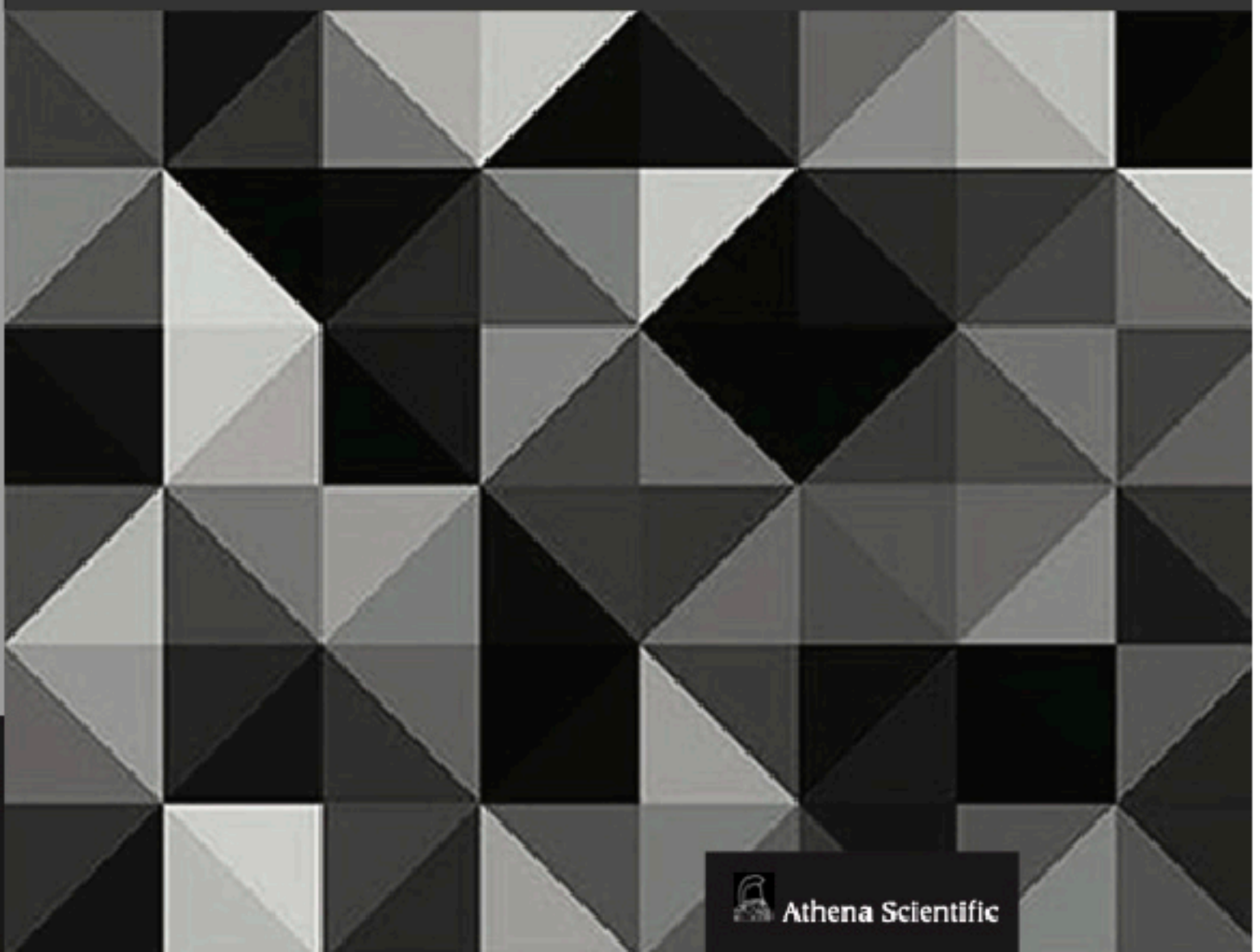
ABSTRACT DYNAMIC PROGRAMMING

3rd Edition

Dimitri P. Bertsekas



Athena Scientific

Abstract Dynamic Programming

THIRD EDITION

Dimitri P. Bertsekas

Arizona State University

Massachusetts Institute of Technology

WWW site for book information and orders

http://www.athenasc.com

Athena Scientific, Belmont, Massachusetts

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

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

Cover design: Dimitri Bertsekas

Publisher's Cataloging-in-Publication Data

Bertsekas, Dimitri P.
Abstract Dynamic Programming: Third Edition
Includes bibliographical references and index
1. Mathematical Optimization. 2. Dynamic Programming. I. Title.
QA402.5 .B465 2022 519.703 01-75941

ISBN-10: 1-886529-47-7, ISBN-13: 978-1-886529-47-2

ABOUT THE AUTHOR

Dimitri Bertsekas studied Mechanical and Electrical Engineering at the National Technical University of Athens, Greece, and obtained his Ph.D. in system science from the Massachusetts Institute of Technology. He has held faculty positions with the Engineering-Economic Systems Department, Stanford University, and the Electrical Engineering Department of the University of Illinois, Urbana. From 1979 to 2019 he was a professor at the Electrical Engineering and Computer Science Department of the Massachusetts Institute of Technology (M.I.T.), where he continues to hold the title of McAfee Professor of Engineering. In 2019, he joined the School of Computing and Augmented Intelligence at the Arizona State University, Tempe, AZ, as Fulton Professor of Computational Decision Making.

Professor Bertsekas' teaching and research have spanned several fields, including deterministic optimization, dynamic programming and stochastic control, large-scale and distributed computation, artificial intelligence, and data communication networks. He has authored or coauthored numerous research papers and twenty books, several of which are currently used as textbooks in MIT classes, including "Dynamic Programming and Optimal Control," "Data Networks," "Introduction to Probability," and "Nonlinear Programming." At ASU, he has been focusing in teaching and research in reinforcement learning, and he has written several textbooks and research monographs in this field since 2019.

Professor Bertsekas was awarded the INFORMS 1997 Prize for Research Excellence in the Interface Between Operations Research and Computer Science for his book "Neuro-Dynamic Programming" (co-authored with John Tsitsiklis), the 2001 AACC John R. Ragazzini Education Award, the 2009 INFORMS Expository Writing Award, the 2014 AACC Richard Bellman Heritage Award, the 2014 INFORMS Khachiyan Prize for Life-Time Accomplishments in Optimization, the 2015 MOS/SIAM George B. Dantzig Prize, and the 2022 IEEE Control Systems Award. In 2018 he shared with his coauthor, John Tsitsiklis, the 2018 INFORMS John von Neumann Theory Prize for the contributions of the research monographs "Parallel and Distributed Computation" and "Neuro-Dynamic Programming." Professor Bertsekas was elected in 2001 to the United States National Academy of Engineering for "pioneering contributions to fundamental research, practice and education of optimization/control theory."

ATHENA SCIENTIFIC

OPTIMIZATION AND COMPUTATION SERIES

1. A Course in Reinforcement Learning by Dimitri P. Bertsekas, 2023, ISBN 978-1-886529-49-6, 424 pages
2. Lessons from AlphaZero for Optimal, Model Predictive, and Adaptive Control by Dimitri P. Bertsekas, 2022, ISBN 978-1-886529-17-5, 245 pages
3. Abstract Dynamic Programming, 3rd Edition, by Dimitri P. Bertsekas, 2022, ISBN 978-1-886529-47-2, 420 pages
4. Rollout, Policy Iteration, and Distributed Reinforcement Learning, by Dimitri P. Bertsekas, 2020, ISBN 978-1-886529-07-6, 480 pages
5. Reinforcement Learning and Optimal Control, by Dimitri P. Bertsekas, 2019, ISBN 978-1-886529-39-7, 388 pages
6. Dynamic Programming and Optimal Control, Two-Volume Set, by Dimitri P. Bertsekas, 2017, ISBN 1-886529-08-6, 1270 pages
7. Nonlinear Programming, 3rd Edition, by Dimitri P. Bertsekas, 2016, ISBN 1-886529-05-1, 880 pages
8. Convex Optimization Algorithms, by Dimitri P. Bertsekas, 2015, ISBN 978-1-886529-28-1, 576 pages
9. Convex Optimization Theory, by Dimitri P. Bertsekas, 2009, ISBN 978-1-886529-31-1, 256 pages
10. Introduction to Probability, 2nd Edition, by Dimitri P. Bertsekas and John N. Tsitsiklis, 2008, ISBN 978-1-886529-23-6, 544 pages
11. Convex Analysis and Optimization, by Dimitri P. Bertsekas, Angelia Nedić, and Asuman E. Ozdaglar, 2003, ISBN 1-886529-45-0, 560 pages
12. Network Optimization: Continuous and Discrete Models, by Dimitri P. Bertsekas, 1998, ISBN 1-886529-02-7, 608 pages
13. Network Flows and Monotropic Optimization, by R. Tyrrell Rockafellar, 1998, ISBN 1-886529-06-X, 634 pages
14. Introduction to Linear Optimization, by Dimitris Bertsimas and John N. Tsitsiklis, 1997, ISBN 1-886529-19-1, 608 pages
15. Parallel and Distributed Computation: Numerical Methods, by Dimitri P. Bertsekas and John N. Tsitsiklis, 1997, ISBN 1-886529-01-9, 718 pages
16. Neuro-Dynamic Programming, by Dimitri P. Bertsekas and John N. Tsitsiklis, 1996, ISBN 1-886529-10-8, 512 pages
17. Constrained Optimization and Lagrange Multiplier Methods, by Dimitri P. Bertsekas, 1996, ISBN 1-886529-04-3, 410 pages
18. Stochastic Optimal Control: The Discrete-Time Case, by Dimitri P. Bertsekas and Steven E. Shreve, 1996, ISBN 1-886529-03-5, 330 pages

Contents

Preface of the First Edition

This book aims at a unified and economical development of the core theory and algorithms of total cost sequential decision problems, based on the strong connections of the subject with fixed point theory. The analysis focuses on the abstract mapping that underlies dynamic programming (DP for short) and defines the mathematical character of the associated problem. Our discussion centers on two fundamental properties that this mapping may have: *monotonicity* and (weighted sup-norm) *contraction*. It turns out that the nature of the analytical and algorithmic DP theory is determined primarily by the presence or absence of these two properties, and the rest of the problem's structure is largely inconsequential.

In this book, with some minor exceptions, we will assume that monotonicity holds. Consequently, we organize our treatment around the contraction property, and we focus on four main classes of models:

(a) **Contractive models**, discussed in Chapter 2, which have the richest and strongest theory, and are the benchmark against which the theory of other models is compared. Prominent among these models are discounted stochastic optimal control problems. The development of these models is quite thorough and includes the analysis of recent approximation algorithms for large-scale problems (neuro-dynamic programming, reinforcement learning).

(b) **Semicontractive models**, discussed in Chapter 3 and parts of Chapter 4. The term "semicontractive" is used qualitatively here, to refer to a variety of models where some policies have a regularity/contraction-like property but others do not. A prominent example is stochastic shortest path problems, where one aims to drive the state of a Markov chain to a termination state at minimum expected cost. These models also have a strong theory under certain conditions, often nearly as strong as those of the contractive models.

(c) **Noncontractive models**, discussed in Chapter 4, which rely on just monotonicity. These models are more complex than the preceding ones and much of the theory of the contractive models generalizes in weaker form, if at all. For example, in general the associated Bellman equation need not have a unique solution, the value iteration method may work starting with some functions but not with others, and the policy iteration method may not work at all. Infinite horizon examples of these models are the classical positive and negative DP problems, first analyzed by Dubins and Savage, Blackwell, and

Strauch, which are discussed in various sources. Some new semicontractive models are also discussed in this chapter, further bridging the gap between contractive and noncontractive models.

