geq 0. \tag{5.13}$$

If $x + u \leq 1/2$, then by Eq. (5.5)

$$\begin{aligned} J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u) \\ = p\big(J_{\mu^*}(2x) - pJ_{\mu^*}(2x+2u) - (1-p)J_{\mu^*}(2x-2u)\big) \end{aligned}$$

and using Eq. (5.13), the desired relation Eq. (5.12) is proved for the case under consideration.

Case 2. If $x, u \in S_{n+1}$, then $(2x-1) \in S_n$ and $2u \in S_n$, and by the induction hypothesis

$$J_{\mu^*}(2x-1) - pJ_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u-1) \geq 0.$$

If $x - u \geq 1/2$, then by Eq. (5.5)

$$\begin{aligned} &J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u) \\ &\quad = p + (1-p)J_{\mu^*}(2x-1) - p\big(p + (1-p)J_{\mu^*}(2x+2u-1)\big) \\ &\qquad - (1-p)\big(p + (1-p)J_{\mu^*}(2x-2u-1)\big) \\ &\quad = (1-p)\big(J_{\mu^*}(2x-1) - pJ_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u-1)\big) \\ &\quad \geq 0, \end{aligned}$$

and Eq. (5.12) follows from the preceding relations.

Case 3. Using Eq. (5.5), we have

$$\begin{aligned} J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u) \\ = pJ_{\mu^*}(2x) - p\big(p + (1-p)J_{\mu^*}(2x+2u-1)\big) - p(1-p)J_{\mu^*}(2x-2u) \\ = p\big(J_{\mu^*}(2x) - p - (1-p)J_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u)\big). \end{aligned}$$

Now we must have $x \geq \frac{1}{4}$, for otherwise $u < \frac{1}{4}$ and $x + u < 1/2$. Hence $2x \geq 1/2$ and the sequence of equalities can be continued as follows:

$$\begin{aligned} J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u) \\ = p\big(p + (1-p)J_{\mu^*}(4x-1) - p \\ \qquad - (1-p)J_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u)\big) \\ = p(1-p)\big(J_{\mu^*}(4x-1) - J_{\mu^*}(2x+2u-1) - J_{\mu^*}(2x-2u)\big) \\ = (1-p)\big(J_{\mu^*}(2x-1/2) - pJ_{\mu^*}(2x+2u-1) - pJ_{\mu^*}(2x-2u)\big). \end{aligned}$$

Since $p \leq (1-p)$, the last expression is greater than or equal to both

$$(1-p)\big(J_{\mu^*}(2x-1/2) - pJ_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u)\big)$$

and

$$(1-p)\big(J_{\mu^*}(2x-1/2) - (1-p)J_{\mu^*}(2x+2u-1) - pJ_{\mu^*}(2x-2u)\big).$$

Now for $x, u \in S_{n+1}$, and $n \geq 1$, we have $(2x-1/2) \in S_n$ and $(2u-1/2) \in S_n$ if $(2u-1/2) \in [0,1]$, and $(1/2-2u) \in S_n$ if $(1/2-2u) \in [0,1]$. By the induction hypothesis, the first or the second of the preceding expressions is nonnegative, depending on whether $2x+2u-1 \geq 2x-1/2$ or $2x-2u \geq 2x-1/2$ (i.e., $u \geq \frac{1}{4}$ or $u \leq \frac{1}{4}$). Hence Eq. (5.12) is proved for case 3.

Case 4. The proof resembles the one for case 3. Using Eq. (5.5), we have

$$\begin{aligned}
&J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u)\\
&\quad = p + (1-p)J_{\mu^*}(2x-1) - p\big(p + (1-p)J_{\mu^*}(2x+2u-1)\big)\\
&\qquad - (1-p)pJ_{\mu^*}(2x-2u)\\
&\quad = p(1-p)\\
&\qquad + (1-p)\big(J_{\mu^*}(2x-1) - pJ_{\mu^*}(2x+2u-1) - pJ_{\mu^*}(2x-2u)\big).
\end{aligned}$$

We must have $x \leq \frac{3}{4}$ for otherwise $u < \frac{1}{4}$ and $x-u > \frac{1}{2}$. Hence $0 \leq 2x-1 \leq 1/2 \leq 2x-1/2 \leq 1$, and using Eq. (5.5) we have

$$(1-p)J_{\mu^*}(2x-1) = (1-p)pJ_{\mu^*}(4x-2) = p\big(J_{\mu^*}(2x-1/2) - p\big).$$

Using the preceding relations, we obtain

$$\begin{aligned}
&J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u)\\
&\quad = p(1-p) + p\big(J_{\mu^*}(2x-1/2) - p\big) - p(1-p)J_{\mu^*}(2x+2u-1)\\
&\qquad - p(1-p)J_{\mu^*}(2x-2u)\\
&\quad = p\big((1-2p) + J_{\mu^*}(2x-1/2) - (1-p)J_{\mu^*}(2x+2u-1)\\
&\qquad - (1-p)J_{\mu^*}(2x-2u)\big).
\end{aligned}$$

These relations are equal to both

$$p\big((1-2p)\big(1 - J_{\mu^*}(2x+2u-1)\big) + J_{\mu^*}(x-1/2) - pJ_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u)\big)$$

and

$$p\big((1-2p)\big(1 - J_{\mu^*}(2x-2u)\big) + J_{\mu^*}(2x-1/2) - (1-p)J_{\mu^*}(2x+2u-1) - pJ_{\mu^*}(2x-2u)\big).$$

Since $0 \leq J_{\mu^*}(2x+2u-1) < 1$ and $0 \leq J_{\mu^*}(2x-2u) < 1$, these expressions are greater than or equal to both

$$p\big(J_{\mu^*}(2x-1/2) - pJ_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u)\big)$$

and

$$p\big(J_{\mu^*}(2x-1/2) - (1-p)J_{\mu^*}(2x+2u-1) - pJ_{\mu^*}(2x-2u)\big)$$

and the result follows as in case 3. **Q.E.D.**

We note that the bold strategy is not the unique optimal stationary gambling strategy. For a characterization of all optimal strategies, see [DuS65], p. 90. Several other gambling problems where strategies of the bold type are optimal are described in [DuS65], Chapters 5 and 6.

3.6 NONSTATIONARY AND PERIODIC PROBLEMS

The standing assumption so far in this chapter has been that the problem involves a stationary system and a stationary cost per stage (except for the presence of the discount factor). Problems with nonstationary system or cost per stage arise occasionally in practice or in theoretical studies and are thus of some interest. It turns out that such problems can be converted to stationary ones by a simple reformulation. We can then obtain results analogous to those obtained earlier for stationary problems.

Consider a nonstationary system of the form

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

and a cost function of the form

$$J_\pi(x_0) = \lim_{N\to\infty} \mathop{E}_{\substack{w_k\\ k=0,1,\ldots,N-1}} \left\{ \sum_{k=0}^{N-1} \alpha^k g_k\big(x_k, \mu_k(x_k), w_k\big) \right\}. \tag{6.1}$$

In these equations, for each k, x_k belongs to a space S_k, u_k belongs to a space C_k and satisfies $u_k \in U_k(x_k)$ for all $x_k \in S_k$, and w_k belongs to a countable space D_k. The sets S_k, C_k, $U_k(x_k)$, D_k may differ from one stage to the next. The random disturbances w_k are characterized by probabilities $P_k(\cdot \mid x_k, u_k)$, which depend on x_k and u_k as well as the time index k. The set of admissible policies Π is the set of all sequences $\pi = \{\mu_0, \mu_1, \ldots\}$ with $\mu_k : S_k \mapsto C_k$ and $\mu_k(x_k) \in U_k(x_k)$ for all $x_k \in S_k$ and $k = 0, 1, \ldots$. The functions $g_k : S_k \times C_k \times D_k \mapsto \Re$ are given and are assumed to satisfy one of the following three assumptions:

Assumption D′: We have $\alpha < 1$, and the functions g_k satisfy, for all $k = 0, 1, \ldots$,

$$\left|g_k(x_k, u_k, w_k)\right| \leq M, \qquad \text{for all } (x_k, u_k, w_k) \in S_k \times C_k \times D_k,$$

where M is some scalar.

Assumption P′: The functions g_k satisfy, for all $k = 0, 1, \ldots$,

$$0 \leq g_k(x_k, u_k, w_k), \qquad \text{for all } (x_k, u_k, w_k) \in S_k \times C_k \times D_k.$$

Assumption N′: The functions g_k satisfy, for all $k = 0, 1, \ldots$,

$$g_k(x_k, u_k, w_k) \leq 0, \qquad \text{for all } (x_k, u_k, w_k) \in S_k \times C_k \times D_k.$$

We will refer to the problem formulated as the *nonstationary problem* (NSP for short). We can get an idea on how the NSP can be converted to a stationary problem by considering the special case where the state space is the same for each stage (i.e., $S_k = S$ for all k). We consider an augmented state

$$\tilde{x} = (x, k),$$

where $x \in S$, and k is the time index. The new state space is $\tilde{S} = S \times K$, where K denotes the set of nonnegative integers. The augmented system evolves according to

$$(x, k) \to \big(f_k(x, u_k, w_k), k + 1\big), \qquad (x, k) \in \tilde{S}.$$

Similarly, we can define a cost per stage as

$$\tilde{g}\big((x, k), u_k, w_k\big) = g_k(x, u_k, w_k), \qquad (x, k) \in \tilde{S}.$$

It is evident that the problem corresponding to the augmented system is stationary. If we restrict attention to initial states $\tilde{x}_0 \in S \times \{0\}$, it can be seen that this stationary problem is equivalent to the NSP.

Let us now consider the more general case. To simplify notation, we will assume that the state spaces S_i, $i = 0, 1, \ldots$, the control spaces C_i, $i = 0, 1, \ldots$, and the disturbance spaces D_i, $i = 0, 1, \ldots$, are all mutually disjoint. This assumption does not involve a loss of generality since, if necessary, we may relabel the elements of S_i, C_i, and D_i without affecting the structure of the problem. Define now a new state space S, a new control space C, and a new (countable) disturbance space D by

$$S = \cup_{i=0}^{\infty} S_i, \qquad C = \cup_{i=0}^{\infty} C_i, \qquad D = \cup_{i=0}^{\infty} D_i.$$

Introduce a new (stationary) system

$$\tilde{x}_{k+1} = f(\tilde{x}_k, \tilde{u}_k, \tilde{w}_k), \qquad k = 0, 1, \ldots, \tag{6.2}$$

where $\tilde{x}_k \in S$, $\tilde{u}_k \in C$, $\tilde{w}_k \in D$, and the system function $f : S \times C \times D \mapsto S$ is defined by

$$f(\tilde{x}, \tilde{u}, \tilde{w}) = f_i(\tilde{w}, \tilde{u}, w), \qquad \text{if } x \in S_i, \quad u \in C_i, \quad w \in D_i, \quad i = 0, 1, \ldots$$

For triplets $(\tilde{x}, \tilde{u}, \tilde{w})$, where for some $i = 0, 1, \ldots$, we have $\tilde{x} \in S_i$, but $\tilde{u} \notin C_i$ or $\tilde{w} \notin D_i$, the definition of f is immaterial; any definition is adequate for our purposes in view of the control constraints to be introduced. The control constraint is taken to be $\tilde{u} \in U(\tilde{x})$ for all $\tilde{x} \in S$, where $U(\cdot)$ is defined by

$$U(\tilde{x}) = U_i(\tilde{x}), \qquad \text{if } \tilde{x} \in S_i, \quad i = 0, 1, \ldots$$

The disturbance $\tilde{w}$ is characterized by probabilities $P(\tilde{w} \mid \tilde{x}, \tilde{u})$ such that

$$P(\tilde{w} \in D_i \mid \tilde{x} \in S_i, \tilde{u} \in C_i) = 1, \qquad i = 0, 1, \ldots,$$

$$P(\tilde{w} \notin D_i \mid \tilde{x} \in S_i, \tilde{u} \in C_i) = 0, \qquad i = 0, 1, \ldots$$

Furthermore, for any $w_i \in D_i$, $x_i \in S_i$, $u_i \in C_i$, $i = 0, 1, \ldots$, we have

$$P(w_i \mid x_i, u_i) = P_i(w_i \mid x_i, u_i).$$

We also introduce a new cost function

$$\tilde{J}_{\tilde{\pi}}(\tilde{x}_0) = \lim_{N \to \infty} \mathop{E}_{\substack{\tilde{w}_k \\ k = 0, 1, \ldots, N-1}} \left\{ \sum_{k=0}^{N-1} \alpha^k g\big(\tilde{x}_k, \mu_k(\tilde{x}_k), \tilde{w}_k\big) \right\}, \tag{6.3}$$

where the (stationary) cost per stage $g : S \times C \times D \mapsto \Re$ is defined for all $i = 0, 1, \ldots$ by

$$g(x, u, w) = g_i(x, \tilde{u}, \tilde{w}), \qquad \text{if } x \in S_i, \quad u \in C_i, \quad w \in D_i.$$

For triplets $(\tilde{x}, \tilde{u}, \tilde{w})$, where for some $i = 0, 1, \ldots$, we have $\tilde{x} \in S_i$ but $\tilde{u} \notin C_i$ or $\tilde{w} \notin D_i$, any definition of g is adequate provided $\left|g(\tilde{x}, \tilde{u}, \tilde{w})\right| \leq M$ for all $(\tilde{x}, \tilde{u}, \tilde{w})$ when Assumption D′ holds, $0 \leq g(\tilde{x}, \tilde{u}, \tilde{w})$ when P′ holds, and

$g(\tilde{x}, \tilde{u}, \tilde{w}) \leq 0$ when N′ holds. The set of admissible policies $\tilde{\Pi}$ for the new problem consists of all sequences $\tilde{\pi} = \{\tilde{\mu}_0, \tilde{\mu}_1, \ldots\}$, where $\tilde{\mu}_k : S \mapsto C$ and $\tilde{\mu}_k(\tilde{x}) \in U(\tilde{x})$ for all $\tilde{x} \in S$ and $k = 0, 1, \ldots$.

The construction given defines a problem that clearly fits the framework of the infinite horizon total cost problem. We will refer to this problem as the *stationary problem* (SP for short).

It is important to understand the nature of the intimate connection between the NSP and the SP formulated here. Let $\pi = \{\mu_0, \mu_1 \ldots\}$ be an admissible policy for the NSP. Also, let $\tilde{\pi} = \{\tilde{\mu}_0, \tilde{\mu}_1, \ldots\}$ be an admissible policy for the SP such that

$$\tilde{\mu}_i(\tilde{x}) = \mu_i(\tilde{x}), \qquad \text{if } \tilde{x} \in S_i, \quad i = 0, 1, \ldots \tag{6.4}$$

Let $x_0 \in S_0$ be the initial state for the NSP and consider the same initial state for the SP (i.e., $\tilde{x}_0 = x_0 \in S_0$). Then the sequence of states $\{\tilde{x}_i\}$ generated in the SP will satisfy $\tilde{x}_i \in S_i$, $i = 0, 1, \ldots$, with probability 1 (i.e., the system will move from the set S_0 to the set S_1, then to S_2, etc., just as in the NSP). Furthermore, the probabilistic law of generation of states and costs is identical in the NSP and the SP. As a result, it is easy to see that for any admissible policies π and $\tilde{\pi}$ satisfying Eq. (6.4) and initial states x_0, $\tilde{x}_0$ satisfying $x_0 = \tilde{x}_0 \in S_0$, the sequence of generated states in the NSP and the SP is the same ($x_i = \tilde{x}_i$, for all i) provided the generated disturbances w_i and $\tilde{w}_i$ are also the same for all i ($w_i = \tilde{w}_i$, for all i). Furthermore, if π and $\tilde{\pi}$ satisfy Eq. (6.4), we have $J_\pi(x_0) = \tilde{J}_{\tilde{\pi}}(\tilde{x}_0)$ if $x_0 = \tilde{x}_0 \in S_0$. Let us also consider the optimal cost functions for the NSP and the SP:

$$J^*(x_0) = \min_{\pi \in \Pi} J_\pi(x_0), \qquad x_0 \in S_0,$$

$$\tilde{J}^*(\tilde{x}_0) = \min_{\tilde{\pi} \in \tilde{\Pi}} J_{\tilde{\pi}}(\tilde{x}_0), \qquad \tilde{x}_0 \in S_0.$$

Then it follows from the construction of the SP that

$$\tilde{J}^*(\tilde{x}_0) = \tilde{J}^*(\tilde{x}_0, i), \qquad \text{if } \tilde{x}_0 \in S_i, \qquad i = 0, 1, \ldots, \tag{6.5}$$

where, for all $i = 0, 1, \ldots$,

$$\tilde{J}^*(\tilde{x}_0, i) = \min_{\pi \in \Pi} \lim_{N \to \infty} \mathop{E}_{\substack{w_k \\ k = 0, 1, \ldots, N-1}} \left\{ \sum_{k=i}^{N-1} \alpha^{k-i} g_k\big(x_k, \mu_k(x_k), w_k\big) \right\}, \tag{6.6}$$

if $\tilde{x}_0 = x_i \in S_i$. Note that in this equation, the right-hand side is defined in terms of the data of the NSP. As a special case of this equation, we obtain

$$\tilde{J}^*(\tilde{x}_0) = \tilde{J}^*(\tilde{x}_0, 0) = J^*(x_0), \qquad \text{if } \tilde{x}_0 = x_0 \in S_0. \tag{6.7}$$

Thus *the optimal cost function J^* of the NSP can be obtained from the optimal cost function $\tilde{J}^*$ of the SP*. Furthermore, if $\tilde{\pi}^* = \{\tilde{\mu}_0^*, \tilde{\mu}_1^*, \ldots\}$ is an optimal policy for the SP, then the policy $\pi^* = \{\mu_0^*, \mu_1^*, \ldots\}$ defined by

$$\mu_i^*(x_i) = \tilde{\mu}_i^*(x_i), \qquad \text{for all } x_i \in S_i, \qquad i = 0, 1, \ldots, \tag{6.8}$$

is an optimal policy for the NSP. *Thus optimal policies for the SP yield optimal policies for the NSP via Eq. (6.8).* Another point to be noted is that *if Assumption D′ (P′, N′) is satisfied for the NSP, then Assumption D (P, N) introduced earlier in this chapter is satisfied for the SP.*

These observations show that one may analyze the NSP by means of the SP. Every result given in the preceding sections when applied to the SP yields a corresponding result for the NSP. We will just provide the form of the optimality equation for the NSP in the following proposition.

Proposition 6.1: Under Assumption D′ (P′, N′), there holds

$$J^*(x_0) = \tilde{J}^*(x_0, 0), \qquad x_0 \in S_0,$$

where for all $i = 0, 1, \ldots$, the functions $\tilde{J}^*(\cdot, i)$ map S_i into $\Re$ ($[0, \infty]$, $[-\infty, 0]$), are given by Eq. (6.6), and satisfy for all $x_i \in S_i$ and $i = 0, 1, \ldots$,

$$\tilde{J}^*(x_i, i) = \min_{u_i \in U_i(x_i)} \mathop{E}_{w_i} \big\{ g_i(x_i, u_i, w_i) + \alpha \tilde{J}^*\big(f_i(x_i, u_i, w_i), i+1\big) \big\}. \tag{6.9}$$

Under Assumption D′ the functions $\tilde{J}^*(\cdot, i)$, $i = 0, 1, \ldots$, are the unique bounded solutions of the set of equations Eq. (6.9). Furthermore, under Assumption D′ or P′, if $\mu_i^*(x_i) \in U_i(x_i)$ attains the minimum in Eq. (6.9) for all $x_i \in S_i$ and i, then the policy $\pi^* = \{\mu_0^*, \mu_1^*, \ldots\}$ is optimal for the NSP.

Periodic Problems

Assume within the framework of the NSP that there exists an integer $p \geq 2$ (called the *period*) such that for all integers i and j with $|i - j| = mp$, $m = 1, 2, \ldots$, we have

$$S_i = S_j, \qquad C_i = C_j, \qquad D_i = D_j, \qquad U_i(\cdot) = U_j(\cdot),$$

$$f_i = f_j, \qquad g_i = g_j, \qquad P_i(\cdot \mid x, j) = P_j(\cdot \mid x, u), \qquad (x, u) \in S_i \times C_i.$$

We assume that the spaces S_i, C_i, D_i, $i = 0, 1, \ldots, p-1$, are mutually disjoint. We define new state, control, and disturbance spaces by

$$S = \cup_{i=0}^{p-1} S_i, \qquad C = \cup_{i=0}^{p-1} C_i, \qquad D = \cup_{i=0}^{p-1} D_i.$$

The optimality equation for the equivalent stationary problem reduces to the system of p equations

$$\tilde{J}^*(x_0,0) = \min_{u_0 \in U_0(x_0)} E_{w_0}\{g_0(x_0,u_0,w_0) + \alpha \tilde{J}^*\big(f_0(x_0,u_0,w_0),1\big)\},$$

$$\tilde{J}^*(x_1,1) = \min_{u_1 \in U_1(x_1)} E_{w_1}\{g(x_1,u_1,w_1) + \alpha \tilde{J}^*\big(f_1(x_1,u_1,w_1),2\big)\}.$$

$$\vdots$$

$$\tilde{J}^*(x_{p-1},p-1) = \min_{u_{p-1} \in U_{p-1}(x_{p-1})} E_{w_{p-1}}\{g_{p-1}(x_{p-1},u_{p-1},w_{p-1}) + \alpha \tilde{J}^*\big(f_{p-1}(x_{p-1},u_{p-1},w_{p-1}),0\big)\}.$$

These equations may be used to obtain (under Assumption D′ or P′) a periodic policy of the form $\{\mu_0^*,\ldots,\mu_{p-1}^*,\mu_0^*,\ldots,\mu_{p-1}^*,\ldots\}$ whenever the minimum of the right-hand side is attained for all x_i, $i = 0,1,\ldots,p-1$. When all spaces involved are finite, an optimal policy may be found by means of the algorithms of Section 1.3, appropriately adapted to the corresponding SP.

3.7 NOTES, SOURCES, AND EXERCISES

Undiscounted problems and discounted problems with unbounded cost per stage were first analyzed systematically in [DuS65], [Bla65], and [Str66]. An extensive treatment, which also resolves the associated measurability questions, is [BeS78]. Sufficient conditions for convergence of the value iteration method under Assumption P (cf. Props. 1.6 and 1.7) were derived independently in [Ber77] and [Sch75]. The former reference also derives necessary conditions for convergence. Problems involving convexity assumptions are analyzed in [Ber73b].

We have bypassed a number of complex theoretical issues relating to stationary policies that historically have played an important role in the development of the subject of this chapter. The main question is to what extent is it possible to restrict attention to stationary policies. Much theoretical work has been done on this question [BeS79], [Bla65], [Bla70], [DuS65], [Fei78], [FeS83], [Fei92a], [Fei92b], [Orn69], and some aspects are still open. Suppose, for example, that we are given an $\epsilon > 0$. One issue is whether there exists an ϵ-optimal stationary policy, that is, a stationary policy μ such that

$$J_\mu(x) \le J^*(x) + \epsilon, \qquad \text{for all } x \in S \text{ with } J^*(x) > -\infty,$$

$$J_\mu(x) \le -\frac{1}{\epsilon}, \qquad \text{for all } x \in S \text{ with } J^*(x) = -\infty.$$

The answer is positive under any one of the following conditions:

1. Assumption P holds and $\alpha < 1$ (see Exercise 3.8).
2. Assumption N holds, S is a finite set, $\alpha = 1$, and $J^*(x) > -\infty$ for all $x \in S$ (see Exercise 3.11 or [Bla65], [Bla70], and [Orn69]).
3. Assumption N holds, S is a countable set, $\alpha = 1$, and the problem is deterministic (see [BeS79]).

The answer can be negative under any one of the following conditions:

1. Assumption P holds and $\alpha = 1$ (see Exercise 3.8).
2. Assumption N holds and $\alpha < 1$ (see Exercise 3.11 or [BeS79]).

The existence of an ϵ-optimal stationary policy for stochastic shortest path problems with a finite state space, but under somewhat different assumptions than the ones of Section 2.1 is established in [Fei92b].

Another issue is whether there exists an optimal stationary policy whenever there exists an optimal policy for each initial state. This is true under Assumption P (see Exercise 3.9). It is also true (but very hard to prove) under Assumption N if $J^*(x) > -\infty$ for all $x \in S$, $\alpha = 1$, and the disturbance space D is countable [Bla70], [DuS65], [Orn69]. Simple two-state examples can be constructed showing that the result fails to hold if $\alpha = 1$ and $J^*(x) = -\infty$ for some state x (see Exercise 3.10). However, these examples rely on the presence of a stochastic element in the problem. If the problem is deterministic, stronger results are available; one can find an optimal stationary policy if there exists an optimal policy at each initial state and either $\alpha = 1$ or $\alpha < 1$ and $J^*(x) > -\infty$ for all $x \in S$. These results also require a difficult proof [BeS79].

The gambling problem and its solution are taken from [DuS65]. In [Bil83], a surprising property of the optimal reward function J^* for this problem is shown: J^* is almost everywhere differentiable with derivative zero, yet it is strictly increasing, taking values that range from 0 to 1.

EXERCISES

3.1

Let $S = [0,\infty)$ and $C = U(x) = (0,\infty)$ be the state and control spaces,

respectively, let the system equation be

$$x_{k+1} = \left(\frac{2}{\alpha}\right) x_k + u_k, \qquad k = 0, 1, \ldots,$$

where α is the discount factor, and let

$$g(x_k, u_k) = x_k + u_k$$

be the cost per stage. Show that for this deterministic problem, Assumption P holds and that $J^*(x) = \infty$ for all $x \in S$, but $(T^k J_0)(0) = 0$ for all k [J_0 is the zero function, $J_0(x) = 0$, for all $x \in S$].

3.2

Let Assumption P hold and consider the finite-state case $S = D = \{1, 2, \ldots, n\}$, $\alpha = 1$, $x_{k+1} = w_k$. The mapping T is represented as

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

where $p_{ij}(u)$ denotes the transition probability that the next state will be j when the current state is i and control u is applied. Assume that the sets $U(i)$ are compact subsets of $\Re^n$ for all i, and that $p_{ij}(u)$ and $g(i, u)$ are continuous on $U(i)$ for all i and j. Show that $\lim_{k \to \infty} (T^k J_0)(i) = J^*(i)$, where $J_0(i) = 0$ for all $i = 1, \ldots, n$. Show also that there exists an optimal stationary policy.

3.3

Consider a deterministic problem involving a linear system

$$x_{k+1} = Ax_k + Bu_k, \qquad k = 0, 1, \ldots,$$

where the pair (A, B) is controllable and $x_k \in \Re^n$, $u_k \in \Re^m$. Assume no constraints on the control and a cost per stage g satisfying

$$0 \leq g(x, u), \qquad (x, u) \in \Re^n \times \Re^m.$$

Assume furthermore that g is continuous in x and u, and that $g(x_n, u_n) \to \infty$ if $\{x_n\}$ is bounded and $\|u_n\| \to \infty$.

(a) Show that for a discount factor $\alpha < 1$, the optimal cost satisfies $0 \leq J^*(x) < \infty$, for all $x \in \Re^n$. Furthermore, there exists an optimal stationary policy and

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

(b) Show that the same is true, except perhaps for $J^*(x) < \infty$, when the system is of the form $x_{k+1} = f(x_k, u_k)$, with $f : \Re^n \times \Re^m \mapsto \Re^n$ being a continuous function.

(c) Prove the same results assuming that the control is constrained to lie in a compact set $U \in \Re^m$ [$U(x) = U$ for all x] in place of the assumption $g(x_n, u_n) \to \infty$ if $\{x_n\}$ is bounded and $\|u_n\| \to \infty$. *Hint*: Show that $T^k J_0$ is real valued and continuous for every k, and use Prop. 1.7.

3.4

Under Assumption P, let μ be such that for all $x \in S$, $\mu(x) \in U(x)$ and

$$(T_\mu J^*)(x) \leq (TJ^*)(x) + \epsilon,$$

where ϵ is some positive scalar. Show that, if $\alpha < 1$,

$$J_\mu(x) \leq J^*(x) + \frac{\epsilon}{1 - \alpha}, \qquad x \in S.$$

Hint: Show that $(T_\mu^k J^*)(x) \leq J^*(x) + \sum_{t=0}^{k-1} \alpha^t \epsilon$. Alternatively, let $J' = J^* + \big(\epsilon/(1-\alpha)\big)e$, show that $T_\mu J' \leq J'$, and use Cor. 7.1.1.

3.5

Under Assumption P or N, show that if $\alpha < 1$ and $J' : S \mapsto \Re$ is a *bounded* function satisfying $J' = TJ'$, then $J' = J^*$. *Hint*: Under P, let r be a scalar such that $J^* + re \geq J'$. Argue that $J^* \geq J'$ and use Prop. 1.2(a).

3.6

We want to find a scalar sequence $\{u_0, u_1, \ldots\}$ that satisfies $\sum_{k=0}^{\infty} u_k \leq c$, $u_k \geq 0$, for all k, and maximizes $\sum_{k=0}^{\infty} g(u_k)$, where $c > 0$ and $g(u) \geq 0$ for all $u \geq 0$, $g(0) = 0$. Assume that g is monotonically nondecreasing on $[0, \infty)$. Show that the optimal value of the problem is $J^*(c)$, where J^* is a monotonically nondecreasing function on $[0, \infty)$ satisfying $J^*(0) = 0$ and

$$J^*(x) = \max_{0 \leq u \leq x} \big\{ g(u) + J^*(x - u) \big\}, \qquad x \in [0, \infty).$$

3.7

Let Assumption P hold and assume that $\pi^* = \{\mu_0^*, \mu_1^*, \ldots\} \in \Pi$ satisfies $J^* = T_{\mu_k^*} J^*$ for all k. Show that π^* is optimal, i.e., $J_{\pi^*} = J^*$.

3.8

Under Assumption P, show that, given $\epsilon > 0$, there exists a policy $\pi_\epsilon \in \Pi$ such that $J_{\pi_\epsilon}(x) \leq J^*(x) + \epsilon$ for all $x \in S$, and that for $\alpha < 1$ the policy π_ϵ can be taken stationary. Give an example where $\alpha = 1$ and for each stationary policy π we have $J_\pi(x) = \infty$, while $J^*(x) = 0$ for all x. *Hint*: See the proof of Prop. 1.1.

3.9

Under Assumption P, show that if there exists an optimal policy (a policy $\pi^* \in \Pi$ such that $J_{\pi^*} = J^*$), then there exists an optimal stationary policy.

3.10

Use the following counterexample to show that the result of Exercise 3.9 may fail to hold under Assumption N if $J^*(x) = -\infty$ for some $x \in S$. Let $S = D = \{0, 1\}$, $f(x, u, w) = w$, $g(x, u, w) = u$, $U(0) = (-\infty, 0]$, $U(1) = \{0\}$, $p(w = 0 \mid x = 0, u) = \frac{1}{2}$, and $p(w = 1 \mid x = 1, u) = 1$. Show that $J^*(0) = -\infty$, $J^*(1) = 0$ and that the admissible nonstationary policy $\{\mu_0^*, \mu_1^*, \ldots\}$ with $\mu_k^*(0) = -(2/\alpha)^k$ is optimal. Show that every stationary policy μ satisfies $J_\mu(0) = \big(2/(2 - \alpha)\big)\mu(0)$, $J_\mu(1) = 0$ (see [Bla70], [DuS65], and [Orn69] for related analysis).

3.11

Show that the result of Exercise 3.8 holds under Assumption N if S is a finite set, $\alpha = 1$, and $J^*(x) > -\infty$ for all $x \in S$. Construct a counterexample to show that the result can fail to hold if S is countable and $\alpha < 1$ [even if $J^*(x) > -\infty$ for all $x \in S$]. *Hint*: Consider an integer N such that the N-stage optimal cost J_N satisfies $J_N(x) \leq J^*(x) + \epsilon$ for all x. For a counterexample, see [BeS79].

3.12 (Deterministic Linear-Quadratic Problems)

Consider the deterministic linear-quadratic problem involving the system

$$x_{k+1} = Ax_k + Bu_k$$

and the cost

$$J_\pi(x_0) = \sum_{k=0}^{\infty} \big(x_k' Q x_k + \mu_k(x_k)' R \mu_k(x_k)\big).$$

We assume that R is positive definite symmetric, Q is of the form $C'C$, and the pairs (A, B), (A, C) are controllable and observable, respectively. Use the theory of Sections 4.1 of Vol. I and 8.1 to show that the stationary policy μ^* with

$$\mu^*(x) = -(B'KB + R)^{-1} B'KAx$$

is optimal, where K is the unique positive semidefinite symmetric solution of the algebraic Riccati equation (cf. Section 4.1 of Vol. I):

$$K = A'\big(K - KB(B'KB + R)^{-1}B'K\big)A + Q.$$

Provide a similar result under an appropriate controllability assumption for the case of a periodic deterministic linear system and a periodic quadratic cost (cf. Section 3.6).

3.13

Consider the linear-quadratic problem of Section 3.2 with the only difference that the disturbances w_k have zero mean, but their covariance matrices are nonstationary and uniformly bounded over k. Show that the optimal control law remains unchanged.

3.14 (Periodic Linear-Quadratic Problems)

Consider the linear system

$$x_{k+1} = A_k x_k + B_k u_k + w_k, \qquad k = 0, 1, \ldots,$$

and the quadratic cost

$$J_\pi(x_0) = \lim_{N \to \infty} \mathop{E}_{\substack{w_k \\ k=0,\ldots,N-1}} \left\{ \sum_{k=0}^{N-1} \alpha^k (x_k' Q_k x_k + u_k' R_k u_k) \right\},$$

where the matrices have appropriate dimensions, Q_k and R_k are positive semidefinite and positive definite symmetric, respectively, for all k, and $0 < \alpha < 1$. Assume that the system and cost are periodic with period p (cf. Section 3.6), that the controls are unconstrained, and that the disturbances are independent, and have zero mean and finite covariance. Assume further that the following (controllability) condition is in effect.