(d) **Restricted policies and Borel space models**, which are discussed in Chapter 5. These models are motivated in part by the complex measurability questions that arise in mathematically rigorous theories of stochastic optimal control involving continuous probability spaces. Within this context, the admissible policies and DP mapping are restricted to have certain measurability properties, and the analysis of the preceding chapters requires modifications. Restricted policy models are also useful when there is a special class of policies with favorable structure, which is "closed" with respect to the standard DP operations, in the sense that analysis and algorithms can be confined within this class.

We do not consider average cost DP problems, whose character bears a much closer connection to stochastic processes than to total cost problems. We also do not address specific stochastic characteristics underlying the problem, such as for example a Markovian structure. Thus our results apply equally well to Markovian decision problems and to sequential minimax problems. While this makes our development general and a convenient starting point for the further analysis of a variety of different types of problems, it also ignores some of the interesting characteristics of special types of DP problems that require an intricate probabilistic analysis.

Let us describe the research content of the book in summary, deferring a more detailed discussion to the end-of-chapter notes. A large portion of our analysis has been known for a long time, but in a somewhat fragmentary form. In particular, the contractive theory, first developed by Denardo [Den67], has been known for the case of the unweighted sup-norm, but does not cover the important special case of stochastic shortest path problems where all policies are proper. Chapter 2 transcribes this theory to the weighted sup-norm contraction case. Moreover, Chapter 2 develops extensions of the theory to approximate DP, and includes material on asynchronous value iteration (based on the author's work [Ber82], [Ber83]), and asynchronous policy iteration algorithms (based on the author's joint work with Huizhen (Janey) Yu [BeY10a], [BeY10b], [YuB11a]). Most of this material is relatively new, having been presented in the author's recent book [Ber12a] and survey paper [Ber12b], with detailed references given there. The analysis of infinite horizon noncontractive models in Chapter 4 was first given in the author's paper [Ber77], and was also presented in the book by Bertsekas and Shreve [BeS78], which in addition contains much of the material on finite horizon problems, restricted policies models, and Borel space models. These were the starting point and main sources for our development.

The new research presented in this book is primarily on the semi-

contractive models of Chapter 3 and parts of Chapter 4. Traditionally, the theory of total cost infinite horizon DP has been bordered by two extremes: discounted models, which have a contractive nature, and positive and negative models, which do not have a contractive nature, but rely on an enhanced monotonicity structure (monotone increase and monotone decrease models, or in classical DP terms, positive and negative models). Between these two extremes lies a gray area of problems that are not contractive, and either do not fit into the categories of positive and negative models, or possess additional structure that is not exploited by the theory of these models. Included are stochastic shortest path problems, search problems, linear-quadratic problems, a host of queueing problems, multiplicative and exponential cost models, and others. Together these problems represent an important part of the infinite horizon total cost DP landscape. They possess important theoretical characteristics, not generally available for positive and negative models, such as the uniqueness of solution of Bellman's equation within a subset of interest, and the validity of useful forms of value and policy iteration algorithms.

Our semicontractive models aim to provide a unifying abstract DP structure for problems in this gray area between contractive and noncontractive models. The analysis is motivated in part by stochastic shortest path problems, where there are two types of policies: *proper*, which are the ones that lead to the termination state with probability one from all starting states, and *improper*, which are the ones that are not proper. Proper and improper policies can also be characterized through their Bellman equation mapping: for the former this mapping is a contraction, while for the latter it is not. In our more general semicontractive models, policies are also characterized in terms of their Bellman equation mapping, through a notion of *regularity*, which generalizes the notion of a proper policy and is related to classical notions of asymptotic stability from control theory.

In our development a policy is regular within a certain set if its cost function is the unique asymptotically stable equilibrium (fixed point) of the associated DP mapping within that set. *We assume that some policies are regular while others are not*, and impose various assumptions to ensure that attention can be focused on the regular policies. From an analytical point of view, this brings to bear the theory of fixed points of monotone mappings. From the practical point of view, this allows application to a diverse collection of interesting problems, ranging from stochastic shortest path problems of various kinds, where the regular policies include the proper policies, to linear-quadratic problems, where the regular policies include the stabilizing linear feedback controllers.

The definition of regularity is introduced in Chapter 3, and its theoretical ramifications are explored through extensions of the classical stochastic shortest path and search problems. In Chapter 4, semicontractive models are discussed in the presence of additional monotonicity structure, which brings to bear the properties of positive and negative DP models. With the

aid of this structure, the theory of semicontractive models can be strengthened and can be applied to several additional problems, including risk-sensitive/exponential cost problems.

The book has a theoretical research monograph character, but requires a modest mathematical background for all chapters except the last one, essentially a first course in analysis. Of course, prior exposure to DP will definitely be very helpful to provide orientation and context. A few exercises have been included, either to illustrate the theory with examples and counterexamples, or to provide applications and extensions of the theory. Solutions of all the exercises can be found in Appendix D, at the book's internet site

http://www.athenasc.com/abstractdp.html

and at the author's web site

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

Additional exercises and other related material may be added to these sites over time.

I would like to express my appreciation to a few colleagues for interactions, recent and old, which have helped shape the form of the book. My collaboration with Steven Shreve on our 1978 book provided the motivation and the background for the material on models with restricted policies and associated measurability questions. My collaboration with John Tsitsiklis on stochastic shortest path problems provided inspiration for the work on semicontractive models. My collaboration with Janey (Huizhen) Yu played an important role in the book's development, and is reflected in our joint work on asynchronous policy iteration, on perturbation models, and on risk-sensitive models. Moreover Janey contributed significantly to the material on semicontractive models with many insightful suggestions. Finally, I am thankful to Mengdi Wang, who went through portions of the book with care, and gave several helpful comments.

Dimitri P. Bertsekas

Spring 2013

Preface to the Second Edition

The second edition aims primarily to amplify the presentation of the semicontractive models of Chapter 3 and Chapter 4, and to supplement it with a broad spectrum of research results that I obtained and published in journals and reports since the first edition was written. As a result, the size of this material more than doubled, and the size of the book increased by about 40%.

In particular, I have thoroughly rewritten Chapter 3, which deals with semicontractive models where stationary regular policies are sufficient. I expanded and streamlined the theoretical framework, and I provided new analyses of a number of shortest path-type applications (deterministic, stochastic, affine monotonic, exponential cost, and robust/minimax), as well as several types of optimal control problems with continuous state space (including linear-quadratic, regulation, and planning problems).

In Chapter 4, I have extended the notion of regularity to nonstationary policies (Section 4.4), aiming to explore the structure of the solution set of Bellman's equation, and the connection of optimality with other structural properties of optimal control problems. As an application, I have discussed in Section 4.5 the relation of optimality with classical notions of stability and controllability in continuous-spaces deterministic optimal control. In Section 4.6, I have similarly extended the notion of a proper policy to continuous-spaces stochastic shortest path problems.

I have also revised Chapter 1 a little (mainly with the addition of Section 1.2.5 on the relation between proximal algorithms and temporal difference methods), added to Chapter 2 some analysis relating to λ-policy iteration and randomized policy iteration algorithms (Section 2.5.3), and I have also added several new exercises (with complete solutions) to Chapters 1-4. Additional material relating to various applications can be found in some of my journal papers, reports, and video lectures on semicontractive models, which are posted at my web site.

In addition to the changes in Chapters 1-4, I have also eliminated from the second edition the analysis that deals with restricted policies (Chapter 5 and Appendix C of the first edition). This analysis is motivated in part by the complex measurability questions that arise in mathematically rigorous theories of stochastic optimal control with Borel state and control spaces. This material is covered in Chapter 6 of the monograph by Bertsekas and Shreve [BeS78], and followup research on the subject has been limited. Thus, I decided to just post Chapter 5 and Appendix C of the first

edition at the book's web site (40 pages), and omit them from the second edition. As a result of this choice, the entire book now requires only a modest mathematical background, essentially a first course in analysis and in elementary probability.

The range of applications of dynamic programming has grown enormously in the last 25 years, thanks to the use of approximate simulation-based methods for large and challenging problems. Because approximations are often tied to special characteristics of specific models, their coverage in this book is limited to general discussions in Chapter 1 and to error bounds given in Chapter 2. However, much of the work on approximation methods so far has focused on finite-state discounted, and relatively simple deterministic and stochastic shortest path problems, for which there is solid and robust analytical and algorithmic theory (part of Chapters 2 and 3 in this monograph). As the range of applications becomes broader, I expect that the level of mathematical understanding projected in this book will become essential for the development of effective and reliable solution methods. In particular, much of the new material in this edition deals with infinite-state and/or complex shortest path type-problems, whose approximate solution will require new methodologies that transcend the current state of the art.

Dimitri P. Bertsekas

January 2018

Preface to the Third Edition

The third edition is based on the same theoretical framework as the second edition, but contains two major additions. The first is to highlight the central role of abstract DP methods in the conceptualization of reinforcement learning and approximate DP methods, as described in the author's recent book "Lessons from AlphaZero for Optimal, Model Predictive, and Adaptive Control," Athena Scientific, 2022. The main idea here is that approximation in value space with one-step lookahead amounts to a step of Newton's method for solving the abstract Bellman's equation. This material is included in summary form in view of its strong reliance on abstract DP visualization. Our presentation relies primarily on geometric illustrations rather than mathematical analysis, and is given in Section 1.3.

The second addition is a new Chapter 5 on abstract DP methods for minimax and zero sum game problems, which is based on the author's recent paper [Ber21c]. A primary motivation here is the resolution of some long-standing convergence difficulties of the "natural" policy iteration algorithm, which have been known since the Pollatschek and Avi-Itzhak method [PoA69] for finite-state Markov games. Mathematically, this "natural" algorithm is a form of Newton's method for solving the corresponding Bellman's equation, but Newton's method, contrary to the case of single-player DP problems, is not globally convergent in the case of a minimax problem, because the Bellman operator may have components that are neither convex nor concave. Our approach in Chapter 5 has been to introduce a special type of abstract Bellman operator for minimax problems, and modify the standard PI algorithm along the lines of the asynchronous optimistic PI algorithm of Section 2.6.3, which involves a parametric contraction mapping with a uniform fixed point.

The third edition also contains a number of small corrections and editorial changes. The author wishes to thank the contributions of several colleagues in this regard, and particularly Yuchao Li, who proofread with care large portions of the book.

Dimitri P. Bertsekas

February 2022

1

Introduction

Contents

1.1 STRUCTURE OF DYNAMIC PROGRAMMING PROBLEMS

Dynamic programming (DP for short) is the principal method for analysis of a large and diverse class of sequential decision problems. Examples are deterministic and stochastic optimal control problems with a continuous state space, Markov and semi-Markov decision problems with a discrete state space, minimax problems, and sequential zero-sum games. While the nature of these problems may vary widely, their underlying structures turn out to be very similar. In all cases there is an underlying mapping that depends on an associated controlled dynamic system and corresponding cost per stage. This mapping, the DP (or Bellman) operator, provides a compact "mathematical signature" of the problem. It defines the cost function of policies and the optimal cost function, and it provides a convenient shorthand notation for algorithmic description and analysis.

More importantly, the structure of the DP operator defines the mathematical character of the associated problem. The purpose of this book is to provide an analysis of this structure, centering on two fundamental properties: *monotonicity* and (weighted sup-norm) *contraction*. It turns out that the nature of the analytical and algorithmic DP theory is determined primarily by the presence or absence of one or both of these two properties, and the rest of the problem's structure is largely inconsequential.