For any state $\bar{x}_0$, there exists a finite sequence of controls $\{\bar{u}_0, \bar{u}_1, \ldots, \bar{u}_r\}$ such that $\bar{x}_{r+1} = 0$, where $\bar{x}_{r+1}$ is generated by

$$\bar{x}_{k+1} = A_k \bar{x}_k + B_k \bar{u}_k, \qquad k = 0, 1, \ldots, r.$$

Show that there is an optimal periodic policy π^* of the form

$$\pi^* = \{\mu_0^*, \mu_1^*, \ldots, \mu_{p-1}^*, \mu_0^*, \mu_1^*, \ldots, \mu_{p-1}^*, \ldots\},$$

where $\mu_0^*, \mu_1^*, \ldots, \mu_{p-1}^*$ are given by

$$\mu_i^*(x) = -\alpha(\alpha B_i' K_{i+1} B_i + R_i)^{-1} B_i' K_{i+1} A_i x, \qquad i = 0, \ldots, p-2,$$

$$\mu_{p-1}^*(x) = -\alpha(\alpha B_{p-1}' K_0 B_{p-1} + R_{p-1})^{-1} B_{p-1}' K_0 A_{p-1} x,$$

and the matrices $K_0, K_1, \ldots, K_{p-1}$ satisfy the coupled set of p algebraic Riccati equations given for $i = 0, 1, \ldots, p-1$ by

$$K_i = A_i'\big(\alpha K_{i+1} - \alpha^2 K_{i+1} B_i (\alpha B_i' K_{i+1} B_i + R_i)^{-1} B_i' K_{i+1}\big) A_i + Q_i,$$

with

$$K_p = K_0.$$

3.15 (Linear-Quadratic Problems - Imperfect State Information)

Consider the linear-quadratic problem of Section 3.2 with the difference that the controller, instead of having perfect state information, has access to measurements of the form

$$z_k = Cx_k + v_k, \qquad k = 0, 1, \ldots$$

As in Section 5.2 of Vol. I, the disturbances v_k are independent and have identical statistics, zero mean, and finite covariance matrix. Assume that for every admissible policy π the matrices

$$E\Big\{\big(x_k - E\{x_k \mid I_k\}\big)\big(x_k - E\{x_k \mid I_k\}\big)' \mid \pi\Big\}$$

are uniformly bounded over k, where I_k is the information vector defined in Section 5.2 of Vol. I. Show that the stationary policy μ^* given by

$$\mu^*(I_k) = -\alpha(\alpha B'KB + R)^{-1}B'KAE\{x_k \mid I_k\}, \quad \text{for all } I_k,\ k = 0, 1, \ldots$$

is optimal. Show also that the same is true if w_k and v_k are nonstationary with zero mean and covariance matrices that are uniformly bounded over k. *Hint*: Combine the theory of Sections 5.2 of Vol. I and 3.2.

3.16 (Policy Iteration for Linear-Quadratic Problems [Kle68])

Consider the problem of Section 3.2 and let L_0 be an $m \times n$ matrix such that the matrix $(A + BL_0)$ has eigenvalues strictly within the unit circle.

(a) Show that the cost corresponding to the stationary policy μ_0, where $\mu_0(x) = L_0x$ is of the form

$$J_{\mu_0}(x) = x'K_0x + \text{constant},$$

where K_0 is a positive semidefinite symmetric matrix satisfying the (linear) equation

$$K_0 = \alpha(A + BL_0)'K_0(A + BL_0) + Q + L_0'RL_0.$$

(b) Let $\mu_1(x)$ attain the minimum for each x in the expression

$$\min_u \big\{u'Ru + \alpha(Ax + Bu)'K_0(Ax + bu)\big\}.$$

Show that for all x we have

$$J_{\mu_1}(x) = x'K_1x + \text{constant} \ \leq J_{\mu_0}(x),$$

where K_1 is some positive semidefinite symmetric matrix.

(c) Show that the policy iteration process described in parts (a) and (b) yields a sequence $\{K_k\}$ such that

$$K_k \to K,$$

where K is the optimal cost matrix of the problem.

3.17 (Periodic Inventory Control Problems)

In the inventory control problem of Section 3.3, consider the case where the statistics of the demands w_k, the prices c_k, and the holding and the shortage costs are periodic with period p. Show that there exists an optimal periodic policy of the form $\pi^* = \{\mu_0^*, \ldots, \mu_{p-1}^*, \mu_0^*, \ldots, \mu_{p-1}^*, \ldots\}$,

$$\mu_i^*(x) = \begin{cases} S_i^* - x & \text{if } x \leq S_i^*, \\ 0 & \text{if otherwise,} \end{cases} \qquad i = 0, 1, \ldots, p-1,$$

where $S_0^*, \ldots, S_{p-1}^*$ are appropriate scalars.

3.18 [HeS84]

Show that the critical level S^* for the inventory problem with zero fixed cost of Section 3.3 minimizes $(1-\alpha)cy + L(y)$ over y. *Hint*: Show that the cost can be expressed as

$$J_\pi(x_0) = E\left\{\sum_{k=0}^{\infty} \alpha^k \big((1-\alpha)cy_k + L(y_k)\big) + \frac{c\alpha}{1-\alpha}E\{w\} - cx_0\right\},$$

where $y_k = x_k + \mu_k(x_k)$.

3.19

Consider a machine that may break down and can be repaired. When it operates over a time unit, it costs -1 (that is, it produces a benefit of 1 unit), and it may break down with probability 0.1. When it is in the breakdown mode, it may be repaired with an effort u. The probability of making it operative over one time unit is then u, and the cost is Cu^2. Determine the optimal repair effort over an infinite time horizon with discount factor $\alpha < 1$.

3.20

Let $z_0, z_1, \ldots$ be a sequence of independent and identically distributed random variables taking values on a finite set Z. We know that the probability distribution of the z_k's is one out of n distributions $f_1, \ldots, f_n$, and we are trying to decide which distribution is the correct one. At each time k after observing $z_1, \ldots, z_k$, we may either stop the observations and accept one of the n distributions as correct, or take another observation at a cost $c > 0$. The cost for accepting f_i given that f_j is correct is L_{ij}, $i, j = 1, \ldots, n$. We assume $L_{ij} > 0$ for $i \neq j$, $L_{ii} = 0$, $i = 1, \ldots, n$. The a priori distribution of $f_1, \ldots, f_n$ is denoted

$$P_0 = \{p_0^1, p_0^2, \ldots, p_0^n\}, \qquad p_0^i \geq 0, \qquad \sum_{i=1}^{n} p_0^i = 1.$$

Show that the optimal cost $J^*(P_0)$ is a concave function of P_0. Characterize the optimal acceptance regions and show how they can be obtained in the limit by means of a value iteration method.

3.21 (Gambling Strategies for Favorable Games)

A gambler plays a game such as the one of Section 3.5, but where the probability of winning p satisfies $1/2 < p < 1$. His objective is to reach a final fortune n, where n is an integer with $n \geq 2$. His initial fortune is an integer i with $0 < i < n$, and his stake at time k can take only integer values u_k satisfying $0 \leq u_k \leq x_k$, $0 \leq u_k \leq n - x_k$, where x_k is his fortune at time k. Show that the strategy that always stakes one unit is optimal [i.e., $\mu^*(x) = 1$ for all integers x with $0 < x < n$ is optimal]. *Hint*: Show that if $p \in (1/2, 1)$,

$$J_{\mu^*}(i) = \left[\left(\frac{1-p}{p}\right)^i - 1\right]\left[\left(\frac{1-p}{p}\right)^n - 1\right]^{-1}, \qquad 0 \leq i \leq n,$$

and if $p = 1/2$,

$$J_{\mu^*}(i) = \frac{i}{n}, \qquad 0 \leq i \leq n,$$

(or see [Ash70], p. 182, for a proof). Then use the sufficiency condition of Prop. 1.1 in Section 3.1.

3.22 [Sch81]

Consider a network of n queues whereby a customer at queue i upon completion of service is routed to queue j with probability p_{ij}, and exits the network with probability $1 - \sum_j p_{ij}$. For each queue i denote:

r_i: the external customer arrival rate,

$\frac{1}{\mu_i}$: the average customer service time,

λ_i: the customer departure rate,

a_i: the total customer arrival rate (sum of external rate and departure rates from upstream queues weighted by the corresponding probabilities).

We have

$$a_i = r_i + \sum_{j=1}^{n} \lambda_j p_{ji}, \qquad \text{for all } i,$$

and we assume that any portion of the arrival rate a_i in excess of the service rate μ_i is lost; so the departure rate at queue i satisfies

$$\lambda_i = \min[\mu_i, a_i] = \min\left[\mu_i, r_i + \sum_{j=1}^{n} \lambda_j p_{ji}\right].$$

Assume that $r_i > 0$ for at least one i, and that for every queue i_1 with $r_{i_1} > 0$, there is a queue i with $1 - \sum_j p_{ij} > 0$, and a sequence $i_1, i_2, \ldots, i_k, i$ such that $p_{i_1 i_2} > 0, \ldots, p_{i_k i} > 0$. Show that the departure rates λ_i satisfying the preceding equations are unique and can be found by value iteration or policy iteration. *Hint*: This problem does not quite fit our framework because we may have $\sum_j p_{ij} > 1$ for some i. However, it is possible to carry out an analysis based on m-stage contraction mappings.

3.23 (Infinite Time Reachability [Ber71], [Ber72])

Consider the stationary system

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

where the disturbance space D is an arbitrary (not necessarily countable) set. The disturbances w_k can take values in a subset $W(x_k, u_k)$ of D that may depend on x_k and u_k. This problem deals with the following question: Given a nonempty subset X of the state space S, under what conditions does there exist an admissible policy that keeps the state of the (closed-loop) system

$$x_{k+1} = f\big(x_k, \mu_k(x_k), w_k\big) \tag{7.1}$$

in the set X for all k and all possible values $w_k \in W\big(x_k, \mu_k(x_k)\big)$, that is,

$$x_k \in X, \qquad \text{for all } w_k \in W\big(x_k, \mu_k(x_k)\big), \qquad k = 0, 1, \ldots. \tag{7.2}$$

The set X is said to be *infinitely reachable* if there exists an admissible policy $\{\mu_0, \mu_1, \ldots\}$ and *some* initial state $x_0 \in X$ for which the above relations are satisfied. It is said to be *strongly reachable* if there exists an admissible policy $\{\mu_0, \mu_1, \ldots\}$ such that for *all* initial states $x_0 \in X$ the above relations are satisfied.

Consider the function R mapping any subset Z of the state space S into a subset $R(Z)$ of S defined by

$$R(Z) = \Big\{x \mid \text{for some } u \in U(x),\ f(x, u, w) \in Z, \text{ for all } w \in W(x, u)\Big\} \cap Z.$$

(a) Show that the set X is strongly reachable if and only if $R(X) = X$.

(b) Given X, consider the set X^* defined as follows: $x_0 \in X^*$ if and only if $x_0 \in X$ and there exists an admissible policy $\{\mu_0, \mu_1, \ldots\}$ such that that Eqs. (7.1) and (7.2) are satisfied when x_0 is taken as the initial state of the system. Show that a set X is infinitely reachable if and only if it contains a nonempty strongly reachable set. Furthermore, the largest such set is X^* in the sense that X^* is strongly reachable whenever nonempty, and if $\tilde{X} \in X$ is another strongly reachable set, then $\tilde{X} \subset X^*$.

(c) Show that if X is infinitely reachable, there exists an admissible stationary policy μ such that if the initial state x_0 belongs to X^*, then all subsequent states of the closed-loop system $x_{k+1} = f\big(x_k, \mu(x_k), w_k\big)$ are guaranteed to belong to X^*.

(d) Given X, consider the sets $R^k(X)$, $k = 1, 2, \ldots$, where $R^k(X)$ denotes the set obtained after k applications of the mapping R on X. Show that

$$X^* \subset \cap_{k=1}^{\infty} R^k(X).$$

(e) Given X, consider for each $x \in X$ and $k = 1, 2, \ldots$ the set

$$U_k(x) = \Big\{u \mid f(x, u, w) \in R^k(X) \text{ for all } w \in W(x, u)\Big\}.$$

Show that, if there exists an index $\bar{k}$ such that for all $x \in X$ and $k \geq \bar{k}$ the set $U_k(x)$ is a compact subset of a Euclidean space, then $X^* = \cap_{k=1}^{\infty} R^k(X)$.

3.24 (Infinite Time Reachability for Linear Systems)

Consider the linear stationary system

$$x_{k+1} = Ax_k + Bu_k + Gw_k,$$

where $x_k \in \Re^n$, $u_k \in \Re^m$, and $w_k \in \Re^r$, and the matrices A, B, and G are known and have appropriate dimensions. The matrix A is assumed invertible. The controls u_k and the disturbances w_k are restricted to take values in the ellipsoids $U = \{u \mid u'Ru \leq 1\}$ and $W = \{w \mid w'Qw \leq 1\}$, respectively, where R and Q are positive definite symmetric matrices of appropriate dimensions. Show that in order for the ellipsoid $X = \{x \mid x'Kx \leq 1\}$, where K is a positive definite symmetric matrix, to be strongly reachable (in the terminology of Exercise 3.23), it is sufficient that for some positive definite symmetric matrix M and for some scalar $\beta \in (0, 1)$ we have

$$K = A' \left[(1 - \beta)K^{-1} - \frac{1-\beta}{\beta} GQ^{-1}G' + BR^{-1}B' \right]^{-1} A + M,$$

$$K^{-1} - \frac{1}{\beta} GQ^{-1}G' : \text{positive definite.}$$

Show also that if the above relations are satisfied, the linear stationary policy μ^*, where $\mu^*(x) = Lx$ and

$$L = -(R + B'FB)^{-1}B'FA,$$

$$F = \left[(1-\beta)K^{-1} - \frac{1-\beta}{\beta} GQ^{-1}G' \right]^{-1},$$

achieves reachability of the ellipsoid $X = \{x \mid x'Kx \leq 1\}$. Furthermore, the matrix $(A + BL)$ has all its eigenvalues strictly within the unit circle. (For a proof together with a computational procedure for finding matrices K satisfying the above, see [Ber71] and [Ber72b].)

3.25 (The Blackmailer's Dilemma)

Consider Example 1.1 of Section 2.1. Here, there are two states, state 1 and a termination state t. At state 1, we can choose a control u with $0 < u \leq 1$; we then move to state t at no cost with probability $p(u)$, and stay in state 1 at a cost $-u$ with probability $1 - p(u)$.

(a) Let $p(u) = u^2$. For this case it was shown in Example 1.1 of Section 2.1, that the optimal costs are $J^*(1) = -\infty$ and $J^*(t) = 0$. Furthermore, it was shown that there is no optimal stationary policy, although there is an optimal nonstationary policy. Find the set of solutions to Bellman's equation and verify the result of Prop. 1.2(b).

(b) Let $p(u) = u$. Find the set of solutions to Bellman's equation and use Prop. 1.2(b) to show that the optimal costs are $J^*(1) = -1$ and $J^*(t) = 0$. Show that there is no optimal policy (stationary or not).

4

Average Cost per Stage Problems

Contents

The results of the preceding chapters apply mainly to problems where the optimal total expected cost is finite either because of discounting or because of a cost-free absorbing state that the system eventually enters. In many situations, however, discounting is inappropriate and there is no natural cost-free absorbing state. In such situations it is often meaningful to optimize the average cost per stage, to be defined shortly. In this chapter, we discuss this type of optimization, with an emphasis on the case of a finite-state Markov chain.

An introductory analysis of the problem of this chapter was given in Section 7.4 of Vol. I. That analysis was based on a connection between the average cost per stage and the stochastic shortest path problem. While this connection can be further extended to obtain more powerful results (see Exercises 4.13-4.16), we develop here an alternative line of analysis that is based on a relation with the discounted cost problem. This relation allows us to use discounted cost results, derived in Sections 1.2 and 1.3, in order to conjecture and prove results for the average cost problem.

4.1 PRELIMINARY ANALYSIS

Let us formulate the problem of this chapter for the case of finite state and control spaces. We adopt the Markov chain notation used in Section 1.3. In particular, we denote the states by $1,\ldots,n$. To each state i and control u there corresponds a set of transition probabilities $p_{ij}(u)$, $j = 1,\ldots,n$. Each time the system is in state i and control u is applied, we incur an expected cost $g(i,u)$, and the system moves to state j with probability $p_{ij}(u)$. The objective is to minimize over all policies $\pi = \{\mu_0,\mu_1,\ldots\}$ with $\mu_k(i) \in U(i)$, for all i and k, the average cost per stage †

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

for any given initial state x_0.

† When the limit defining the average cost is not known to exist, we use instead the definition

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

We will show, however, as part of our subsequent analysis that the limit exists at least for those policies π that are of interest.

As in Section 1.3, we use the following shorthand notation for a stationary policy μ:

$$g_\mu = \begin{pmatrix} g\big(1,\mu(1)\big) \\ \vdots \\ g\big(n,\mu(n)\big) \end{pmatrix}, \qquad P_\mu = \begin{pmatrix} p_{11}\big(\mu(1)\big) \ldots p_{1n}\big(\mu(1)\big) \\ \cdots \\ p_{n1}\big(\mu(n)\big) \ldots p_{nn}\big(\mu(n)\big) \end{pmatrix},$$

$$J_\mu = \begin{pmatrix} J_\mu(1) \\ \vdots \\ J_\mu(n) \end{pmatrix}.$$

Since the (i,j)th element of the matrix P_μ^k (P_μ to the kth power) is the k-step transition probability $P(x_k = j \mid x_0 = i)$ corresponding to μ, it can be seen that

$$J_\mu = \left(\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P_\mu^k\right) g_\mu.$$

An important result regarding transition probability matrices is that the limit in the preceding equation exists. We show this fact shortly in the context of a more general result, which establishes the connection between the average cost per stage problem and the discounted cost problem.

An Overview of Results

While the material of this chapter does not rely on the analysis of the average cost problem of Section 7.4 in Vol. I, it is worth summarizing some of the salient features of that analysis (see also Exercises 4.13-4.16). We assumed there that there is a special state, by convention state n, which is recurrent in the Markov chain corresponding to each stationary policy. If we consider a sequence of generated states, and divide it into cycles marked by successive visits to the special state n, we see that each of the cycles can be viewed as a state trajectory of a corresponding stochastic shortest path problem with the termination state being essentially n. More precisely, this stochastic shortest path problem has states $1,2,\ldots,n$, plus an artificial termination state t to which we move from state i with transition probability $p_{in}(u)$. The transition probabilities from a state i to a state $j \neq n$ are the same as those of the original problem, while $p_{in}(u)$ is zero. For any scalar λ, we considered the stochastic shortest path problem with expected stage cost $g(i,u) - \lambda$ for each state $i = 1,\ldots,n$. We then argued that if we fix the expected stage cost incurred at state i to be

$$g(i,u) - \lambda^*,$$

where λ^* is the optimal average cost per stage starting from the special state n, then the associated stochastic shortest path problem becomes essentially equivalent to the original average cost per stage problem. Furthermore, Bellman's equation for the associated stochastic shortest path

problem can be viewed as Bellman's equation for the original average cost per stage problem. Based on this line of analysis, we showed a number of results, which will be strengthened in the present chapter by using different methods. In summary, these results are the following:

(a) The optimal average cost per stage is independent of the initial state. This property is a generic feature for almost all average cost problems of practical interest.

(b) Bellman's equation takes the form

$$\lambda^* + h^*(i) = \min_{u \in U(i)} \left[g(i,u) + \sum_{j=1}^{n} p_{ij}(u) h^*(j) \right], \qquad i = 1, \ldots, n,$$

where $h^*(n) = 0$, λ^* is the optimal average cost per stage, and $h^*(i)$ has the interpretation of a relative or differential cost for each state i (it is the minimum of the difference between the expected cost to reach n from i for the first time and the cost that would be incurred if the cost per stage was the average λ^*).

(c) There are versions of the value iteration, policy iteration, adaptive aggregation, and linear programming methods that can be used for computational solution under reasonable conditions.

We will now provide the foundation for the analysis of this chapter by developing the connection between average cost and discounted problems.

Relation with the Discounted Cost Problem

Let us consider the cost of a stationary policy μ for the corresponding α-discounted problem. It is given by

$$J_{\alpha,\mu} = \sum_{k=0}^{\infty} \alpha^k P_\mu^k g_\mu = \left(\sum_{k=0}^{\infty} \alpha^k P_\mu^k \right) g_\mu = (I - \alpha P_\mu)^{-1} g_\mu, \qquad \alpha \in (0,1). \tag{1.1}$$

To get a sense of the relation with the average cost of μ, we note that this latter cost is written as

$$\begin{aligned} J_\mu(i) &= \lim_{N \to \infty} \frac{1}{N} E\left\{ \sum_{k=0}^{N-1} g\big(x_k, \mu(x_k)\big) \right\} \\ &= \lim_{N \to \infty} \lim_{\alpha \to 1} \frac{E\left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu(x_k)\big) \right\}}{\sum_{k=0}^{N-1} \alpha^k}. \end{aligned}$$

Assuming that the order of the two limits in the right-hand side above can be interchanged, we obtain

$$\begin{aligned} J_\mu(i) &= \lim_{\alpha \to 1} \lim_{N \to \infty} \frac{E\left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu(x_k)\big) \right\}}{\sum_{k=0}^{N-1} \alpha^k} \\ &= \lim_{\alpha \to 1} \frac{\lim_{N \to \infty} E\left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu(x_k)\big) \right\}}{\lim_{N \to \infty} \sum_{k=0}^{N-1} \alpha^k} \\ &= \lim_{\alpha \to 1} (1-\alpha) J_{\alpha,\mu}(i). \end{aligned}$$

The formal proof of the above relation will follow as a corollary to the next proposition.

Proposition 1.1: For any stochastic matrix P and $\alpha \in (0,1)$, there holds

$$(I - \alpha P)^{-1} = (1-\alpha)^{-1} P^* + H + O(|1-\alpha|), \tag{1.2}$$

where $O(|1-\alpha|)$ is an α-dependent matrix such that

$$\lim_{\alpha \to 1} O(|1-\alpha|) = 0, \tag{1.3}$$

and the matrices P^* and H are given by

$$P^* = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k, \tag{1.4}$$

$$H = (I - P + P^*)^{-1} - P^*. \tag{1.5}$$

[It will be shown as part of the proof that the limit in Eq. (1.4) and the inverse in Eq. (1.5) exist.] Furthermore, P^* and H satisfy the following equations:

$$P^* = PP^* = P^*P = P^*P^*, \tag{1.6}$$

$$P^*H = 0, \tag{1.7}$$

$$P^* + H = I + PH. \tag{1.8}$$

Proof: From the matrix inversion formula that expresses each entry of the inverse as a ratio of two determinants, it is seen that the matrix

$$M(\alpha) = (1-\alpha)(I - \alpha P)^{-1}$$

can be expressed as a matrix with elements that are either zero or fractions whose numerator and denominator are polynomials in α with no common divisor. The denominator polynomials of the nonzero elements of $M(\alpha)$ cannot have 1 as a root, since otherwise some elements of $M(\alpha)$ would tend to infinity as $\alpha \to 1$; this is not possible, because from Eq. (1.1) for any μ, we have $(1-\alpha)^{-1}M(\alpha)g_\mu = (I-\alpha P)^{-1}g_\mu = J_{\alpha,\mu}$ and $|J_{\alpha,\mu}(j)| \leq (1-\alpha)^{-1}\max_i |g_\mu(i)|$, implying that the absolute values of the coordinates of $M(\alpha)g_\mu$ are bounded by $\max_i |g_\mu(i)|$ for all $\alpha < 1$. Therefore, the (i,j)th element of the matrix $M(\alpha)$ is of the form

$$m_{ij}(\alpha) = \frac{\gamma(\alpha-\zeta_1)\cdots(\alpha-\zeta_p)}{(\alpha-\xi_1)\cdots(\alpha-\xi_q)}$$

where γ, ζ_i, $i=1,\ldots,p$, and ξ_i, $i=1,\ldots,q$, are scalars such that $\zeta_i \neq 1$ for $i=1,\ldots,q$.

Define

$$P^* = \lim_{\alpha \to 1} M(\alpha). \tag{1.9}$$

and let H be the matrix having as (i,j)th element the 1st derivative of $-m_{ij}(\alpha)$ evaluated at $\alpha = 1$. By the 1st order Taylor expansion of the elements of $m_{ij}(\alpha)$ of $M(\alpha)$, we have for all α in a neighborhood of $\alpha = 1$

$$M(\alpha) = P^* + (1-\alpha)H + O\big((1-\alpha)^2\big), \tag{1.10}$$

where $O\big((1-\alpha)^2\big)$ is an α-dependent matrix such that

$$\lim_{\alpha \to 1} \frac{O\big((1-\alpha)^2\big)}{(1-\alpha)} = 0.$$

Multiplying Eq. (1.10) with $(1-\alpha)^{-1}$, we obtain the desired relation (1.2) [although, we have yet to show that P^* and H are also given by Eqs. (1.4) and (1.5), respectively].

We will now show that P^* as defined by Eq. (1.9), satisfies Eqs. (1.6), (1.5), (1.7), (1.8), and (1.4), in that order.

We have

$$(I-\alpha P)(I-\alpha P)^{-1} = I \tag{1.11}$$

and

$$\alpha(I-\alpha P)(I-\alpha P)^{-1} = \alpha I. \tag{1.12}$$

Subtracting these two equations and rearranging terms, we obtain

$$\alpha P(1-\alpha)(I-\alpha P)^{-1} = (1-\alpha)(I-\alpha P)^{-1} + (\alpha-1)I.$$

By taking the limit as $\alpha \to 1$ and using the definition (1.9), it follows that

$$PP^* = P^*.$$

Also, by reversing the order of $(I-\alpha P)$ and $(I-\alpha P)^{-1}$ in Eqs. (1.11) and (1.12), it follows similarly that $P^*P = P^*$. From $PP^* = P^*$, we also obtain $(I-\alpha P)P^* = (1-\alpha)P^*$ or $P^* = (1-\alpha)(I-\alpha P)^{-1}P^*$, and by taking the limit as $\alpha \to 1$ and by using Eq. (1.9), we have $P^* = P^*P^*$. Thus Eq. (1.6) has been proved.

We have, using Eq. (1.6), $(P-P^*)^2 = P^2 - P^*$, and similarly

$$(P-P^*)^k = P^k - P^*, \qquad k > 0.$$

Therefore,

$$\begin{aligned}(I-\alpha P)^{-1} - (1-\alpha)^{-1}P^* &= \sum_{k=0}^{\infty} \alpha^k (P^k - P^*) \\ &= I - P^* + \sum_{k=1}^{\infty} \alpha^k (P-P^*)^k \\ &= \big(I - \alpha(P-P^*)\big)^{-1} - P^*.\end{aligned}$$

On the other hand, from Eq. (1.10), we have

$$\begin{aligned}H &= \lim_{\alpha \to 1}\big((1-\alpha)^{-1}M(\alpha) - (1-\alpha)^{-1}P^*\big) \\ &= \lim_{\alpha \to 1}\big((I-\alpha P)^{-1} - (1-\alpha)^{-1}P^*\big).\end{aligned}$$

By combining the last two equations, we obtain

$$H = \lim_{\alpha \to 1}\big(I - \alpha(P-P^*)\big)^{-1} - P^* = (I - P + P^*)^{-1} - P^*,$$

which is Eq. (1.5).

From Eq. (1.5), we obtain

$$(I - P + P^*)H = I - (I - P + P^*)P^*$$

or, using Eq. (1.6),

$$H - PH + P^*H = I - P^*. \tag{1.13}$$

Multiplying this relation by P^* and using Eq. (1.6), we obtain $P^*H = 0$, which is Eq. (1.7). Equation (1.8) then follows from Eq. (1.13).

Multiplying Eq. (1.8) with P^k and using Eq. (1.6), we obtain

$$P^* + P^kH = P^k + P^{k+1}H, \qquad k = 0, 1, \ldots$$

Adding this relation over $k = 0, \ldots, N-1$, we have

$$NP^* + H = \sum_{k=0}^{N-1} P^k + P^N H.$$

Dividing by N and taking the limit as $N \to \infty$, we obtain Eq. (1.4). **Q.E.D.**

Note that the matrix P^* of Eq. (1.4) can be used to express concisely the average cost vector J of any Markov chain with transition probability matrix P and cost vector g as

$$J = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k g = \left(\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k \right) g = P^* g.$$

To interpret this equation, note that we may view the ith row of P^* as a vector of steady-state occupancy probabilities corresponding to starting at state i; that is, the ijth element p^*_{ij} of P^* represents the long-term fraction of time that the Markov chain spends at state j given that it starts at state i. Thus the above equation gives the average cost per stage $J(i)$, starting from state i, as the sum $\sum_{j=1}^n p^*_{ij} g_j$ of all the single-stage costs g_j weighted by the corresponding occupancy probabilities.

From Eq. (1.1) and Prop. 1.1, we obtain the following relation between α-discounted and average cost corresponding to a stationary policy.

Proposition 1.2: For any stationary policy μ and $\alpha \in (0,1)$, we have

$$J_{\alpha,\mu} = (1-\alpha)^{-1} J_\mu + h_\mu + O(|1-\alpha|), \tag{1.14}$$

where

$$J_\mu = P^*_\mu g_\mu = \left(\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k_\mu \right) g_\mu$$

is the average cost vector corresponding to μ, and h_μ is a vector satisfying

$$J_\mu + h_\mu = g_\mu + P_\mu h_\mu. \tag{1.15}$$

Proof: Equation (1.14) follows from Eqs. (1.1) and (1.2) with the identifications $P = P_\mu$, $P^* = P^*_\mu$, and $h_\mu = Hg_\mu$. Equation (1.15) follows by multiplying Eq. (1.8) with g_μ and by using the same identifications. **Q.E.D.**

In the next section we use the preceding results to establish Bellman's equation for the average cost per stage problem. As in the earlier chapters, this equation involves the mappings T and T_μ, which take the form

$$(TJ)(i) = \min_{u\in U(i)} \left[g(i,u) + \sum_{j=1}^n p_{ij}(u) J(j) \right], \qquad i = 1, \ldots, n, \tag{1.16}$$

$$(T_\mu J)(i) = g\big(i, \mu(i)\big) + \sum_{j=1}^n p_{ij}\big(\mu(i)\big) J(j), \qquad i = 1, \ldots, n. \tag{1.17}$$

4.2 OPTIMALITY CONDITIONS

Our first result introduces the analog of Bellman's equation for the case of equal optimal cost for each initial state. This is the case that normally appears in practice, as discussed in Section 7.1 of Vol. I. The proposition shows that all solutions of this equation can be identified with the optimal average cost and an associated differential cost. However, it provides no assurance that the equation has a solution. For this we need further assumptions, which will be given in the sequel (see Prop. 2.6).

Proposition 2.1: If a scalar λ and an n-dimensional vector h satisfy

$$\lambda + h(i) = \min_{u\in U(i)} \left[g(i,u) + \sum_{j=1}^n p_{ij}(u) h(j) \right], \qquad i = 1, \ldots, n, \tag{2.1}$$

or equivalently

$$\lambda e + h = Th, \tag{2.2}$$

then λ is the optimal average cost per stage $J^*(i)$ for all i,

$$\lambda = \min_\pi J_\pi(i) = J^*(i), \qquad i = 1, \ldots, n. \tag{2.3}$$

Furthermore, if $\mu^*(i)$ attains the minimum in Eq. (2.1) for each i, the stationary policy μ^* is optimal, that is, $J_{\mu^*}(i) = \lambda$ for all i.

Proof: Let $\pi = \{\mu_0, \mu_1, \ldots\}$ be any admissible policy and let N be a positive integer. We have, from Eq. (2.2),

$$T_{\mu_{N-1}} h \geq \lambda e + h.$$

By applying $T_{\mu_{N-2}}$ to both sides of this relation, and by using the monotonicity of $T_{\mu_{N-2}}$ and Eq. (2.2), we see that

$$T_{\mu_{N-2}} T_{\mu_{N-1}} h \geq T_{\mu_{N-2}}(\lambda e + h) = \lambda e + T_{\mu_{N-2}} h \geq 2\lambda e + h.$$

Continuing in the same manner, we finally obtain

$$T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} h \geq N\lambda e + h, \tag{2.4}$$

with equality if each μ_k, $k = 0, 1, \ldots, N-1$, attains the minimum in Eq. (2.1). As discussed in Section 1.1, $(T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} h)(i)$ is equal to the N-stage cost corresponding to initial state i, policy $\{\mu_0, \mu_1, \ldots, \mu_{N-1}\}$, and terminal cost function h; that is,

$$(T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} h)(i) = E\left\{ h(x_N) + \sum_{k=0}^{N-1} g\big(x_k, \mu_k(x_k)\big) \;\Big|\; x_0 = i, \pi \right\}.$$

Using this relation in Eq. (2.1) and dividing by N, we obtain for all i

$$\frac{1}{N} E\big\{h(x_N) \mid x_0 = i, \pi\big\} + \frac{1}{N} E\left\{ \sum_{k=0}^{N-1} g\big(x_k, \mu_k(x_k)\big) \;\Big|\; x_0 = i, \pi \right\} \geq \lambda + \frac{1}{N} h(i).$$

By taking the limit as $N \to \infty$, we see that

$$J_\pi(i) \geq \lambda, \qquad i = 1, \ldots, n,$$

with equality if $\mu_k(i)$, $k = 0, 1, \ldots$, attains the minimum in Eq. (2.1). **Q.E.D.**

Note that the proof of Prop. 2.1 carries through even if the state space and control space are infinite as long as the function h is bounded and the minimum in the optimality equation (2.1) is attained for each i.

In order to interpret the vector h in Bellman's equation $\lambda e + h = Th$, note that by iterating this equation N times (see also the proof of the preceding proposition), we obtain $N\lambda e + h = T^N h$. Thus for any two states i and j we have

$$\lambda + h(i) = (T^N h)(i), \qquad \lambda + h(j) = (T^N h)(j),$$

which by subtraction yields

$$h(i) - h(j) = (T^N h)(i) - (T^N h)(j), \qquad \text{for all } i, j.$$

For any i, $(T^N h)(i)$ is the optimal N-stage expected cost starting at i when the terminal cost function is h. Thus, according to the preceding equation, $h(i) - h(j)$ represents, for every N, the difference in optimal N-stage expected cost due to starting at state i rather that starting at state j. Based on this interpretation, we refer to h as the *differential* or *relative* cost vector. (An alternative but similar interpretation is given in Section 7.4 of Vol. I.)

Now given a stationary policy μ, we may consider, as in Section 1.2, a problem where the constraint set $U(i)$ is replaced by the set $\tilde{U}(i) = \{\mu(i)\}$;

that is, $\tilde{U}(i)$ contains a single element, the control $\mu(i)$. Since then we would have only one admissible policy, the policy μ, application of Prop. 2.1 yields the following corollary.

Corollary 2.1.1: Let μ be a stationary policy. If a scalar λ_μ and an n-dimensional vector h_μ satisfy, for all i,

$$\lambda_\mu + h_\mu(i) = g\big(i, \mu(i)\big) + \sum_{j=1}^{n} p_{ij}\big(\mu(i)\big) h_\mu(j), \tag{2.5}$$

or equivalently

$$\lambda_\mu e + h_\mu = T_\mu h_\mu,$$

then

$$\lambda_\mu = J_\mu(i), \qquad i = 1, \ldots, n.$$

Blackwell Optimal Policies

It turns out that the converse of Prop. 2.1 also holds; that is, if for some scalar λ we have $J^*(i) = \lambda$ for all $i = 1, \ldots, n$, then λ together with a vector h satisfies Bellman's equation (2.1). We show this by introducing the notion of a Blackwell optimal policy, which was first formulated in [Bla62], together with the line of analysis of the present section.