A Deterministic Optimal Control Example

To illustrate our viewpoint, let us consider a discrete-time deterministic optimal control problem described by a system equation

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

Here x_k is the state of the system taking values in a set X (the state space), and u_k is the control taking values in a set U (the control space). † At stage k, there is a cost

$$\alpha^k g(x_k, u_k)$$

incurred when u_k is applied at state x_k, where α is a scalar in $(0, 1]$ that has the interpretation of a discount factor when $\alpha < 1$. The controls are chosen as a function of the current state, subject to a constraint that depends on that state. In particular, at state x the control is constrained to take values in a given set $U(x) \subset U$. Thus we are interested in optimization over the set of (nonstationary) policies

$$\Pi = \big\{\{\mu_0, \mu_1, \ldots\} \mid \mu_k \in \mathcal{M},\ k = 0, 1, \ldots\big\},$$

† Our discussion of this section is somewhat informal, without strict adherence to mathematical notation and rigor. We will introduce a rigorous mathematical framework later.

where $\mathcal{M}$ is the set of functions $\mu : X \mapsto U$ defined by

$$\mathcal{M} = \big\{\mu \mid \mu(x) \in U(x),\ \forall\ x \in X\big\}.$$

The total cost of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ over an infinite number of stages (an infinite horizon) and starting at an initial state x_0 is the limit superior of the N-step costs

$$J_\pi(x_0) = \limsup_{N\to\infty} \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k)\big), \tag{1.2}$$

where the state sequence $\{x_k\}$ is generated by the deterministic system (1.1) under the policy π:

$$x_{k+1} = f\big(x_k, \mu_k(x_k)\big), \quad k = 0, 1, \ldots.$$

(We use limit superior rather than limit to cover the case where the limit does not exist.) The optimal cost function is

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

For any policy $\pi = \{\mu_0, \mu_1, \ldots\}$, consider the policy $\pi_1 = \{\mu_1, \mu_2, \ldots\}$ and write by using Eq. (1.2),

$$J_\pi(x) = g\big(x, \mu_0(x)\big) + \alpha J_{\pi_1}\big(f(x, \mu_0(x))\big).$$

We have for all $x \in X$

$$\begin{aligned} J^*(x) &= \inf_{\pi=\{\mu_0,\pi_1\}\in\Pi} \Big\{g\big(x, \mu_0(x)\big) + \alpha J_{\pi_1}\big(f(x, \mu_0(x))\big)\Big\} \\ &= \inf_{\mu_0\in\mathcal{M}} \Big\{g\big(x, \mu_0(x)\big) + \alpha \inf_{\pi_1\in\Pi} J_{\pi_1}\big(f(x, \mu_0(x))\big)\Big\} \\ &= \inf_{\mu_0\in\mathcal{M}} \Big\{g\big(x, \mu_0(x)\big) + \alpha J^*\big(f(x, \mu_0(x))\big)\Big\}. \end{aligned}$$

The minimization over $\mu_0 \in \mathcal{M}$ can be written as minimization over all $u \in U(x)$, so we can write the preceding equation as

$$J^*(x) = \inf_{u\in U(x)} \Big\{g(x, u) + \alpha J^*\big(f(x, u)\big)\Big\}, \qquad \forall\ x \in X. \tag{1.3}$$

This equation is an example of *Bellman's equation*, which plays a central role in DP analysis and algorithms. If it can be solved for J^*, an optimal stationary policy $\{\mu^*, \mu^*, \ldots\}$ may typically be obtained by minimization of the right-hand side for each x, i.e.,

$$\mu^*(x) \in \arg\min_{u\in U(x)} \Big\{g(x, u) + \alpha J^*\big(f(x, u)\big)\Big\}, \qquad \forall\ x \in X. \tag{1.4}$$

We now note that both Eqs. (1.3) and (1.4) can be stated in terms of the expression

$$H(x,u,J) = g(x,u) + \alpha J\big(f(x,u)\big), \qquad x \in X,\ u \in U(x).$$

Defining

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

and

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

we see that Bellman's equation (1.3) can be written compactly as

$$J^* = TJ^*,$$

i.e., J^* is the fixed point of T, viewed as a mapping from the set of functions on X into itself. Moreover, it can be similarly seen that J_μ, the cost function of the stationary policy $\{\mu,\mu,\ldots\}$, is a fixed point of T_μ. In addition, the optimality condition (1.4) can be stated compactly as

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

We will see later that additional properties, as well as a variety of algorithms for finding J^* can be stated and analyzed using the mappings T and T_μ.

The mappings T_μ can also be used in the context of DP problems with a finite number of stages (a finite horizon). In particular, for a given policy $\pi = \{\mu_0,\mu_1,\ldots\}$ and a terminal cost $\alpha^N \bar{J}(x_N)$ for the state x_N at the end of N stages, consider the N-stage cost function

$$J_{\pi,N}(x_0) = \alpha^N \bar{J}(x_N) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k,\mu_k(x_k)\big). \tag{1.5}$$

Then it can be verified by induction that for all initial states x_0, we have

$$J_{\pi,N}(x_0) = (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} \bar{J})(x_0). \tag{1.6}$$

Here $T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}}$ is the composition of the mappings $T_{\mu_0}, T_{\mu_1}, \ldots T_{\mu_{N-1}}$, i.e., for all J,

$$(T_{\mu_0} T_{\mu_1} J)(x) = \big(T_{\mu_0}(T_{\mu_1} J)\big)(x), \qquad x \in X,$$

and more generally

$$(T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} J)(x) = \big(T_{\mu_0}(T_{\mu_1}(\cdots(T_{\mu_{N-1}} J)))\big)(x), \qquad x \in X,$$

(our notational conventions are summarized in Appendix A). Thus the finite horizon cost functions $J_{\pi,N}$ of π can be defined in terms of the mappings T_μ [cf. Eq. (1.6)], and so can the infinite horizon cost function J_π:

$$J_\pi(x) = \limsup_{N \to \infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} \bar{J})(x), \qquad x \in X, \tag{1.7}$$

where $\bar{J}$ is the zero function, $\bar{J}(x) = 0$ for all $x \in X$.

Connection with Fixed Point Methodology

The Bellman equation (1.3) and the optimality condition (1.4), stated in terms of the mappings T_μ and T, highlight a central theme of this book, which is that DP theory is intimately connected with the theory of abstract mappings and their fixed points. Analogs of the Bellman equation, $J^* = TJ^*$, optimality conditions, and other results and computational methods hold for a great variety of DP models, and can be stated compactly as described above in terms of the corresponding mappings T_μ and T. The gain from this abstraction is greater generality and mathematical insight, as well as a more unified, economical, and streamlined analysis.

1.2 ABSTRACT DYNAMIC PROGRAMMING MODELS

In this section we formally introduce and illustrate with examples an abstract DP model, which embodies the ideas just discussed in Section 1.1.

1.2.1 Problem Formulation

Let X and U be two sets, which we loosely refer to as a set of "states" and a set of "controls," respectively. For each $x \in X$, let $U(x) \subset U$ be a nonempty subset of controls that are feasible at state x. We denote by $\mathcal{M}$ the set of all functions $\mu : X \mapsto U$ with $\mu(x) \in U(x)$, for all $x \in X$.

In analogy with DP, we refer to sequences $\pi = \{\mu_0, \mu_1, \ldots\}$, with $\mu_k \in \mathcal{M}$ for all k, as "nonstationary policies," and we refer to a sequence $\{\mu, \mu, \ldots\}$, with $\mu \in \mathcal{M}$, as a "stationary policy." In our development, stationary policies will play a dominant role, and with slight abuse of terminology, we will also refer to any $\mu \in \mathcal{M}$ as a "policy" when confusion cannot arise.

Let $\mathcal{R}(X)$ be the set of real-valued functions $J : X \mapsto \Re$, and let $H : X \times U \times \mathcal{R}(X) \mapsto \Re$ be a given mapping.† For each policy $\mu \in \mathcal{M}$, we consider the mapping $T_\mu : \mathcal{R}(X) \mapsto \mathcal{R}(X)$ defined by

$$(T_\mu J)(x) = H\big(x, \mu(x), J\big), \qquad \forall\ x \in X,\ J \in \mathcal{R}(X),$$

and we also consider the mapping T defined by‡

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

† Our notation and mathematical conventions are outlined in Appendix A. In particular, we denote by $\Re$ the set of real numbers, and by $\Re^n$ the space of n-dimensional vectors with real components.

‡ We assume that H, $T_\mu J$, and TJ are real-valued for $J \in \mathcal{R}(X)$ in the present chapter and in Chapter 2. In Chapters 3 and 4 we will allow $H(x, u, J)$, and hence also $(T_\mu J)(x)$ and $(TJ)(x)$, to take the values ∞ and $-\infty$.

We will generally refer to T and T_μ as the (abstract) *DP mappings* or *DP operators* or *Bellman operators* (the latter name is common in the artificial intelligence and reinforcement learning literature).

Similar to the deterministic optimal control problem of the preceding section, the mappings T_μ and T serve to define a multistage optimization problem and a DP-like methodology for its solution. In particular, for some function $\bar{J} \in \mathcal{R}(X)$, and nonstationary policy $\pi = \{\mu_0, \mu_1, \ldots\}$, we define for each integer $N \geq 1$ the functions

$$J_{\pi,N}(x) = (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} \bar{J})(x), \qquad x \in X,$$

where $T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}}$ denotes the composition of the mappings T_{μ_0}, T_{μ_1}, $\ldots$, $T_{\mu_{N-1}}$, i.e.,

$$T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} J = \big(T_{\mu_0}(T_{\mu_1}(\cdots(T_{\mu_{N-2}}(T_{\mu_{N-1}} J)))\cdots)\big), \quad J \in \mathcal{R}(X).$$

We view $J_{\pi,N}$ as the "N-stage cost function" of π [cf. Eq. (1.5)]. Consider also the function

$$J_\pi(x) = \limsup_{N\to\infty} J_{\pi,N}(x) = \limsup_{N\to\infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} \bar{J})(x), \qquad x \in X,$$

which we view as the "infinite horizon cost function" of π [cf. Eq. (1.7); we use lim sup for generality, since we are not assured that the limit exists]. We want to minimize J_π over π, i.e., to find

$$J^*(x) = \inf_\pi J_\pi(x), \qquad x \in X,$$

and a policy π^* that attains the infimum, if one exists.

The key connection with fixed point methodology is that J^* "typically" (under mild assumptions) can be shown to satisfy

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

i.e., it is a fixed point of T. We refer to this as *Bellman's equation* [cf. Eq. (1.3)]. Another fact is that if an optimal policy π^* exists, it "typically" can be selected to be stationary, $\pi^* = \{\mu^*, \mu^*, \ldots\}$, with $\mu^* \in \mathcal{M}$ satisfying an optimality condition, such as for example

$$(T_{\mu^*} J^*)(x) = (T J^*)(x), \qquad x \in X,$$

[cf. Eq. (1.4)]. Several other results of an analytical or algorithmic nature also hold under appropriate conditions, which will be discussed in detail later.

However, Bellman's equation and other related results may not hold without T_μ and T having some special structural properties. Prominent among these are a monotonicity assumption that typically holds in DP problems, and a contraction assumption that holds for some important classes of problems. We describe these assumptions next.

1.2.2 Monotonicity and Contraction Properties

Let us now formalize the monotonicity and contraction assumptions. We will require that both of these assumptions hold for most of the next chapter, and we will gradually relax the contraction assumption in Chapters 3 and 4. Recall also our assumption that T_μ and T map $\mathcal{R}(X)$ (the space of real-valued functions over X) into $\mathcal{R}(X)$. In Chapters 3 and 4 we will relax this assumption as well.