Definition 1.1: A stationary policy μ is said to be *Blackwell optimal* if it is simultaneously optimal for all the α-discounted problems with α in an interval $(\overline{\alpha}, 1)$, where $\overline{\alpha}$ is some scalar with $0 < \overline{\alpha} < 1$.

The following proposition provides a useful characterization of Blackwell optimal policies, and essentially shows the converse of Prop. 2.1.

Proposition 2.2: The following hold true:

(a) A Blackwell optimal policy is optimal for the average cost problem within the class of all stationary policies.

(b) There exists a Blackwell optimal policy.

Proof: (a) If μ^* is Blackwell optimal, then for all stationary policies μ

and α in an interval $(\overline{\alpha}, 1)$ we have $J_{\alpha,\mu^*} \leq J_{\alpha,\mu}$. Equivalently, using Eq. (1.11),

$$(1-\alpha)^{-1}J_{\mu^*}+h_{\mu^*}+O(|1-\alpha|) \leq (1-\alpha)^{-1}J_{\mu}+h_{\mu}+O(|1-\alpha|), \quad \alpha \in (\overline{\alpha}, 1)$$

or

$$J_{\mu^*} \leq J_{\mu} + (1-\alpha)(h_{\mu} - h_{\mu^*}) + (1-\alpha)O(|1-\alpha|), \qquad \alpha \in (\overline{\alpha}, 1).$$

By taking the limit as $\alpha \to 1$, we obtain $J_{\mu^*} \leq J_{\mu}$.

(b) From Eq. (1.1), we know that, for each μ and state i, $J_{\alpha,\mu}(i)$ is a rational function of α, that is, a ratio of two polynomials in α. Therefore, for any two policies μ and μ' the graphs of $J_{\alpha,\mu}(i)$ and $J_{\alpha,\mu'}(i)$ either coincide or cross only a finite number of times in the interval $(0, 1)$. Since there are only a finite number of policies, we conclude that for each state i there is a policy μ' and a scalar $\overline{\alpha}_i \in (0, 1)$ such that μ' is optimal for the α-discounted problem for $\alpha \in (\overline{\alpha}_i, 1)$ when the initial state is i. Consider the stationary policy defined for each i by $\mu^*(i) = \mu'(i)$. Then $\mu^*(i)$ attains the minimum in Bellman's equation for the α-discounted problem

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

for all i and for all α in the interval $\left(\max_i \overline{\alpha}_i, 1\right)$. Therefore, μ^* is a stationary optimal policy for the α-discounted problem for all α in $\left(\max_i \overline{\alpha}_i, 1\right)$, implying that μ^* is Blackwell optimal. **Q.E.D.**

We note that the converse of Prop. 2.2(a) is not true; it is possible that a stationary average cost optimal policy is not Blackwell optimal (see Exercise 4.6). We mention also that one can show a stronger result than Prop. 2.2(b), namely that a Blackwell optimal policy is average cost optimal within the class of all policies (not just those that are stationary; see Exercise 4.7).

The next proposition provides a useful characterization of Blackwell optimal policies.

Proposition 2.3: If μ^* is Blackwell optimal, then for all stationary policies μ we have

$$J_{\mu^*} = P_{\mu^*} J_{\mu^*} \leq P_{\mu} J_{\mu^*}. \tag{2.6}$$

Furthermore, for all μ such that $P_{\mu^*} J_{\mu^*} = P_{\mu} J_{\mu^*}$, we have

$$J_{\mu^*} + h_{\mu^*} = g_{\mu^*} + P_{\mu^*} h_{\mu^*} \leq g_{\mu} + P_{\mu} h_{\mu^*}, \tag{2.7}$$

where h_{μ^*} is a vector corresponding to μ^* as in Prop. 1.2.

Proof: Since μ^* is optimal for the α-discounted problem for all α in an interval $(\overline{\alpha}, 1)$, we must have, for every μ and $\alpha \in (\overline{\alpha}, 1)$,

$$g_{\mu^*} + \alpha P_{\mu^*} J_{\alpha,\mu^*} \leq g_{\mu} + \alpha P_{\mu} J_{\alpha,\mu^*}. \tag{2.8}$$

From Prop. 1.2, we have, for all $\alpha \in (\overline{\alpha}, 1)$,

$$J_{\alpha,\mu^*} = (1-\alpha)^{-1} J_{\mu^*} + h_{\mu^*} + O(|1-\alpha|).$$

Substituting this expression in Eq. (2.8), we obtain

$$0 \leq g_{\mu} - g_{\mu^*} + \alpha(P_{\mu} - P_{\mu^*})\big((1-\alpha)^{-1} J_{\mu^*} + h_{\mu^*} + O(|1-\alpha|)\big), \tag{2.9}$$

or equivalently

$$0 \leq (1-\alpha)(g_{\mu} - g_{\mu^*}) + \alpha(P_{\mu} - P_{\mu^*})\big(J_{\mu^*} + (1-\alpha)h_{\mu^*} + O\big((1-\alpha)^2\big)\big).$$

By taking the limit as $\alpha \to 1$, we obtain the desired relation $P_{\mu^*} J_{\mu^*} \leq P_{\mu} J_{\mu^*}$.

If μ is such that $P_{\mu^*} J_{\mu^*} = P_{\mu} J_{\mu^*}$, then from Eq. (2.9) we obtain

$$0 \leq g_{\mu} - g_{\mu^*} + \alpha(P_{\mu} - P_{\mu^*})\big(h_{\mu^*} + O(|1-\alpha|)\big).$$

By taking the limit as $\alpha \to 1$ and by using also the relation $J_{\mu^*} + h_{\mu^*} = g_{\mu^*} + P_{\mu^*} h_{\mu^*}$ [cf. Eq. (1.15)], we obtain the desired relation (2.7). **Q.E.D.**

As a consequence of the preceding proposition, we obtain a converse of Prop. 2.1.

Proposition 2.4: If the optimal average cost over the class of stationary policies is equal to λ for all initial states, then there exists a vector h such that

$$\lambda + h(i) = \min_{u \in U(i)} \left[g(i,u) + \sum_{j=1}^{n} p_{ij}(u) h(j) \right], \qquad i = 1, \ldots, n, \tag{2.10}$$

or equivalently

$$\lambda e + h = Th.$$

Proof: Let μ^* be a Blackwell optimal policy. We then have $J_{\mu^*}(i) = \lambda$ for all i. For every μ, each element of the vector $P_{\mu} J_{\mu^*}$ is equal to λ, so that $P_{\mu^*} J_{\mu^*} = P_{\mu} J_{\mu^*}$. From Eq. (2.7), we then obtain the desired relation (2.10) with $h = h_{\mu^*}$. **Q.E.D.**

Bellman's Equation for a Unichain Policy

We recall from Appendix D of Vol. I that in a finite-state Markov chain, a recurrent class is a set of states that communicate in the sense that from every state of the set, there is a probability of 1 to eventually go to all other states of the set and a probability of 0 to ever go to any state outside the set. There are two kinds of states: those that belong to some recurrent class (these are the states that after they are visited once, they will be visited an infinite number of times with probability 1), and those that are transient (these are the states that with probability 1 will be visited only a finite number of times regardless of the initial state).

Stationary policies whose associated Markov chains have a single recurrent class and a possibly empty set of transient states will play an important role in our development. Such policies are called *unichain*. The state trajectory of the Markov chain corresponding to a unichain policy, is eventually (with probability 1) confined to the recurrent class of states, so the average cost per stage corresponding to all initial states as well as the differential costs of the recurrent states are independent of the stage costs of the transient states. The next proposition shows that for a unichain policy μ, the average cost per stage is the same for all initial states, and that Bellman's equation $\lambda_\mu e + h_\mu = T_\mu h_\mu$ holds. Furthermore, we show that Bellman's equation has a unique solution, provided we fix the differential cost of some state at some arbitrary value (0, for example). This is necessary, since if λ_μ and h_μ satisfy Bellman's equation (2.5), the same is true for λ_μ and $h_\mu + \gamma e$, where γ is any scalar.

Proposition 2.5: Let μ be a unichain policy. Then:

(a) There exists a constant λ_μ and a vector h_μ such that

$$J_\mu(i) = \lambda_\mu, \qquad i = 1, \ldots, n, \tag{2.11}$$

and

$$\lambda_\mu + h_\mu(i) = g\big(i, \mu(i)\big) + \sum_{j=1}^{n} p_{ij}\big(\mu(i)\big) h_\mu(j), \qquad i = 1, \ldots, n. \tag{2.12}$$

(b) Let t be a fixed state. The system of the $n+1$ linear equations

$$\lambda + h(i) = g\big(i, \mu(i)\big) + \sum_{j=1}^{n} p_{ij}\big(\mu(i)\big) h(j), \qquad i = 1, \ldots, n, \tag{2.13}$$

$$h(t) = 0, \tag{2.14}$$

in the $n+1$ unknowns $\lambda, h(1), \ldots, h(n)$ has a unique solution.

Proof: (a) Let t be a recurrent state under μ. For each state $i \neq t$ let C_i and N_i be the expected cost and the expected number of stages, respectively, to reach t for the first time starting from i under policy μ. Let also C_t and N_t be the expected cost and expected number of stages, respectively, to return to t for the first time starting from t under policy μ. From Prop. 1.1 in Section 2.1, we have that C_i and N_i solve uniquely the systems of equations

$$C_i = g\big(i, \mu(i)\big) + \sum_{j=1, j\neq t}^{n} p_{ij}\big(\mu(i)\big) C_j, \qquad i = 1, \ldots, n, \tag{2.15}$$

$$N_i = 1 + \sum_{j=1, j\neq t}^{n} p_{ij}\big(\mu(i)\big) N_j, \qquad i = 1, \ldots, n. \tag{2.16}$$

Let

$$\lambda_\mu = \frac{C_t}{N_t}. \tag{2.17}$$

Multiplying Eq. (2.16) by λ_μ and subtracting it from Eq. (2.15), we obtain

$$C_i - \lambda_\mu N_i = g\big(i, \mu(i)\big) - \lambda_\mu + \sum_{j=1, j\neq t}^{n} p_{ij}\big(\mu(i)\big)(C_j - \lambda_\mu N_j), \qquad i = 1, \ldots, n.$$

By defining

$$h_\mu(i) = C_i - \lambda_\mu N_i, \qquad i = 1, \ldots, n, \tag{2.18}$$

and by noting that from Eq. (2.17), we have

$$h_\mu(t) = 0,$$

we obtain

$$\lambda_\mu + h_\mu(i) = g\big(i, \mu(i)\big) + \sum_{j=1}^{n} p_{ij}\big(\mu(i)\big) h_\mu(j), \qquad i = 1, \ldots, n,$$

which is Eq. (2.12). Equation (2.11) follows from Eq. (2.12) and Cor. 2.1.1.

(b) By part (a), for any solution (λ, h) of the system of equations (2.13) and (2.14), we have $\lambda = \lambda_\mu$, as well as $h(t) = 0$. Suppose that t belongs to

the recurrent class of states of the Markov chain corresponding to μ. Then, in view of Eq. (2.14), the system of equations (2.13) can be written as

$$h(i) = g\big(i,\mu(i)\big) - \lambda_\mu + \sum_{j=1,\, j\neq t}^{n} p_{ij}\big(\mu(i)\big)h(j), \qquad i = 1,\ldots,n,\ i \neq t,$$

and is the same as Bellman's equation for a corresponding stochastic shortest path problem where t is the termination state, $g\big(i,\mu(i)\big) - \lambda_\mu$ is the expected stage cost at state i, and $h(i)$ is the average cost starting from i up to reaching t. By Prop. 1.2 in Section 2.1, this system has a unique solution, so $h(i)$ is uniquely defined by Eq. (2.13) for all $i \neq t$.

Suppose now that t is a transient state of the Markov chain corresponding to μ. Then we choose another state $\bar{t}$ that belongs to the recurrent class and make the transformation of variables $\overline{h}(i) = h(i) - h(\bar{t})$. The system of equations (2.13) and (2.14) can be written in terms of the variables λ and $\overline{h}(i)$ as

$$\overline{h}(i) = g\big(i,\mu(i)\big) - \lambda + \sum_{j=1,\, j\neq \bar{t}}^{n} p_{ij}\big(\mu(i)\big)\overline{h}(j), \qquad i = 1,\ldots,n,\ i \neq \bar{t},$$

$$\overline{h}(\bar{t}) = 0,$$

so by the stochastic shortest path argument given earlier, it has a unique solution, implying that the solution of the system of equations (2.13) and (2.14) is also unique. **Q.E.D.**

Conditions for Equal Optimal Cost for All Initial States

We now turn to the case of multiple policies, and we provide conditions under which Bellman's equation $\lambda e + h = Th$ has a solution, and by Prop. 2.1, the optimal cost is independent of the initial state.

Proposition 2.6: Assume any one of the following three conditions:

(1) Every policy that is optimal within the class of stationary policies is unichain.

(2) For every two states i and j, there exists a stationary policy π (depending on i and j) such that, for some k,

$$P(x_k = j \mid x_0 = i, \pi) > 0.$$

(3) There exists a state t, and constants $L > 0$ and $\overline{\alpha} \in (0,1)$ such that

$$|J_\alpha(i) - J_\alpha(t)| \leq L, \qquad \text{for all } i = 1,\ldots,n, \text{ and } \alpha \in (\overline{\alpha},1), \tag{2.19}$$

where J_α is the α-discounted optimal cost vector.

Then the optimal average cost per stage has the same value λ for all initial states i. Furthermore, λ satisfies

$$\lambda = \lim_{\alpha\to 1}(1-\alpha)J_\alpha(i), \qquad i = 1,\ldots,n, \tag{2.20}$$

and for any state t, the vector h given by

$$h(i) = \lim_{\alpha\to 1}\big(J_\alpha(i) - J_\alpha(t)\big), \qquad i = 1,\ldots,n, \tag{2.21}$$

satisfies Bellman's equation

$$\lambda e + h = Th. \tag{2.22}$$

together with λ.

Proof: Assume condition (1). Proposition 2.2 asserts that a Blackwell optimal policy exists and is optimal within the class of stationary policies. Therefore, by condition (1), this policy is unichain, and by Prop. 2.5, the corresponding average cost is independent of the initial state. The result follows from Prop. 2.4.

Assume condition (2). Consider a Blackwell optimal policy μ^*. If it yields average cost that is independent of the initial state, we are done, as earlier. Assume the contrary; that is, both the set

$$M = \left\{ i \;\middle|\; J_{\mu^*}(i) = \max_j J_{\mu^*}(j) \right\}$$

and its complement $\overline{M}$ are nonempty. The idea now is to use the hypothesis that every pair of states communicates under some stationary policy, in order to show that the average cost of states in M can be reduced by opening communication to the states in $\overline{M}$, thereby creating a contradiction. Take any states $i \in M$ and $j \in \overline{M}$, and a stationary policy μ such that, for some k, $P(x_k = j \mid x_0 = i, \mu) > 0$. Then there must exist states $m \in M$ and $\overline{m} \in \overline{M}$ such that there is a positive transition probability from m to $\overline{m}$ under μ; that is, $[P_\mu]_{m\overline{m}} = P(x_{k+1} = \overline{m} \mid x_k = m, \mu) > 0$. It can thus be seen that the mth component of $P_\mu J_{\mu^*}$ is strictly less than

$\max_i J_{\mu^*}(i)$, which is equal to the mth component of J_{μ^*}. This contradicts the necessary condition (2.6).

Finally, assume condition (3). Let μ^* be a Blackwell optimal policy. By Eq. (1.14), we have for all states i and α in some interval $(\overline{\alpha}, 1)$

$$J_\alpha(i) = (1-\alpha)^{-1} J_{\mu^*}(i) + h_{\mu^*}(i) + O(|1-\alpha|). \tag{2.23}$$

Writing this equation for state i and for state t, and subtracting, we obtain for all $i \neq t$,

$$\big|J_{\mu^*}(i) - J_{\mu^*}(t)\big| \leq (1-\alpha)\big|J_\alpha(i) - J_\alpha(t)\big| + (1-\alpha)\big|h_{\mu^*}(i) - h_{\mu^*}(t)\big| + O\big((1-\alpha)^2\big).$$

Taking the limit as $\alpha \to 1$ and using the hypothesis that $\big|J_\alpha(i) - J_\alpha(t)\big| \leq L$ for all $\alpha \in (0,1)$, we obtain that $J_{\mu^*}(i) = J_{\mu^*}(t)$ for all i. Thus the average cost of the Blackwell optimal policy is independent of the initial state, and we are done.

To show Eqs. (2.20)-(2.22), we note that the relation $\lim_{\alpha \to 1}(1-\alpha)J_\alpha(i) = \lambda$ for all i follows from Eq. (2.23) and the fact $J_{\mu^*}(i) = \lambda$ for all i. Also, from Eq. (2.23), we have

$$J_\alpha(i) - J_\alpha(t) = h_{\mu^*}(i) - h_{\mu^*}(t) + O(|1-\alpha|),$$

so that

$$\lim_{\alpha \to 1}\big(J_\alpha(i) - J_\alpha(t)\big) = h_{\mu^*}(i) - h_{\mu^*}(t).$$

Setting $h(i) = h_{\mu^*}(i) - h_{\mu^*}(t)$ for all i, and using the fact $J_\mu(i) = \lambda$ for all i and Eq. (2.7), we see that the condition $\lambda e + h = Th$ is satisfied. **Q.E.D.**

The conditions of the preceding proposition are among the weakest guaranteeing that the optimal average cost per stage is independent of the initial state. In particular, it is clear that some sort of accessibility condition must be satisfied by the transition probability matrices corresponding to stationary policies or at least to optimal stationary policies. For if there existed two states neither of which could be reached from the other no matter which policy we use, then it can be only by accident that the same optimal cost per stage will correspond to each one. An extreme example is a problem where the state is forced to stay the same regardless of the control applied (i.e., each state is absorbing). Then the optimal average cost per stage for each state i is $\min_{u \in U(i)} g(i,u)$, and this cost may be different for different states.

Example 2.1: (Machine Replacement)

Consider a machine that can be in any one of n states, $1, 2, \ldots, n$. There is a cost $g(i)$ for operating for one time period the machine when it is in state i. The options at the start of each period are to (a) let the machine

operate one more period in the state it currently is, or (b) repair the machine at a positive cost R and bring it to state 1 (corresponding to a machine in perfect condition). The transitions between different states over each time period are governed by given probabilities p_{ij}. Once repaired, the machine is guaranteed to stay in state 1 for one period, and in subsequent periods, it may deteriorate to states $j \geq 1$ according to the transition probabilities p_{1j}. The problem is to find a policy that minimizes the average cost per stage. Note that we have analyzed the discounted cost version of this problem in Example 2.1 of Section 1.2. As in that example, we will assume that $g(i)$ is nondecreasing in i, and that the transition probabilities satisfy

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

for all functions $J(i)$, which are monotonically nondecreasing in i.

Note that not all policies are unichain here. For example, consider the stationary policy that replaces at every state except the worst state n (a poor but legitimate choice). The corresponding Markov chain has two recurrent classes, $\{1, 2, \ldots, n-1\}$ and $\{n\}$ (assuming that $p_{1n} = 0$). It can also be seen that condition (2) of Prop. 2.6 is not guaranteed in the absence of further assumptions. [This condition is satisfied if we assume in addition that for all i we have $p_{i(i+1)} > 0$, because, by replacing, we can bring the system to state 1, from where, by not replacing, we can reach every other state.]

We can show, however, that condition (3) of Prop. 2.6 is satisfied. Indeed, consider the corresponding discounted problem with a discount factor $\alpha < 1$. We have for all i

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

and in particular,

$$J_\alpha(i) \leq R + g(1) + \alpha J_\alpha(1),$$

$$J_\alpha(1) = \min\left[R + g(1) + \alpha J_\alpha(1),\ g(1) + \alpha \sum_{j=1}^{n} p_{1j} J_\alpha(j)\right].$$

From the last two equations, by subtraction we obtain

$$J_\alpha(i) - J_\alpha(1) \leq \max\left[0,\ R + \alpha\left(J_\alpha(1) - \sum_{j=1}^{n} p_{1j} J_\alpha(j)\right)\right] \leq R,$$

where the last inequality follows from the fact

$$0 \leq J_\alpha(i) - J_\alpha(1), \qquad i = 1, \ldots, n,$$

which holds since $J_\alpha(i) - J_\alpha(1)$ is nondecreasing in i, as shown in Example 2.1 of Section 1.2. The last two relations imply that condition (3) of Prop.

2.6 is satisfied, and it follows that there exists a scalar λ and a vector h, such that for all i,

$$\lambda + h(i) = \min\left[R + g(1) + h(1),\ g(i) + \sum_{j=1}^{n} p_{ij}h(j)\right],$$

while the policy that chooses the minimizing action above is average cost optimal.

By Prop. 2.6, we can take $h(i) = \lim_{\alpha\to 1}\big(J_\alpha(i) - J_\alpha(1)\big)$, and since $J_\alpha(i) - J_\alpha(1)$ is nondecreasing in i, it follows that $h(i)$ is also nondecreasing in i. Similar to Example 2.1 of Section 1.2, this implies that an optimal policy takes the form

replace if and only if $i \geq i^*$,

where

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

and

$$S_R = \left\{ i \;\middle|\; R + g(1) + h(1) \leq g(i) + \sum_{j=1}^{n} p_{ij}h(j) \right\}.$$

4.3 COMPUTATIONAL METHODS

All the computational methods developed for discounted and stochastic shortest path problems (cf. Sections 1.3 and 2.2) have average cost per stage counterparts, which we discuss in this section. However, the derivations of these methods are often intricate, and have no direct analogs in the discounted and stochastic shortest path context. In fact, the validity of these methods may depend on assumptions that relate to the structure of the underlying Markov chains, something that we have not encountered so far.

4.3.1 Value Iteration

The natural version of the value iteration method for the average cost problem is simply to generate successively the finite horizon optimal costs T^kJ_0, $k = 1, 2, \ldots$, starting with the zero function J_0. It is then natural to speculate that the k-stage average costs T^kJ_0/k converge to the optimal average cost vector as $k \to \infty$ (this is in fact proved under natural conditions in Section 7.4 of Vol. I). This method has two drawbacks. First, some of the components of T^kJ_0 typically diverge to ∞ or $-\infty$, so

direct calculation of $\lim_{k\to\infty} T^kJ_0/k$ is numerically impractical. Second, this method will not provide us with a corresponding differential cost vector h.

We can bypass both difficulties by subtracting a multiple of the unit vector e from T^kJ_0, so that the difference, call it h^k, remains bounded. In particular, we consider methods of the form

$$h^k = T^kJ_0 - \delta^k e, \tag{3.1}$$

where δ^k is some scalar satisfying

$$\min_{i=1,\ldots,n}(T^kJ_0)(i) \leq \delta^k \leq \max_{i=1,\ldots,n}(T^kJ_0)(i),$$

such as for example the average of $(T^kJ_0)(i)$

$$\delta^k = \frac{1}{n}\sum_{i=1}^{n}(T^kJ_0)(i),$$

or

$$\delta^k = (T^kJ_0)(t),$$

where t is some fixed state. Then if the differences $\max_i(T^kJ_0)(i) - \min_i(T^kJ_0)(i)$ remain bounded as $k \to \infty$ (this can be guaranteed under the assumptions of the subsequent Prop. 3.1), the vectors h^k also remain bounded, and we will see that with a proper choice of the scalar δ^k, the vectors h^k converge to a differential cost vector.

Let us now restate the algorithm $h^k = T^kJ_0 - \delta^k e$ in a form that is suitable for iterative calculation. We have

$$h^{k+1} = T^{k+1}J_0 - \delta^{k+1}e,$$

and since

$$T^{k+1}J_0 = T(T^kJ_0) = T(h^k + \delta^k e) = Th^k + \delta^k e,$$

we obtain

$$h^{k+1} = Th^k + (\delta^k - \delta^{k+1})e. \tag{3.2}$$

In the case where δ^k is given by the average of $(T^kJ_0)(i)$, we have

$$\delta^{k+1} = \frac{1}{n}\sum_{i=1}^{n}(T^{k+1}J_0)(i) = \frac{1}{n}\sum_{i=1}^{n}\big(T(h^k + \delta^k e)\big)(i) = \frac{1}{n}\sum_{i=1}^{n}(Th^k)(i) + \delta^k,$$

so that the iteration (3.2) is written as

$$h^{k+1} = Th^k - \frac{1}{n}\sum_{i=1}^{n}(Th^k)(i)e. \tag{3.3}$$

Similarly, in the case where we fix a state t and we choose $\delta^k = (T^k J_0)(t)$, we have

$$\delta^{k+1} = (T^{k+1}J_0)(t) = \big(T(h^k + \delta^k e)\big)(t) = (Th^k)(t) + \delta^k,$$

and the iteration (3.2) is written as

$$h^{k+1} = Th^k - (Th^k)(t)e. \tag{3.4}$$

We will henceforth restrict attention to the case where $\delta^k = (T^k J_0)(t)$, and we will call the corresponding algorithm (3.4) *relative value iteration*, since the iterate h^k is equal to $T^k J_0 - (T^k J_0)(t)e$ and may be viewed as a k-stage optimal cost vector *relative to state* t. The following results also apply to other versions of the algorithm (see Exercises 4.4 and 4.5). Note that relative value iteration, which generates h^k, is not really different than ordinary value iteration, which generates $T^k J_0$. The vectors generated by the two methods merely differ by a multiple of the unit vector, and the minimization problems involved in the corresponding iterations of the two methods are mathematically equivalent.

It can be seen that if the relative value iteration (3.4) converges to some vector h, then

$$(Th)(t)e + h = Th,$$

which by Prop. 2.1, implies that $(Th)(t)$ is the optimal average cost per stage for all initial states, and h is an associated differential cost vector. Thus convergence can only be expected when the optimal average cost per stage is independent of the initial state, indicating that at least one of the conditions of Prop. 2.6 is required. However, it turns out that a stronger hypothesis is needed for convergence. The following example illustrates the reason.

Example 3.1:

Consider the iteration

$$h^{k+1} = T_\mu h^k - (T_\mu h^k)(t)e,$$

which is the relative value iteration (3.4) for the case of a fixed μ. Using the expressions $T_\mu h^k = g_\mu + P_\mu h^k$ and $(T_\mu h^k)(t) = c_t'(g_\mu + P_\mu h^k)$, where c_t' is the row vector having all coordinates equal to 0 except for coordinate t which is equal to 1, this iteration can be written as

$$h^{k+1} = g_\mu + P_\mu h^k - ec_t'(g_\mu + P_\mu h^k).$$

Equivalently, we have

$$h^{k+1} = (I - ec_t')g_\mu + \hat{P}_\mu h^k, \tag{3.5}$$

where

$$\hat{P}_\mu = (I - ec_t')P_\mu. \tag{3.6}$$

Convergence of iteration (3.5) depends on whether all the eigenvalues of $\hat{P}_\mu$ lie strictly within the unit circle. We have for any eigenvalue γ of P_μ with corresponding eigenvector v,

$$P_\mu v = (I - ec_t')P_\mu v = \gamma(v - ec_t'v),$$

and in particular, for the eigenvalue $\gamma = 1$ and the corresponding eigenvector $v = e$ we obtain using the fact $c_t'e = 1$,

$$\hat{P}_\mu e = 0.$$

Therefore, we have

$$\hat{P}_\mu(v - ec_t'v) = \gamma(v - ec_t'v),$$

and it follows that each eigenvalue γ of P_μ with corresponding eigenvector v, which is not a scalar multiple of e, is also an eigenvalue of $\hat{P}_\mu$ with corresponding eigenvector $(v - ec_t'v)$. Thus, if P_μ has an eigenvalue $\gamma \neq 1$ that is *on* the unit circle, the iteration (3.5) is not convergent. This occurs when P_μ has a periodic structure and some of its nonunity eigenvalues are on the unit circle. For example, suppose that

$$P_\mu = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

which has eigenvalues 1 and -1. Then taking $t = 1$, the matrix $\hat{P}_\mu$ of Eq. (3.6) is given by

$$\hat{P}_\mu = \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix}(1 \quad 0)\right)\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix},$$

and has eigenvalues 0 and -1. As a result, iteration (3.5) does not converge even though μ is a unichain policy.

The following proposition shows convergence of the relative value iteration (3.4) under a technical condition that excludes situations such as the one of the preceding example. When there is only one control available per state, that is, there is only one stationary policy μ, the condition of the following proposition requires that for some positive integer m, the matrix P_μ^m has at least one column all the components of which are positive. As can be seen from the preceding example, this condition need not hold if μ is unichain. However, we will later provide a variant of the relative value iteration (3.4) that converges under the weaker condition that all stationary policies are unichain (see Prop. 3.3).

Proposition 3.1: Assume that there exists a positive integer m such that for every admissible policy $\pi = \{\mu_0, \mu_1, \ldots\}$, there exists an $\epsilon > 0$ and a state s such that

$$[P_{\mu_m} P_{\mu_{m-1}} \cdots P_{\mu_1}]_{is} \geq \epsilon, \qquad i = 1, \ldots, n, \tag{3.7}$$

$$[P_{\mu_{m-1}} P_{\mu_{m-2}} \cdots P_{\mu_0}]_{is} \geq \epsilon, \qquad i = 1, \ldots, n, \tag{3.8}$$

where $[\cdot]_{is}$ denotes the element of the ith row and sth column of the corresponding matrix. Fix a state t and consider the relative value iteration algorithm

$$h^{k+1}(i) = (Th^k)(i) - (Th^k)(t), \qquad i = 1, \ldots, n, \tag{3.9}$$

where $h^0(i)$ are arbitrary scalars. Then the sequence $\{h^k\}$ converges to a vector h satisfying $(Th)(t)e + h = Th$, so that by Prop. 2.1, $(Th)(t)$ is equal to the optimal average cost per stage for all initial states and h is an associated differential cost vector.

Proof: Denote

$$q^k(i) = h^{k+1}(i) - h^k(i), \qquad i = 1, 2, \ldots, n.$$

We will show that for all i and $k \geq m$ we have

$$\max_i q^k(i) - \min_i q^k(i) \leq (1 - \epsilon) \left(\max_i q^{k-m}(i) - \min_i q^{k-m}(i) \right), \tag{3.10}$$

where m and ϵ are as stated in the hypothesis. From this relation we then obtain, for some $B > 0$ and all k,

$$\max_i q^k(i) - \min_i q^k(i) \leq B(1 - \epsilon)^{k/m}.$$

Since $q^k(t) = 0$, it follows that, for all i,

$$\left| h^{k+1}(i) - h^k(i) \right| = |q^k(i)| \leq \max_j q^k(j) - \min_j q^k(j) \leq B(1 - \epsilon)^{k/m}.$$

Therefore, for every $r > 1$ and i we have

$$\begin{aligned} \left| h^{k+r}(i) - h^k(i) \right| &\leq \sum_{l=0}^{r-1} \left| h^{k+l+1}(i) - h^{k+l}(i) \right| \\ &\leq B(1 - \epsilon)^{k/m} \sum_{l=0}^{r-1} (1 - \epsilon)^{l/m} \\ &= \frac{B(1 - \epsilon)^{k/m} \left(1 - (1 - \epsilon)^{r/m}\right)}{1 - (1 - \epsilon)^{1/m}}, \end{aligned} \tag{3.11}$$

so that $\{h^k(i)\}$ is a Cauchy sequence and converges to a limit $h(i)$. From Eq. (3.9) we see then that the equation $(Th)(t) + h(i) = (Th)(i)$ holds for all i. It will thus be sufficient to prove Eq. (3.10).

To prove Eq. (3.10), we denote by $\mu_k(i)$ the control that attains the minimum in the relation

$$(Th^k)(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^{n} p_{ij}(u) h^k(j) \right], \tag{3.12}$$

for every k and i. Denote

$$\lambda_k = (Th^k)(t).$$

Then we have

$$h^{k+1} = g_{\mu_k} + P_{\mu_k} h^k - \lambda_k e \leq g_{\mu_{k-1}} + P_{\mu_{k-1}} h^k - \lambda_k e,$$

$$h^k = g_{\mu_{k-1}} + P_{\mu_{k-1}} h^{k-1} - \lambda_{k-1} e \leq g_{\mu_k} + P_{\mu_k} h^{k-1} - \lambda_{k-1} e,$$

where $e = (1, \ldots, 1)'$ is the unit vector. From these relations, using the definition $q^k = h^{k+1} - h^k$, we obtain

$$P_{\mu_k} q^{k-1} + (\lambda_{k-1} - \lambda_k) e \leq q^k \leq P_{\mu_{k-1}} q^{k-1} + (\lambda_{k-1} - \lambda_k) e.$$

Since this relation holds for every $k \geq 1$, by iterating we obtain

$$\begin{aligned} P_{\mu_k} \cdots P_{\mu_{k-m+1}} q^{k-m} + (\lambda_{k-m} - \lambda_k) e \leq q^k \\ \leq P_{\mu_{k-1}} \cdots P_{\mu_{k-m}} q^{k-m} + (\lambda_{k-m} - \lambda_k) e. \end{aligned} \tag{3.13}$$

First, let us assume that the special state s corresponding to $\mu_{k-m}, \ldots, \mu_k$ as in Eqs. (3.7) and (3.8) is the fixed state t used in iteration (3.9); that is,

$$[P_{\mu_k} \cdots P_{\mu_{k-m+1}}]_{it} \geq \epsilon, \qquad i = 1, \ldots, n, \tag{3.14}$$

$$[P_{\mu_{k-1}} \cdots P_{\mu_{k-m}}]_{it} \geq \epsilon, \qquad i = 1, \ldots, n. \tag{3.15}$$

The right-hand side of Eq. (3.13) yields

$$q^k(i) \leq \sum_{j=1}^{n} [P_{\mu_{k-1}} \cdots P_{\mu_{k-m}}]_{ij} q^{k-m}(j) + \lambda_{k-m} - \lambda_k,$$

so using Eq. (3.15) and the fact $q^{k-m}(t) = 0$, we obtain

$$q^k(i) \leq (1 - \epsilon) \max_j q^{k-m}(j) + \lambda_{k-m} - \lambda_k, \qquad i = 1, \ldots, n,$$

implying that

$$\max_j q^k(j) \leq (1-\epsilon)\max_j q^{k-m}(j) + \lambda_{k-m} - \lambda_k. \tag{3.16}$$

Similarly, from the left-hand side of Eq. (3.13) we obtain

$$\min_j q^k(j) \geq (1-\epsilon)\min_j q^{k-m}(j) + \lambda_{k-m} - \lambda_k, \tag{3.17}$$

and by subtracting the last two relations, we obtain the desired Eq. (3.10).