Assumption 1.2.1: (Monotonicity) If $J, J' \in \mathcal{R}(X)$ and $J \leq J'$, then

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

Note that by taking infimum over $u \in U(x)$, we have

$$J(x) \leq J'(x),\ \ \forall\ x \in X \quad \Rightarrow \quad \inf_{u \in U(x)} H(x,u,J) \leq \inf_{u \in U(x)} H(x,u,J'),\ \ \forall\ x \in X,$$

or equivalently, †

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

Another way to arrive at this relation, is to note that the monotonicity assumption is equivalent to

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

and to use the simple but important fact

$$\inf_{u \in U(x)} H(x,u,J) = \inf_{\mu \in \mathcal{M}} (T_\mu J)(x), \qquad \forall\ x \in X,\ J \in \mathcal{R}(X),$$

i.e., for a fixed $x \in X$, *infimum over u is equivalent to infimum over μ*. This is true because for any μ, there is no coupling constraint between the controls $\mu(x)$ and $\mu(x')$ that correspond to two different states x and x', i.e., the set $\mathcal{M} = \big\{\mu \mid \mu(x) \in U(x),\ \forall\ x \in X\big\}$ can be viewed as the Cartesian product $\Pi_{x \in X} U(x)$. We will be writing this relation as $TJ = \inf_{\mu \in \mathcal{M}} T_\mu J$.

For the contraction assumption, we introduce a function $v : X \mapsto \Re$ with

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

† Unless otherwise stated, in this book, inequalities involving functions, minima and infima of a collection of functions, and limits of function sequences are meant to be pointwise; see Appendix A for our notational conventions.

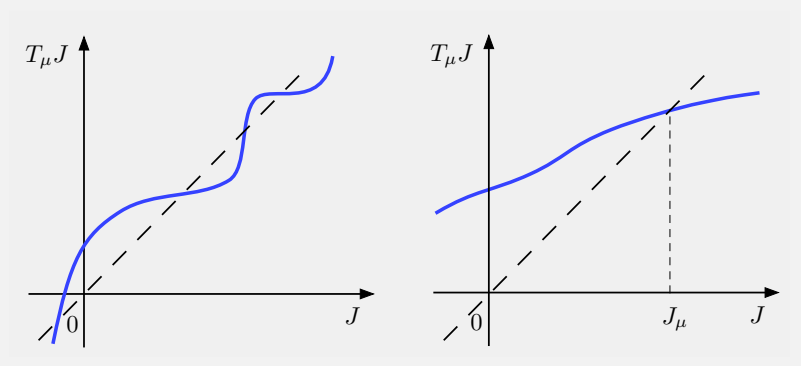


Figure 1.2.1. Illustration of the monotonicity and the contraction assumptions in one dimension. The mapping T_μ on the left is monotone but is not a contraction. The mapping T_μ on the right is both monotone and a contraction. It has a unique fixed point at J_μ.

Let us denote by $\mathcal{B}(X)$ the space of real-valued functions J on X such that $J(x)/v(x)$ is bounded as x ranges over X, and consider the weighted sup-norm

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

on $\mathcal{B}(X)$. The properties of $\mathcal{B}(X)$ and some of the associated fixed point theory are discussed in Appendix B. In particular, as shown there, $\mathcal{B}(X)$ is a complete normed space, so any mapping from $\mathcal{B}(X)$ to $\mathcal{B}(X)$ that is a contraction or an m-stage contraction for some integer $m > 1$, with respect to $\|\cdot\|$, has a unique fixed point (cf. Props. B.1 and B.2).

Assumption 1.2.2: (Contraction) For all $J \in \mathcal{B}(X)$ and $\mu \in \mathcal{M}$, the functions $T_\mu J$ and TJ belong to $\mathcal{B}(X)$. Furthermore, for some $\alpha \in (0,1)$, we have

$$\|T_\mu J - T_\mu J'\| \leq \alpha \|J - J'\|, \qquad \forall\ J, J' \in \mathcal{B}(X),\ \mu \in \mathcal{M}. \tag{1.8}$$

Figure 1.2.1 illustrates the monotonicity and the contraction assumptions. It can be shown that the contraction condition (1.8) implies that

$$\|TJ - TJ'\| \leq \alpha \|J - J'\|, \qquad \forall\ J, J' \in \mathcal{B}(X), \tag{1.9}$$

so that T is also a contraction with modulus α. To see this we use Eq. (1.8) to write

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

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

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

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

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

and combining the preceding two relations, and taking the supremum of the left side over $x \in X$, we obtain Eq. (1.9).

Nearly all mappings related to DP satisfy the monotonicity assumption, and many important ones satisfy the weighted sup-norm contraction assumption as well. When both assumptions hold, the most powerful analytical and computational results can be obtained, as we will show in Chapter 2. These are:

(a) Bellman's equation has a unique solution, i.e., T and T_μ have unique fixed points, which are the optimal cost function J^* and the cost functions J_μ of the stationary policies $\{\mu, \mu, \ldots\}$, respectively [cf. Eq. (1.3)].

(b) A stationary policy $\{\mu^*, \mu^*, \ldots\}$ is optimal if and only if

$$T_{\mu^*} J^* = TJ^*,$$

[cf. Eq. (1.4)].

(c) J^* and J_μ can be computed by the *value iteration* method,

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

starting with any $J \in \mathcal{B}(X)$.

(d) J^* can be computed by the *policy iteration* method, whereby we generate a sequence of stationary policies via

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

starting from some initial policy μ^0 [here J_{μ^k} is obtained as the fixed point of T_{μ^k} by several possible methods, including value iteration as in (c) above].

These are the most favorable types of results one can hope for in the DP context, and they are supplemented by a host of other results, involving approximate and/or asynchronous implementations of the value and policy iteration methods, and other related methods that combine features of both. As the contraction property is relaxed and is replaced by various weaker assumptions, some of the preceding results may hold in weaker form. For example J^* turns out to be a solution of Bellman's equation in most of the models to be discussed, but it may not be the unique solution. The interplay between the monotonicity and contraction-like properties, and the associated results of the form (a)-(d) described above is a recurring analytical theme in this book.

1.2.3 Some Examples

In what follows in this section, we describe a few special cases, which indicate the connections of appropriate forms of the mapping H with the most popular total cost DP models. In all these models the monotonicity Assumption 1.2.1 (or some closely related version) holds, but the contraction Assumption 1.2.2 may not hold, as we will indicate later. Our descriptions are by necessity brief, and the reader is referred to the relevant textbook literature for more detailed discussion.

Example 1.2.1 (Stochastic Optimal Control - Markovian Decision Problems)

Consider the stationary discrete-time dynamic system

$$x_{k+1} = f(x_k, u_k, w_k), \qquad k = 0, 1, \ldots, \tag{1.10}$$

where for all k, the state x_k is an element of a space X, the control u_k is an element of a space U, and w_k is a random "disturbance," an element of a space W. We consider problems with infinite state and control spaces, as well as problems with discrete (finite or countable) state space (in which case the underlying system is a Markov chain). However, for technical reasons that relate to measure-theoretic issues, *we assume that W is a countable set*.

The control u_k is constrained to take values in a given nonempty subset $U(x_k)$ of U, which depends on the current state x_k [$u_k \in U(x_k)$, for all $x_k \in X$]. The random disturbances w_k, $k = 0, 1, \ldots$, are characterized by probability distributions $P(\cdot \mid x_k, u_k)$ that are identical for all k, where $P(w_k \mid x_k, u_k)$ is the probability of occurrence of w_k, when the current state and control are x_k and u_k, respectively. Thus the probability of w_k may depend explicitly on x_k and u_k, but not on values of prior disturbances $w_{k-1}, \ldots, w_0$.

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

$$J_\pi(x_0) = \limsup_{N \to \infty} \mathop{E}_{\substack{w_k \\ k=0,1,\ldots}} \left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\}, \tag{1.11}$$

where $\alpha \in (0, 1]$ is a discount factor, subject to the system equation constraint

$$x_{k+1} = f\big(x_k, \mu_k(x_k), w_k\big), \qquad k = 0, 1, \ldots.$$

This is a classical problem, which is discussed extensively in various sources, including the author's text [Ber12a]. It is usually referred to as the *stochastic optimal control problem* or the *Markovian Decision Problem* (MDP for short).

Note that the expected value of the N-stage cost of π,

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

is defined as a (possibly countably infinite) sum, since the disturbances w_k, $k = 0, 1, \ldots$, take values in a countable set. Indeed, the reader may verify that all the subsequent mathematical expressions that involve an expected value can be written as summations over a finite or a countable set, so they make sense without resort to measure-theoretic integration concepts. †

In what follows we will often impose appropriate assumptions on the cost per stage g and the scalar α, which guarantee that the infinite horizon cost $J_\pi(x_0)$ is defined as a limit (rather than as a lim sup):

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

In particular, it can be shown that the limit exists if $\alpha < 1$ and the expected value of $|g|$ is uniformly bounded, i.e., for some $B > 0$,

$$E\big\{\big|g(x,u,w)\big|\big\} \le B, \qquad \forall\ x \in X,\ u \in U(x). \tag{1.12}$$

In this case, we obtain the classical discounted infinite horizon DP problem, which generally has the most favorable structure of all infinite horizon stochastic DP models (see [Ber12a], Chapters 1 and 2).

To make the connection with abstract DP, let us define

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

so that

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

and

$$(TJ)(x) = \inf_{u\in U(x)} E\big\{g(x,u,w) + \alpha J\big(f(x,u,w)\big)\big\}.$$

Similar to the deterministic optimal control problem of Section 1.1, the N-stage cost of π, can be expressed in terms of T_μ:

$$(T_{\mu_0}\cdots T_{\mu_{N-1}}\bar{J})(x_0) = \mathop{E}_{\substack{w_k \\ k=0,1,\ldots}} \left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\},$$

† As noted in Appendix A, the formula for the expected value of a random variable w defined over a space Ω is

$$E\{w\} = E\{w^+\} + E\{w^-\},$$

where w^+ and w^- are the positive and negative parts of w,

$$w^+(\omega) = \max\big\{0, w(\omega)\big\}, \qquad w^-(\omega) = \min\big\{0, w(\omega)\big\}, \qquad \forall\ \omega \in \Omega.$$

In this way, taking also into account the rule $\infty - \infty = \infty$ (see Appendix A), $E\{w\}$ is well-defined as an extended real number if Ω is finite or countably infinite.

where $\bar{J}$ is the zero function, $\bar{J}(x) = 0$ for all $x \in X$. The same is true for the infinite-stage cost [cf. Eq. (1.11)]:

$$J_\pi(x_0) = \limsup_{N\to\infty} (T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(x_0).$$

It can be seen that the mappings T_μ and T are monotone, and it is well-known that if $\alpha < 1$ and the boundedness condition (1.12) holds, they are contractive as well (under the unweighted sup-norm); see e.g., [Ber12a], Chapter 1. In this case, the model has the powerful analytical and algorithmic properties (a)-(d) mentioned at the end of the preceding subsection. In particular, the optimal cost function J^* [i.e., $J^*(x) = \inf_\pi J_\pi(x)$ for all $x \in X$] can be shown to be the unique solution of the fixed point equation $J^* = TJ^*$, also known as Bellman's equation, which has the form

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

and parallels the one given for deterministic optimal control problems [cf. Eq. (1.3)].

These properties can be expressed and analyzed in an abstract setting by using just the mappings T_μ and T, both when T_μ and T are contractive (see Chapter 2), and when they are only monotone and not contractive while either $g \geq 0$ or $g \leq 0$ (see Chapter 4). Moreover, under some conditions, it is possible to analyze these properties in cases where T_μ is contractive for some but not all μ (see Chapter 3, and Section 4.4).

Example 1.2.2 (Finite-State Discounted Markovian Decision Problems)

In the special case of the preceding example where the number of states is finite, the system equation (1.10) may be defined in terms of the transition probabilities

$$p_{xy}(u) = \text{Prob}\big(y = f(x,u,w) \mid x\big), \qquad x, y \in X,\ u \in U(x),$$

so H takes the form

$$H(x,u,J) = \sum_{y\in X} p_{xy}(u)\big(g(x,u,y) + \alpha J(y)\big).$$

When $\alpha < 1$ and the boundedness condition

$$\big|g(x,u,y)\big| \leq B, \qquad \forall\ x, y \in X,\ u \in U(x),$$

[cf. Eq. (1.12)] holds (or more simply, when U is a finite set), the mappings T_μ and T are contraction mappings with respect to the standard (unweighted) sup-norm. This is a classical model, referred to as *discounted finite-state MDP*, which has a favorable theory and has found extensive applications (cf. [Ber12a], Chapters 1 and 2). The model is additionally important, because it is often used for computational solution of continuous state space problems via discretization.

Example 1.2.3 (Discounted Semi-Markov Problems)

With x, y, and u as in Example 1.2.2, consider a mapping of the form

$$H(x,u,J) = G(x,u) + \sum_{y \in X} m_{xy}(u) J(y),$$

where G is some function representing expected cost per stage, and $m_{xy}(u)$ are nonnegative scalars with

$$\sum_{y \in X} m_{xy}(u) < 1, \qquad \forall\, x \in X,\, u \in U(x).$$

The equation $J^* = TJ^*$ is Bellman's equation for a finite-state continuous-time semi-Markov decision problem, after it is converted into an equivalent discrete-time problem (cf. [Ber12a], Section 1.4). Again, the mappings T_μ and T are monotone and can be shown to be contraction mappings with respect to the unweighted sup-norm.

Example 1.2.4 (Discounted Zero-Sum Dynamic Games)

Let us consider a zero-sum game analog of the finite-state MDP Example 1.2.2. Here there are two players that choose actions at each stage: the first (called the *minimizer*) may choose a move i out of n moves and the second (called the *maximizer*) may choose a move j out of m moves. Then the minimizer gives a specified amount a_{ij} to the maximizer, called a *payoff*. The minimizer wishes to minimize a_{ij}, and the maximizer wishes to maximize a_{ij}.

The players use mixed strategies, whereby the minimizer selects a probability distribution $u = (u_1, \ldots, u_n)$ over his n possible moves and the maximizer selects a probability distribution $v = (v_1, \ldots, v_m)$ over his m possible moves. Thus the probability of selecting i and j is $u_i v_j$, and the expected payoff for this stage is $\sum_{i,j} a_{ij} u_i v_j$ or $u'Av$, where A is the $n \times m$ matrix with components a_{ij}.

In a single-stage version of the game, the minimizer must minimize $\max_{v \in V} u'Av$ and the maximizer must maximize $\min_{u \in U} u'Av$, where U and V are the sets of probability distributions over $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$, respectively. A fundamental result (which will not be proved here) is that these two values are equal:

$$\min_{u \in U} \max_{v \in V} u'Av = \max_{v \in V} \min_{u \in U} u'Av. \tag{1.13}$$

Let us consider the situation where a separate game of the type just described is played at each stage. The game played at a given stage is represented by a "state" x that takes values in a finite set X. The state evolves according to transition probabilities $q_{xy}(i,j)$ where i and j are the moves selected by the minimizer and the maximizer, respectively (here y represents

the next game to be played after moves i and j are chosen at the game represented by x). When the state is x, under $u \in U$ and $v \in V$, the one-stage expected payoff is $u'A(x)v$, where $A(x)$ is the $n \times m$ payoff matrix, and the state transition probabilities are

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

where Q_{xy} is the $n \times m$ matrix that has components $q_{xy}(i,j)$. Payoffs are discounted by $\alpha \in (0,1)$, and the objectives of the minimizer and maximizer, roughly speaking, are to minimize and to maximize the total discounted expected payoff. This requires selections of u and v to strike a balance between obtaining favorable current stage payoffs and playing favorable games in future stages.

We now introduce an abstract DP framework related to the sequential move selection process just described. We consider the mapping G given by

$$\begin{aligned} G(x,u,v,J) &= u'A(x)v + \alpha \sum_{y \in X} p_{xy}(u,v) J(y) \\ &= u' \left(A(x) + \alpha \sum_{y \in X} Q_{xy} J(y) \right) v, \end{aligned} \tag{1.14}$$

where $\alpha \in (0,1)$ is discount factor, and the mapping H given by

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

The corresponding mappings T_μ and T are

$$(T_\mu J)(x) = \max_{v \in V} G\big(x, \mu(x), v, J\big), \qquad x \in X,$$

and

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

It can be shown that T_μ and T are monotone and (unweighted) sup-norm contractions. Moreover, the unique fixed point J^* of T satisfies

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

(see [Ber12a], Section 1.6.2).

We now note that since

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

[cf. Eq. (1.14)] is a matrix that is independent of u and v, we may view $J^*(x)$ as the value of a static game (which depends on the state x). In particular, from the fundamental minimax equality (1.13), we have

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

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

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

where

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

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

In the preceding development, we have introduced J^* as the unique fixed point of the mappings T and $\overline{T}$. However, J^* also has an interpretation in game theoretic terms. In particular, it can be shown that $J^*(x)$ is the value of a dynamic game, whereby at state x the two opponents choose multistage (possibly nonstationary) policies that consist of functions of the current state, and continue to select moves using these policies over an infinite horizon. For further discussion of this interpretation, we refer to [Ber12a] and to books on dynamic games such as [FiV96]; see also [PaB99] and [Yu14] for an analysis of the undiscounted case ($\alpha = 1$) where there is a termination state, as in the stochastic shortest path problems of the subsequent Example 1.2.6. An alternative and more general formulation of sequential zero-sum games, which allows for an infinite state space, will be given in Chapter 5.

Example 1.2.5 (Minimax Problems)

Consider a minimax version of Example 1.2.1, where w is not random but is rather chosen from within a set $W(x, u)$ by an antagonistic opponent. Let

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

Then the equation $J^* = TJ^*$ is Bellman's equation for an infinite horizon minimax DP problem. A special case of this mapping arises in zero-sum dynamic games (cf. Example 1.2.4). We will also discuss alternative and more general abstract DP formulations of minimax problems in Chapter 5.

Example 1.2.6 (Stochastic Shortest Path Problems)

The stochastic shortest path (SSP for short) problem is the special case of the stochastic optimal control Example 1.2.1 where:

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

(b) The state space is $X = \{t, 1, \ldots, n\}$ and we are given transition probabilities, denoted by

$$p_{xy}(u) = P(x_{k+1} = y \mid x_k = x, u_k = u), \qquad x, y \in X,\ u \in U(x).$$

(c) The control constraint set $U(x)$ is finite for all $x \in X$.

(d) A cost $g(x,u)$ is incurred when control $u \in U(x)$ is selected at state x.

(e) State t is a special termination state, which is cost-free and absorbing, i.e., for all $u \in U(t)$,

$$g(t,u) = 0, \qquad p_{tt}(u) = 1.$$

To simplify the notation, we have assumed that the cost per stage does not depend on the successor state, which amounts to using expected cost per stage in all calculations.

Since the termination state t is cost-free, 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

$$H(x,u,J) = g(x,u) + \sum_{y=1}^{n} p_{xy}(u)J(y), \quad x = 1,\ldots,n,\ u \in U(x),\ J \in \Re^n.$$

The mappings T_μ and T are defined by

$$(T_\mu J)(x) = g\big(x,\mu(x)\big) + \sum_{y=1}^{n} p_{xy}\big(\mu(x)\big)J(y), \qquad x = 1,\ldots,n,$$

$$(TJ)(x) = \min_{u\in U(x)} \left[g(x,u) + \sum_{y=1}^{n} p_{xy}(u)J(y) \right], \qquad x = 1,\ldots,n.$$

Note that the matrix that has components $p_{xy}(u)$, $x,y = 1,\ldots,n$, is sub-stochastic (some of its row sums may be less than 1) because there may be a positive transition probability from a state x to the termination state t. Consequently T_μ may be a contraction for some μ, but not necessarily for all $\mu \in \mathcal{M}$.

The SSP problem has been discussed in many sources, including the books [Pal67], [Der70], [Whi82], [Ber87], [BeT89], [HeL99], [Ber12a], and [Ber17a], where it is sometimes referred to by earlier names such as "first passage problem" and "transient programming problem." In the framework that is most relevant to our purposes, given in the paper by Bertsekas and Tsitsiklis [BeT91], there is a classification of stationary policies for SSP into *proper* and *improper*. We say that $\mu \in \mathcal{M}$ is proper if, when using μ, there is positive probability that termination will be reached after at most n stages, regardless of the initial state; i.e., if

$$\rho_\mu = \max_{x=1,\ldots,n} P\{x_n \neq 0 \mid x_0 = x, \mu\} < 1.$$

Otherwise, we say that μ is improper. It can be seen that μ is proper if and only if in the Markov chain corresponding to μ, each state x is connected to the termination state with a path of positive probability transitions.

For a proper policy μ, it can be shown that T_μ is a weighted sup-norm contraction, as well as an n-stage contraction with respect to the unweighted

sup-norm. For an improper policy μ, T_μ is not a contraction with respect to any norm. Moreover, T also need not be a contraction with respect to any norm (think of the case where there is only one policy, which is improper). However, T is a weighted sup-norm contraction in the important special case where all policies are proper (see [BeT96], Prop. 2.2, or [Ber12a], Chapter 3).

Nonetheless, even in the case where there are improper policies and T is not a contraction, results comparable to the case of discounted finite-state MDP are available for SSP problems assuming that:

(a) There exists at least one proper policy.

(b) For every improper policy there is an initial state that has infinite cost under this policy.

Under the preceding two assumptions, referred to as the *strong SSP conditions* in Section 3.5.1, it was shown in [BeT91] that T has a unique fixed point J^*, the optimal cost function of the SSP problem. Moreover, a policy $\{\mu^*, \mu^*, \ldots\}$ is optimal if and only if

$$T_{\mu^*} J^* = T J^*.$$

In addition, J^* and J_μ can be computed by value iteration,

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

starting with any $J \in \Re^n$ (see [Ber12a], Chapter 3, for a textbook account). These properties are in analogy with the desirable properties (a)-(c), given at the end of the preceding subsection in connection with contractive models.

Regarding policy iteration, it works in its strongest form when there are no improper policies, in which case the mappings T_μ and T are weighted sup-norm contractions. When there are improper policies, modifications to the policy iteration method are needed; see [Ber12a], [YuB13a], and also Section 3.6.2, where these modifications will be discussed in an abstract setting.

In Section 3.5.1 we will also consider SSP problems where the strong SSP conditions (a) and (b) above are not satisfied. Then we will see that unusual phenomena can occur, including that J^* may not be a solution of Bellman's equation. Still our line of analysis of Chapter 3 will apply to such problems.