When the special state s corresponding to $\mu_{k-m}, \ldots, \mu_k$ as in Eqs. (3.7) and (3.8) is not equal to t, we define a related iterative process

$$\tilde{h}^{k+1}(i) = (T\tilde{h}^k)(i) - (T\tilde{h}^k)(s), \qquad i = 1, \ldots, n, \tag{3.18}$$

$$\tilde{h}^0(i) = h^0(i), \qquad i = 1, \ldots, n.$$

Then, as earlier, we have

$$\max_i \tilde{q}^k(i) - \min_i q^k(i) \leq (1-\epsilon)\left(\max_i \tilde{q}^{k-m}(i) - \min_i \tilde{q}^{k-m}(i)\right), \tag{3.19}$$

where

$$\tilde{q}^k = \tilde{h}^{k+1} - \tilde{h}^k.$$

It is straightforward to verify, using Eqs. (3.9) and (3.18), that for all i and k we have

$$h^k(i) = \tilde{h}^k(i) + (T\tilde{h}^{k-1})(s) - (T\tilde{h}^{k-1})(t).$$

Therefore, the coordinates of both h^k and q^k differ from the coordinates of $\tilde{h}^k$ and $\tilde{q}^k$, respectively, by a constant. It follows that

$$\max_i q^k(i) - \min_i q^k(i) = \max_i \tilde{q}^k(i) - \min_i \tilde{q}^k(i),$$

and from Eq. (3.19) we obtain the desired Eq. (3.10). **Q.E.D.**

As a by-product of the preceding proof, we obtain a rate of convergence estimate. By taking the limit in Eq. (3.11) as $r \to \infty$, we obtain

$$\max_i \left|h^k(i) - h(i)\right| \leq \frac{B(1-\epsilon)^{k/m}}{1-(1-\epsilon)^{1/m}}, \qquad k = 0, 1, \ldots,$$

so the bound on the error is reduced by $(1-\epsilon)^{1/m}$ at each iteration. A sharper rate of convergence result can be obtained if we assume that there exists a unique optimal stationary policy μ^*. Then, it is possible to show that the minimum in Eq. (3.12) is attained by $\mu^*(i)$ for all i and all k after a certain index, so for such k, the relative value iteration takes the form $h^{k+1} = T_{\mu^*}h^k - (T_{\mu^*}h^k)(t)e$, and is governed by the largest eigenvalue modulus of the matrix P_{μ^*} given by Eq. (3.6).

Note that contrary to the case of a discounted or a stochastic shortest path problem, the Gauss-Seidel version of the relative value iteration method need not converge. Indeed, the reader can construct examples of such behavior involving two-state systems and a single policy.

Error Bounds

Similar to discounted problems, the relative value iteration method can be strengthened by the calculation of monotonic error bounds.

Proposition 3.2: Under the assumption of Prop. 3.1, the iterates h^k of the relative value iteration method (3.9) satisfy

$$\underline{c}_k \leq \underline{c}_{k+1} \leq \lambda \leq \overline{c}_{k+1} \leq \overline{c}_k, \tag{3.20}$$

where λ is the optimal average cost per stage for all initial states, and

$$\underline{c}_k = \min_i\left[(Th^k)(i) - h^k(i)\right],$$

$$\overline{c}_k = \max_i\left[(Th^k)(i) - h^k(i)\right].$$

Proof: Let $\mu_k(i)$ attain the minimum in

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

for each k and i. We have, using Eq. (3.9),

$$\begin{aligned}(Th^k)(i) &= g\big(i,\mu_k(i)\big) + \sum_{j=1}^n p_{ij}\big(\mu_k(i)\big)h^k(j)\\ &= g\big(i,\mu_k(i)\big) + \sum_{j=1}^n p_{ij}\big(\mu_k(i)\big)(Th^{k-1})(j) - (Th^{k-1})(t),\end{aligned}$$

and

$$h^k(i) \leq g\big(i,\mu_k(i)\big) + \sum_{j=1}^n p_{ij}\big(\mu_k(i)\big)h^{k-1}(j) - (Th^{k-1})(t).$$

Subtracting the last two relations, we obtain

$$(Th^k)(i) - h^k(i) \geq \sum_{j=1}^n p_{ij}\big(\mu_k(i)\big)\big((Th^{k-1})(j) - h^{k-1}(j)\big),$$

and it follows that

$$\min_i\left[(Th^k)(i) - h^k(i)\right] \geq \min_i\left[(Th^{k-1})(i) - h^{k-1}(i)\right],$$

or equivalently

$$\underline{c}_{k-1} \leq \underline{c}_k.$$

A similar argument shows that

$$\overline{c}_k \leq \overline{c}_{k-1}.$$

By Prop. 3.1 we have $h^k(i) \to h(i)$ and $(Th)(i) - h(i) = \lambda$ for all i, so that $\underline{c}_k \to \lambda$. Since $\{\underline{c}_k\}$ is also nondecreasing, we must have $\underline{c}_k \leq \lambda$ for all k. Similarly, $\overline{c}_k \geq \lambda$ for all k. **Q.E.D.**

We now demonstrate the relative value iteration method and the error bounds (3.20) by means of an example.

Example 3.2:

Consider an undiscounted version of the example of Section 1.3. We have

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

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

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

and

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

The mapping T has the form

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

Letting $t = 1$ be the reference state, the relative value iteration (3.9) takes the form

$$h^{k+1}(1) = 0$$

$$h^{k+1}(2) = (Th^k)(2) - (Th^k)(1).$$

The results of the computation starting with $h^0(1) = h^0(2) = 0$ are shown in the table of Fig. 4.3.1.

k	$h^k(1)$	$h^k(2)$	$\underline{c}_k$	$\overline{c}_k$
0	0	0		
1	0	0.500	0.625	0.875
2	0	0.250	0.687	0.812
3	0	0.375	0.719	0.781
4	0	0.312	0.734	0.765
5	0	0.344	0.742	0.758
6	0	0.328	0.746	0.754
7	0	0.336	0.748	0.752
8	0	0.332	0.749	0.751
9	0	0.334	0.749	0.750
10	0	0.333	0.750	0.750

Figure 4.3.1 Iterates and error bounds generated by the relative value iteration method for the problem of Example 3.2.

We note an interesting application of the error bounds of Prop. 3.2. Suppose that for some vector h, we calculate a μ such that

$$T_\mu h = Th.$$

Then by applying Prop. 3.2 to the original problem and also to the modified problem where the only stationary policy is μ, we obtain

$$\underline{c} \leq J^*(i) \leq J_\mu(i) \leq \overline{c},$$

where

$$\underline{c} = \min_i \big[(Th)(i) - h(i)\big], \qquad \overline{c} = \max_i \big[(Th)(i) - h(i)\big].$$

We thus obtain a bound on the degree of suboptimality of μ. This bound can be proved in a more general setting, where $J^*(i)$ is not necessarily independent of the initial state i (see Exercise 4.10).

Other Versions of the Relative Value Iteration Method

As mentioned earlier, the condition for convergence of the relative value iteration method given in Prop. 3.1 is stronger than the conditions of Prop. 2.6 for the optimal average cost per stage to be independent of the initial state. We now show that we can bypass this difficulty by modifying

the problem without affecting either the optimal cost or the optimal policies and by applying the relative value iteration method to the modified problem.

Let τ be any scalar with

$$0 < \tau < 1,$$

and consider the problem that results when each transition matrix P_μ corresponding to a stationary policy μ is replaced by

$$\tilde{P}_\mu = \tau P_\mu + (1-\tau)I, \tag{3.21}$$

where I is the identity matrix. Note that $\tilde{P}_\mu$ is a transition probability matrix with the property that, at every state, a self-transition occurs with probability at least $(1-\tau)$. This destroys any periodic character that P_μ may have. For another view of the same point, note that each eigenvalue of $\tilde{P}_\mu$ is of the form $\tau\gamma + (1-\tau)$, where γ is an eigenvalue of P_μ. Therefore, all eigenvalues $\gamma \neq 1$ of P_μ that lie on the unit circle are mapped into eigenvalues of $\tilde{P}_\mu$ strictly inside the unit circle.

Bellman's equation for the modified problem is

$$\tilde{\lambda}_\mu e + \tilde{h}_\mu = g_\mu + \tilde{P}_\mu \tilde{h}_\mu = g_\mu + \big(\tau P_\mu + (1-\tau)I\big)\tilde{h}_\mu,$$

which can be written as

$$\tilde{\lambda}_\mu e + \tau \tilde{h}_\mu = g_\mu + P_\mu(\tau \tilde{h}_\mu).$$

We observe that this equation is the same as Bellman's equation for the original problem,

$$\lambda_\mu e + h_\mu = g_\mu + P_\mu h_\mu,$$

with the identification

$$h_\mu = \tau \tilde{h}_\mu.$$

It follows from Cor. 2.1.1 that if the average cost per stage for the original problem is independent of i for every μ, then the same is true for the modified problem. Furthermore, the costs of all stationary policies, as well as the optimal cost, are equal for both the original and the modified problem.

Consider now the relative value iteration method (3.9) for the modified problem. A straightforward calculation shows that it takes the form

$$\begin{aligned} h^{k+1}(i) = (1-\tau)h^k(i) &+ \min_{u \in U(i)} \left[g(i,u) + \tau \sum_{j=1}^{n} p_{ij}(u) h^k(j) \right] \\ &- \min_{u \in U(t)} \left[g(t,u) + \tau \sum_{j=1}^{n} p_{tj}(u) h^k(j) \right], \end{aligned} \tag{3.22}$$

where t is some fixed state with $h^0(t) = 0$. Note that this iteration is as easy to execute as the original version. It is convergent, however, under weaker conditions than those required in Prop. 3.1.

Proposition 3.3: Assume that each stationary policy is unichain. Then, for $0 < \tau < 1$, the sequences $\{h^k(i)\}$ generated by the modified relative value iteration (3.22) satisfy

$$\lim_{k \to \infty} h^k(i) = \frac{h(i)}{\tau},$$

$$\lim_{k \to \infty} \min_{u \in U(i)} \left[g(t,u) + \tau \sum_{j=1}^{n} p_{tj}(u) h^k(j) \right] = \lambda, \tag{3.23}$$

where λ is the optimal average cost per stage and h is a differential cost vector.

Proof: The proof consists of showing that the conditions of Prop. 3.1 are satisfied for the modified problem involving the transition probability matrices $\tilde{P}_\mu$ of Eq. (3.21).

Indeed, let $m > nn_M$, where n is the number of states and n_M is the number of distinct stationary policies. Consider a set of control functions $\mu_0, \mu_1, \ldots, \mu_m$. Then at least one μ is repeated n times within the subset $\mu_1, \ldots, \mu_{m-1}$. Let s be a state belonging to the recurrent class of the Markov chain corresponding to μ. Then the conditions

$$[\tilde{P}_{\mu_m} \cdots \tilde{P}_{\mu_1}]_{is} \geq \epsilon, \qquad i = 1, \ldots, n,$$

$$[\tilde{P}_{\mu_{m-1}} \cdots \tilde{P}_{\mu_0}]_{is} \geq \epsilon, \qquad i = 1, \ldots, n,$$

are satisfied for some ϵ because, in view of Eq. (3.21), when there is a positive probability of reaching s from i at some stage, there is also a positive probability of reaching it at any subsequent stage. **Q.E.D.**

Note that, since the modified value iteration method is nothing but the ordinary method applied to a modified problem, the error bounds of Prop. 3.2 apply in appropriately modified form.

4.3.2 Policy Iteration

The policy iteration algorithm for the average cost problem is similar to those described in Sections 1.3 and 2.2. Given a stationary policy, one obtains an improved policy by means of a minimization process until no further improvement is possible. *We will assume throughout this section*

that every stationary policy encountered in the course of the algorithm is unichain.

At the kth step of the policy iteration algorithm, we have a stationary policy μ^k. We then perform a *policy evaluation* step; that is, we obtain corresponding average and differential costs λ^k and $h^k(i)$ satisfying

$$\lambda^k + h^k(i) = g\big(i,\mu^k(i)\big) + \sum_{j=1}^{n} p_{ij}\big(\mu^k(i)\big)h^k(j), \qquad i = 1,\ldots,n, \quad (3.24)$$

or equivalently

$$\lambda^k e + h^k = T_{\mu^k}h^k = g_{\mu^k} + P_{\mu^k}h^k.$$

Note that λ^k and h^k can be computed as the unique solution of the linear system of equations (3.24) together with the normalizing equation $h^k(t) = 0$, where t is any state (cf. Prop. 2.5). This system can be solved either directly or iteratively using the relative value iteration method or by an adaptive aggregation method, as discussed later.

We subsequently perform a *policy improvement* step; that is, we find a stationary policy μ^{k+1}, where for all i, $\mu^{k+1}(i)$ is such that

$$g\big(i,\mu^{k+1}(i)\big) + \sum_{j=1}^{n} p_{ij}\big(\mu^{k+1}(i)\big)h^k(j) = \min_{u\in U(i)}\left[g(i,u) + \sum_{j=1}^{n} p_{ij}(u)h^k(j)\right], \quad (3.25)$$

or equivalently

$$T_{\mu^{k+1}}h^k = Th^k.$$

If $\mu^{k+1} = \mu^k$, the algorithm terminates; otherwise, the process is repeated with μ^{k+1} replacing μ^k.

There is an easy proof, given in Exercise 4.11, that the policy iteration algorithm terminates finitely if we assume that the Markov chain corresponding to each μ^k is irreducible (is unichain and has no transient states). To prove the result without this assumption, we impose the following restriction in the way the algorithm is operated: *if $\mu^k(i)$ attains the minimum in Eq. (3.25), we choose $\mu^{k+1}(i) = \mu^k(i)$ even if there are other controls attaining the minimum in addition to $\mu^k(i)$.* We then have:

Proposition 3.4: If all the generated policies are unichain, the policy iteration algorithm terminates finitely with an optimal stationary policy.

It is convenient to state the main argument needed for the proof of Prop. 3.4 as a lemma:

Lemma 3.1: Let μ be a unichain stationary policy, and let λ and h be corresponding average and differential costs satisfying

$$\lambda e + h = T_\mu h, \quad (3.26)$$

as well as the normalization condition

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1} P_\mu^k h = 0. \quad (3.27)$$

[The above limit and the limit in the following Eq. (3.29) are shown to exist in Prop. 1.1.] Let $\{\overline{\mu},\overline{\mu},\ldots\}$ be the policy obtained from μ via the policy iteration step described previously, and let $\overline{\lambda}$ and $\overline{h}$ be corresponding average and differential cost satisfying

$$\overline{\lambda}e + \overline{h} = T_{\overline{\mu}}(\overline{h}) \quad (3.28)$$

and

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1} P_{\overline{\mu}}^k\overline{h} = 0. \quad (3.29)$$

Then if $\overline{\mu}\neq\mu$, we must have either (1) $\overline{\lambda}<\lambda$, or (2) $\overline{\lambda}=\lambda$ and $\overline{h}(i)\leq h(i)$ for all $i=1,\ldots,n$, with strict inequality for at least one state i.

We note that, once Lemma 3.1 is established, it can be shown that the policy iteration algorithm will terminate finitely. The reason is that the vector h corresponding to μ via Eq. (3.26) and (3.27) is unique by Prop. 2.5(b), and therefore the conclusion of Lemma 3.1 guarantees that no policy will be encountered more than once in the course of the algorithm. Since the number of stationary policies is finite, the algorithm must terminate finitely. If the algorithm stops at the kth step with $\mu^{k+1}=\mu^k$, we see from Eqs. (3.24) and (3.25) that

$$\lambda^k e + h^k = Th^k,$$

which by Prop. 2.1 implies that μ^k is an optimal stationary policy. So to prove Prop. 3.4 there remains to prove Lemma 3.1.

Proof of Lemma 3.1: For notational convenience, denote

$$P = P_\mu, \quad \overline{P} = P_{\overline{\mu}}, \quad P^* = \lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1}P^k, \quad \overline{P}^* = \lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1}\overline{P}^k.$$

$$g = g_\mu, \qquad \overline{g} = g_{\overline{\mu}}.$$

Define the vector δ by

$$\delta = \lambda e + h - \overline{g} - \overline{P}h. \tag{3.30}$$

We have, by assumption, $T_{\overline{\mu}}h = Th \leq T_\mu h = \lambda e + h$, or equivalently

$$\overline{g} + \overline{P}h \leq g + Ph = \lambda e + h, \tag{3.31}$$

from which we obtain

$$\delta(i) \geq 0, \qquad i = 1, \ldots, n. \tag{3.32}$$

Define also

$$\Delta = h - \overline{h}.$$

By combining Eq. (3.30) with the equation $\overline{\lambda}e + \overline{h} = \overline{g} + \overline{P}\overline{h}$, we obtain

$$\delta = (\lambda - \overline{\lambda})e + \Delta - \overline{P}\Delta.$$

Multiplying this relation with $\overline{P}^k$ and adding from 0 to $N-1$, we obtain

$$\sum_{k=0}^{N-1} \overline{P}^k \delta = N(\lambda - \overline{\lambda})e + \Delta - \overline{P}^N \Delta. \tag{3.33}$$

Dividing by N and taking the limit as $N \to \infty$, we obtain

$$\overline{P}^* \delta = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \overline{P}^k \delta = (\lambda - \overline{\lambda})e. \tag{3.34}$$

In view of the fact $\delta \geq 0$ [cf. Eq. (3.32)], we see that

$$\lambda \geq \overline{\lambda}.$$

If $\lambda > \overline{\lambda}$, we are done, so assume that $\lambda = \overline{\lambda}$. A state i is called *$\overline{P}$-recurrent ($\overline{P}$-transient)* if i belongs (does not belong, respectively) to the single recurrent class of the Markov chain corresponding to $\overline{P}^*$. From Eq. (3.34), $\overline{P}^*\delta = 0$ and since $\delta \geq 0$ and the elements of $\overline{P}^*$ that are positive are those columns corresponding to $\overline{P}$-recurrent states, we obtain

$$\delta(i) = 0, \qquad \text{for all } i \text{ that are } \overline{P}\text{-recurrent.} \tag{3.35}$$

It follows by construction of the algorithm that if i is $\overline{P}$-recurrent, then the ith rows of P and $\overline{P}$ are identical [since $\overline{\mu}(i) = \mu(i)$ for all i with $\delta(i) = 0$]. Since P and $\overline{P}$ have a single recurrent class, it follows that this class is identical for both P and $\overline{P}$. From the normalization conditions (3.27) and (3.29), we then obtain $h(i) = \overline{h}(i)$ for all i that are $\overline{P}$-recurrent. Equivalently,

$$\Delta(i) = 0, \qquad \text{for all } i \text{ that are } \overline{P}\text{-recurrent.} \tag{3.36}$$

From Eq. (3.33) we obtain

$$\lim_{N\to\infty} \overline{P}^N \Delta = \Delta - \lim_{N\to\infty} \sum_{k=0}^{N-1} P^k \delta \leq \Delta - \delta.$$

In view of Eq. (3.36), the coordinates of $\overline{P}^N\Delta$ corresponding to P-transient states tend to zero. Therefore, we have

$$\delta(i) \leq \Delta(i), \qquad \text{for all } i \text{ that are } \overline{P}\text{-transient.} \tag{3.37}$$

From Eqs. (3.32) and (3.35) to (3.37), we see how that either $\delta = 0$, in which case $\mu = \overline{\mu}$, or else $\Delta \geq 0$ with strict inequality $\Delta(i) > 0$ for at least one $\overline{P}$-transient state i. **Q.E.D.**

We now demonstrate the policy iteration algorithm by means of the example of the previous section.

Example 3.2: (continued)

Let

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

We take $t = 1$ as a reference state and obtain λ_{μ^0}, $h_{\mu^0}(1)$, and $h_{\mu^0}(2)$ from the system of equations

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

$$\lambda_{\mu^0} + h_{\mu^0}(2) = g(2, u^2) + p_{21}(u^2)h_{\mu^0}(1) + p_{22}(u^2)h_{\mu^0}(2),$$

$$h_{\mu^0}(1) = 0.$$

Substituting the data of the problem,

$$\lambda_{\mu^0} = 2 + \frac{1}{4}h_{\mu^0}(2), \qquad \lambda_{\mu^0} + h_{\mu^0}(2) = 3 + \frac{3}{4}h_{\mu^0}(2),$$

from which

$$\lambda_{\mu^0} = \frac{5}{2}, \qquad h_{\mu^0}(1) = 0, \qquad h_{\mu^0}(2) = 2.$$

We now find $\mu^1(1)$ and $\mu^1(2)$ by the minimization indicated in Eq. (3.25). We determine

$$\begin{aligned}\min\big[&g(1,u^1)+p_{11}(u^1)h_{\mu^0}(1)+p_{12}(u^1)h_{\mu^0}(2),\\ &g(1,u^2)+p_{11}(u^2)h_{\mu^0}(1)+p_{12}(u^2)h_{\mu^0}(2)\big]\\ &=\min\left[2+\frac{1}{4}\cdot 2,\ 0.5+\frac{3}{4}\cdot 2\right]\\ &=\min[2.5,2]\end{aligned}$$

and

$$\begin{aligned}\min\big[&g(2,u^1)+p_{21}(u^1)h_{\mu^0}(1)+p_{22}(u^1)h_{\mu^0}(2),\\ &g(2,u^2)+p_{21}(u^1)h_{\mu^0}(1)+p_{22}(u^2)h_{\mu^0}(2)\big]\\ &=\min\left[1+\frac{1}{4}\cdot 2,\ 3+\frac{3}{4}\cdot 2\right]\\ &=\min[1.5,\ 4.5].\end{aligned}$$

The minimization yields

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

We obtain λ_{μ^1}, $h_{\mu^1}(1)$, and $h_{\mu^1}(2)$ from the system of equations

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

$$\lambda_{\mu^1}+h_{\mu^1}(2)=g(2,u^1)+p_{21}(u^1)h_{\mu^1}(1)+p_{22}(u^1)h_{\mu^1}(2),$$

$$h_{\mu^1}(1)=0.$$

By substituting the data of the problem, we obtain

$$\lambda_{\mu^1}=\frac{3}{4},\qquad h_{\mu^1}(1)=0,\qquad h_{\mu^1}(2)=\frac{1}{3}.$$

We find $\mu^2(1)$ and $\mu^2(2)$ by determining the minimum in

$$\begin{aligned}\min\big[&g(1,u^1)+p_{11}(u^1)h_{\mu^1}(1)+p_{12}(u^1)h_{\mu^1}(2),\\ &g(1,u^2)+p_{11}(u^2)h_{\mu^1}(1)+p_{12}(u^2)h_{\mu^1}(2)\big]\\ &=\min\left[2+\frac{1}{4}\cdot\frac{1}{3},\ 0.5+\frac{3}{4}\cdot\frac{1}{3}\right]\\ &=\min[2.08,\ 0.75],\end{aligned}$$

and

$$\begin{aligned}\min\big[&g(2,u^1)+p_{21}(u^1)h_{\mu^1}(1)+p_{22}(u^1)h_{\mu^1}(2),\\ &g(2,u^2)+p_{21}(u^2)h_{\mu^1}(1)+p_{22}(u^2)h_{\mu^1}(2)\big]\\ &=\min\left[1+\frac{1}{4}\cdot\frac{1}{3},\ 3+\frac{3}{4}\cdot\frac{1}{3}\right]\\ &=\min[1.08,\ 3.25].\end{aligned}$$

The minimization yields

$$\mu^2(1)=\mu^1(1)=u^2,\qquad \mu^2(2)=\mu^1(2)=u^1,$$

and hence the preceding policy is optimal and the optimal average cost is $\lambda_{\mu^1}=3/4$.

The algorithm of this section shares some of the features of other types of policy iteration algorithms. In particular, it is possible to carry out policy evaluation approximately by using a few relative value iterations; see [Put94] for an analysis. Note also that in specially structured problems one may be able to confine policy iteration within a convenient subset of policies, for which policy evaluation is facilitated.

Adaptive Aggregation

Consider now an extension to the average cost problem of the aggregation method described in Section 1.3.3. For a given unichain stationary policy μ, we want to calculate an approximation to the pair (λ_μ, h_μ) satisfying Bellman's equation $\lambda_\mu e+h_\mu=T_\mu h_\mu$ and $h_\mu(n)=0$, where the state n is viewed as the reference state. By expressing λ_μ as $\lambda_\mu=(T_\mu h_\mu)(n)$, we can eliminate it from this system of equations, and obtain $h_\mu=T_\mu h_\mu-(T_\mu h_\mu)(n)e$. This equation is written compactly as

$$h_\mu=\hat{T}h_\mu,$$

where the mapping $\hat{T}$ is defined by

$$\hat{T}h=g_r+P_r h,$$

and

$$g_r=(I-ee_n')g_\mu,\qquad P_r=(I-ee_n')P_\mu,$$

with $e_n=(0,0,\ldots,0,1)'$.

We partition the set of states into disjoint subsets $S_1,S_2,\ldots,S_m$ that are viewed as aggregate states. We assume that one of the subsets, say S_m, consists of just the reference state n; that is, $S_m=\{n\}$. As in Section 1.3.3, consider the $n\times m$ matrix W whose ith column has unit entries at coordinates corresponding to states in S_i and all other entries equal to zero. Consider also an $m\times n$ matrix Q such that the ith row of Q is a probability distribution $(q_{i1},\ldots,q_{in})$ with $q_{is}=0$ if $s\notin S_i$. Note that $QW=I$, and that the $m\times m$ matrix $R=QP_\mu W$ is the transition probability matrix of the aggregate Markov chain, whose states are the m aggregate states.

Suppose now that we have an estimate h of h_μ and that *we postulate that over the states s of every aggregate state S_i the variation $h_\mu(s)-h(s)$*

is constant. This amounts to hypothesizing that for some m-dimensional vector y we have

$$h_\mu - h = Wy.$$

By combining the equations $\hat{T}h = g_r + P_r h$ and $h_\mu = g_r + P_r h_\mu$, we have

$$(I - P_r)(h_\mu - h) = \hat{T}h - h.$$

By multiplying both sides of this equation with Q, and by using the relations $h_\mu - h = Wy$ and $QW = I$, we obtain

$$(I - QP_r W)y = Q(\hat{T}h - h).$$

Assuming that the matrix $I - QP_r W$ is invertible, this equation can be solved for y. Also, by applying $\hat{T}$ to both sides of the equation $h_\mu = h + Wy$, we obtain

$$h_\mu = \hat{T}h_\mu = \hat{T}h + P_r Wy.$$

Thus the aggregation iteration for average cost problems is as follows:

Aggregation Iteration

Step 1: Compute $\hat{T}h = g_r + P_r h$, where

$$g_r = (I - ce_n')g_\mu, \qquad P_r = (I - ce_n')P_\mu.$$

Step 2: Delineate the aggregate states (i.e., define W) and specify the matrix Q.

Step 3: Solve for y the system

$$(I - QP_r W)y = Q\hat{T}h,$$

and approximate h_μ using

$$h := \hat{T}h + P_r Wy.$$

For the iteration to be valid, the matrix $I - QP_r W$ must be invertible. We will show that this is guaranteed under an aperiodicity assumption such as the one used to prove convergence of the relative value iteration method (cf. Prop. 3.1). In particular, we assume that all the eigenvalues of the transition matrix $R = QP_\mu W$ of the aggregate Markov chain, except for a single unity eigenvalue, lie strictly within the unit circle. Let us denote by e the n-dimensional vector of all 1's, and by e_m the m-dimensional vector

with last coordinate 1, and all other coordinates 0. Then using the easily verified relations $Qe = e$ and $e_m'Q = e_n'$, we see that

$$QP_r W = (I - ee_m')R.$$

From the analysis of Example 3.1 in Section 4.3, we have that $QP_r W$ has $m - 1$ eigenvalues that are equal to the $m - 1$ nonunity eigenvalues of R and has 0 as its mth eigenvalue. Thus $QP_r W$ must have all its eigenvalues strictly within the unit circle, and it follows that the matrix $I - QP_r W$ is invertible.

In an adaptive aggregation method, a key issue is how to identify the aggregate states $S_1, \ldots, S_m$ in a way that the error $h_\mu - h$ is of similar magnitude in each aggregate state. Similar to Section 4.3.3, one way to do this is to view $\hat{T}h$ as an approximation to h_μ and to group together states i with comparable magnitudes of $(\hat{T}h)(i) - h(i)$. As discussed in Section 4.3.3, this type of aggregation method can be greatly improved by interleaving each aggregation iteration with multiple relative value iterations (applications of the mapping $\hat{T}$ on the current iterate). We refer to [BeC89] for further experimentation, analysis, and discussion.

4.3.3 Linear Programming

Let us now develop a linear programming-based solution method, assuming that any one of the conditions of Prop. 3.3 holds, so that the optimal average cost λ^* is independent of the initial state, and together with an associated differential cost vector h^*, satisfies $\lambda^* e + h^* = Th^*$. Consider the following optimization problem in the variables λ and $h(i)$, $i = 1, \ldots, n$,

$$\begin{aligned} &\text{maximize} \quad \lambda \\ &\text{subject to} \quad \lambda + h(i) \leq (Th)(i), \qquad i = 1, \ldots, n, \end{aligned}$$

which is equivalent to the linear program

$$\text{maximize} \quad \lambda$$

$$\text{subject to} \quad \lambda + h(i) \leq g(i,u) + \sum_{j=1}^{n} p_{ij}(u)h(j), \qquad i = 1, \ldots, n, \quad u \in U(i). \tag{3.38}$$

A nearly verbatim repetition of the proof of Prop. 2.1 shows that if (λ, h) is a feasible solution, that is, $\lambda e + h \leq Th$, then $\lambda \leq \lambda^*$, which implies that (λ^*, h^*) is an optimal solution of the linear program (3.38). Furthermore, in any optimal solution $(\overline{\lambda}, \overline{h})$ of the linear program (3.38), we have $\overline{\lambda} = \lambda^*$.

There is a linear program, which is dual to the above and which admits an interesting interpretation. In particular, the duality theory of

linear programming (see e.g., [Dan63]) asserts that the following (dual) linear program

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{n} \sum_{u \in U(i)} q(i,u) g(i,u) \\ &\text{subject to} \quad \sum_{u \in U(j)} q(j,u) = \sum_{i=1}^{n} \sum_{u \in U(i)} q(i,u) p_{ij}(u), \quad j = 1, \ldots, n, \\ &\qquad \sum_{i=1}^{n} \sum_{u \in U(i)} q(i,u) = 1, \\ &\qquad q(i,u) \geq 0, \quad i = 1, \ldots, n, \quad u \in U(i), \end{aligned} \tag{3.39}$$

has the same optimal value as the (primal) program (3.38). The variables $q(i,u)$, $i = 1, \ldots, n$, $u \in U(i)$, of the dual program can be interpreted as the steady-state probabilities that state i will be visited at the typical transition and that control u will then be applied. The constraints of the dual program are the constraints that $q(i,u)$ must satisfy in order to be feasible steady-state probabilities under some *randomized* stationary policy, that is, a policy that chooses at state i the control u probabilistically, by sampling the constraint set $U(i)$ according to the probabilities $q(i,u)$, $u \in U(i)$. The cost function

$$\sum_{i=1}^{n} \sum_{u \in U(i)} q(i,u) g(i,u)$$

of the dual problem is the steady-state average cost per transition. Duality theory asserts that the minimal value of this cost is λ^*, thus implying that the optimal average cost per stage that can be obtained using randomized stationary policies is no better than what can be achieved with ordinary (deterministic) stationary policies. Indeed, it can be verified that if μ^* is an optimal (deterministic) stationary policy that is unichain, and p_i^* is the steady-state probability of state i in the corresponding Markov chain, then

$$q^*(i,u) = \begin{cases} p_i^* & \text{if } u = \mu^*(i), \\ 0 & \text{otherwise,} \end{cases}$$

is an optimal solution of the dual problem (3.39).

4.3.4 Simulation-Based Methods

We now describe briefly how the simulation-based methods of Section 2.3 can be adapted to work for average cost problems. We make a slight change in the problem definition to make the notation better suited for the simulation context. In particular, instead of considering the expected cost $g(i,u)$ at state i under control u, we allow the cost g to depend on the next state j. Thus our notation for the cost per stage is now $g(i,u,j)$, as in the simulation-related material for stochastic shortest path and discounted problems (cf. Section 2.3). All the results and the entire analysis of the preceding sections can be rewritten in terms of the new notation by replacing $g(i,u)$ with $\sum_{j=1}^{n} p_{ij}(u) g(i,u,j)$.

Policy Iteration

In order to implement a simulation-based policy iteration algorithm like the one of Section 2.3.1, we need to be able to carry out the policy evaluation step for a given unichain policy μ. This can be done by using the connection with the stochastic shortest path formulation described in Section 4.1. We fix a state t, and we evaluate the cost of the given policy μ for two stochastic shortest path problems whose termination state is (essentially) t. In particular, we evaluate by Monte-Carlo simulation or TD(λ) the expected cost C_i from each state i up to reaching t [cf. Eq. (2.15)]. This requires the generation of many trajectories terminating at state t and the corresponding sample costs. Simultaneously with the evaluation of the costs C_i, we evaluate the expected number of transitions N_i from each state i up to reaching t [cf. Eq. (2.16)]. Then the average cost λ_μ of the policy is obtained as

$$\lambda_\mu = \frac{C_t}{N_t}, \tag{3.40}$$

[cf. Eq. 2.17)], and the associated differential costs are obtained as

$$h_\mu(i) = C_i - \lambda_\mu N_i, \qquad i = 1, \ldots, n, \tag{3.41}$$

[cf. Eq. (2.18)].

To implement a simulation-based approximate policy iteration algorithm, a similar procedure can be used. In particular, one can obtain by Monte-Carlo simulation or TD(1) functions $\tilde{C}_i(r)$ and $\tilde{N}_i(r)$ that depend on a parameter vector r and approximate the costs C_i and N_i of the corresponding stochastic shortest path problems, as described in Section 2.3.3. Then, one can use

$$\tilde{\lambda}_\mu(r) = \frac{\tilde{C}_t(r)}{\tilde{N}_t(r)}$$

as an approximation to the average cost per stage of the policy and also use

$$\tilde{h}_\mu(i) = \tilde{C}_i(r) - \tilde{\lambda}_\mu(r) \tilde{N}_i(r), \qquad i = 1, \ldots, n,$$

as approximations to the corresponding differential costs [cf. Eqs. (3.40) and (3.41)].

Note here that because the approximations $\tilde{C}_t(r)$ and $\tilde{N}_t(r)$ play an important role in the calculations, it may be worth doing some extra simulations starting from the reference state t to ensure that these approximations are nearly exact.

Value Iteration and Q-Learning

To derive the appropriate form of the Q-learning algorithm of Section 2.3.2, we form an auxiliary average cost problem by augmenting the original system with one additional state for each possible pair (i, u) with $u \in U(i)$. The probabilistic transition mechanism from the original states is the same as for the original problem, while the probabilistic transition mechanism from an auxiliary state (i, u) is that we move only to states j of the original problem with corresponding probabilities $p_{ij}(u)$ and costs $g(i, u, j)$. It can be seen that the auxiliary problem has the same optimal average cost per stage λ as the original, and that the corresponding Bellman's equation is

$$\lambda + h(i) = \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u)\big(g(i, u, j) + h(j)\big), \qquad i = 1, \ldots, n, \tag{3.42}$$

$$\lambda + Q(i, u) = \sum_{j=1}^{n} p_{ij}(u)\big(g(i, u, j) + h(j)\big), \qquad i = 1, \ldots, n, \ u \in U(i), \tag{3.43}$$

where $Q(i, u)$ is the differential cost corresponding to (i, u). Taking the minimum over u in Eq. (3.43) and substituting in Eq. (3.42), we obtain

$$h(i) = \min_{u \in U(i)} Q(i, u), \qquad i = 1, \ldots, n. \tag{3.44}$$

Substituting the above form of $h(i)$ in Eq. (3.43), we obtain Bellman's equation in a form that exclusively involves the Q-factors:

$$\lambda + Q(i, u) = \sum_{j=1}^{n} p_{ij}(u)\left(g(i, u, j) + \min_{u' \in U(j)} Q(j, u')\right), \ i = 1, \ldots, n, \ u \in U(i). \tag{3.45}$$

Let us now apply to the auxiliary problem the following variant of the relative value iteration

$$h^{k+1} = Th^k - h^k(t)e,$$

(see Exercise 4.5 for the case where $e = 0$, $p_t = 1$, and $p_j = 0$ for $j \neq t$). We then obtain the iteration

$$h^{k+1}(i) = \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u)\big(g(i, u, j) + h^k(j)\big) - h^k(t), \qquad i = 1, \ldots, n, \tag{3.46}$$

$$Q^{k+1}(i, u) = \sum_{j=1}^{n} p_{ij}(u)\big(g(i, u, j) + h^k(j)\big) - h^k(t), \qquad i = 1, \ldots, n, \ u \in U(i). \tag{3.47}$$

From these equations, we have that

$$h^k(i) = \min_{u \in U(i)} Q^k(i, u), \qquad i = 1, \ldots, n, \tag{3.48}$$

and by substituting the above form of h^k in Eq. (3.47), we obtain the following relative value iteration for the Q-factors

$$Q^{k+1}(i, u) = \sum_{j=1}^{n} p_{ij}(u)\left(g(i, u, j) + \min_{u' \in U(j)} Q^k(j, u')\right) - \min_{u' \in U(t)} Q^k(t, u'). \tag{3.49}$$

This iteration is analogous to the value iteration for Q-factors in the stochastic shortest path context. The sequence of values $\min_{u \in U(t)} Q^k(t, u)$ is expected to converge to the optimal average cost per stage and the sequences of values $\min_{u \in U(i)} Q(i, u)$ are expected to converge to differential costs $h(i)$.

An incremental version of the preceding iteration that involves a positive stepsize γ is given by

$$Q(i, u) := (1 - \gamma)Q(i, u) + \gamma\left(\sum_{j=1}^{n} p_{ij}(u)\left(g(i, u, j) + \min_{u' \in U(j)} Q(j, u')\right) - \min_{u' \in U(t)} Q(t, u')\right), \tag{3.50}$$

[compare with Eq. (3.8) in Section 2.3]. The natural form of the Q-learning method for the average cost problem is an approximate version of this iteration, whereby the expected value is replaced by a single sample, i.e.,

$$Q(i, u) := Q(i, u) + \gamma\Big(g(i, u, j) + \min_{u' \in U(j)} Q(j, u') - \min_{u' \in U(t)} Q(t, u') - Q(i, u)\Big), \tag{3.51}$$

where j and $g(i, u, j)$ are generated from the pair (i, u) by simulation.

Minimization of the Bellman Equation Error

There is a straightforward extension of the method of Section 2.3.3 for obtaining an approximate representation of the average cost λ and associated differential costs $h(i)$, based on minimizing the squared error in Bellman's equation. Here we approximate λ by $\tilde{\lambda}(r)$ and $h(i)$ by $\tilde{h}(i, r)$.

where r is a vector of unknown parameters/weights. We impose a normalization constraint such as $\tilde{h}(t,r) = 0$, where t is a fixed state, and we minimize the error in Bellman's equation by solving the problem

$$\min_r \sum_{i \in S} \left| \tilde{\lambda}(r) + \tilde{h}(i,r) - \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u)\big(g(i,u,j) + \tilde{h}(j,r)\big) \right|^2,$$

where S is a suitably chosen subset of "representative" states. This minimization may be attempted by using some gradient method of the type discussed in Section 2.3.3.

4.4 INFINITE STATE SPACE

The standing assumption in the preceding sections has been that the state space is finite. Without finiteness of the state space, many of the results presented in the past three sections no longer hold. For example, whereas one could restrict attention to stationary policies for finite state systems, this is no longer true when the state space is infinite. The following example from [Ros83a] shows that if the state space is countable, there may not exist an optimal policy.

Example 4.1:

Let the state space be $\{1, 1', 2, 2', 3, 3', \ldots\}$, and let there be two controls, u^1 and u^2. The transition probabilities and costs per stage are

$$p_{i(i+1)}(u^1) = 1, \quad p_{ii'}(u^2) = 1, \qquad i = 1, 2, \ldots,$$

$$p_{i'i'}(u^1) = p_{i'i'}(u^2) = 1, \qquad i = 1, 2, \ldots,$$

$$g(i, u^1) = g(i, u^2) = 0, \qquad i = 1, 2, \ldots,$$

$$g(i', u^1) = g(i', u^2) = -1 + \frac{1}{i}, \qquad i = 1, 2, \ldots.$$

In words, at state i we may, at a cost 0, either move to state $(i+1)$ or move to state i', where we stay thereafter at a cost $-1 + 1/i$ per stage.

It can be seen that for every policy π and state $i = 1, 2, \ldots$, we have $J_\pi(i) > -1$. However, for every state i, we can obtain an average cost per stage $-1 + 1/j$, where $j \geq i$, by moving to state j' once we get to state j. Hence, for every initial state $i = 1, 2, \ldots$, an average cost per stage of -1 can be approached arbitrarily closely, but cannot be attained by any policy.

Here is another example, from [Ros70], which shows that for a countable state space there may exist an optimal nonstationary policy, but not an optimal stationary policy.

Example 4.2:

Let the state space be $\{1, 2, 3, \ldots\}$, and let there be two controls, u^1 and u^2. The transition probabilities and costs per stage are

$$p_{i(i+1)}(u^1) = p_{ii}(u^2) = 1,$$

$$g(i, u^1) = 1, \qquad g(i, u^2) = \frac{1}{i}, \qquad i = 1, 2, \ldots.$$

In words, at state i we may either move to state $(i+1)$ at a cost 1 or stay at i at a cost $1/i$.

For any stationary policy μ other than the policy for which $\mu(i) = u^1$ for all i, let $n(\mu)$ be the smallest integer for which

$$\mu\big(n(\mu)\big) = u^2.$$

Then the corresponding average cost per stage satisfies

$$J_\mu(i) = \frac{1}{n(\pi)} > 0, \qquad \text{for all } i \text{ with } i \leq n(\pi).$$

For the policy where $\mu(i) = u^1$ for all i, we have $J_\pi(i) = 1$ for all i. Since the optimal cost per stage cannot be less than zero, it is clear that

$$\min_\pi J_\pi(i) = 0, \qquad i = 1, 2, \ldots$$

However, the optimal cost is not attained by any stationary policy, so no stationary policy is optimal. On the other hand, consider the nonstationary policy π^* that on entering state i chooses u^2 for i consecutive times and then chooses u^1. If the starting state is i, the sequence of costs incurred is

$$\underbrace{\frac{1}{i}, \frac{1}{i}, \ldots, \frac{1}{i}}_{i \text{ times}}, \quad 1, \quad \underbrace{\frac{1}{i+1}, \frac{1}{i+1}, \ldots, \frac{1}{i+1}}_{(i+1) \text{ times}}, \quad 1, \quad \frac{1}{i+2}, \frac{1}{i+2}, \ldots.$$

The average cost corresponding to this policy is

$$J_{\pi^*}(i) = \lim_{m \to \infty} \frac{2m}{\sum_{k=1}^{m} (i+k)} = 0, \qquad i = 1, 2, 3, \ldots$$

Hence the nonstationary policy π^* is optimal while, as shown previously, no stationary policy is optimal.

Generally, the analysis of average cost problems with an infinite state space is difficult, although there has been considerable progress (see the references). An important tool is Prop. 2.1, which admits a straightforward extension to the case where the state and control spaces are infinite. In particular, if we can find a scalar λ and a bounded function h such that Bellman's equation (2.1) holds, then by repeating the proof of Prop. 2.1, we can show that λ must be the optimal average cost per stage for all initial states. Among other situations, this result is useful when we can guess the right λ and h, and verify that they satisfy Bellman's equation. Some important special cases can be satisfactorily analyzed in this way (see the references). We describe one such case, the average cost version of the linear-quadratic problem examined in Chapters 4, and 5 of Vol. I.

Linear Systems with Quadratic Cost

Consider the linear-quadratic problem involving the system

$$x_{k+1} = Ax_k + Bu_k + w_k, \qquad k = 0, 1, \ldots, \tag{4.1}$$

and the cost function

$$J_\pi(x_0) = \lim_{N\to\infty} \frac{1}{N} \mathop{E}_{\substack{w_k \\ k=0,1,\ldots}} \left\{ \sum_{k=0}^{N-1} \big(x_k' Q x_k + \mu_k(x_k)' R \mu_k(x_k)\big) \right\}. \tag{4.2}$$

We make the same assumptions as in Section 8.1, that is, Q is positive semidefinite symmetric, R is positive definite symmetric, and w_k are independent, and have zero mean and finite second moments. We also assume that the pair (A, B) is controllable and that the pair (A, C), where $Q = C'C$, is observable. Under these assumptions, it was shown in Section 4.1 of Vol. I that the Riccati equation

$$K_0 = 0, \tag{4.3}$$

$$K_{k+1} = A'\big(K_k - K_k B(B'K_k B + R)^{-1} B'K_k\big)A + Q \tag{4.4}$$

yields in the limit a matrix K,

$$K = \lim_{k\to\infty} K_k, \tag{4.5}$$

which is the unique solution of the equation

$$K = A'\big(K - KB(B'KB + R)^{-1} B'K\big)A + Q \tag{4.6}$$

within the class of positive semidefinite symmetric matrices.

The optimal value of the N-stage costs

$$\frac{1}{N} \mathop{E}_{\substack{w_k \\ k=0,1,\ldots,N-1}} \left\{ \sum_{k=0}^{N-1} (x_k' Q x_k + u_k' R u_k) \right\} \tag{4.7}$$

has been derived earlier and was seen to be equal to

$$\frac{1}{N} \left(x_0' K_N x_0 + \sum_{k=0}^{N-1} E\{w' K_k w\} \right).$$

Since $K = \lim_{k\to\infty} K_k$ and

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} E\{w' K_k w\} = E\{w' K w\},$$

we see that the optimal N-stage costs tend to

$$\lambda = E\{w' K w\} \tag{4.8}$$

as $N \to \infty$. In addition, the N-stage optimal policy in its initial stages tends to the stationary policy

$$\mu^*(x) = -(B'KB + R)^{-1} B'KAx. \tag{4.9}$$

Furthermore, a simple calculation shows that, by the definition of λ, K, and $\mu^*(x)$, we have

$$\lambda + x'Kx = \min_u E\big\{x'Qx + u'Ru + (Ax + Bu + w)'K(Ax + Bu + w)\big\},$$

while the minimum in the right-hand side of this equation is attained at $u^* = \mu^*(x)$ as given by Eq. (4.9).

By repeating the proof of Prop. 2.1, we obtain

$$\begin{aligned} \lambda \le{}& \frac{1}{N} E\{x_N' K x_N \mid x_0, \pi\} \\ &- \frac{1}{N} x_0' K x_0 + \frac{1}{N} E\left\{ \sum_{k=0}^{N-1} (x_k' Q x_k + u_k' R u_k) \;\middle|\; x_0, \pi \right\}, \end{aligned}$$

with equality if $\pi = \{\mu^*, \mu^*, \ldots\}$. Hence, if π is such that $E\{x_N' K x_N \mid x_0, \pi\}$ is uniformly bounded over N, we have, by taking the limit as $N \to \infty$ in the preceding relation,

$$\lambda \le J_\pi(x), \qquad x \in \Re^n,$$

with equality if $\pi = \{\mu^*, \mu^*, \ldots\}$. Thus the linear stationary policy given by Eq. (4.9) is optimal over all policies π with $E\{x_N' K x_N \mid x_0, \pi\}$ bounded uniformly over N.

4.5 NOTES, SOURCES, AND EXERCISES

Several authors have contributed to the average cost problem ([How60], [Bro65], [Ros70], [Sch68], [Vei66], [Vei69]), most notably Blackwell ([Bla62]). An alternative detailed treatment to ours is given in [Put94]. An extensive survey containing many references is given in [ABF93].

The result of Prop. 2.6 under conditions (2) and (3) was shown in [Bat73] and [Ros70], respectively. The relative value iteration method of Section 4.3 is due to [Whi63], and its modified version of Eq. (3.22) is due

to [Sch71]. The error bounds of Prop. 3.2 are due to Odoni ([Odo69]). The value iteration method has been analyzed exhaustively in [Sch71], [ScF77], and [ScF78]. Convergence under slightly weaker conditions than those given here is shown in [Pla77]. The error bounds of Exercise 4.10 are due to Varaiya ([Var78]), who used them to construct a differential form of the value iteration method. Discrete-time versions of Varaiya's method are given in [PBW79]. The value iteration method based on stochastic shortest paths of Exercise 4.15 is new (see [Ber95c]).

The policy iteration algorithm can be generalized for problems where the optimal average cost per stage is not the same for every initial state (see [Bla62], [Put94], [Vei66], and [Der70]). The adaptive aggregation method is due to [BeC89].

The approximation procedures of Section 4.3.4, and the Q-learning algorithms of Section 4.3.4 and Exercise 4.16 are new. Alternative algorithms of the Q-learning type are given in [Sch93] and [Sin94].

For analysis of infinite horizon versions of inventory control problems, such as the ones of Section 4.2 of Vol. I, see [Igl63a], [Igl63b], and [HoT74]. Infinite state space models are discussed in [Kus78], [Sen86], [Las88], [Bor89], [Cav89a], [Cav89b], [Her89], [Sen89a], [Sen89b], [FAM90], [FAM91], [Cav91], [HHL91], [Sen91], [CaS92], [RiS92], [ABF93], [Sen93a], [Sen93b], and [Put94].

EXERCISES

4.1

Solve the average cost version ($\alpha = 1$) of the computer manufacturer's problem (Exercise 7.3, Vol. I).

4.2

Consider a stationary inventory control problem of the type considered in Section 4.2 of Vol. I but with the difference that the stock x_k can only take integer values from 0 to some integer M. The order u_k can take integer values with $0 \leq u_k \leq M - x_k$, and the random demand w_k can only take nonnegative integer values with $P(w_k = 0) > 0$ and $P(w_k = 1) > 0$. Unsatisfied demand is lost, so stock evolves according to the equation $x_{k+1} = \max(0, x_k + u_k - w_k)$. The problem is to find an inventory policy that minimizes the average cost per stage. Show that there exists an optimal stationary policy and that the optimal cost is independent of the initial stock x_0.

4.3 [LiR71]

Consider a person providing a certain type of service to customers. The person receives at the beginning of each time period with probability p_i a proposal by a customer of type i, where $i = 1, 2, \ldots, n$, who offers an amount of money M_i. We assume that $\sum_{i=1}^{n} p_i \leq 1$. The person may reject the offer, in which case the customer leaves and the person remains idle during that period, or the person may accept the offer in which case the person spends some time with that customer determined according to a Markov process with transition probabilities β_{ik}, where, for $k = 1, 2, \ldots$,

β_{ik} = probability that the type i customer will leave after k periods, given that the customer has already stayed with the person for $k-1$ periods.

The problem is to determine an acceptance-rejection policy that maximizes

$$\lim_{N \to \infty} \frac{1}{N}\{\text{Expected payment over } N \text{ periods}\}.$$

Consider two cases:

1. $\beta_{ik} = \beta_i \in (0, 1)$ for all k.
2. For each i there exists $\overline{k}_i$ such that $\beta_{i\overline{k}_i} = 1$.

(a) Formulate the person's problem as an average cost per stage problem, and show that the optimal cost is independent of the initial state.

(b) Show that there exists a scalar λ^* and an optimal policy that accepts the offer of a type i customer if and only if

$$\lambda^* T_i \leq M_i,$$

where T_i is the expected time spent with the type i customer given by

$$T_i = \beta_{i1} + \sum_{k=2}^{\infty} k \beta_{ik} (1 - \beta_{ik-1}) \cdots (1 - \beta_{i0}).$$

4.4

Let h^0 be an arbitrary vector in $\Re^n$, and define for all i and $k \geq 1$

$$h_i^k = T^k h^0 - (T^k h^0)(i)e,$$

$$h^k = T^k h^0 - \frac{1}{n} \sum_{i=1}^{n} (T^k h^0)(i)e,$$

$$\tilde{h}^k = T^k h^0 - \min_{i=1,\ldots,n} (T^k h^0)(i)e.$$

Let also $h_i^0 = h^0 = \tilde{h}^0 = h^0$.

(a) Show that the sequences $\{h_i^k\}$, $\{\hat{h}^k\}$, and $\{\tilde{h}^k\}$ are generated by the algorithms

$$h_i^{k+1} = Th_i^k - (Th_i^k)(i)e,$$

$$\hat{h}_i^{k+1} = T\hat{h}^k - \frac{1}{n}\sum_{i=1}^{n}(T\hat{h}_i^k)(i)e,$$

$$\tilde{h}_i^{k+1} = T\tilde{h}^k - \min_{i=1,\ldots,n}(T\tilde{h}^k)(i)e.$$

(b) Show that the convergence result of Prop. 3.1 holds for the algorithms of part (a). *Hint*: Proposition 3.1 applies to the algorithms that generate $\{h_i^k\}$. Express $\hat{h}^k$ and $\tilde{h}^k$ as continuous functions of $\{h_i^k\}$, $i = 1, \ldots, n$.

4.5 (Variants of Relative Value Iteration)

Consider the following two variants of the relative value iteration algorithm:

$$h^{k+1}(i) = (Th^k)(i) - \lambda^k, \qquad i = 1, \ldots, n,$$

where

$$\lambda^k = c + \sum_{j=1}^{n} p_j h^k(j),$$

or

$$\lambda^k = c + \sum_{j=1}^{n} p_j h^{k-1}(j).$$

Here c is an arbitrary scalar and $(p_1, \ldots, p_n)$ is an arbitrary probability distribution over the states of the system. Under the assumptions of Prop. 3.1, show that the sequence $\{h^k\}$ converges to a vector h and the sequence $\{\lambda^k\}$ converges to a scalar λ satisfying $\lambda e + h = Th$, so that by Prop. 2.1, λ is equal to the optimal average cost per stage for all initial states and h is an associated differential cost vector. *Hint*: Modify the problem by introducing an artificial state t' from which the system moves at a cost c to state j with probability p_j, for all u. Apply Prop. 3.1.

4.6

Consider a deterministic system with two states 0 and 1. Upon entering state 0, the system stays there permanently at no cost. In state 1 there is a choice of staying there at no cost or moving to state 0 at cost 1. Show that every policy is average cost optimal, but the only stationary policy that is Blackwell optimal is the one that keeps the system in the state it currently is.

4.7

Show that a Blackwell optimal policy is optimal over all policies (not just those that are stationary). *Hint*: Use the following result: If $\{c_n\}$ is a nonnegative bounded sequence, then

$$\begin{aligned}\liminf_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1}c_k &\le \liminf_{\alpha\uparrow 1}(1-\alpha)\sum_{k=0}^{\infty}\alpha^k c_k \\ &\le \limsup_{\alpha\uparrow 1}(1-\alpha)\sum_{k=0}^{\infty}\alpha^k c_k \\ &\le \limsup_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1}c_k.\end{aligned}$$

A proof of this result can be found in [Put94], p. 417.

4.8 (Reduction to the Discounted Case)

For the finite-state average cost problem suppose there is a state t such that for some $\beta > 0$ we have $p_{it}(u) \ge \beta$ for all states i and controls u. Consider the $(1-\beta)$-discounted problem with the same state space, control space, and transition probabilities

$$\overline{p}_{ij}(u) = \begin{cases} (1-\beta)^{-1}p_{ij}(u) & \text{if } j \neq t, \\ (1-\beta)^{-1}\big(p_{ij}(u) - \beta\big) & \text{if } j = t. \end{cases}$$

Show that $\beta\overline{J}(t)$ and $\overline{J}(i)$ are optimal average and differential costs, respectively, where $\overline{J}$ is the optimal cost function of the $(1-\beta)$-discounted problem.

4.9 (Deterministic Finite-State Systems)

Consider a deterministic finite-state system. Suppose that the system is controllable in the sense that given any two states i and j, there exists a sequence of admissible controls that drives the state of the system from i to j. Consider the problem of finding an admissible control sequence $\{u_0, u_1, \ldots\}$ that minimizes

$$J_\pi(i) = \lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1}g(x_k, u_k).$$

Show that the optimal cost is independent of the initial state, and that there exist optimal control sequences, which after a certain time index are periodic.

4.10 (Generalized Error Bounds)

Let h be any n-dimensional vector and μ be such that

$$T_\mu h = Th.$$

Show that, for all i,

$$\min_j \big[(Th)(j) - h(j)\big] \le J^*(i) \le J_\mu(i) \le \max_j \big[(Th)(j) - h(j)\big],$$

regardless of whether $J^*(i)$ is independent of the initial state i. *Hint*: Complete the details of the following argument. Let

$$\delta(i) = (Th)(i) - h(i), \qquad i = 1, \ldots, n,$$

and let δ be the vector with coordinates $\delta(i)$. We have

$$T_\mu h = \delta + h, \qquad T_\mu^2 h = T_\mu h + P_\mu \delta = \delta + P_\mu \delta + h$$

and, continuing in the same manner,

$$T_\mu^N h = \sum_{k=0}^{N-1} P_\mu^k \delta + h, \qquad N = 1, 2, \ldots$$

Hence

$$J_\mu = \lim_{N \to \infty} \frac{1}{N} T_\mu^N h = P_\mu^* \delta,$$

where

$$P_\mu^* = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} P_\mu^k,$$

proving the right-hand side of the desired relation. Also, let $\pi = \{\mu_0, \mu_1, \ldots\}$ be any admissible policy. We have

$$T_{\mu_N} h \ge \delta + h$$

from which we obtain

$$T_{\mu_{N-1}} T_{\mu_N} h \ge P_{\mu_{N-1}} \delta + T_{\mu_{N-1}} h \ge P_{\mu_{N-1}} \delta + \delta + h \ge 2 \min_j \delta(j) e + h.$$

Thus, for all i,

$$\frac{1}{N+1} (T_{\mu_0} \cdots T_{\mu_N} h)(i) \ge \min_j \delta(j) + \frac{h(i)}{N+1}$$

and, taking the limit as $N \to \infty$, we obtain

$$J_\pi(i) \ge \min_j \delta(j).$$

Since π is arbitrary, we obtain the left-hand side of the desired relation.

4.11

Use Prop. 1.1 to show that in the policy iteration algorithm we have for all k,

$$\lambda^{k+1} e = \lambda^k e + P_{\mu^{k+1}}^* (T h^k - h^k - \lambda^k e),$$

where

$$P_{\mu^{k+1}}^* = \lim_{N \to \infty} \frac{1}{N} \sum_{m=0}^{N-1} P_{\mu^{k+1}}^m.$$

Use this fact to show that if the Markov chain corresponding to μ^{k+1} has no transient states and μ^{k+1} is not optimal, then $\lambda^{k+1} < \lambda^k$.

4.12 (Policy Iteration for Linear-Quadratic Problems)

The purpose of this problem is to show that policy iteration works for linear-quadratic problems (even though neither the state space nor the control space are finite). Consider the problem of Section 4.4 under the usual controllability, observability, and positive (semi)definiteness assumptions. Let L_0 be an $m \times n$ matrix such that the matrix $(A + BL_0)$ is stable.

(a) Show that the average cost per stage corresponding to the stationary policy μ^0, where $\mu^0(x) = L_0 x$, is of the form

$$J_{\mu^0} = E\{w' K_0 w\},$$

where K_0 is a positive semidefinite symmetric matrix satisfying the (linear) equation

$$K_0 = (A + BL_0)' K_0 (A + BL_0) + Q + L_0' R L_0.$$

(b) Let $\mu^1(x) = L_1 x = -(R + B'K_0 B)^{-1} B' K_0 A x$ be the control function attaining the minimum for each x in the expression

$$\min_u \big\{ u' R u + (Ax + Bu)' K_0 (Ax + Bu) \big\}.$$

Show that

$$J_{\mu^1} = E\{w' K_1 w\} \le J_{\mu^0},$$

where K_1 is some positive semidefinite symmetric matrix.

(c) Consider repeating the (policy iteration) process described in parts (a) and (b), thereby obtaining a sequence of positive semidefinite symmetric matrices $\{K_k\}$. Show that

$$K_k \to K,$$

where K is the optimal cost matrix of the problem.

4.13 (Alternative Analysis for the Unichain Case)

The purpose of this exercise is to show how to extend the average cost problem analysis based on the connection with the stochastic shortest path problem, which is given in Section 7.4 of Vol. I. In particular, here this connection is used to show that there exists a solution (λ, h) to Bellman's equation $\lambda e + h = Th$ if every policy that is optimal within the class of stationary policies is unichain, without resorting to the use of Blackwell optimal policies (cf. Prop. 2.6). For this we will use the stochastic shortest path theory of Section 2.1, and from the present chapter, Prop. 2.1 and Prop. 2.5 (which is proved using a stochastic shortest path argument). Complete the details of the following proof:

For any stationary policy μ, let λ_μ be the average cost per stage as defined by Eq. (2.17), let $\lambda = \min_\mu \lambda_\mu$, and let $M = \{\mu \mid \lambda_\mu = \lambda\}$. Suppose that there is a state s that is simultaneously recurrent in the Markov chains corresponding to all $\mu \in M$. Similar to Section 7.4 in Vol. I, consider an associated stochastic shortest path problem with states $1, 2, \ldots, n$ and an artificial termination state t to which we move from state i with transition probability $p_{is}(u)$. The stage costs in this problem are $g(i,u) - \lambda$ for $i = 1, \ldots, n$, and the transition probabilities from a state i to a state $j \neq s$ are the same as those of the original problem, while $p_{is}(u)$ is zero. Show that in this stochastic shortest path problem, every improper policy has infinite cost for some initial state, and use this fact to conclude that if $h(i)$ is the optimal cost starting at state $i = 1, \ldots, n$, then λ and h satisfy $\lambda e + h = Th$. If there is no state s that is simultaneously recurrent for all $\mu \in M$, select a $\overline{\mu} \in M$ such that there is no $\mu \in M$ whose recurrent class is a strict subset of the recurrent class of $\overline{\mu}$ (it is sufficient that $\overline{\mu}$ has minimal number of recurrent states over all $\mu \in M$), change the stage cost of all states i that are not recurrent under $\overline{\mu}$ to $g(i,u) + \epsilon$, where $\epsilon > 0$, use as state s in the preceding argument any state that is recurrent under $\overline{\mu}$, and take $\epsilon \to 0$.

4.14 (Stochastic Shortest Path Solution Method)

The purpose of this exercise is to show how the average cost problem can be solved by solving a finite sequence of stochastic shortest path problems. As in Section 7.4 of Vol. I, we assume that a special state, by convention state n, is recurrent in the Markov chain corresponding to each stationary policy. For a stationary policy μ, let

$C_\mu(i)$: expected cost starting from i up to the first visit to n,

$N_\mu(i)$: expected number of stages starting from i up to the first visit to n.

The proof of Prop. 2.5 shows that $\lambda_\mu = C_\mu(n)/N_\mu(n)$. Let $\lambda^* = \min_\mu \lambda_\mu$ be the corresponding optimal cost.

Consider the stochastic shortest path problem obtained by leaving unchanged all transition probabilities $p_{ij}(u)$ for $j \neq n$, by setting all transition probabilities $p_{in}(u)$ to 0, and by introducing an artificial termination state t to which we move from each state i with probability $p_{in}(u)$. The expected stage

cost at state i is $g(i,u) - \lambda$, where λ is a scalar parameter. Let $h_{\mu,\lambda}(i)$ be the cost of stationary policy μ for this stochastic shortest path problem, starting from state i, and let $h_\lambda(i) = \min_\mu h_{\mu,\lambda}(i)$ be the corresponding optimal cost.

(a) Show that for all scalars λ and λ', we have

$$h_{\mu,\lambda}(i) = h_{\mu,\lambda'}(i) + (\lambda' - \lambda)N_\mu(i), \qquad i = 1, \ldots, n.$$

(b) Define

$$h_\lambda(i) = \min_\mu h_{\mu,\lambda}(i), \qquad i = 1, \ldots, n.$$

Show that $h_\lambda(i)$ is concave, monotonically decreasing, and piecewise linear as a function of λ, and that

$$h_\lambda(n) = 0 \qquad \text{if and only if} \qquad \lambda = \lambda^*.$$

Figure 4.5.1 illustrates these relations.

(c) Consider a one-dimensional search procedure that finds a zero of the function $h_\lambda(n)$ of λ. This procedure brackets λ^* from above and below, and is illustrated in Fig. 4.5.2. Show that this procedure solves the average cost problem by solving a finite number of stochastic shortest path problems.

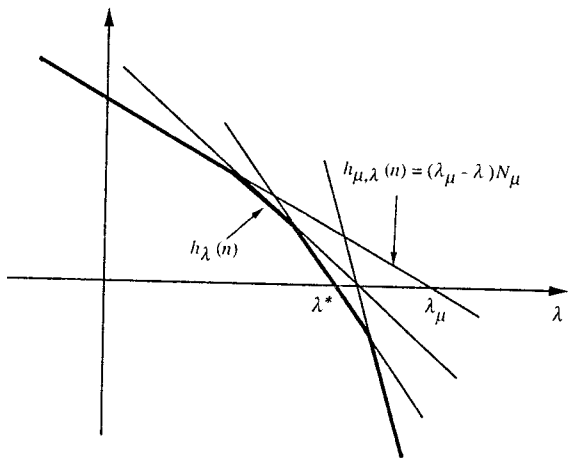


Figure 4.5.1 Relation of the costs of stationary policies in the average cost problem and the associated stochastic shortest path problem.

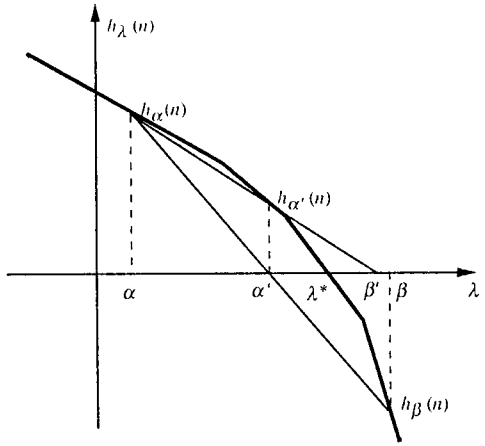


Figure 4.5.2 One dimensional iterative search procedure to find λ such that $h_\lambda(n) = 0$ [cf. Exercise 4.14(c)]. Each value $h_\lambda(n)$ is obtained by solving the associated stochastic shortest path problem with stage cost $g(i,u) - \lambda$. At the start of the typical iteration, we have scalars α and β such that $\alpha < \lambda^* < \beta$, together with the corresponding nonzero values $h_\alpha(n)$ and $h_\beta(n)$. We find α' such that

$$\frac{\alpha' - \alpha}{\alpha' - \beta} = \frac{h_\alpha(n)}{h_\beta(n)},$$

and we calculate $h_{\alpha'}(n)$. Let β' be such that

$$\frac{\beta' - \alpha'}{\beta' - \alpha} = \frac{h_{\alpha'}(n)}{h_\alpha(n)}.$$

We then replace α by α', and if $\beta' < \beta$, we also calculate $h_{\beta'}(n)$ and we replace β by β'. We then perform another iteration. The algorithm stops if either $h_\alpha(n) = 0$ or $h_\beta(n) = 0$.

4.15 (Stochastic Shortest Path-Based Value Iteration [Ber95c])

The purpose of this exercise is to provide a value iteration method for average cost problems, which is based on the connection with the stochastic shortest path problem. Let the assumptions of Exercise 4.14 hold. Consider the algorithm

$$h^{k+1}(i) = \min_{u \in U(i)} \left[g(i,u) + \sum_{j=1}^{n-1} p_{ij}(u) h^k(j) \right] - \lambda^k, \qquad i = 1, \ldots, n,$$

$$\lambda^{k+1} = \lambda^k + \delta^k h^{k+1}(n),$$

where n is the special state that is recurrent for all unichain policies and δ^k is a positive stepsize.

(a) Interpret the algorithm as a value iteration algorithm for a slowly varying stochastic shortest path problem of the type considered in Exercise 4.14. Given that, for small δ, the iteration of h is faster that the iteration of λ, speculate on the convergence properties of the algorithm. [It can be proved that there exists a positive constant $\bar{\delta}$ such that we have $h^k(n) \to 0$ and $\lambda^k \to \lambda^*$ if $\underline{\delta} \leq \delta^k \leq \bar{\delta}$, where $\underline{\delta}$ is some positive constant. Another interesting possibility for which convergence can be proved is to select δ^k as a constant divided by 1 plus the number of times that $h^k(n)$ has changed sign.]