Example 1.2.7 (Deterministic Shortest Path Problems)

The special case of the SSP problem where the state transitions are deterministic is the classical shortest path problem. Here, we have a graph of n nodes $x = 1, \ldots, n$, plus a destination t, and an arc length a_{xy} for each directed arc (x, y). At state/node x, a policy μ chooses an outgoing arc from x. Thus the controls available at x can be identified with the outgoing neighbors of x [the nodes u such that (x, u) is an arc]. The corresponding mapping H is

$$H(x, u, J) = \begin{cases} a_{xu} + J(u) & \text{if } u \neq t, \\ a_{xt} & \text{if } u = t, \end{cases} \qquad x = 1, \ldots, n.$$

A stationary policy μ defines a graph whose arcs are $\big(x, \mu(x)\big)$, $x = 1, \ldots, n$. The policy μ is proper if and only if this graph is acyclic (it consists of

a tree of directed paths leading from each node to the destination). Thus there exists a proper policy if and only if each node is connected to the destination with a directed path. Furthermore, an improper policy has finite cost starting from every initial state if and only if all the cycles of the corresponding graph have nonnegative cycle cost. It follows that the favorable analytical and algorithmic results described for SSP in the preceding example hold if the given graph is connected and the costs of all its cycles are positive. We will see later that significant complications result if the cycle costs are allowed to be zero, even though the shortest path problem is still well posed in the sense that shortest paths exist if the given graph is connected (see Section 3.1).

Example 1.2.8 (Multiplicative and Risk-Sensitive Models)

With x, y, u, and transition probabilities $p_{xy}(u)$, as in the finite-state MDP of Example 1.2.2, consider the mapping

$$H(x,u,J) = \sum_{y \in X} p_{xy}(u) g(x,u,y) J(y) = E\Big\{ g(x,u,y) J(y) \mid x, u \Big\}, \tag{1.15}$$

where g is a scalar function satisfying $g(x,u,y) \geq 0$ for all x, y, u (this is necessary for H to be monotone). This mapping corresponds to the multiplicative model of minimizing over all $\pi = \{\mu_0, \mu_1, \ldots\}$ the cost

$$\begin{aligned} J_\pi(x_0) = \limsup_{N \to \infty} E\Big\{ & g\big(x_0, \mu_0(x_0), x_1\big) g\big(x_1, \mu_1(x_1), x_2\big) \cdots \\ & g\big(x_{N-1}, \mu_{N-1}(x_{N-1}), x_N\big) \mid x_0 \Big\}, \end{aligned} \tag{1.16}$$

where the state sequence $\{x_0, x_1, \ldots\}$ is generated using the transition probabilities $p_{x_k x_{k+1}}\big(\mu_k(x_k)\big)$.

To see that the mapping H of Eq. (1.15) corresponds to the cost function (1.16), let us consider the unit function

$$\bar{J}(x) \equiv 1, \qquad x \in X,$$

and verify that for all $x_0 \in X$, we have

$$\begin{aligned} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} \bar{J})(x_0) = E\Big\{ & g\big(x_0, \mu_0(x_0), x_1\big) g\big(x_1, \mu_1(x_1), x_2\big) \cdots \\ & g\big(x_{N-1}, \mu_{N-1}(x_{N-1}), x_N\big) \mid x_0 \Big\}, \end{aligned} \tag{1.17}$$

so that

$$J_\pi(x) = \limsup_{N \to \infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} \bar{J})(x), \qquad x \in X.$$

Indeed, taking into account that $\bar{J}(x) \equiv 1$, we have

$$\begin{aligned} (T_{\mu_{N-1}} \bar{J})(x_{N-1}) &= E\Big\{ g\big(x_{N-1}, \mu_{N-1}(x_{N-1}), x_N\big) \bar{J}(x_N) \mid x_{N-1} \Big\} \\ &= E\Big\{ g\big(x_{N-1}, \mu_{N-1}(x_{N-1}), x_N\big) \mid x_{N-1} \Big\}, \end{aligned}$$

$$(T_{\mu_{N-2}}T_{\mu_{N-1}}\bar{J})(x_{N-2}) = \big((T_{\mu_{N-2}}(T_{\mu_{N-1}}\bar{J})\big)(x_{N-2})$$
$$= E\Big\{g\big(x_{N-2},\mu_{N-2}(x_{N-2}),x_{N-1}\big)$$
$$\cdot E\Big\{g\big(x_{N-1},\mu_{N-1}(x_{N-1}),x_N\big) \mid x_{N-1}\Big\} \mid x_{N-2}\Big\},$$

and continuing similarly,

$$(T_{\mu_0}T_{\mu_1}\cdots T_{\mu_{N-1}}\bar{J})(x_0) = E\Big\{g\big(x_0,\mu_0(x_0),x_1\big)E\Big\{g\big(x_1,\mu_1(x_1),x_2\big)\cdots$$
$$E\Big\{g\big(x_{N-1},\mu_{N-1}(x_{N-1}),x_N\big) \mid x_{N-1}\Big\} \mid x_{N-2}\Big\}\cdots\Big\} \mid x_0\Big\},$$

which by using the iterated expectations formula (see e.g., [BeT08]) proves the expression (1.17).

An important special case of a multiplicative model is when g has the form

$$g(x,u,y) = e^{h(x,u,y)}$$

for some one-stage cost function h. We then obtain a finite-state MDP with an exponential cost function,

$$J_\pi(x_0) = \limsup_{N\to\infty} E\Big\{e^{\left(h(x_0,\mu_0(x_0),x_1)+\cdots+h(x_{N-1},\mu_{N-1}(x_{N-1}),x_N)\right)}\Big\},$$

which is often used to introduce risk aversion in the choice of policy through the convexity of the exponential.

There is also a multiplicative version of the infinite state space stochastic optimal control problem of Example 1.2.1. The mapping H takes the form

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

where $x_{k+1} = f(x_k,u_k,w_k)$ is the underlying discrete-time dynamic system; cf. Eq. (1.10).

Multiplicative models and related risk-sensitive models are discussed extensively in the literature, mostly for the exponential cost case and under different assumptions than ours; see e.g., [HoM72], [Jac73], [Rot84], [ChS87], [Whi90], [JBE94], [FlM95], [HeM96], [FeM97], [BoM99], [CoM99], [BoM02], [BBB08], [Ber16a]. The works of references [DeR79], [Pat01], and [Pat07] relate to the stochastic shortest path problems of Example 1.2.6, and are the closest to the semicontractive models discussed in Chapters 3 and 4, based on the author's paper [Ber16a]; see the next example and Section 3.5.2.

Example 1.2.9 (Affine Monotonic Models)

Consider a finite state space $X = \{1,\ldots,n\}$ and a (possibly infinite) control constraint set $U(x)$ for each state x. For each policy μ, let the mapping T_μ be given by

$$T_\mu J = b_\mu + A_\mu J, \tag{1.18}$$

where b_μ is a vector of $\Re^n$ with components $b\big(x,\mu(x)\big)$, $x = 1,\ldots,n$, and A_μ is an $n\times n$ matrix with components $A_{xy}\big(\mu(x)\big)$, $x,y = 1,\ldots,n$. We assume that $b(x,u)$ and $A_{xy}(u)$ are nonnegative,

$$b(x,u) \geq 0, \qquad A_{xy}(u) \geq 0, \qquad \forall\ x,y = 1,\ldots,n,\ u\in U(x).$$

Thus T_μ and T map nonnegative functions to nonnegative functions $J : X \mapsto [0, \infty]$.

This model was introduced in the first edition of this book, and was elaborated on in the author's paper [Ber16a]. Special cases of the model include the finite-state Markov and semi-Markov problems of Examples 1.2.1-1.2.3, and the stochastic shortest path problem of Example 1.2.6, with A_μ being the transition probability matrix of μ (perhaps appropriately discounted), and b_μ being the cost per stage vector of μ, which is assumed nonnegative. An interesting affine monotonic model of a different type is the multiplicative cost model of the preceding example, where the initial function is $\bar{J}(x) \equiv 1$ and the cost accumulates multiplicatively up to reaching a termination state t. In the exponential case of this model, the cost of a generated path starting from some initial state accumulates additively as in the SSP case, up to reaching t. However, the cost of the model is the expected value of the *exponentiated* cost of the path up to reaching t. It can be shown then that the mapping T_μ has the form

$$(T_\mu J)(x) = p_{xt}\big(\mu(x)\big)\exp\big(g(x,\mu(x),t)\big) + \sum_{y=1}^{n} p_{xy}(\mu(x))\exp\big(g(x,\mu(x),y)\big)J(y), \qquad x \in X,$$

where $p_{xy}(u)$ is the probability of transition from x to y under u, and $g(x, u, y)$ is the cost of the transition; see Section 3.5.2 for a detailed derivation. Clearly T_μ has the affine monotonic form (1.18).

Example 1.2.10 (Aggregation)

Aggregation is an approximation approach that simplifies a large dynamic programming (DP) problem by "combining" multiple states into aggregate states. This results in a reduced or "aggregate" problem with fewer states, which can often be solved using exact DP methods. The optimal cost-to-go function derived from this aggregate problem then serves as an approximation of the optimal cost function for the original problem.

Consider an n-state Markovian decision problem with transition probabilities $p_{ij}(u)$. To construct an aggregation framework, we introduce a finite set $\mathcal{A}$ of aggregate states. We generically denote the aggregate states by letters such as x and y, and the original system states by letters such as i and j. The approximation framework is constructed by combining in various ways the aggregate states and the original system states to form a larger system (see Fig. 1.2.2). To specify the probabilistic structure of this system, we introduce two (somewhat arbitrary) choices of probability distributions, which relate the original system states with the aggregate states:

(1) For each aggregate state x and original system state i, we specify the *disaggregation probability* d_{xi}. We assume that $d_{xi} \geq 0$ and

$$\sum_{i=1}^{n} d_{xi} = 1, \qquad \forall\ x \in \mathcal{A}.$$

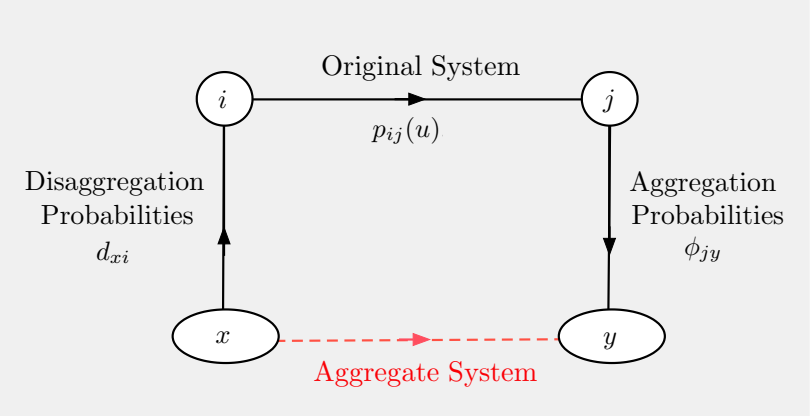


Figure 1.2.2 Illustration of the relation between aggregate and original system states.

Roughly, d_{xi} may be interpreted as the "degree to which x is represented by i."

(2) For each aggregate state y and original system state j, we specify the *aggregation probability* ϕ_{jy}. We assume that $\phi_{jy} \geq 0$ and

$$\sum_{y \in \mathcal{A}} \phi_{jy} = 1, \qquad \forall\, j = 1, \ldots, n.$$

Roughly, ϕ_{jy} may be interpreted as the "degree of membership of j in the aggregate state y."

The aggregation and disaggregation probabilities specify a dynamic system involving both aggregate and original system states (cf. Fig. 1.2.2). In this system:

(i) From aggregate state x, we generate original system state i according to d_{xi}.

(ii) We generate transitions from original system state i to original system state j according to $p_{ij}(u)$, with cost $g(i, u, j)$.

(iii) From original system state j, we generate aggregate state y according to ϕ_{jy}.