(b) Use the error bounds of Prop. 3.2 to justify the iteration

$$h^{k+1}(i) = \min_{u \in U(i)} \left[g(i,u) + \sum_{j=1}^{n-1} p_{ij}(u) h^k(j) \right] - \lambda^k, \qquad i = 1, \ldots, n,$$

$$\lambda^{k+1} = \left[\lambda^k + \delta^k h^{k+1}(n) \right]^+,$$

where $[c]^+$ denotes the projection of a scalar c on the interval

$$\left[\max_{m=0,\ldots,k} \underline{\beta}^m, \min_{m=0,\ldots,k} \bar{\beta}^m \right],$$

with

$$\underline{\beta}^k = \lambda^k + \min \left[\min_{i \neq n} \left[h^{k+1}(i) - h^k(i) \right], h^{k+1}(n) \right],$$

$$\bar{\beta}^k = \lambda^k + \max \left[\max_{i \neq n} \left[h^{k+1}(i) - h^k(i) \right], h^{k+1}(n) \right].$$

4.16 (Q-Learning Based on Stochastic Shortest Paths)

The purpose of this exercise is to provide a Q-learning method for average cost problems, which is based on the value iteration method of Exercise 4.15. Let the assumptions of Exercise 4.14 hold. Speculate on the convergence properties of the following Q-learning algorithm

$$Q(i,u) := Q(i,u) + \gamma \left(g(i,u,j) + \min_{u' \in U(j)} \hat{Q}(j,u') - Q(i,u) \right) - \lambda,$$

$$i = 1, \ldots, n, \ u \in U(i),$$

$$\lambda := \lambda + \delta \min_{u' \in U(n)} Q(n,u'),$$

where

$$\hat{Q}(j,u') = \begin{cases} Q(j,u') & \text{if } j \neq n, \\ 0 & \text{otherwise,} \end{cases}$$

and j and $g(i,u,j)$ are generated from the pair (i,u) by simulation. Here the stepsizes γ and δ should be diminishing, but δ should diminish "faster" than γ [for example $\gamma = c_1/k$ and $\delta = c_2/k \log k$, where c_1 and c_2 are positive constants and k is the number of iterations performed on the corresponding pair (i,u) or λ].

5

Continuous-Time Problems

Contents

We have considered so far problems where the cost per stage does not depend on the time required for transition from one state to the next. Such problems have a natural discrete-time representation. On the other hand, there are situations where controls are applied at discrete times but cost is continuously accumulated. Furthermore, the time between successive control choices is variable; it may be random or it may depend on the current state and the choice of control. For example, in queueing systems state transitions correspond to arrivals or departures of customers, and the corresponding times of transition are random. This chapter primarily discusses problems of this type. We restrict attention to continuous-time systems with a finite or countable number of states. Many of the practical systems of this type involve the Poisson process, so for many of the examples discussed, we assume that the reader is familiar with this process at the level of textbooks such as [Ros83b] and [Gal95].

In Section 5.1, we concentrate on an important class of continuous-time optimization models of the discounted type, where the times between successive transitions have an *exponential probability distribution*. We show that by using a conversion process called *uniformization*, discounted versions of these models can be analyzed within the discrete-time framework discussed up to now.

In Section 5.2, we discuss applications of uniformization. We concentrate on queueing models arising in various communications and scheduling contexts.

In Section 5.3, we discuss more general continuous-time models, called *semi-Markov problems*, where the times between successive transitions need not have an exponential distribution.

5.1 UNIFORMIZATION

In this chapter, we restrict ourselves to continuous-time systems with a finite or a countable number of states. Here state transitions and control selections take place at discrete times, but the time from one transition to the next is random. In this section, we assume that:

1. If the system is in state i and control u is applied, the next state will be j with probability $p_{ij}(u)$.
2. The time interval τ between the transition to state i and the transition to the next state is exponentially distributed with parameter $\nu_i(u)$; that is,

$$P\{\text{transition time interval } \leq \tau \mid i, u\} \leq 1 - e^{-\nu_i(u)\tau},$$

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 *rate of transition* associated with state i and control u. It can be verified that the corresponding average transition time is

$$E\{\tau\} = \int_0^\infty \tau \nu_i(u) e^{-\nu_i(u)\tau} d\tau = \frac{1}{\nu_i(u)},$$

so $\nu_i(u)$ can be interpreted as the average number of transitions per unit time.

The state and control at any time t are denoted by $x(t)$ and $u(t)$, respectively, and stay constant between transitions. We use the following notation:

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

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

$x_k = x(t_k)$: We have $x(t) = x_k$ for $t_k \leq t < t_{k+1}$.

$u_k = u(t_k)$: We have $u(t) = u_k$ for $t_k \leq t < t_{k+1}$.

We consider a cost function of the form

$$\lim_{N \to \infty} E\left\{ \int_0^{t_N} e^{-\beta t} g\big(x(t), u(t)\big) dt \right\}, \tag{1.1}$$

where g is a given function and β is a given positive discount parameter. Similar to discrete-time problems, an admissible policy is a sequence $\pi = \{\mu_0, \mu_1, \ldots\}$, 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 $\mu_k(x_k)$. Because states stay constant between transitions, the cost function of π is given by

$$J_\pi(x_0) = \sum_{k=0}^{\infty} E\left\{ \int_{t_k}^{t_{k+1}} e^{-\beta t} g\big(x_k, \mu_k(x_k)\big) \,\Big|\, x_0 \right\}.$$

We first consider the case where *the rate of transition is the same for all states and controls*; that is,

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

A little thought shows that the problem is then essentially the same as the one where transition times are fixed, because the control cannot influence the cost of a stage by affecting the length of the next transition time interval.

Indeed, the cost (1.1) corresponding to a sequence $\{(x_k, u_k)\}$ can be expressed as

$$\sum_{k=0}^{\infty} E\left\{\int_{t_k}^{t_{k+1}} e^{-\beta t} g\big(x(t), u(t)\big)dt\right\} = \sum_{k=0}^{\infty} E\left\{\int_{t_k}^{t_{k+1}} e^{-\beta t}dt\right\} E\big\{g(x_k, u_k)\big\} \tag{1.2}$$

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

$$\begin{aligned} E\left\{\int_{t_k}^{t_{k+1}} e^{-\beta t}dt\right\} &= \frac{E\{e^{-\beta t_k}\}\big(1 - E\{e^{-\beta\tau_{k+1}}\}\big)}{\beta} \\ &= \frac{E\{e^{-\beta(\tau_1+\cdots+\tau_k)}\}\big(1 - E\{e^{-\beta\tau_{k+1}}\}\big)}{\beta} \\ &= \frac{\alpha^k(1-\alpha)}{\beta}, \end{aligned} \tag{1.3}$$

where

$$\alpha = E\{e^{-\beta\tau}\} = \int_0^{\infty} e^{-\beta\tau}\nu e^{-\nu\tau}d\tau = \frac{\nu}{\beta+\nu}.$$

The above expression for α yields $(1-\alpha)/\beta = 1/(\beta+\nu)$, so that from Eq. (1.3), we have

$$E\left\{\int_{t_k}^{t_{k+1}} e^{-\beta t}dt\right\} = \frac{\alpha^k}{\beta+\nu}.$$

From this equation together with Eq. (1.2) it follows that the cost of the problem can be expressed as

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

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

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

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

$$\alpha = \frac{\nu}{\beta+\nu}, \tag{1.4}$$

and cost per stage given by

$$\hat{g}(i,u) = \frac{g(i,u)}{\beta+\nu}. \tag{1.5}$$

In particular, Bellman's equation takes the form

$$J(i) = \min_{u\in U(i)}\left[\frac{g(i,u)}{\beta+\nu} + \alpha\sum_j p_{ij}(u)J(j)\right], \tag{1.6}$$

or equivalently,

$$J(i) = \frac{1}{\beta+\nu}\min_{u\in U(i)}\left[g(i,u) + \nu\sum_j p_{ij}(u)J(j)\right]. \tag{1.7}$$

In some problems, in addition to the cost (1.1), there is an extra expected stage cost $\hat{g}(i,u)$ that is incurred at the time the control u is chosen at state i, and is independent of the length of the transition interval. In that case the expected stage cost (1.5) should be changed to

$$\hat{g}(i,u) + \frac{g(i,u)}{\beta+\nu}, \tag{1.8}$$

and Bellman's equation (1.6) becomes

$$J(i) = \min_{u\in U(i)}\left[\hat{g}(i,u) + \frac{g(i,u)}{\beta+\nu} + \alpha\sum_j p_{ij}(u)J(j)\right]. \tag{1.9}$$

Example 1.1

A manufacturer of a specialty item processes orders in batches. Orders arrive according to a Poisson process with rate ν per unit time; that is, the successive interarrival intervals are independent and exponentially distributed with parameter ν. For each order there is a positive cost c per unit time that the order is unfilled. Costs are discounted with a discount parameter $\beta > 0$. The setup cost for processing the orders is K. Upon arrival of a new order, the manufacturer must decide whether to process the current batch or to wait for the next order.

Here the state is the number i of unfilled orders. If the decision to fill the orders at state i is made, the cost is K and the next transition will be to state 1. Otherwise, there will be an average cost $(ci)/(\beta+\nu)$ up to the transition to the next state $i+1$ [cf. Eq. (1.5)], as shown in Fig. 5.1.1. [Note that the setup cost K is incurred immediately after a decision to process the orders is made, so K is not discounted over the time interval up to the next

transition; cf. Eq. (1.9).] We are in effect faced with a discounted discrete-time problem with positive but unbounded cost per stage. (We could also consider an alternative model where an upper bound is placed on the number of unfilled orders. We would then have a discounted discrete-time problem with bounded cost per stage.)

Since Assumption P is satisfied (cf. Section 3.1), Bellman's equation holds and takes the form

$$J(i) = \min\left[K + \alpha J(1), \frac{ci}{\beta+\nu} + \alpha J(i+1)\right], \qquad i = 1, 2, \ldots, \tag{1.10}$$

where $\alpha = \nu/(\beta+\nu)$ is the effective discount factor [cf. Eq. (1.1)]. Reasoning from first principles, we see that $J(i)$ is a monotonically nondecreasing function of i, so from Bellman's equation it follows that there exists a threshold i^* such that it is optimal to process the orders if and only if their number exceeds i^*.

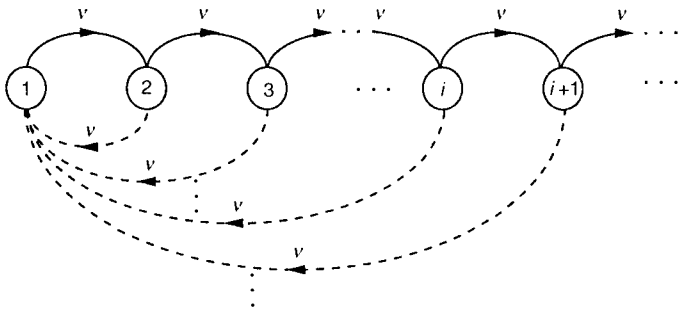


Figure 5.1.1 Transition diagram for the continuous-time Markov chain of Example 1.1. The transitions associated with the first control (do not fill the orders) are shown with solid lines, and the transitions associated with the second control (fill the orders) are shown with broken lines.

Nonuniform Transition Rates

We now argue that the more general case where the transition rate $\nu_i(u)$ depends on the state and the control can be converted to the previous case of uniform transition rate by using the trick of *allowing fictitious transitions from a state to itself*. Roughly, transitions that are slow on the average are speeded up with the understanding that sometimes after a transition the state may stay unchanged. To see how this works, let ν be a new uniform transition rate with $\nu_i(u) \leq \nu$ for all i and u, and define new

transition probabilities by

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

We refer to this process as the *uniform* version of the original (see Fig. 5.1.2). We argue now that leaving state i at a rate $\nu_i(u)$ in the original process is statistically identical to leaving state i at the faster rate ν, but returning back to i with probability $1 - \nu_i(u)/\nu$ in the new process. Equivalently, transitions are real (lead to a different state) with probability $\nu_i(u)/\nu < 1$. By statistical equivalence, we mean that, for any given policy π, initial state x_0, time t, and state i, the probability $P\{x(t) = i \mid x_0, \pi\}$ is identical for the original process and its uniform version. We give a proof of this fact in Exercise 5.1 for the case of a finite number of states (see also [Lip75], [Ser79], 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

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

into a discrete-time Markov chain problem with discount factor

$$\alpha = \frac{\nu}{\beta+\nu}, \tag{1.11}$$

where ν is a uniform transition rate chosen so that

$$\nu_i(u) \leq \nu, \qquad \text{for all } i, u. \tag{1.12}$$

The transition probabilities are

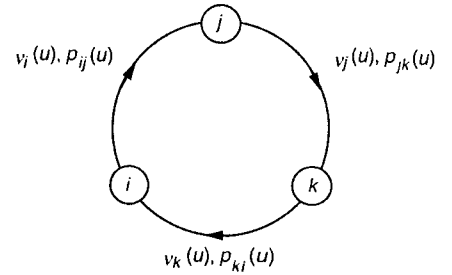
$$\tilde{p}_{ij}(u) = \begin{cases} \frac{\nu_i(u)}{\nu} p_{ij}(u) & \text{if } i \neq j, \\ \frac{\nu_i(u)}{\nu} p_{ii}(u) + 1 - \frac{\nu_i(u)}{\nu} & \text{if } i = j, \end{cases} \tag{1.13}$$

and the cost per stage is

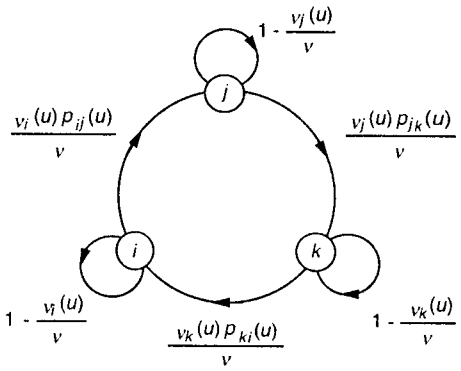
$$\tilde{g}(i, u) = \frac{g(i, u)}{\beta+\nu}, \qquad \text{for all } i, u.$$

In particular, Bellman's equation takes the form

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



Transition rates and probabilities for continuous-time chain



Transition probabilities for uniform version

Figure 5.1.2 Transforming a continuous-time Markov chain into its uniform version through the use of fictitious self-transitions. The uniform version has a uniform transition rate ν, which is an upper bound for all transition rates $\nu_i(u)$ of the original, and transition probabilities $\tilde{p}_{ij}(u) = \big(\nu_i(u)/\nu\big)p_{ij}(u)$ for $i \neq j$, and $\tilde{p}_{ii}(u) = \big(\nu_i(u)/\nu\big)p_{ii}(u) + 1 - \nu_i(u)/\nu$ for $j = i$. In the example of the figure we have $p_{ii}(u) = 0$ for all i and u.

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

$$J(i) = \frac{1}{\beta + \nu} \min_{u \in U(i)} \left[g(i,u) + \big(\nu - \nu_i(u)\big)J(i) + \nu_i(u) \sum_j p_{ij}(u)J(j) \right]. \tag{1.14}$$

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. (1.9)]

$$J(i) = \frac{1}{\beta + \nu} \min_{u \in U(i)} \Big[(\beta + \nu)\hat{g}(i,u) + g(i,u) + \big(\nu - \nu_i(u)\big)J(i) + \nu_i(u) \sum_j p_{ij}(u)J(j) \Big]. \tag{1.15}$$

Undiscounted and Average Cost Problems

When the discount parameter β is zero in the preceding problem formulation, we obtain a continuous-time version of the undiscounted cost problem considered in Chapter 3. If in addition, the number of states is finite and there is a cost-free and absorbing state, we obtain a continuous-time analog of the stochastic shortest path problem considered of Chapter 2. However, when $\beta = 0$, it is unnecessary to resort to uniformization. It can be seen that the problem is essentially the same as the discrete-time problem with the same transition probabilities but where the average transition cost at state i under u is the average cost per unit time $g(i,u)$ multiplied with the expected length $1/\nu_i(u)$ of the transition interval. Thus Bellman's equation has the form

$$J(i) = \min_{u \in U(i)} \left[\frac{g(i,u)}{\nu_i(u)} + \sum_j p_{ij}(u)J(j) \right]. \tag{1.16}$$

After some calculation, it can be seen that the above equation can also be obtained from Eq. (1.14) by setting $\beta = 0$.

In fact for undiscounted problems, the preceding argument does not depend on the character of the probability distributions of the transition times. Regardless of whether these distributions are exponential or not, one simply needs to multiply $g(i,u)$ with the average transition time corresponding to (i,u) and then treat the problem as if it were a discrete-time problem.

There is also a continuous-time version of the average cost per stage problem of Chapter 4. The cost function has the form

$$\lim_{N \to \infty} \frac{1}{E\{t_N\}} E\left\{ \int_0^{t_N} g\big(x(t), u(t)\big)dt \right\}.$$

We will consider this problem in Section 5.3 in a more general context where the probability distributions of the transition times need not be exponential.

5.2 QUEUEING APPLICATIONS

We now illustrate the theory of the preceding section through some applications involving the control of queues.

Example 2.1 (M/M/1 Queue with Controlled Service Rate)

Consider a single-server queueing system where customers arrive according to a Poisson process with rate λ. The service time of a customer is exponentially distributed with parameter μ (called the service rate). Service times of customers are independent and are also independent of customer interarrival times. The service rate μ can be selected from a closed subset M of an interval $[0, \bar{\mu}]$ and can be changed at the times when a customer arrives or when a customer departs from the system. There is a cost $q(\mu)$ per unit time for using rate μ and a waiting cost $c(i)$ per unit time when there are i customers in the system (waiting in queue or undergoing service). The idea is that one should be able to cut down on the customer waiting costs by choosing a faster service rate, which presumably costs more. The problem, roughly, is to select the service rate so that the service cost is optimally traded off with the customer waiting cost.

We assume the following:

1. For some $\mu \in M$ we have $\mu > \lambda$. (In words, there is available a service rate that is fast enough to keep up with the arrival rate, thereby maintaining the queue length bounded.)

2. The waiting cost function c is nonnegative, monotonically nondecreasing, and "convex" in the sense

$$c(i+2) - c(i+1) \geq c(i+1) - c(i), \qquad i = 0, 1, \ldots.$$

3. The service rate cost function q is nonnegative, and continuous on $[0, \bar{\mu}]$, with $q(0) = 0$.

The problem fits the framework of this section. The state is the number of customers in the system, and the control is the choice of service rate following a customer arrival or departure. The transition rate at state i is

$$\nu_i(\mu) = \begin{cases} \lambda & \text{if } i = 0, \\ \lambda + \mu & \text{if } i \geq 1. \end{cases}$$

The transition probabilities of the Markov chain and its uniform version for the choice

$$\nu = \lambda + \bar{\mu}$$

are shown in Fig. 5.2.1.

The effective discount factor is

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

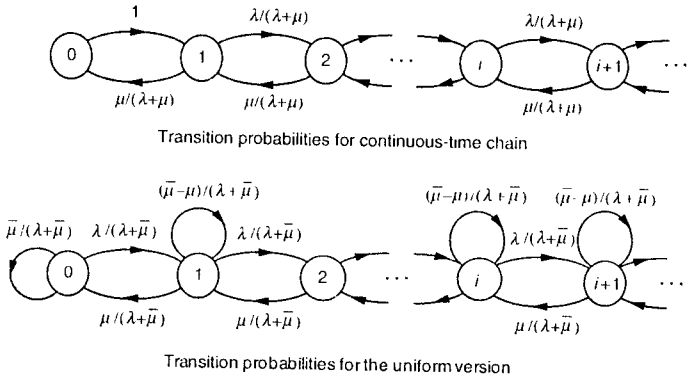


Figure 5.2.1 Continuous-time Markov chain and uniform version for Example 2.1 when the service rate is equal to μ. The transition rates of the original Markov chain are $\nu_i(\mu) = \lambda + \mu$ for states $i \geq 1$, and $\nu_0(\mu) = \lambda$ for state 0. The transition rate for the uniform version is $\nu = \lambda + \bar{\mu}$.

and the cost per stage is

$$\frac{1}{\beta + \nu}\big(c(i) + q(\mu)\big).$$

The form of Bellman's equation is [cf. Eq. (1.14)]

$$J(0) = \frac{1}{\beta + \nu}\big(c(0) + (\nu - \lambda)J(0) + \lambda J(1)\big)$$

and for $i = 1, 2, \ldots$,

$$J(i) = \frac{1}{\beta + \nu} \min_{\mu \in M}\big[c(i) + q(\mu) + \mu J(i-1) + (\nu - \lambda - \mu)J(i) + \lambda J(i+1)\big]. \quad (2.1)$$

An optimal policy is to use at state i the service rate that minimizes the expression on the right. Thus it is optimal to use at state i the service rate

$$\mu^*(i) = \arg\min_{\mu \in M}\big\{q(\mu) - \mu\Delta(i)\big\}, \quad (2.2)$$

where $\Delta(i)$ is the differential of the optimal cost

$$\Delta(i) = J(i) - J(i-1), \qquad i = 1, 2, \ldots$$

[When the minimum in Eq. (2.2) is attained by more than one service rate μ we choose by convention the smallest.] We will demonstrate shortly that $\Delta(i)$ *is monotonically nondecreasing*. It will then follow from Eq. (2.2) (see

Fig. 5.2.2) that *the optimal service rate $\mu^*(i)$ is monotonically nondecreasing*. Thus, as the queue length increases, it is optimal to use a faster service rate.

To show that $\Delta(i)$ is monotonically nondecreasing, we use the value iteration method to generate a sequence of functions J_k from the starting function

$$J_0(i) = 0, \qquad i = 0, 1, \ldots$$

For $k = 0, 1, \ldots$, [cf. Eq. (2.1)], we have

$$J_{k+1}(0) = \frac{1}{\beta + \nu}\big(c(0) + (\nu - \lambda)J_k(0) + \lambda J_k(1)\big),$$

and for $i = 1, 2, \ldots$,

$$J_{k+1}(i) = \frac{1}{\beta + \nu} \min_{\mu \in M}\big[c(i) + q(\mu) + \mu J_k(i-1) + (\nu - \lambda - \mu)J_k(i) + \lambda J_k(i+1)\big]. \tag{2.3}$$

For $k = 0, 1, \ldots$ and $i = 1, 2, \ldots$, let

$$\Delta_k(i) = J_k(i) - J_k(i-1).$$

For completeness of notation, define also $\Delta_k(0) = 0$. From the theory of Section 3.1 (see Prop. 1.7 of that section), we have $J_k(i) \to J(i)$ as $k \to \infty$. It follows that we have

$$\lim_{k \to \infty} \Delta_k(i) = \Delta(i), \qquad i = 1, 2, \ldots$$

Therefore, it will suffice to show that $\Delta_k(i)$ is monotonically nondecreasing for every k. For this we use induction. The assertion is trivially true for $k = 0$. Assuming that $\Delta_k(i)$ is monotonically nondecreasing, we show that the same is true for $\Delta_{k+1}(i)$. Let

$$\mu^k(0) = 0,$$

$$\mu^k(i) = \arg\min_{\mu \in M}\big[q(\mu) - \mu\Delta_k(i)\big], \qquad i = 1, 2, \ldots$$

From Eq. (2.3) we have, for all $i = 0, 1, \ldots$,

$$\begin{aligned}
\Delta_{k+1}(i+1) &= J_{k+1}(i+1) - J_{k+1}(i) \\
&= \frac{1}{\beta + \nu}\big(c(i+1) + q\big(\mu^k(i+1)\big) + \mu^k(i+1)J_k(i) \\
&\quad + \big(\nu - \lambda - \mu^k(i+1)\big)J_k(i+1) \\
&\quad + \lambda J_k(i+2) - c(i) - q\big(\mu^k(i+1)\big) - \mu^k(i+1)J_k(i-1) \\
&\quad - \big(\nu - \lambda - \mu^k(i+1)\big)J_k(i) - \lambda J_k(i+1)\big) \\
&= \frac{1}{\beta + \nu}\big(c(i+1) - c(i) + \lambda\Delta_k(i+2) + (\nu - \lambda)\Delta_k(i+1) \\
&\quad - \mu^k(i+1)\big(\Delta_k(i+1) - \Delta_k(i)\big)\big).
\end{aligned} \tag{2.4}$$

Similarly, we obtain, for $i = 1, 2, \ldots$,

$$\Delta_{k+1}(i) \le \frac{1}{\beta + \nu}\big(c(i) - c(i-1) + \lambda\Delta_k(i+1) + (\nu - \lambda)\Delta_k(i) - \mu^k(i-1)\big(\Delta_k(i) - \Delta_k(i-1)\big)\big).$$

Subtracting the last two inequalities, we obtain, for $i = 1, 2, \ldots$,

$$\begin{aligned}
(\beta + \nu)\big(\Delta_{k+1}(i+1) - \Delta_{k+1}(i)\big) &\ge \big(c(i+1) - c(i)\big) - \big(c(i) - c(i-1)\big) \\
&\quad + \lambda\big(\Delta_k(i+2) - \Delta_k(i+1)\big) \\
&\quad + \big(\nu - \lambda - \mu^k(i+1)\big)\big(\Delta_k(i+1) - \Delta_k(i)\big) \\
&\quad + \mu^k(i-1)\big(\Delta_k(i) - \Delta_k(i-1)\big).
\end{aligned}$$

Using our convexity assumption on $c(i)$, the fact $\nu - \lambda - \mu^k(i+1) = \overline{\mu} - \mu^k(i+1) \ge 0$, and the induction hypothesis, we see that every term on the right-hand side of the preceding inequality is nonnegative. Therefore, $\Delta_{k+1}(i+1) \ge \Delta_{k+1}(i)$ for $i = 1, 2, \ldots$. From Eq. (2.4) we can also show that $\Delta_{k+1}(1) \ge 0 = \Delta_{k+1}(0)$, and the induction proof is complete.

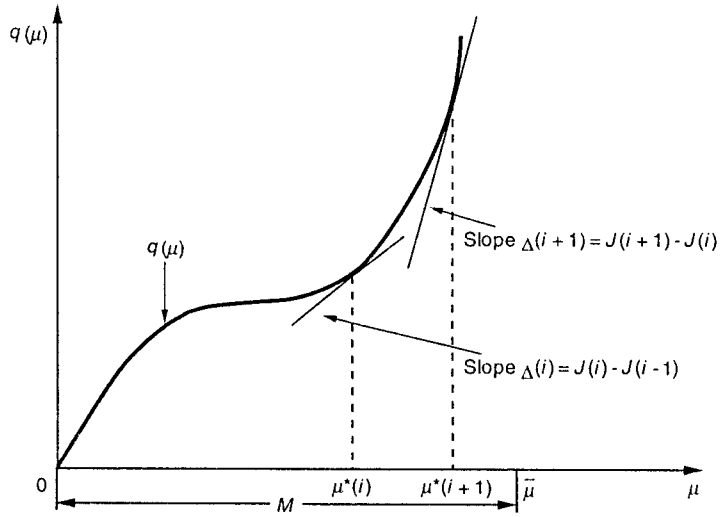


Figure 5.2.2 Determining the optimal service rate at states i and $(i+1)$ in Example 2.1. The optimal service rate $\mu^*(i)$ tends to increase as the system becomes more crowded (i increases).

To summarize, the optimal service rate $\mu^*(i)$ is given by Eq. (2.2) and tends to become faster as the system becomes more crowded (i increases).

Example 2.2 (M/M/1 Queue with Controlled Arrival Rate)

Consider the same queueing system as in the previous example with the difference that the service rate μ is fixed, but the arrival rate λ can be controlled. We assume that λ is chosen from a closed subset Λ of an interval $[0, \overline{\lambda}]$, and there is a cost $q(\lambda)$ per unit time. All other assumptions of Example 2.1 are also in effect. What we have here is a problem of flow control, whereby we want to trade off optimally the cost for throttling the arrival process with the customer waiting cost.

This problem is very similar to the one of Example 2.1. We choose as uniform transition rate

$$\nu = \overline{\lambda} + \mu$$

and construct the uniform version of the Markov chain. Bellman's equation takes the form

$$J(0) = \frac{1}{\beta + \nu} \min_{\lambda \in \Lambda} \Big[c(0) + q(\lambda) + (\nu - \lambda) J(0) + \lambda J(1) \Big],$$

$$J(i) = \frac{1}{\beta + \nu} \min_{\lambda \in \Lambda} \Big[c(i) + q(\lambda) + \mu J(i-1) + (\nu - \lambda - \mu) J(i) + \lambda J(i+1) \Big].$$

An optimal policy is to use at state i the arrival rate

$$\lambda^*(i) = \arg\min_{\lambda \in \Lambda} \Big[q(\lambda) + \lambda \Delta(i+1) \Big], \tag{2.5}$$

where, as before, $\Delta(i)$ is the differential of the optimal cost

$$\Delta(i) = J(i) - J(i-1), \qquad i = 1, 2, \ldots$$

As in Example 2.1, we can show that $\Delta(i)$ is monotonically nondecreasing; so from Eq. (2.5) we see that *the optimal arrival rate tends to decrease as the system becomes more crowded* (i increases).

Example 2.3 (Priority Assignment and the μc Rule)

Consider n queues that share a single server. There is a positive cost c_i per unit time and per customer in each queue i. The service time of a customer of queue i is exponentially distributed with parameter μ_i, and all customer service times are independent. Assuming that we start with a given number of customers in each queue and no further arrivals occur, what is the optimal order for serving the customers? The cost here is

$$\lim_{N \to \infty} E \left\{ \int_0^{t_N} e^{-\beta t} \sum_{i=1}^{n} c_i x_i(t) dt \right\},$$

where $x_i(t)$ is the number of customers in the ith queue at time t, and β is a positive discount parameter.

We first construct the uniform version of the problem. The construction is shown in Fig. 5.2.3. The discount factor is

$$\alpha = \frac{\mu}{\beta + \mu}, \tag{2.6}$$

where

$$\mu = \max_i \{\mu_i\},$$

and the corresponding cost is

$$\frac{1}{\beta + \mu} \sum_{k=0}^{\infty} \alpha^k E \left\{ \sum_{i=1}^{n} c_i x_k^i \right\}, \tag{2.7}$$

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

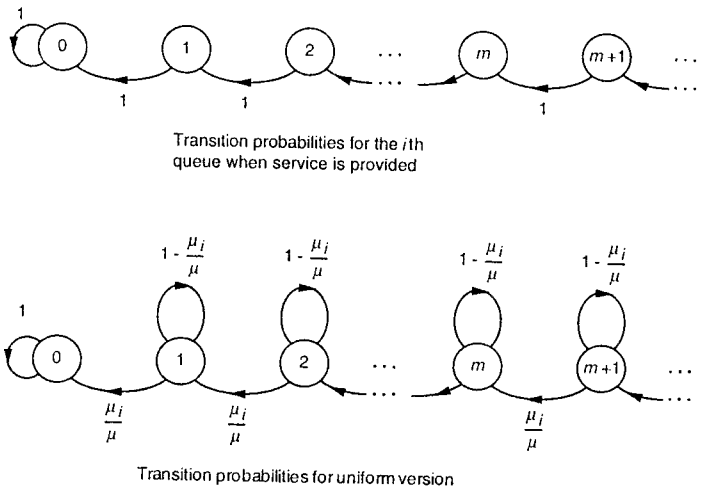


Figure 5.2.3 Continuous-time Markov chain and uniform version for the ith queue of Example 2.3 when service is provided. The transition rate for the uniform version is $\mu = \max_i \{\mu_i\}$.

We now rewrite the cost in a way that is more convenient for analysis. The idea is to transform the problem from one of minimizing waiting costs to one of maximizing savings in waiting costs through customer service. For $k = 0, 1, \ldots$, define

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

Denote also

$$c_{i_0} = 0,$$

x_0^i : the initial number of customers in queue i.

Then the cost (2.7) can also be written as

$$\begin{aligned}
\frac{1}{\beta+\mu}\left[\sum_{i=1}^{n} c_i x_0^i + \sum_{k=1}^{\infty} \alpha^k E\left\{\sum_{i=1}^{n} c_i x_0^i - \sum_{m=0}^{k-1} c_{i_m}\right\}\right] \\
= \frac{1}{\beta+\mu}\left[\sum_{k=0}^{\infty} \alpha^k \left(\sum_{i=1}^{n} c_i x_0^i\right) - E\left\{\sum_{m=0}^{\infty}\sum_{k=m+1}^{\infty} \alpha^k c_{i_m}\right\}\right] \\
= \frac{1}{(\beta+\mu)(1-\alpha)}\sum_{i=1}^{n} c_i x_0^i - \frac{\alpha}{(\beta+\mu)(1-\alpha)}\sum_{k=0}^{\infty} \alpha^k E\{c_{i_k}\} \\
= \frac{1}{\beta}\sum_{i=1}^{n} c_i x_0^i - \frac{\alpha}{\beta}\sum_{k=0}^{\infty} \alpha^k E\{c_{i_k}\}.
\end{aligned}$$

Therefore, instead of minimizing the cost (2.7), we can equivalently

$$\text{maximize} \sum_{k=0}^{\infty} \alpha^k E\{c_{i_k}\}, \tag{2.8}$$

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

We now recognize problem (2.8) as a multiarmed bandit problem. The n queues can be viewed as separate projects. At each time, a nonempty queue, say i, is selected and served. Since a customer departure occurs with probability μ_i/μ, and a fictitious transition that leaves the state unchanged occurs with probability $1 - \mu_i/\mu$, the corresponding expected reward is

$$\frac{\mu_i}{\mu} c_i. \tag{2.9}$$

It is evident that the problem falls in the deteriorating case examined at the end of Section 1.5. Therefore, after each customer departure, it is optimal to serve the queue with maximum expected reward per stage (i.e., engage the project with maximal index; cf. the end of Section 1.5). Equivalently [cf. Eq. (2.9)], *it is optimal to serve the nonempty queue i for which $\mu_i c_i$ is maximum*. This policy is known as the *μc rule*. It plays an important role in several other formulations of the priority assignment problem (see [BDM83], [Har75a], and [Har75b]). We can view $\mu_i c_i$ as the ratio of the waiting cost rate c_i by the average time $1/\mu_i$ needed to serve a customer. Therefore, the μc rule amounts to serving the queue for which the savings in waiting cost rate per unit average service time are maximized.

Example 2.4 (Routing Policies for a Two-Station System)

Consider the system consisting of two queues shown in Fig. 5.2.4. Customers arrive according to a Poisson process with rate λ and are routed upon arrival to one of the two queues. Service times are independent and exponentially distributed with parameter μ_1 in the first queue and μ_2 in the second queue. The cost is

$$\lim_{N\to\infty} E\left\{\int_0^{t_N} e^{-\beta t}\big(c_1 x_1(t) + c_2 x_2(t)\big)dt\right\},$$

where β, c_1, and c_2 are given positive scalars, and $x_1(t)$ and $x_2(t)$ denote the number of customers at time t in queues 1 and 2, respectively.

As earlier, we construct the uniform version of this problem with uniform rate

$$\nu = \lambda + \mu_1 + \mu_2 \tag{2.10}$$

and the transition probabilities shown in Fig. 5.2.5. We take as state space the set of pairs (i, j) of customers in queues 1 and 2. Bellman's equation takes the form

$$\begin{aligned}
J(i,j) = \frac{1}{\beta+\nu}\big(c_1 i + c_2 j + \mu_1 J\big((i-1)^+, j\big) + \mu_2 J\big(i, (j-1)^+\big)\big) \\
+ \frac{\lambda}{\beta+\nu}\min\big[J(i+1,j), J(i,j+1)\big],
\end{aligned} \tag{2.11}$$

where for any x we denote

$$(x)^+ = \max(0, x).$$

From this equation we see that an optimal policy is to route an arriving customer to queue 1 if and only if the state (i, j) at the time of arrival belongs to the set

$$S_1 = \big\{(i,j) \mid J(i+1,j) \le J(i,j+1)\big\}. \tag{2.12}$$

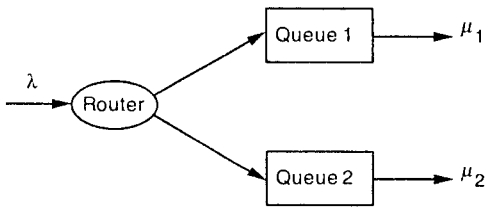


Figure 5.2.4 Queueing system of Example 2.4. The problem is to route each arriving customer to queue 1 or 2 so as to minimize the total average discounted waiting cost.

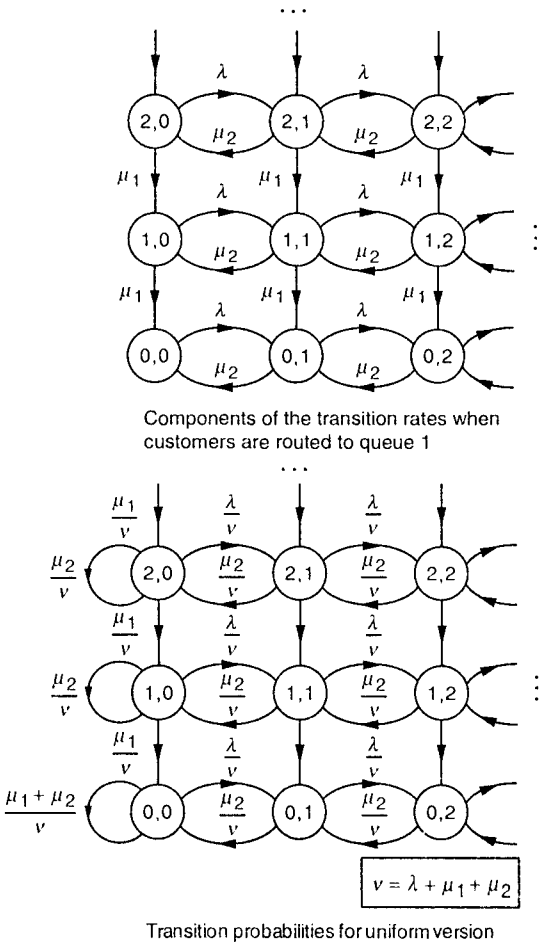


Figure 5.2.5 Continuous-time Markov chain and uniform version for Example 2.1 when customers are routed to the first queue. The states are the pairs of customer numbers in the two queues.

This optimal policy can be characterized better by some further analysis. Intuitively, one expects that optimal routing can be achieved by sending a customer to the queue that is "less crowded" in some sense. It is therefore natural to conjecture that, if it is optimal to route to the first queue when the state is (i, j), it must be optimal to do the same when the first queue is even less crowded; that is, the state is $(i - m, j)$ with $m \geq 1$. This is equivalent

to saying that the set of states S_1 for which it is optimal to route to the first queue is characterized by a monotonically nondecreasing *threshold function* F by means of

$$S_1 = \{(i, u) \mid i = F(j)\} \tag{2.13}$$

(see Fig. 5.2.6). Accordingly, we call the corresponding optimal policy a *threshold policy*.

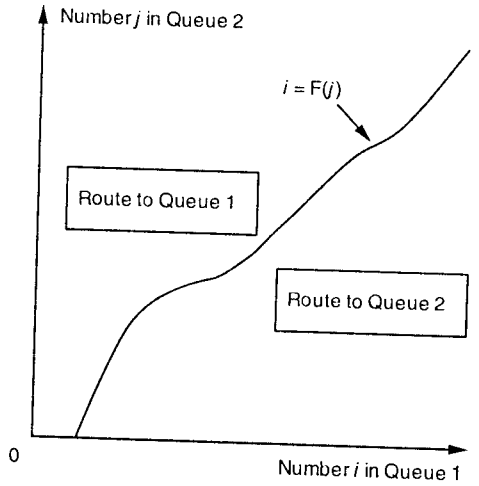


Figure 5.2.6 Typical threshold policy characterized by a threshold function F.

We will demonstrate the existence of a threshold optimal policy by showing that the functions

$$\Delta_1(i, j) = J(i + 1, j) - J(i, j + 1),$$

$$\Delta_2(i, j) = J(i, j + 1) - J(i + 1, j)$$

are monotonically nondecreasing in i for each fixed j, and in j for each fixed i, respectively. We will show this property for Δ_1; the proof for Δ_2 is analogous. It will be sufficient to show that for all $k = 0, 1, \ldots$, the functions

$$\Delta_1^k(i, j) = J_k(i + 1, j) - J_k(i, j + 1) \tag{2.14}$$

are monotonically nondecreasing in i for each fixed j, where J_k is generated by the value iteration method starting from the zero function; that is, $J_{k+1}(i, j) = (TJ_k)(i, j)$, where T is the DP mapping defining Bellman's equation (2.11) and $J_0 = 0$. This is true because $J_k(i, j) \to J(i, j)$ for all i, j as $k \to \infty$ (Prop. 1.6 in Section 3.1). To prove that $\Delta_1^k(i, j)$ has the desired

property, it is useful to first verify that $J_k(i,j)$ is monotonically nondecreasing in i (or j) for fixed j (or i). This is simple to show by induction or by arguing from first principles using the fact that $J_k(i,j)$ has a k-stage optimal cost interpretation. Next we use Eqs. (2.11) and (2.14) to write

$$\begin{aligned}(\beta+\nu)\Delta_1^{k+1}(i,j) = c_1 - c_2 \\ + \mu_1\big(J_k(i,j) - J_k\big((i-1)^+, j+1\big)\big) \\ + \mu_2\big(J_k\big(i+1,(j-1)^+\big) - J_k(i,j)\big) \\ + \lambda\Big(\min\big[J_k(i+2,j), J_k(i+1,j+1)\big] \\ - \min\big[J_k(i+1,j+1), J_k(i,j+2)\big]\Big).\end{aligned} \tag{2.15}$$

We now argue by induction. We have $\Delta_1^0(i,j) = 0$ for all (i,j). We assume that $\Delta_1^k(i,j)$ is monotonically nondecreasing in i for fixed j, and show that the same is true for $\Delta_1^{k+1}(i,j)$. This can be verified by showing that each of the terms in the right-hand side of Eq. (2.15) is monotonically nondecreasing in i for fixed j. Indeed, the first term is constant, and the second and third terms are seen to be monotonically nondecreasing in i using the induction hypothesis for the case where $i,j>0$ and the earlier shown fact that $J_k(i,j)$ is monotonically nondecreasing in i for the case where $i=0$ or $j=0$. The last term on the right-hand side of Eq. (2.15) can be written as

$$\begin{aligned}&\lambda\Big(J_k(i+1,j+1) + \min\big[J_k(i+2,j) - J_k(i+1,j+1), 0\big] \\ &\quad - J_k(i+1,j+1) - \min\big[0, J_k(i,j+2) - J_k(i+1,j+1)\big]\Big) \\ &= \lambda\Big(\min\big[0, J_k(i+1,j) - J_k(i+1,j+1)\big] \\ &\quad + \max\big[0, J_k(i+1,j+1) - J_k(i,j+2)\big]\Big) \\ &= \lambda\Big(\min\big[0, \Delta_1^k(i+1,j)\big] + \max\big[0, \Delta_1^k(i,j+1)\big]\Big).\end{aligned}$$

Since $\Delta_1^k(i+1,j)$ and $\Delta_1^k(i,j+1)$ are monotonically nondecreasing in i by the induction hypothesis, the same is true for the preceding expression. Therefore, each of the terms on the right-hand side of Eq. (2.15) is monotonically nondecreasing in i, and the induction proof is complete. Thus the existence of an optimal threshold policy is established.

There are a number of generalizations of the routing problem of this example that admit a similar analysis and for which there exist optimal policies of the threshold type. For example, suppose that there are additional Poisson arrival processes with rates λ_1 and λ_2 at queues 1 and 2, respectively. The existence of an optimal threshold policy can be shown by a nearly verbatim repetition of our analysis. A more substantive extension is obtained when there is additional service capacity μ that can be switched at the times of transition due to an arrival or service completion to serve a customer in queue 1 or 2. Then we can similarly prove that it is optimal to route to queue 1 if and only if $(i,j) \in S_1$ and to switch the additional service capacity to queue 2 if and only if $(i+1,j+1) \in S_1$, where S_1 is given by Eq. (2.12) and is characterized by a threshold function as in Eq. (2.13). For a proof of this and further extensions, we refer to [Haj84], which generalizes and unifies several earlier results on the subject.

5.3 SEMI-MARKOV PROBLEMS

We now consider a more general version of the continuous-time problem of Section 5.1. We still have a finite or a countable number of states, but we replace transition probabilities with *transition distributions* $Q_{ij}(\tau,u)$ that, for a given pair (i,u), specify the joint distribution of the transition interval and the next state:

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

We assume that for all states i and j, and controls $u \in U(i)$, $Q_{ij}(\tau,u)$ is known and that the average transition time is finite:

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

Note that the transition distributions specify the ordinary transition probabilities via

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

The difference from the model of Section 5.1 is that $Q_{ij}(\tau,u)$ need not be an exponential distribution.

Continuous-time problems with general transition distributions as described above are called *semi-Markov problems* because, for any given policy, while at a transition time t_k the future of the system statistically depends only on the current state, at other times it may depend in addition on the time elapsed since the preceding transition. By contrast, when the transition distributions are exponential, the future of the system depends only on its current state at all times. This is a consequence of the so called *memoryless property* of the exponential distribution. In our context, this property implies that, for any time t between the transition times t_k and t_{k+1}, the additional time $t_{k+1} - t$ needed to effect the next transition is independent of the time $t - t_k$ that the system has been in the current state [to see this, use the following generic calculation

$$\begin{aligned}P\{\tau > r_1 + r_2 \mid \tau > r_1\} &= \frac{P\{\tau > r_1 + r_2\}}{P\{\tau > r_1\}} \\ &= \frac{e^{-\nu(r_1+r_2)}}{e^{-\nu r_1}} \\ &= e^{-\nu r_2} \\ &= P\{\tau > r_2\},\end{aligned}$$

where $r_1 = t - t_k$, $r_2 = t_{k+1} - t$, and ν is the transition rate]. Thus, when the transition distributions are exponential, the state evolves in continuous time as a Markov process, but this need not be true for a more general distribution.

Discounted Problems

Let us first consider a cost function of the form

$$\lim_{N\to\infty} E\left\{\int_0^{t_N} e^{-\beta t} g\big(x(t),u(t)\big)dt\right\}, \tag{3.1}$$

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

$$J_\pi^N(i)=\sum_{k=0}^{N-1} E\left\{\int_{t_k}^{t_{k+1}} e^{-\beta t} g\big(x_k,\mu_k(x_k)\big)dt\,\Big|\, x_0=i\right\}.$$

We see that for all states i we have

$$J_\pi^N(i)=G\big(i,\mu_0(i)\big)+\sum_j \int_0^\infty e^{-\beta\tau} Q_{ij}\big(d\tau,\mu(i)\big)J_{\pi_1}^{N-1}(j), \tag{3.2}$$

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 $\int_0^\tau e^{-\beta t}dt=(1-e^{-\beta\tau})/\beta$,

$$G(i,u)=g(i,u)\sum_j \int_0^\infty \frac{1-e^{-\beta\tau}}{\beta} Q_{ij}(d\tau,u). \tag{3.3}$$

If we denote

$$m_{ij}(u)=\int_0^\infty e^{-\beta\tau} Q_{ij}(d\tau,u), \tag{3.4}$$

we see that Eq. (3.2) can be written in the form

$$J_\pi^N(i)=G\big(i,\mu_0(i)\big)+\sum_j m_{ij}\big(\mu_0(i)\big)J_{\pi_1}^{N-1}(j), \tag{3.5}$$

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 (3.5) motivates the use of mappings T and T_μ that are similar to those used in Chapter 1 for discounted problems. Let us define for a function J and a stationary policy μ,

$$(T_\mu J)(i)=G\big(i,\mu(i)\big)+\sum_j m_{ij}\big(\mu(i)\big)J(j), \tag{3.6}$$

$$(TJ)(i)=\min_{u\in U(i)}\left[G(i,u)+\sum_j m_{ij}(u)J(j)\right]. \tag{3.7}$$

Then by using Eq. (3.5), 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

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

The discounted cost analysis of Section 1.2 carries through in its entirety, provided we assume that:

(a) $g(i,u)$ [and hence also $G(i,u)$] is a bounded function of i and u.

(b) The maximum over (i,u) of the sum $\sum_j m_{ij}(u)$ is less than one; that is,

$$\rho=\max_{i,\,u\in U(i)}\sum_j m_{ij}(u)<1. \tag{3.8}$$

Under these circumstances, the mappings T and T_μ can be shown to be contraction mappings with modulus of contraction ρ [compare also with Prop. 2.4 in Section 1.2]. Using this fact, analogs of Props. 2.1-2.3 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

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

In addition, there are analogs of several of the computational methods of Section 1.3, including policy iteration and linear programming.

What is happening here is that essentially we have the equivalent of a discrete-time discounted problem where the discount factor depends on i and u. In fact, in Exercise 1.12 of Chapter 1, a data transformation is given, which converts such a problem to an ordinary discrete-time discounted problem where the discount factor is the same for all i and u. With a little thought it can be seen that this data transformation is very similar to the uniformization process we discussed in Section 5.1.

We note that for the contraction property $\rho<1$ [cf. Eq. (3.8)] to hold, it is sufficient that there exist $\tau>0$ and $\epsilon>0$ such that the transition time is greater than $\bar\tau$ with probability greater than $\epsilon>0$; that is, we have for all i and $u\in U(i)$,

$$1-\sum_j Q_{ij}(\bar\tau,u)=\sum_j P\{\tau\ge\bar\tau\mid i,u,j\}\ge\epsilon. \tag{3.9}$$

In the case where the state space is countably infinite and the function $g(i,u)$ is not bounded, the mappings T and T_μ are not contraction mappings, and a discounted cost analysis that parallels the one of Section 1.2 is not possible. Even in this case, however, analogs of the results of Section 3.1 can often be shown under appropriate conditions that parallel Assumptions P and N of that section.

We finally note that in some problems, in addition to the cost (3.1), there is an extra expected stage cost $\hat{g}(i,u)$ that is incurred at the time the control u is chosen at state i, and is independent of the length of the transition interval. In that case the mappings T and T_μ should be changed to

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

$$(TJ)(i) = \min_{u\in U(i)} \left[\hat{g}(i,u) + G(i,u) + \sum_j m_{ij}(u)J(j)\right]. \qquad (3.10)$$

Another problem variation arises when the cost per unit time g depends on the next state j. In this problem formulation, once the system goes into state i, a control $u \in U(i)$ is selected, the next state is determined to be j with probability $p_{ij}(u)$, and the cost of the next transition is $g(i,u,j)\tau_{ij}(u)$ where $\tau_{ij}(u)$ is random with distribution $Q_{ij}(\tau,u)/p_{ij}(u)$. In this case, $G(i,u)$ should be defined by

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

[cf. Eq. (3.3)] and the preceding development goes through without modification.

Example 3.1

Consider the manufacturer's problem of Example 1.1, with the only difference that the times between the arrivals of successive orders are uniformly distributed in a given interval $[0,\tau_{\max}]$ instead of being exponentially distributed. Let F and NF denote the choices of filling and not filling the orders, respectively. The transition distributions are

$$Q_{ij}(\tau,F) = \begin{cases} \min\left[1,\frac{\tau}{\tau_{\max}}\right] & \text{if } j=1,\\ 0 & \text{otherwise,}\end{cases}$$

and

$$Q_{ij}(\tau,NF) = \begin{cases} \min\left[1,\frac{\tau}{\tau_{\max}}\right] & \text{if } j=i+1,\\ 0 & \text{otherwise.}\end{cases}$$

The effective cost per stage G of Eq. (3.3) is given by

$$G(i,F) = 0, \qquad G(i,NF) = \gamma ci,$$

where

$$\gamma = \int_0^{\tau_{\max}} \frac{1-e^{-\beta\tau}}{\beta\tau_{\max}}d\tau.$$

The scalars m_{ij} of Eq. (3.1) that are nonzero are

$$m_{i1}(F) = m_{i(i+1)}(NF) = \alpha,$$

where

$$\alpha = \int_0^{\tau_{\max}} \frac{e^{-\beta\tau}}{\tau_{\max}}d\tau = \frac{1-e^{-\beta\tau_{\max}}}{\beta\tau_{\max}}.$$

Bellman's equation has the form

$$J(i) = \min\big[K + \alpha J(1),\ \gamma ci + \alpha J(i+1)\big], \qquad i = 1,2,\ldots.$$

As in Example 1.1, we can conclude that there exists a threshold i^* such that it is optimal to fill the orders if and only if their number i exceeds i^*.

Example 3.2 (Control of an M/D/1 Queue)

Consider a single server queue where customers arrive according to a Poisson process with rate λ. The service time of a customer is deterministic and is equal to $1/\mu$ where μ is the service rate provided. The arrival and service rates λ and μ can be selected from given subsets Λ and M, and can be changed only when a customer departs from the system. There are costs $q(\lambda)$ and $r(\mu)$ per unit time for using rates λ and μ, respectively, and there is a waiting cost $c(i)$ per unit time when there are i customers in the system (waiting in queue or undergoing service). We wish to find a rate-setting policy that minimizes the total cost when there is a positive discount parameter β.

This problem bears similarity with Examples 2.1 and 2.2 of Section 5.2. Note, however, that while in those examples the rates can be changed both when a customer arrives and when a customer departs, here the rates can be changed only when a customer departs. Because the service time distribution is not exponential, it is necessary to make this restriction in order to be able to use as state the number of customers in the system; if we allowed the arrival rate to also change when a customer arrives, the time already spent in service by the customer found in service by the arriving customer would have to be part of the state.

The transition distributions are given by

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

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

where $p_{ij}(\lambda,\mu)$ are the state transition probabilities. It can be seen that for $i \ge 1$ and $j \ge i-1$, $p_{ij}(\lambda,\mu)$ can be calculated as the probability that $j-i+1$ arrivals will occur in an interval of length $[0,1/\mu]$. In particular, we have

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

Using the above formulas and Eqs. (3.3)-(3.4) and (3.6)-(3.7), one can write Bellman's equation and solve the problem as if it were essentially a discrete-time discounted problem.

Average Cost Problems

A natural cost function for the continuous-time average cost problem would be

$$\lim_{T\to\infty}\frac{1}{T}E\left\{\int_0^T g\big(x(t),u(t)\big)dt\right\}. \tag{3.11}$$

However, we will use instead the cost function

$$\lim_{N\to\infty}\frac{1}{E\{t_N\}}E\left\{\int_0^{t_N} g\big(x(t),u(t)\big)dt\right\}, \tag{3.12}$$

where t_N is the completion time of the Nth transition. This cost function is also reasonable and turns out to be analytically convenient. We note, however, that the cost functions (3.11) and (3.12) are equivalent under the conditions of the subsequent analysis, although a rigorous justification of this is beyond our scope (see [Ros70], p. 52 and p. 160 for related analysis).

We assume that there are n states, denoted $1,\ldots,n$, and that the control constraint set $U(i)$ is finite for each state i. For each pair (i,u), we denote by $G(i,u)$ the single stage expected cost corresponding to state i and control u. We have

$$G(i,u)=g(i,u)\overline{\tau}_i(u), \tag{3.13}$$

where $\overline{\tau}_i(u)$ is the expected value of the transition time corresponding to (i,u):

$$\overline{\tau}_i(u)=\sum_{j=1}^n\int_0^\infty \tau Q_{ij}(d\tau,u). \tag{3.14}$$

[If the cost per unit time g depends on the next state j, the expected transition cost $G(i,u)$ should be defined by

$$G(i,u)=\sum_{j=1}^n\int_0^\infty g(i,u,j)\,\tau\,Q_{ij}(d\tau,u),$$

and the following analysis and results go through without modification.] We assume throughout the remainder of this section that

$$0<\overline{\tau}_i(u)<\infty,\qquad i=1,\ldots,n,\ u\in U(i). \tag{3.15}$$

The cost function of an admissible policy $\pi=\{\mu_0,\mu_1,\ldots\}$ is given by

$$J_\pi(i)=\lim_{N\to\infty}\frac{1}{E\{t_N\mid x_0=i,\pi\}}E\left\{\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} g\big(x_k,\mu_k(x_k)\big)dt\;\Big|\;x_0=i\right\}.$$

Our earlier analysis of the discrete-time average cost problem in Chapter 4 suggests that under assumptions similar to those of Section 4.2, the cost $J_\mu(i)$ of a stationary policy μ, as well as the optimal average cost per stage $J^*(i)$, are independent of the initial state i. Indeed, we will see that the character of the solution of the problem is determined by the structure of the *embedded Markov chain*, which is the controlled discrete-time Markov chain whose transition probabilities are

$$p_{ij}(u)=\lim_{\tau\to\infty}Q_{ij}(\tau,u).$$

In particular, we will show that *$J_\mu(i)$ and $J^*(i)$ are independent of i if and only if the same is true for the embedded Markov chain problem.* For example, we will show that $J_\mu(i)$ and $J^*(i)$, are independent of i if all stationary policies μ are *unichain*; that is, the Markov chain with transition probabilities $p_{ij}\big(\mu(i)\big)$ has a single recurrent class.

We will also show that Bellman' s equation for average cost semi-Markov problems resembles the corresponding discrete-time equation, and takes the form

$$h(i)=\min_{u\in U(i)}\left[G(i,u)-\lambda\overline{\tau}_i(u)+\sum_{j=1}^n p_{ij}(u)h(j)\right]. \tag{3.16}$$

As a special case, when $\overline{\tau}_i(u)=1$ for all (i,u), we obtain the corresponding discrete-time equation of Chapter 4. We illustrate Bellman's equation (3.16) for the case of a single unichain policy with the stochastic shortest path argument that we used to prove Prop. 2.5 in Section 4.2.

Consider a unichain policy μ and without loss of generality assume that state n is a recurrent state in the Markov chain corresponding to μ. For each state $i\neq n$ let C_i and T_i be the expected cost and the expected time, respectively, up to reaching state n for the first time starting from i. Let also C_n and T_n be the expected cost and the expected time, respectively, up to returning to n for the first time starting from n. We can view C_i as the costs corresponding to μ in a stochastic shortest path problem where n is a termination state and the costs are $G\big(i,\mu(i)\big)$. Since μ is a proper policy for this problem, from Prop. 1.1 in Section 2.1, we have that the scalars C_i solve uniquely the system of equations

$$C_i=G\big(i,\mu(i)\big)+\sum_{j=1,\,j\neq n}^n p_{ij}\big(\mu(i)\big)C_j,\qquad i=1,\ldots,n. \tag{3.17}$$

Similarly, we can view T_i as the costs corresponding to μ in a stochastic shortest path problem where n is a termination state and the costs are $\overline{\tau}_i(\mu(i))$, so that the T_i solve uniquely the system of equations

$$T_i = \overline{\tau}_i(\mu(i)) + \sum_{j=1, j\neq n}^{n} p_{ij}(\mu(i))T_j, \qquad i = 1, \ldots, n. \tag{3.18}$$

Denote

$$\lambda_\mu = \frac{C_n}{T_n}. \tag{3.19}$$

Multiplying Eq. (3.18) by λ_μ and subtracting it from Eq. (3.17), we obtain for all $i = 1, \ldots, n$,

$$C_i - \lambda_\mu T_i = G(i, \mu(i)) - \lambda_\mu \overline{\tau}_i(\mu(i)) + \sum_{j=1, j\neq n}^{n} p_{ij}(\mu(i))(C_j - \lambda_\mu T_j).$$

By defining

$$h_\mu(i) = C_i - \lambda_\mu T_i, \qquad i = 1, \ldots, n, \tag{3.20}$$

and by noting that from Eq. (3.19) we have

$$h_\mu(n) = 0,$$

we obtain for all $i = 1, \ldots, n$,

$$h_\mu(i) = G(i, \mu(i)) - \lambda_\mu \overline{\tau}_i(\mu(i)) + \sum_{j=1}^{n} p_{ij}(\mu(i))h_\mu(j), \tag{3.21}$$

which is Bellman's equation (3.16) for the case of a single stationary policy μ.

We have not yet proved that the scalar λ_μ of Eq. (3.19) is the average cost per stage corresponding to μ. This fact will follow from the following proposition, which parallels Prop. 2.1 in Section 4.2 and shows that if Bellman's equation (3.16) has a solution (λ, h), then the optimal average cost is equal to λ and is independent of the initial state.

Proposition 3.1: If a scalar λ and an n-dimensional vector h satisfy

$$h(i) = \min_{u \in U(i)} \left[G(i,u) - \lambda \overline{\tau}_i(u) + \sum_{j=1}^{n} p_{ij}(u)h(j) \right], \qquad i = 1, \ldots, n, \tag{3.22}$$

then λ is the optimal average cost per stage $J^*(i)$ for all i,

$$\lambda = \min_{\pi} J_\pi(i) = J^*(i), \qquad i = 1, \ldots, n. \tag{3.23}$$

Furthermore, if $\mu^*(i)$ attains the minimum in Eq. (3.22) for each i, the stationary policy μ^* is optimal; that is, $J_{\mu^*}(i) = \lambda$ for all i.

Proof: For any μ consider the mapping $T_\mu : \Re^n \mapsto \Re^n$ given by

$$(T_\mu h)(i) = G(i, \mu(i)) + \sum_{j=1}^{n} p_{ij}(\mu(i))h(j), \qquad i = 1, \ldots, n,$$

and the vector $\overline{\tau}(\mu)$ and matrix P_μ given by

$$\overline{\tau}(\mu) = \begin{pmatrix} \overline{\tau}_1(\mu(1)) \\ \vdots \\ \overline{\tau}_n(\mu(n)) \end{pmatrix}, \qquad P_\mu = \begin{pmatrix} p_{11}(\mu(1)) \ldots p_{1n}(\mu(1)) \\ \cdots \\ p_{n1}(\mu(n)) \ldots p_{nn}(\mu(n)) \end{pmatrix}.$$

Let $\pi = \{\mu_0, \mu_1, \ldots\}$ be any admissible policy and N be a positive integer. We have from Eq. (3.22),

$$T_{\mu_{N-1}} h \geq \lambda \overline{\tau}(\mu_{N-1}) + h.$$

By applying $T_{\mu_{N-2}}$ to both sides of this relation, and by using the monotonicity of $T_{\mu_{N-2}}$ and Eq. (3.22), we see that

$$\begin{aligned} T_{\mu_{N-2}} T_{\mu_{N-1}} h &\geq T_{\mu_{N-2}} (\lambda \overline{\tau}(\mu_{N-1}) + h) \\ &= \lambda P_{\mu_{N-2}} \overline{\tau}(\mu_{N-1}) + T_{\mu_{N-2}} h \\ &\geq \lambda P_{\mu_{N-2}} \overline{\tau}(\mu_{N-1}) + \lambda \overline{\tau}(\mu_{N-2}) + h. \end{aligned}$$

Continuing in the same manner, we finally obtain

$$T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} h \geq \lambda \overline{t}_N(\pi) + h, \tag{3.24}$$

where $\overline{t}_N(\pi)$ is given by

$$\begin{aligned} \overline{t}_N(\pi) = {} & P_{\mu_0} \cdots P_{\mu_{N-2}} \overline{\tau}(\mu_{N-1}) \\ & + P_{\mu_0} \cdots P_{\mu_{N-3}} \overline{\tau}(\mu_{N-2}) + \cdots + \overline{\tau}(\mu_0). \end{aligned}$$

Note that the ith component of the vector $\overline{t}_N(\pi)$ is $E\{t_N \mid x_0 = i, \pi\}$, the expected value of the completion time of the Nth transition when the initial state is i and π is used. Note also that equality holds in Eq. (3.24) if $\mu_k(i)$ attains the minimum in Eq. (3.22) for all k and i. It can be seen that

$$(T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} h)(i) = E\left\{ h(x_N) + \int_0^{t_N} g(x(t), u(t))dt \;\middle|\; x_0 = i, \pi \right\}.$$

Using this relation in Eq. (3.24) and dividing by $E\{t_N \mid x_0 = i, \pi\}$, we obtain for all i

$$\begin{aligned} \frac{E\{h(x_N) \mid x_0 = i, \pi\}}{E\{t_N \mid x_0 = i, \pi\}} + \frac{E\left\{\int_0^{t_N} g(x(t), u(t))dt \;\middle|\; x_0 = i, \pi\right\}}{E\{t_N \mid x_0 = i, \pi\}} \\ \geq \lambda + \frac{h(i)}{E\{t_N \mid x_0 = i, \pi\}}. \end{aligned}$$

Taking the limit as $N \to \infty$ and using the fact $\lim_{N\to\infty} E\{t_N \mid x_0 = i, \pi\} = \infty$ [cf. Eq. (3.15)], we see that

$$\lim_{N\to\infty} \frac{E\left\{\int_0^{t_N} g\big(x(t),u(t)\big)dt \;\middle|\; x_0 = i, \pi\right\}}{E\{t_N \mid x_0 = i, \pi\}} = J_\pi(i) \geq \lambda, \qquad i = 1, \ldots, n,$$

with equality if $\mu_k(i)$ attains the minimum in Eq. (3.22) for all k and i. **Q.E.D.**

By combining Prop. 3.1 with Eq. (3.21), we obtain the following:

Proposition 3.2: Let μ be a unichain policy. Then:

(a) There exists a scalar λ_μ and a vector h_μ such that

$$J_\mu(i) = \lambda_\mu, \qquad i = 1, \ldots, n, \tag{3.25}$$

and

$$h_\mu(i) = G\big(i,\mu(i)\big) - \lambda_\mu \overline{\tau}_i\big(\mu(i)\big) + \sum_{j=1}^{n} p_{ij}\big(\mu(i)\big) h_\mu(j), \quad i = 1, \ldots, n. \tag{3.26}$$

(b) Let t be a fixed state. The system of the $n+1$ linear equations

$$h(i) = G\big(i,\mu(i)\big) - \lambda_\mu \overline{\tau}_i\big(\mu(i)\big) + \sum_{j=1}^{n} p_{ij}\big(\mu(i)\big) h(j), \quad i = 1, \ldots, n, \tag{3.27}$$

$$h(t) = 0, \tag{3.28}$$

in the $n+1$ unknowns $\lambda, h(1), \ldots, h(n)$ has a unique solution.

Proof: Part (a) follows from Prop. 3.1 and Eq. (3.21). The proof of part (b) is identical to the proof of Prop. 2.5(b) in Section 4.2. **Q.E.D.**

To establish conditions under which there exists a solution (λ, h) to Bellman's equation (3.22), we formulate a corresponding discrete-time average cost problem. Let γ be any scalar such that

$$0 < \gamma < \frac{\overline{\tau}_i(u)}{1 - p_{ii}(u)}$$

for all i and $u \in U(i)$ with $p_{ii}(u) < 1$. Define also for all i and $u \in U(i)$,

$$\tilde{p}_{ij}(u) = \begin{cases} \dfrac{\gamma p_{ij}(u)}{\overline{\tau}_i(u)} & \text{if } j \neq i, \\ 1 - \dfrac{\gamma\big(1 - p_{ii}(u)\big)}{\overline{\tau}_i(u)} & \text{if } j = i. \end{cases} \tag{3.29}$$

It can be seen that we have for all i and j

$$0 \leq \tilde{p}_{ij}(u), \qquad \sum_{j=1}^{n} \tilde{p}_{ij}(u) = 1,$$

$$\tilde{p}_{ij}(u) > 0 \quad \text{if and only if} \quad p_{ij}(u) > 0. \tag{3.30}$$

We view $\tilde{p}_{ij}(u)$ as the transition probabilities of the discrete-time average cost problem whose expected stage cost corresponding to (i,u) is

$$\tilde{G}(i,u) = \frac{G(i,u)}{\overline{\tau}_i(u)}. \tag{3.31}$$

We call this the *auxiliary discrete-time average cost problem*. The following proposition shows that this problem is essentially equivalent with our original semi-Markov average cost problem.

Proposition 3.3 If the scalar λ and the vector $\tilde{h}$ satisfy

$$\tilde{h}(i) = \min_{u\in U(i)} \left[\tilde{G}(i,u) - \lambda + \sum_{j=1}^{n} \tilde{p}_{ij}(u)\tilde{h}(j)\right], \qquad i = 1, \ldots, n, \tag{3.32}$$

then λ and the vector h with components

$$h(i) = \gamma \tilde{h}(i), \qquad i = 1, \ldots, n, \tag{3.33}$$

satisfy Bellman's equation

$$h(i) = \min_{u\in U(i)} \left[G(i,u) - \lambda \overline{\tau}_i(u) + \sum_{j=1}^{n} p_{ij}(u) h(j)\right], \qquad i = 1, \ldots, n, \tag{3.34}$$

for the semi-Markov average cost problem.

Proof: By substituting Eqs. (3.29), (3.31), and (3.33) in Eq. (3.32), we obtain after a straightforward calculation

$$0 = \min_{u\in U(i)} \frac{1}{\overline{\tau}_i(u)} \left[G(i,u) - \lambda \overline{\tau}_i(u) + \sum_{j=1}^{n} p_{ij}(u) h(j) - h(i)\right], \quad i = 1, \ldots, n.$$

This implies that the minimum of the expression within brackets in the right-hand side above is zero, which is equivalent to Bellman's equation (3.34). **Q.E.D.**

Note that in view of Eq. (3.30), the auxiliary discrete-time average cost problem and the semi-Markov average cost problem have the same probabilistic structure. In particular, if all stationary policies are unichain for one problem, the same is true for the other. Thus, the results and algorithms of Sections 4.2 and 4.3, when applied to the auxiliary discrete-time problem, yield results and algorithms for the semi-Markov problem. For example, value iteration, policy iteration, and linear programming can be applied to the auxiliary problem in order to solve the semi-Markov problem. We state a partial analog of Prop. 2.6 from Section 4.2.

Proposition 3.4: Consider the semi-Markov average cost problem, and assume either one of the following two conditions:

(1) Every policy that is optimal within the class of stationary policies is unichain.