Illustrative examples of aggregation frameworks are given in the books [Ber12a] and [Ber17a]. One possibility is *hard aggregation*, where aggregate states are identified with the sets of a partition of the state space. For another type of common scheme, think of the case where the original system states form a fine grid in some space, which is "aggregated" into a much coarser grid. In particular let us choose a collection of "representative" original system states, and associate each one of them with an aggregate state. Thus, each aggregate state x is associated with a unique representative state i_x, and the

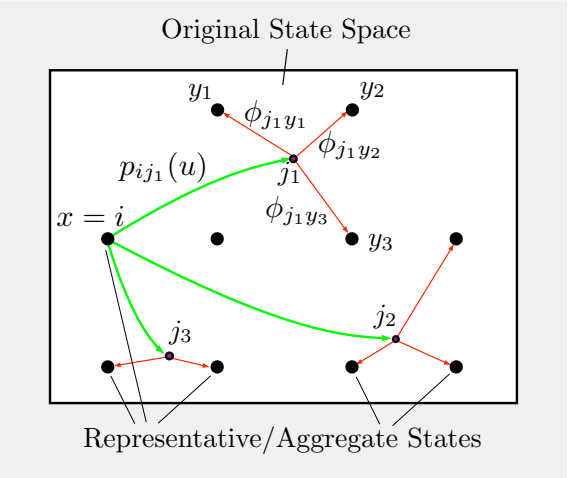


Figure 1.2.3 Aggregation based on a small subset of representative states (these are shown with larger dark circles, while the other (nonrepresentative) states are shown with smaller dark circles). In this figure, from representative state $x = i$, there are three possible transitions, to states j_1, j_2, and j_3, according to $p_{ij_1}(u), p_{ij_2}(u), p_{ij_3}(u)$, and each of these states is associated with a convex combination of representative states using the aggregation probabilities. For example, j_1 is associated with $\phi_{j_1 y_1} y_1 + \phi_{j_1 y_2} y_2 + \phi_{j_1 y_3} y_3$.

disaggregation probabilities are

$$d_{xi} = \begin{cases} 1 & \text{if } i = i_x, \\ 0 & \text{if } i \neq i_x. \end{cases} \tag{1.19}$$

The aggregation probabilities are chosen to represent each original system state j with a convex combination of aggregate/representative states; see Fig. 1.2.3. It is also natural to assume that the aggregation probabilities map representative states to themselves, i.e.,

$$\phi_{jy} = \begin{cases} 1 & \text{if } j = j_y, \\ 0 & \text{if } j \neq j_y. \end{cases}$$

This scheme makes intuitive geometrical sense as an interpolation scheme in the special case where both the original and the aggregate states are associated with points in a Euclidean space. The scheme may also be extended to problems with a continuous state space. In this case, the state space is discretized with a finite grid, and the states of the grid are viewed as the aggregate states. The disaggregation probabilities are still given by Eq. (1.19), while the aggregation probabilities may be arbitrarily chosen to represent each original system state with a convex combination of representative states.

As an extension of the preceding schemes, suppose that through some special insight into the problem's structure or some preliminary calculation, we know some features of the system's state that can "predict well" its cost. Then it seems reasonable to form the aggregate states by grouping together

states with "similar features," or to form aggregate states by using "representative features" instead of representative states. This is called "feature-based aggregation;" see the books [BeT96] (Section 3.1) and [Ber12a] (Section 6.5) for a description and analysis.

Given aggregation and disaggregation probabilities, we may define an *aggregate problem* whose states are the aggregate states. This problem involves an aggregate discrete-time system, which we will describe shortly. We require that the control is applied with knowledge of the current aggregate state only (rather than the original system state).† To this end, we assume that the control constraint set $U(i)$ is independent of the state i, and we denote it by U. Then, by adding the probabilities of all the relevant paths in Fig. 1.2.2, it can be seen that the transition probability from aggregate state x to aggregate state y under control $u \in U$ is

$$\hat{p}_{xy}(u) = \sum_{i=1}^{n} d_{xi} \sum_{j=1}^{n} p_{ij}(u)\phi_{jy}.$$

The corresponding expected transition cost is given by

$$\hat{g}(x,u) = \sum_{i=1}^{n} d_{xi} \sum_{j=1}^{n} p_{ij}(u)g(i,u,j).$$

These transition probabilities and costs define the aggregate problem.

We may compute the optimal costs-to-go $\hat{J}(x)$, $x \in \mathcal{A}$, of this problem by using some exact DP method. Then, the costs-to-go of each state j of the original problem are usually approximated by

$$\tilde{J}(j) = \sum_{y \in \mathcal{A}} \phi_{jy}\hat{J}(y).$$

Example 1.2.11 (Distributed Aggregation)

The abstract DP framework is useful not only in modeling DP problems, but also in modeling algorithms arising in DP and even other contexts. We illustrate this with an example from Bertsekas and Yu [BeY10] that relates to the distributed solution of large-scale discounted finite-state MDP using cost function approximation based on aggregation. ‡ It involves a partition of the n states into m subsets for the purposes of distributed computation, and yields a corresponding approximation $(V_1, \ldots, V_m)$ to the cost vector J^*.

In particular, we have a discounted n-state MDP (cf. Example 1.2.2), and we introduce aggregate states $S_1, \ldots, S_m$, which are disjoint subsets of

† An alternative form of aggregate problem, where the control may depend on the original system state is discussed in Section 6.5.2 of the book [Ber12a].

‡ See [Ber12a], Section 6.5.2, for a more detailed discussion. Other examples of algorithmic mappings that come under our framework arise in asynchronous policy iteration (see Sections 2.6.3, 3.6.2, and [BeY10], [BeY12], [YuB13a]), and in constrained forms of policy iteration (see [Ber11c], or [Ber12a], Exercise 2.7).

the original state space with $S_1 \cup \cdots \cup S_n = \{1, \ldots, n\}$. We envision a network of processors $\ell = 1, \ldots, m$, each assigned to the computation of a local cost function V_ℓ, defined on the corresponding aggregate state/subset S_ℓ:

$$V_\ell = \{V_{\ell y} \mid y \in S_\ell\}.$$

Processor ℓ also maintains a scalar aggregate cost R_ℓ for its aggregate state, which is a weighted average of the detailed cost values $V_{\ell x}$ within S_ℓ:

$$R_\ell = \sum_{x \in S_\ell} d_{\ell x} V_{\ell x},$$

where $d_{\ell x}$ are given probabilities with $d_{\ell x} \geq 0$ and $\sum_{x \in S_\ell} d_{\ell x} = 1$. The aggregate costs R_ℓ are communicated between processors and are used to perform the computation of the local cost functions V_ℓ (we will discuss computation models of this type in Section 2.6).

We denote $J = (V_1, \ldots, V_m, R_1, \ldots, R_m)$. We introduce the mapping $H(x, u, J)$ defined for each of the n states x by

$$H(x, u, J) = W_\ell(x, u, V_\ell, R_1, \ldots, R_m), \qquad \text{if } x \in S_\ell,$$

where for $x \in S_\ell$

$$W_\ell(x, u, V_\ell, R_1, \ldots, R_m) = \sum_{y=1}^{n} p_{xy}(u) g(x, u, y) + \alpha \sum_{y \in S_\ell} p_{xy}(u) V_{\ell y} + \alpha \sum_{y \notin S_\ell} p_{xy}(u) R_{s(y)},$$

and for each original system state y, we denote by $s(y)$ the index of the subset to which y belongs [i.e., $y \in S_{s(y)}$].

We may view H as an abstract mapping on the space of J, and aim to find its fixed point $J^* = (V_1^*, \ldots, V_m^*, R_1^*, \ldots, R_m^*)$. Then, for $\ell = 1, \ldots, m$, we may view V_ℓ^* as an approximation to the optimal cost vector of the original MDP starting at states $x \in S_\ell$, and we may view R_ℓ^* as a form of aggregate cost for S_ℓ. The advantage of this formulation is that it involves significant decomposition and parallelization of the computations among the processors, when performing various DP algorithms. In particular, the computation of $W_\ell(x, u, V_\ell, R_1, \ldots, R_m)$ depends on just the local vector V_ℓ, whose dimension may be potentially much smaller than n.

1.2.4 Reinforcement Learning - Projected and Aggregation Bellman Equations

Given an abstract DP model described by a mapping H, we may be interested in fixed points of related mappings other than T and T_μ. Such mappings may arise in various contexts, such as for example distributed

asynchronous aggregation in Example 1.2.11. An important context is *subspace approximation*, whereby T_μ and T are restricted onto a subspace of functions for the purpose of approximating their fixed points. Much of the theory of approximate DP, neuro-dynamic programming, and reinforcement learning relies on such approximations (there are quite a few books, which collectively contain extensive accounts these subjects, such as Bertsekas and Tsitsiklis [BeT96], Sutton and Barto [SuB98], Gosavi [Gos03], Cao [Cao07], Chang, Fu, Hu, and Marcus [CFH07], Meyn [Mey07], Powell [Pow07], Borkar [Bor08], Haykin [Hay08], Busoniu, Babuska, De Schutter, and Ernst [BBD10], Szepesvari [Sze10], Bertsekas [Ber12a], [Ber17a], [Ber19b], [Ber20], and Vrabie, Vamvoudakis, and Lewis [VVL13]).

For an illustration, consider the approximate evaluation of the cost vector of a discrete-time Markov chain with states $i = 1, \ldots, n$. We assume that state transitions (i, j) occur at time k according to given transition probabilities p_{ij}, and generate a cost $\alpha^k g(i, j)$, where $\alpha \in (0, 1)$ is a discount factor. The cost function over an infinite number of stages can be shown to be the unique fixed point of the Bellman equation mapping $T : \Re^n \mapsto \Re^n$ whose components are given by

$$(TJ)(i) = \sum_{j=1}^{n} p_{ij}(u)\big(g(i, j) + \alpha J(j)\big), \qquad i = 1, \ldots, n, \ J \in \Re^n.$$

This is the same as the mapping T in the discounted finite-state MDP Example 1.2.2, except that we restrict attention to a single policy. Finding the cost function of a fixed policy is the important policy evaluation subproblem that arises prominently within the context of policy iteration. It also arises in the context of a simplified form of policy iteration, the *rollout algorithm*; see e.g., [BeT96], [Ber12a], [Ber17a], [Ber19b], [Ber20]. In some artificial intelligence contexts, policy iteration is referred to as *self-learning*, and in these contexts the policy evaluation is almost always done approximately, sometimes with the use of neural networks.

A prominent approach for approximation of the fixed point of T is based on the solution of lower-dimensional equations defined on the subspace $\{\Phi r \mid r \in \Re^s\}$ that is spanned by the columns of a given $n \times s$ matrix Φ. Two such approximating equations have been studied extensively (see [Ber12a], Chapter 6, for a detailed account and references; also [BeY07], [BeY09], [YuB10], [Ber11a] for extensions to abstract contexts beyond approximate DP). These are:

(a) The *projected equation*

$$\Phi r = \Pi_\xi T(\Phi r), \tag{1.20}$$

where Π_ξ denotes projection onto S with respect to a weighted Euclidean norm

$$\|J\|_\xi = \sqrt{\sum_{i=1}^{n} \xi_i \big(J(i)\big)^2} \tag{1.21}$$

with $\xi = (\xi_1, \ldots, \xi_n)$ being a probability distribution with positive components (sometimes a seminorm projection is used, whereby some of the components ξ_i may be zero; see Yu and Bertsekas [YuB12]).

(b) The *aggregation equation*

$$\Phi r = \Phi DT(\Phi r), \tag{1.22}$$

with D being an $s \times n$ matrix whose rows are restricted to be probability distributions; these are the disaggregation probabilities of Example 1.2.10. Also, in this approach, the rows of Φ are restricted to be probability distributions; these are the aggregation probabilities of Example 1.2.10.

We now see that solving the projected equation (1.20) and the aggregation equation (1.22) amounts to finding a fixed point of the mappings $\Pi_\xi T$ and ΦDT, respectively. These mappings derive their structure from the DP operator T, so they have some DP-like properties, which can be exploited for analysis and computation.