(2) For every two states i and j, there exists a stationary policy π (depending on i and j) such that, for some k,

$$P(x_k = j \mid x_0 = i, \pi) > 0.$$

Then the optimal average cost per stage has the same value λ for all initial states i. Furthermore, λ together with a vector h satisfies Bellman's equation (3.34) for the semi-Markov average cost problem.

Proof: By Prop. 2.6 in Section 4.2, under either one of the conditions stated, Bellman's equation (3.32) for the auxiliary discrete-time average cost problem has a solution $(\lambda, \tilde{h})$, from which a solution to Bellman's equation (3.34) can be extracted according to Prop. 3.3. **Q.E.D.**

Example 3.3:

Consider the average cost version of the manufacturer's problem of Example 3.1. Here we have

$$\overline{\tau}_i(F) = \overline{\tau}_i(NF) = \frac{\overline{\tau}_{\max}}{2},$$

$$G(i, F) = K, \qquad G(i, NF) = \frac{ci\overline{\tau}_{\max}}{2},$$

where F and NF denote the decisions to fill and not fill the orders, respectively. Bellman's equation takes the form

$$h(i) = \min\left[K - \lambda\frac{\overline{\tau}_{\max}}{2} + h(1),\ ci\frac{\overline{\tau}_{\max}}{2} - \lambda\frac{\overline{\tau}_{\max}}{2} + h(i+1)\right].$$

We leave it as an exercise for the reader to show that there exists a threshold i^* such that it is optimal to fill the orders if and only if i exceeds i^*.

Example 3.4: [LiR71]

Consider a person providing a certain type of service to customers. Potential customers arrive according to a Poisson process with rate r; that is the customer's interarrival times are independent and exponentially distributed with parameter r. Each customer offers one of n pairs (m_i, T_i), $i = 1, \ldots, n$, where m_i is the amount of money offered for the service and T_i is the average amount of time that will be required to perform the service. Successive offers are independent and offer (m_i, T_i) occurs with probability p_i, where $\sum_{i=1}^{n} p_i = 1$. An offer may be rejected, in which case the customer leaves, or may be accepted in which case all offers that arrive while the customer is being served are lost. The problem is to determine the acceptance-rejection policy that maximizes the service provider's average income per unit time.

Let us denote by i the state corresponding to the offer (m_i, T_i), and let A and R denote the accept and reject decision, respectively. We have

$$\overline{\tau}_i(A) = T_i + \frac{1}{r}, \qquad \overline{\tau}_i(R) = \frac{1}{r},$$

$$G(i, A) = -m_i, \qquad G(i, R) = 0,$$

$$p_{ij}(A) = p_{ij}(R) = p_j.$$

Bellman's equation is given by

$$h(i) = \min\left[-m_i - \lambda\left(T_i + \frac{1}{r}\right) + \sum_{j=1}^{n} p_j h(j),\ -\lambda\frac{1}{r} + \sum_{j=1}^{n} p_j h(j)\right].$$

It follows that an optimal policy is to accept offer (i, T_i) if

$$\frac{m_i}{T_i} \geq -\lambda,$$

where $-\lambda$ is the optimal average income per unit time.

5.4 NOTES, SOURCES, AND EXERCISES

The idea of using uniformization to convert continuous-time stochastic control problems involving Markov chains into discrete-time problems gained wide attention following [Lip75]; see also [BeR87].

Control of queueing systems has been researched extensively. For additional material on the problem of control of arrival rate or service rate (cf. Examples 2.1 and 2.2 in Section 5.2), see [BWN92], [CoR87], [CoV84], [RVW82], [Sob82], [StP74], and [Sti85]. For more on priority assignment and routing (cf. Examples 2.3, 2.4 in Section 5.2), see [BDM83], [BaD81], [BeT89b], [BhE91], [CoV84], [Har75a], [Har75b], [PaK81], [SuC91], and [AyR91], [CrC91], [EVW80], [EpV89], [Haj84], [LiK84], [TSC92], [ViE88], respectively.

Semi-Markov decision models were introduced in [Jew63] and are also discussed in [Ros70].

EXERCISES

5.1 (Proof of Validity of Uniformization)

Complete the details of the following argument, showing the validity of the uniformization procedure for the case of a finite number of states $i = 1, \ldots, n$. 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 coordinates

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

We have

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

where $p(0)$ is the row vector with ith coordinate equal to one if $x_0 = i$ and zero otherwise, and the matrix A has elements

$$a_{ij} = \begin{cases} \nu_i p_{ij} & \text{if } i \neq j, \\ -\nu_i & \text{if } i = j. \end{cases}$$

From this we obtain

$$p(t) = p(0)e^{At},$$

where

$$e^{At} = \sum_{k=0}^{\infty} \frac{(At)^k}{k!}.$$

Consider the transition probability matrix B of the uniform version

$$B = I + \frac{A}{\nu},$$

where $\nu \geq \nu_i$, $i = 1, \ldots, n$. 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

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

5.2

Consider the $M/M/1$ queueing problem with variable service rate (Example 2.1 in Section 5.2). Assume that no arrivals are allowed ($\lambda = 0$), and one can either serve a customer at rate μ or refuse service ($M = \{0, \mu\}$). Let the cost rates for customer waiting and service be $c(i) = ci$ and $q(\mu)$, respectively, with $q(0) = 0$.

(a) Show that an optimal policy is to always serve an available customer if

$$\frac{q(\mu)}{\mu} \leq \frac{c}{\beta},$$

and to always refuse service otherwise.

(b) Analyze the problem when the cost rate for waiting is instead $c(i) = ci^2$.

5.3

A person has an asset to sell for which she receives offers that can take one of n values. The times between successive offers are random, independent, and identically distributed with given distribution. Find the offer acceptance policy that maximizes $E\{\alpha^T s\}$, where T is the time of sale, s is the sale price, and $\alpha \in (0, 1)$ is a discount factor.

5.4

Analyze the priority assignment problem of Example 2.3 in Section 5.2 within the semi-Markov context of Section 5.3, assuming that the customer service times are independent but not exponentially distributed. Consider both the discounted and the average cost cases.

5.5

An unemployed worker receives job offers according to a Poisson process with rate r, which she may accept or reject. The offered salary (per unit time) takes one of n possible values $w_1, \ldots, w_n$ with given probabilities, independently of preceding offers. If she accepts an offer at salary w_i, she keeps the job for a random amount of time that has expected value t_i. If she rejects the offer, she receives unemployment compensation c (per unit time) and is eligible to accept future offers. Solve the problem of maximizing the worker's average income per unit time.

5.6

Consider a single server queueing system where the server may be either on or off. Customers arrive according to a Poisson process with rate λ, and their service times are independent, identically distributed with given distribution. Each time a customer departs, the server may switch from on to off at a fixed cost C_0 or from off to on at a fixed cost C_1. There is a cost c per unit time and customer residing in the system. Analyze this problem as a semi-Markov problem for the discounted and the average cost cases. In the latter case, assume that the queue has limited storage, and that customers arriving when the queue is full are lost.

5.7

Consider a semi-Markov version of the machine replacement problem of Example 2.1 in Section 1.2. Here, the transition times are random, independent, and have given distributions. Also $g(i)$ is the cost per unit time of operating the machine at state i. Assume that $p_{i(i+1)} > 0$ for all $i < n$. Derive Bellman's equation and analyze the problem.

References

[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.

[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.

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

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

[BBS93] Barto, A. G., Bradtke, S. J., and Singh, S. P., 1993. "Real-Time Learning and Control Using Asynchronous Dynamic Programming," Comp. Science Dept. Tech. Report 91-57, Univ. of Massachusetts, Artificial Intelligence, Vol. 72, 1995, 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.

[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, IEEE Trans. on Automatic Control, to appear.

[BWN92] Blanc, J. P. C., de Waal, P. R., Nain, P., and Towsley, D., 1992. "Optimal Control of Admission to a Multiserver Queue with Two Arrival Streams," IEEE Trans. on Automatic Control, Vol. 37, pp. 785-797.

[BaD81] Baras, J. S., and Dorsey, A. J., 1981. "Stochastic Control of Two Partially Observed Competing Queues," IEEE Trans. Automatic Control, Vol. AC-26, pp. 1106-1117.

[Bai93] Baird, L. C., 1993. "Advantage Updating," Report WL-TR-93-1146, Wright Patterson AFB, OH.

[Bai94] Baird, L. C., 1994. "Reinforcement Learning in Continuous Time: Advantage Updating," International Conf. on Neural Networks, Orlando, Fla.

[Bai95] Baird, L. C., 1995. "Residual Algorithms: Reinforcement Learning with Function Approximation," Dept. of Computer Science Report, U.S. Air Force Academy, CO.

[Bat73] Bather, J., 1973. "Optimal Decision Procedures for Finite Markov Chains," Advances in Appl. Probability, Vol. 5, pp. 328-339, pp. 521-540, 541-553.

[BeC89] Bertsekas, D. P., and Castanon, D. A., 1989. "Adaptive Aggregation Methods for Infinite Horizon Dynamic Programming," IEEE Trans. on Automatic Control, Vol. AC-34, pp. 589-598.

[BeN93] Bertsimas, D., and Nino-Mora, J., 1993. "Conservation Laws, Extended Polymatroids, and the Multiarmed Bandit Problem: A Unified Polyhedral Approach," Mathematics of Operations Research, to appear.

[BeR87] Beutler, F. J., and Ross, K. W., 1987. "Uniformization for Semi-Markov Decision Processes Under Stationary Policies," J. Appl. Prob., Vol. 24, pp. 399-420.

[BeS78] Bertsekas, D. P., and Shreve, S. E., 1978. Stochastic Optimal Control: The Discrete Time Case, Academic Press, N. Y.

[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.

[BeT89a] Bertsekas, D. P., and Tsitsiklis, J. N., 1989. Parallel and Distributed Computation: Numerical Methods, Prentice-Hall, Englewood Cliffs, N. J.

[BeT89b] Beutler, F. J., and Teneketzis, D., 1989. "Routing in Queueing Networks Under Imperfect Information: Stochastic Dominance and Thresholds," Stochastics and Stochastics Reports, Vol. 26, pp. 81-100.

[BeT91a] Bertsekas, D. P., and Tsitsiklis, J. N., 1991. "A Survey of Some Aspects of Parallel and Distributed Iterative Algorithms," Automatica, Vol. 27, pp. 3-21.

[BeT91b] Bertsekas, D. P., and Tsitsiklis, J. N., 1991. "An Analysis of Stochastic Shortest Path Problems," Math. Operations Research, Vol. 16, pp. 580-595.

[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. Dissertation, Massachusetts Institute of Technology.

[Ber72] Bertsekas, D. P., 1972. "Infinite Time Reachability of State Space Regions by Using Feedback Control," IEEE Trans. Automatic Control, Vol. AC-17, pp. 604-613.

[Ber73a] Bertsekas, D. P., 1973. "Stochastic Optimization Problems with 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. Automatic Control, Vol. AC-18, pp. 314-315.

[Ber75] Bertsekas, D. P., 1975. "Convergence of Discretization Procedures in Dynamic Programming," IEEE Trans. Automatic Control, Vol. AC-20, pp. 415-419.

[Ber76] Bertsekas, D. P., 1976. "On Error Bounds for Successive Approximation Methods," IEEE Trans. Automatic Control, Vol. AC-21, pp. 394-396.

[Ber77] Bertsekas, D. P., 1977. "Monotone Mappings with Application in Dynamic Programming," SIAM J. on Control and Optimization, Vol. 15, pp. 438-464.

[Ber82a] Bertsekas, D. P., 1982. "Distributed Dynamic Programming," IEEE Trans. Automatic Control, Vol. AC-27, pp. 610-616.

[Ber82b] Bertsekas, D. P., 1982. Constrained Optimization and Lagrange Multiplier Methods, Academic Press, N. Y.

[Ber93] Bertsekas, D. P., 1993. "A Generic Rank One Correction Algorithm for Markovian Decision Problems," Lab. for Info. and Decision Systems Report LIDS-P-2212, Massachusetts Institute of Technology, Operations Research Letters, to appear.

[Ber95a] Bertsekas, D. P., 1995. Nonlinear Programming, Athena Scientific, Belmont, MA, to appear.

[Ber95b] Bertsekas, D. P., 1995. "A Counterexample to Temporal Differences Learning," Neural Computation, Vol. 7, pp. 270-279.

[Ber95c] Bertsekas, D. P., 1995. "A New Value Iteration Method for the Average Cost Dynamic Programming Problem," Lab. for Info. and Decision Systems Report, Massachusetts Institute of Technology.

[BhE91] Bhattacharya, P. P., and Ephremides, A., 1991. "Optimal Allocations of a Server Between Two Queues with Due Times," IEEE Trans. on Automatic Control, Vol. 36, pp. 1417-1423.

[Bil83] Billingsley, P., 1983. "The Singular Function of Bold Play," American Scientist, Vol. 71, pp. 392-397.

[Bla62] Blackwell, D., 1962. "Discrete Dynamic Programming," Ann. Math. Statist., Vol. 33, pp. 719-726.

[Bla65] Blackwell, D., 1965. "Discounted Dynamic Programming," Ann. Math. Statist., Vol. 36, pp. 226-235.

[Bla70] Blackwell, D., 1970. "On Stationary Policies," J. Roy. Statist. Soc. Ser. A, Vol. 133, pp. 33-38.

[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.

[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.

[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.

[Cav89a] Cavazos-Cadena, R., 1989. "Necessary Conditions for the Optimality Equations in Average-Reward Markov Decision Processes," Sys. Control Letters, Vol. 11, pp. 65-71.

[Cav89b] Cavazos-Cadena, R., 1989. "Weak Conditions for the Existence of Optimal Stationary Policies in Average Markov Decisions Chains with Unbounded Costs," Kybernetika, Vol. 25, pp. 145-156.

[Cav91] Cavazos-Cadena, R., 1991. "Recent Results on Conditions for the Existence of Average Optimal Stationary Policies," Annals of Operations Research, Vol. 28, pp. 3-28.

[ChT89] Chow, C.-S., and Tsitsiklis, J. N., 1989. "The Complexity of Dynamic Programming," Journal of Complexity, Vol. 5, pp. 466-488.

[ChT91] Chow, C.-S., and Tsitsiklis, J. N., 1991. "An Optimal One Way Multigrid Algorithm for Discrete Time Stochastic Control," IEEE Trans. on Automatic Control, Vol. AC-36, pp. 898-914.

[CoR87] Courcoubetis, C. A., and Reiman, M. I., 1987. "Optimal Control of a Queueing System with Simultaneous Service Requirements," IEEE Trans. on Automatic Control, Vol. AC-32, pp. 717-727.

[CoV84] Courcoubetis, C., and Varaiya, P. P., 1984. "The Service Process with Least Thinking Time Maximizes Resource Utilization," IEEE Trans. Automatic Control, Vol. AC-29, pp. 1005-1008.

[CrC91] Cruz, R. L., and Chuah, M. C., 1991. "A Minimax Approach to a Simple Routing Problem," IEEE Trans. on Automatic Control, Vol. 36, pp. 1424-1435.

[D'Ep60] D'Epenoux, F., 1960. "Sur un Probleme de Production et de Stockage Dans l'Aleatoire," Rev. Francaise Aut. Infor. Recherche Operationnelle, Vol. 14, (English Transl.: Management Sci., Vol. 10, 1963, pp. 98-108).

[Dan63] Dantzig, G. B., 1963. Linear Programming and Extensions. Princeton Univ. Press, Princeton, N. J.

[Den67] Denardo, E. V., 1967. "Contraction Mappings in the Theory Underlying Dynamic Programming," SIAM Review, Vol. 9, pp. 165-177.

[Der70] Derman, C., 1970. Finite State Markovian Decision Processes. Academic Press, N. Y.

[DuS65] Dubins, L., and Savage, L. M., 1965. How to Gamble If You Must, McGraw-Hill, N. Y.

[DyY79] Dynkin, E. B., and Yuskevich, A. A., 1979. Controlled Markov Processes, Springer-Verlag, N. Y.

[EVW80] Ephremides, A., Varaiya, P. P., and Walrand, J. C., 1980. "A Simple Dynamic Routing Problem," IEEE Trans. Automatic Control, Vol. AC-25, pp. 690-693.

[EaZ62] Eaton, J. H., and Zadeh, L. A., 1962. "Optimal Pursuit Strategies in Discrete State Probabilistic Systems," Trans. ASME Ser. D. J. Basic Eng., Vol. 84, pp.23-29.

[EpV89] Ephremides, A., and Verdu, S., 1989. "Control and Optimization Methods in Communication Network Problems," IEEE Trans. Automatic Control, Vol. AC-34, pp. 930-942.

[FAM90] Fernández-Gaucherand, E., Arapostathis, A., and Marcus, S. I., 1990. "Remarks on the Existence of Solutions to the Average Cost Optimality Equation in Markov Decision Processes," Systems and Control Letters, Vol. 15, pp. 425-432.

[FAM91] Fernández-Gaucherand, E., Arapostathis, A., and Marcus, S. I., 1991. "On the Average Cost Optimality Equation and the Structure of Optimal Policies for Partially Observable Markov Decision Processes," Annals of Operations Research, Vol. 29, pp. 439-470.

[FeS94] Feinberg, E. A., and Shwartz, A., 1994. "Markov Decision Models

with Weighted Discounted Criteria," Mathematics of Operations Research, Vol. 19, pp. 1-17.

[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," Contemporary Mathematics, Vol. 125, pp. 87-96.

[Fox71] Fox, B. L., 1971. "Finite State Approximations to Denumerable State Dynamic Progams," J. Math. Anal. Appl., Vol. 34, pp. 665-670.

[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.

[Git79] Gittins, J. C., 1979. "Bandit Processes and Dynamic Allocation Indices," J. Roy. Statist. Soc., Vol. B, No. 41, pp. 148-164.

[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.

[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.

[HaL86] Haurie, A., and L'Ecuyer, P., 1986. "Approximation and Bounds in Discrete Event Dynamic Programming," IEEE Trans. Automatic Control, Vol. AC-31, pp. 227-235.

[Haj84] Hajek, B., 1984. "Optimal Control of Two Interacting Service Stations," IEEE Trans. Automatic Control, Vol. AC-29, pp. 491-499.

[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.

[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.

[HoT74] Hordijk, A., and Tijms, H., 1974. "Convergence Results and Approximations for Optimal (s, S) Policies," Management Sci., Vol. 20, pp. 1431-1438.

[How60] Howard, R., 1960. Dynamic Programming and Markov Processes, MIT Press, Cambridge, MA.

[Igl63a] Iglehart, D. L., 1963. "Optimality of (S, s) Policies in the Infinite Horizon Dynamic Inventory Problem," Management Sci., Vol. 9, pp. 259-267.

[Igl63b] Iglehart, D. L., 1963. "Dynamic Programming and Stationary Analysis of Inventory Problems," in Scarf, H., Gilford, D., and Shelly, M., (eds.), Multistage Inventory Models and Techniques, Stanford University Press, Stanford, CA, 1963.

[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.

[Jew63] Jewell, W., 1963. "Markov Renewal Programming I and II," Operations Research, Vol. 2, pp. 938-971.

[KaV87] Katehakis, M., and Veinott, A. F., 1987. "The Multi-Armed Bandit Problem: Decomposition and Computation," Math. of Operations Research, Vol. 12, pp. 262-268.

[Kal83] Kallenberg, L. C. M., 1983. Linear Programming and Finite Markov Control Problems, Mathematical Centre Report, Amsterdam.

[Kel81] Kelly, F. P., "Multi-Armed Bandits with Discount Factor Near One: The Bernoulli Case," The Annals of Statistics, Vol. 9, pp. 987-1001.

[Kle68] Kleinman, D. L., 1968. "On an Iterative Technique for Riccati Equation Computations," IEEE Trans. Automatic Control, Vol. AC-13, pp. 114-115.

[KuV86] Kumar, P. R., and Varaiya, P. P., 1986. Stochastic Systems: Estimation, Identification, and Adaptive Control, Prentice-Hall, Englewood Cliffs, N. J.

[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.

[Kus71] Kushner, H. J., 1971. Introduction to Stochastic Control, Holt, Rinehart and Winston, N. Y.

[Kus78] Kushner, H. J., 1978. "Optimality Conditions for the Average Cost per Unit Time Problem with a Diffusion Model," SIAM J. Control Opti-

mization, Vol. 16, pp. 330-346.

[Las88] Lasserre, J. B., 1988. Conditions for Existence of Average and Blackwell Optimal Stationary Policies in Denumerable Markov Decision Processes," J. Math. Anal. Appl., Vol. 136, pp. 479-490.

[LiK84] Lin, W., and Kumar, P. R., 1984. "Optimal Control of a Queueing System with Two Heterogeneous Servers," IEEE Trans. Automatic Control, Vol. AC-29, pp. 696-703.

[LiR71] Lippman, S. A., and Ross, S. M., 1971. "The Streetwalker's Dilemma: A Job-Shop Model," SIAM J. of Appl. Math., Vol. 20, pp. 336-342.

[LiS61] Liusternik, L., and Sobolev, V., 1961. Elements of Functional Analysis, Ungar, N. Y.

[Lip75] Lippman, S. A., 1975. "Applying a New Device in the Optimization of Exponential Queuing Systems," Operations Research, Vol. 23, pp. 687-710.

[LjS83] Ljung, L., and Soderstrom, T., 1983. Theory and Practice of Recursive Identification, MIT Press, Cambridge, MA.

[Lue69] Luenberger, D. G., 1969. Optimization by Vector Space Methods, Wiley, N. Y.

[McQ66] MacQueen, J., 1966. "A Modified Dynamic Programming Method for Markovian Decision Problems," J. Math. Anal. Appl., Vol. 14, pp. 38-43.

[MoW77] Morton, T. E., and Wecker, W., 1977. "Discounting, Ergodicity and Convergence for Markov Decision Processes," Management Sci., Vol. 23, pp. 890-900.

[Mor71] Morton, T. E., 1971. "On the Asymptotic Convergence Rate of Cost Differences for Markovian Decision Processes," Operations Research, Vol. 19, pp. 244-248.

[NTW89] Nain, P., Tsoucas, P., and Walrand, J., 1989. "Interchange Arguments in Stochastic Scheduling," J. of Appl. Prob., Vol. 27, pp. 815-826.

[NgP86] Nguyen, S., and Pallottino, S., 1986. "Hyperpaths and Shortest Hyperpaths," in Combinatorial Optimization by B. Simeone (ed.), Springer-Verlag, N. Y, pp. 258-271.

[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.

[Orn69] Ornstein, D., 1969. "On the Existence of Stationary Optimal Strategies," Proc. Amer. Math. Soc., Vol. 20, pp. 563-569.

[PBT95] Polymenakos, L. C., Bertsekas, D. P., and Tsitsiklis, J. N., 1995. "Efficient Algorithms for Continuous-Space Shortest Path Problems," Lab. for Info. and Decision Systems Report LIDS-P-2292, Massachusetts Institute of Technology.

[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.

[PaK81] Pattipati, K. R., and Kleinman, D. L., 1981. "Priority Assignment Using Dynamic Programming for a Class of Queueing Systems," IEEE Trans. on Automatic Control, Vol. AC-26, pp. 1095-1106.

[PaT87] Papadimitriou, C. H., and Tsitsiklis, J. N., 1987. "The Complexity of Markov Decision Processes," Math. Operations Research, Vol. 12, pp. 441-450.

[Pla77] Platzman, L., 1977. "Improved Conditions for Convergence in Undiscounted Markov Renewal Programming," Operations Research, Vol. 25, pp. 529-533.

[PoA69] Pollatschek, M., and Avi-Itzhak, B., 1969. "Algorithms for Stochastic Games with Geometrical Interpretation," Man. Sci., Vol. 15, pp. 399-413.

[PoT78] Porteus, E., and Totten, J., 1978. "Accelerated Computation of the Expected Discounted Return in a Markov Chain," Operations Research, Vol. 26, pp. 350-358.

[PoT92] Polychronopoulos, G. H., and Tsitsiklis, J. N., 1992. "Stochastic Shortest Path Problems with Recourse," Lab. for Info. and Decision Systems Report LIDS-P-P-2183, Massachusetts Institute of Technology.

[Por71] Porteus, E., 1971. "Some Bounds for Discounted Sequential Decision Processes," Man. Sci., Vol. 18, pp. 7-11.

[Por75] Porteus, E., 1975. "Bounds and Transformations for Finite Markov Decision Chains," Operations Research, Vol. 23, pp. 761-784.

[Por81] Porteus, E., 1981. "Improved Conditions for Convergence in Undiscounted Markov Renewal Programming," Operations Research, Vol. 25, pp. 529-533.

[PsT93] Psaraftis, H. N., and Tsitsiklis, J. N., 1993. "Dynamic Shortest Paths in Acyclic Networks with Markovian Arc Costs," Operations Research, Vol. 41, pp. 91-101.

[PuB78] Puterman, M. L., and Brumelle, S. L., 1978. "The Analytic Theory of Policy Iteration," in Dynamic Programming and Its Applications, M. L. Puterman (ed.), Academic Press, N. Y.

[PuS78] Puterman, M. L., and Shin, M. C., 1978. "Modified Policy Iteration

Algorithms for Discounted Markov Decision Problems," Management Sci., Vol. 24, pp. 1127-1137.

[PuS82] Puterman, M. L., and Shin, M. C., 1982. "Action Elimination Procedures for Modified Policy Iteration Algorithms," Operations Research, Vol. 30, pp. 301-318.

[Put78] Puterman, M. L. (ed.), 1978. Dynamic Programming and its Applications, Academic Press, N. Y.

[Put94] Puterman, M. L., 1994. "Markovian Decision Problems," J. Wiley, N. Y.

[RVW82] Rosberg, Z., Varaiya, P. P., and Walrand, J. C., 1982. "Optimal Control of Service in Tandem Queues," IEEE Trans. Automatic Control, Vol. AC-27, pp. 600-609.

[RaF91] Raghavan, T. E. S., and Filar, J. A., 1991. "Algorithms for Stochastic Games A Survey," ZOR Methods and Models of Operations Research, Vol. 35, pp. 437-472.

[RiS92] Ritt, R. K., and 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.

[Roc70] Rockafellar, R. T., 1970. Convex Analysis, Princeton University Press, Princeton, N. J.

[Ros70] Ross, S. M., 1970. Applied Probability Models with Optimization Applications, Holden-Day, San Francisco, CA.

[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.

[Rus94] Rust, J., 1994. "Using Randomization to Break the Curse of Dimensionality," Unpublished Report, Dept. of Economics, University of Wisconsin.

[Rus95] Rust, J., 1995. "Numerical Dynamic Programming in Economics," in Handbook of Computational Economics, H. Amman, D. Kendrick, and J. Rust (eds.).

[SPK85] Schweitzer, P. J., Puterman, M. L., and Kindle, K. W., 1985. "Iterative Aggregation-Disaggregation Procedures for Solving Discounted Semi-Markovian Reward Processes," Operations Research, Vol. 33, pp. 589-605.

[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.

[Sch93] Schwartz, A., 1993. "A Reinforcement Learning Method for Maximizing Undiscounted Rewards," Proc. of the Tenth Machine Learning Conference.

[Sen86] Sennott, L. I., 1986. "A New Condition for the Existence of Optimum Stationary Policies in Average Cost Markov Decision Processes," Operations Research Lett., Vol. 5, pp. 17-23.

[Sen89a] Sennott, L. I., 1989. "Average Cost Optimal Stationary Policies in Infinite State Markov Decision Processes with Unbounded Costs," Operations Research, Vol. 37, pp. 626-633.

[Sen89b] Sennott, L. I., 1989. "Average Cost Semi-Markov Decision Processes and the Control of Queueing Systems," Prob. Eng. Info. Sci., Vol. 3, pp. 247-272.

[Sen91] Sennott, L. I., 1991. "Value Iteration in Countable State Average Cost Markov Decision Processes with Unbounded Cost," Annals of Operations Research, Vol. 28, pp. 261-272.

[Sen93a] Sennott, L. I., 1993. "The Average Cost Optimality Equation and Critical Number Policies," Prob. Eng. Info. Sci., Vol. 7, pp. 47-67.

[Sen93b] Sennott, L. I., 1993. "Constrained Average Cost Markov Decision Chains," Prob. Eng. Info. Sci., Vol. 7, pp. 69-83.

[Ser79] Serfozo, R., 1979. "An Equivalence Between Discrete and Continuous Time Markov Decision Processes," Operations Research, Vol. 27, pp. 616-620.

[Sha53] Shapley, L. S., 1953. "Stochastic Games," Proc. Nat. Acad. Sci. U.S.A., Vol. 39.

[Sin94] Singh, S. P., 1994. "Reinforcement Learning Algorithms for Average-Payoff Markovian Decision Processes," Proc. of 12th National Conference on Artificial Intelligence, pp. 202-207.

[Sob82] Sobel, M. J., 1982. "The Optimality of Full-Service Policies," Operations Research, Vol. 30, pp. 636-649.

[StP74] Stidham, S., and Prabhu, N. U., 1974. "Optimal Control of Queueing Systems," in Mathematical Methods in Queueing Theory (Lecture Notes in Economics and Math. Syst., Vol. 98), A. B. Clarke (Ed.), Springer-Verlag, N. Y., pp. 263-294.

[Sti85] Stidham, S. S., 1985. "Optimal Control of Admission to a Queueing System," IEEE Trans. Automatic Control, Vol. AC-30, pp. 705-713.

[Str66] Strauch, R., 1966. "Negative Dynamic Programming," Ann. Math. Statist., Vol. 37, pp. 871-890.

[SuC91] Suk, J.-B., and Cassandras, C. G., 1991. "Optimal Scheduling of Two Competing Queues with Blocking," IEEE Trans. on Automatic Control, Vol. 36, pp. 1086-1091.

[Sut88] Sutton, R. S., 1988. "Learning to Predict by the Methods of Temporal Differences," Machine Learning, Vol. 3, pp. 9-44.

[TSC92] Towsley, D., Sparaggis, P. D., and Cassandras, C. G., 1992. "Optimal Routing and Buffer Allocation for a Class of Finite Capacity Queueing Systems," IEEE Trans. on Automatic Control, Vol. 37, pp. 1446-1451.

[Tes92] Tesauro, G., 1992. "Practical Issues in Temporal Difference Learning," Machine Learning, Vol. 8, pp. 257-277.

[TsV94] Tsitsiklis, J. N., and Van Roy, B., 1994. "Feature-Based Methods for Large-Scale Dynamic Programming," Lab. for Info. and Decision Systems Report LIDS-P-2277, Massachusetts Institute of Technology, Machine Learning, to appear.

[Tse90] Tseng, P., 1990. "Solving H-Horizon, Stationary Markov Decision Problems in Time Proportional to $\log(H)$." Operations Research Letters, Vol. 9, pp. 287-297.

[Tsi86] Tsitsiklis, J. N., 1986. "A Lemma on the Multiarmed Bandit Problem," IEEE Trans. Automatic Control, Vol. AC-31, pp. 576-577.

[Tsi89] Tsitsiklis, J. N., 1989. "A Comparison of Jacobi and Gauss-Seidel Parallel Iterations," Applied Math. Lett., Vol. 2, pp. 167-170.

[Tsi93a] Tsitsiklis, J. N., 1993. "Efficient Algorithms for Globally Optimal Trajectories," Lab. for Info. and Decision Systems Report LIDS-P-2210, Massachusetts Institute of Technology, IEEE Trans. on Automatic Control, to appear.

[Tsi93b] Tsitsiklis, J. N., 1993. "A Short Proof of the Gittins Index Theorem," Lab. for Info. and Decision Systems Report LIDS-P-2171, Massachusetts Institute of Technology; also Annals of Applied Probability, Vol. 4, 1994, pp. 194-199.

[Tsi94] Tsitsiklis, J. N., 1994. "Asynchronous Stochastic Approximation and Q-Learning," Machine Learning, Vol. 16, pp. 185-202.

[Tso91] Tsoukas, P., 1991. "The Region of Achievable Performance in a Model of Klimov." Research Report, I.B.M.

[VWB85] Varaiya, P. P., Walrand, J. C., and Buyukkoc, C., 1985. "Extensions of the Multiarmed Bandit Problem: The Discounted Case," IEEE Trans. Automatic Control, Vol. AC-30, pp. 426-439.

[Var78] Varaiya, P. P., 1978. "Optimal and Suboptimal Stationary Controls of Markov Chains," IEEE Trans. Automatic Control, Vol. AC-23, pp. 388-394.

[VeP84] Verd'u, S., and Poor, H. V., 1984. "Backward, Forward, and Backward-Forward Dynamic Programming Models under Commutativity Conditions," Proc. 1984 IEEE Decision and Control Conference, Las Vegas, NE, pp. 1081-1086.

[VeP87] Verd'u, S., and Poor, H. V., 1987. "Abstract Dynamic Programming Models under Commutativity Conditions," SIAM J. on Control and Optimization, Vol. 25, pp. 990-1006.

[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 Automatic Control, Vol. 33, pp. 104-109.

[Wat89] Watkins, C. J. C. H., "Learning from Delayed Rewards," Ph.D. Thesis, Cambridge Univ., England.

[Web92] Weber, R., 1991. "On the Gittins Index for Multiarmed Bandits," preprint; Annals of Applied Probability, Vol. 3, 1993.

[WhK80] White, C. C., and Kim, K., 1980. "Solution Procedures for Partially Observed Markov Decision Processes," J. Large Scale Systems, Vol. 1, pp. 129-140.

[Whi63] White, D. J., 1963. "Dynamic Programming, Markov Chains, and the Method of Successive Approximations," J. Math. Anal. and Appl., Vol. 6, pp. 373-376.

[Whi78] Whitt, W., 1978. "Approximations of Dynamic Programs I," Math. Operations Research, Vol. 3, pp. 231-243.

[Whi79] Whitt, W., 1979. "Approximations of Dynamic Programs II," Math. Operations Research, Vol. 4, pp. 179-185.

[Whi80a] White, D. J., 1980. "Finite State Approximations for Denumerable State Infinite Horizon Discounted Markov Decision Processes: The Method of Successive Approximations," in Recent Developments in Markov Decision Processes, Hartley, R., Thomas, L. C., and White, D. J. (eds.), Academic Press, N. Y., pp. 57-72.

[Whi80b] Whittle, P., 1980. "Multi-Armed Bandits and the Gittins Index," J. Roy. Statist. Soc. Ser. B, Vol. 42, pp. 143-149.

[Whi81] Whittle, P., 1981. "Arm-Acquiring Bandits," The Annals of Probability, Vol. 9, pp. 284-292.

[Whi82] Whittle, P., 1982. Optimization Over Time, Wiley, N. Y., Vol. 1, 1982, Vol. 2, 1983.

INDEX