An important fact is that the aggregation mapping ΦDT preserves the monotonicity and the sup-norm contraction property of T, while the projected equation mapping $\Pi_\xi T$ generally does not. The reason for preservation of monotonicity is the nonnegativity of the components of the matrices Φ and D (see the author's survey paper [Ber11c] for a discussion of the importance of preservation of monotonicity in various DP operations). The reason for preservation of sup-norm contraction is that the matrices Φ and D are sup-norm nonexpansive, because their rows are probability distributions. In fact, it can be verified that the solution r of Eq. (1.22) can be viewed as the *exact* DP solution of the "aggregate" DP problem that represents a lower-dimensional approximation of the original (see Example 1.2.10). The preceding observations are important for our purposes, as they indicate that much of the theory developed in this book applies to approximation-related mappings based on aggregation.

By contrast, the projected equation mapping $\Pi_\xi T$ need not be monotone, because the components of Π_ξ need not be nonnegative. Moreover while the projection Π_ξ is nonexpansive with respect to the projection norm $\|\cdot\|_\xi$, it need not be nonexpansive with respect to the sup-norm. As a result the projected equation mapping $\Pi_\xi T$ need not be a sup-norm contraction. These facts play a significant role in approximate DP methodology.

1.2.5 Reinforcement Learning - Temporal Difference and Proximal Algorithms

An important possibility for finding a fixed point of T is to replace T with another mapping, say F, such that F and T have the same fixed points. For example, F may offer some advantages in terms of algorithmic convenience or quality of approximation when used in conjunction with

projection or aggregation [cf. Eqs. (1.20) and (1.22)]. Alternatively, F may be the mapping of some iterative method that is suitable for computing fixed points of T.

In this book we will not consider in much detail the possibility of using an alternative mapping F to find a fixed point of a mapping T. We will just mention here some multistep versions of T, which have been used widely for approximations in reinforcement learning. An important example is the mapping $T^{(\lambda)} : \Re^n \mapsto \Re^n$, defined for a given $\lambda \in (0,1)$ as follows: $T^{(\lambda)}$ transforms a vector $J \in \Re^n$ to the vector $T^{(\lambda)}J \in \Re^n$, whose n components are given by

$$\big(T^{(\lambda)}J\big)(i) = (1-\lambda)\sum_{\ell=0}^{\infty}\lambda^{\ell}(T^{\ell+1}J)(i), \qquad i = 1,\ldots,n,\ J \in \Re^n,$$

for $\lambda \in (0,1)$, where T^{ℓ} is the ℓ-fold composition of T with itself ℓ times. Here there should be conditions that guarantee the convergence of the infinite series in the preceding definition. The multistep analog of the projected Eq. (1.20) is

$$\Phi r = \Pi_{\xi} T^{(\lambda)}(\Phi r).$$

The popular temporal difference methods, such as TD(λ), LSTD(λ), and LSPE(λ), aim to solve this equation (see the book references on approximate DP, neuro-dynamic programming, and reinforcement learning cited earlier). The mapping $T^{(\lambda)}$ also forms the basis for the λ-policy iteration method to be discussed in Sections 2.5, 3.2.4, and 4.3.3.

The multistep analog of the aggregation Eq. (1.22) is

$$\Phi r = \Phi D T^{(\lambda)}(\Phi r),$$

and methods that are similar to the temporal difference methods can be used for its solution. In particular, a multistep method based on the mapping $T^{(\lambda)}$ is the, so-called, λ-aggregation method (see [Ber12a], Chapter 6), as well as other forms of aggregation (see [Ber12a], [YuB12]).

In the case where T is a linear mapping of the form

$$TJ = AJ + b,$$

where b is a vector in $\Re^n$, and A is an $n \times n$ matrix with eigenvalues strictly within the unit circle, there is an interesting connection between the multistep mapping $T^{(\lambda)}$ and another mapping of major importance in numerical convex optimization. This is the *proximal mapping*, associated with T and a scalar $c > 0$, and denoted by $P^{(c)}$. In particular, for a given $J \in \Re^n$, the vector $P^{(c)}J$ is defined as the unique vector $Y \in \Re^n$ that solves the equation

$$Y - AY - b = \frac{1}{c}(J - Y).$$

Equivalently,

$$P^{(c)}J = \left(\frac{c+1}{c}I - A\right)^{-1}\left(b + \frac{1}{c}J\right), \tag{1.23}$$

where I is the identity matrix. Then it can be shown (see Exercise 1.2 or the papers [Ber16b], [Ber18c]) that if

$$c = \frac{\lambda}{1-\lambda},$$

we have

$$T^{(\lambda)} = T \cdot P^{(c)} = P^{(c)} \cdot T.$$

Moreover, the vectors J, $P^{(c)}J$, and $T^{(\lambda)}J$ are colinear and satisfy

$$T^{(\lambda)}J = J + \frac{c+1}{c}\big(P^{(c)}J - J\big).$$

The preceding formulas show that $T^{(\lambda)}$ and $P^{(c)}$ are closely related, and that iterating with $T^{(\lambda)}$ is "faster" than iterating with $P^{(c)}$, since the eigenvalues of A are within the unit circle, so that T is a contraction. In addition, methods such as TD(λ), LSTD(λ), LSPE(λ), and their projected versions, which are based on $T^{(\lambda)}$, can be adapted to be used with $P^{(c)}$.

A more general form of multistep approach, introduced and studied in the paper [YuB12], replaces $T^{(\lambda)}$ with a mapping $T^{(w)} : \Re^n \mapsto \Re^n$ that has components

$$\big(T^{(w)}J\big)(i) = \sum_{\ell=1}^{\infty} w_{i\ell}(T^\ell J)(i), \qquad i = 1, \ldots, n,\ J \in \Re^n,$$

where w is a vector sequence whose ith component, $(w_{i1}, w_{i2}, \ldots)$, is a probability distribution over the positive integers. Then the multistep analog of the projected equation (1.20) is

$$\Phi r = \Pi_\xi T^{(w)}(\Phi r), \tag{1.24}$$

while the multistep analog of the aggregation equation (1.22) is

$$\Phi r = \Phi D T^{(w)}(\Phi r). \tag{1.25}$$

The mapping $T^{(\lambda)}$ is obtained for $w_{i\ell} = (1-\lambda)\lambda^{\ell-1}$, independently of the state i. A more general version, where λ depends on the state i, is obtained for $w_{i\ell} = (1-\lambda_i)\lambda_i^{\ell-1}$. The solution of Eqs. (1.24) and (1.25) by simulation-based methods is discussed in the paper [YuB12]; see also Exercise 1.3.

Let us also note that there is a connection between projected equations of the form (1.24) and aggregation equations of the form (1.25). This connection is based on the use of a seminorm [this is given by the same expression as the norm $\|\cdot\|_\xi$ of Eq. (1.21), with some of the components of ξ allowed to be 0]. In particular, the most prominent cases of aggregation equations can be viewed as seminorm projected equations because, for these cases, ΦD is a seminorm projection (see [Ber12a], p. 639, [YuB12], Section 4). Moreover, they can also be viewed as projected equations where the projection is oblique (see [Ber12a], Section 7.3.6).

1.3 REINFORCEMENT LEARNING - APPROXIMATION IN VALUE SPACE

In this section we will use geometric illustrations to obtain insight into Bellman's equation, the algorithms of value iteration (VI) and policy iteration (PI), and an approximation methodology, which is prominent in reinforcement learning and is known as *approximation in value space*.† Throughout this section, we will make use of the following two properties:

(a) T and T_μ are monotone, i.e., they satisfy Assumption 1.2.1.

(b) We have

$$(TJ)(x) = \min_{\mu \in \mathcal{M}} (T_\mu J)(x), \qquad \text{for all } x, \tag{1.26}$$

where $\mathcal{M}$ is the set of stationary policies. This is true because for any policy μ, there is no coupling constraint between the controls $\mu(x)$ and $\mu(x')$ that correspond to two different states x and x'.

We will first focus on the discounted version of the Markovian decision problem of Example 1.2.1, and we will then consider more general cases.

1.3.1 Approximation in Value Space for Markovian Decision Problems

In Markovian decision problems the mappings T_μ and T are given by

$$(T_\mu J)(x) = E\Big\{ g\big(x, \mu(x), w\big) + \alpha J\big(f(x, \mu(x), w)\big) \Big\}, \qquad \text{for all } x, \tag{1.27}$$

and

$$(TJ)(x) = \inf_{u \in U(x)} E\Big\{ g(x, u, w) + \alpha J\big(f(x, u, w)\big) \Big\}, \qquad \text{for all } x, \tag{1.28}$$

where $\alpha \in (0, 1]$; cf. Example 1.2.1.

In addition to monotonicity, we have an additional important property: T_μ *is linear*, in the sense that it has the form

$$T_\mu J = G_\mu + A_\mu J,$$

where $G_\mu \in \mathcal{R}(X)$ is some function and $A_\mu : \mathcal{R}(X) \mapsto \mathcal{R}(X)$ is an operator such that for any functions J_1, J_2, and scalars γ_1, γ_2, we have

$$A_\mu(\gamma_1 J_1 + \gamma_2 J_2) = \gamma_1 A_\mu J_1 + \gamma_2 A_\mu J_2.$$

† The major alternative reinforcement learning approach is *approximation in policy space*, whereby a suboptimal policy is selected from within a class of parametrized policies, usually by means of some optimization procedure, such as random search, or gradient descent; see e.g., the author's reinforcement learning book [Ber19b].

This is true because of the linearity of the expected value operation in Eq. (1.27). The linearity of T_μ implies another important property: $(TJ)(x)$ *is a concave function of* J *for every* x. By this we mean that the set

$$C_x = \big\{(J,\xi) \mid (TJ)(x) \geq \xi,\ J \in \mathcal{R}(X),\ \xi \in \Re\big\} \tag{1.29}$$

is convex for all $x \in X$, where $\mathcal{R}(X)$ is the set of real-valued functions over the state space X, and $\Re$ is the set of real numbers. This follows from the linearity of T_μ, the alternative definition of T given by Eq. (1.26), and the fact that for a fixed x, the minimum of the linear functions $(T_\mu J)(x)$ over $\mu \in \mathcal{M}$ is concave as a function of J.

We illustrate these properties graphically with an example.

Example 1.3.1 (A Two-State and Two-Control Example)

Assume that there are two states 1 and 2, and two controls u and v. Consider the policy μ that applies control u at state 1 and control v at state 2. Then the operator T_μ takes the form

$$(T_\mu J)(1) = \sum_{y=1}^{2} p_{1j}(u)\big(g(1,u,y) + \alpha J(y)\big), \tag{1.30}$$

$$(T_\mu J)(2) = \sum_{y=1}^{2} p_{2y}(v)\big(g(2,v,y) + \alpha J(y)\big), \tag{1.31}$$

where $p_{xy}(u)$ and $p_{xy}(v)$ are the probabilities that the next state will be y, when the current state is x, and the control is u or v, respectively. Clearly, $(T_\mu J)(1)$ and $(T_\mu J)(2)$ are linear functions of J. Also the operator T of the Bellman equation $J = TJ$ takes the form

$$\begin{aligned}(TJ)(1) = \min \Bigg[&\sum_{y=1}^{2} p_{1y}(u)\big(g(1,u,y) + \alpha J(y)\big), \\ &\sum_{y=1}^{2} p_{1y}(v)\big(g(1,v,y) + \alpha J(y)\big)\Bigg],\end{aligned} \tag{1.32}$$

$$\begin{aligned}(TJ)(2) = \min \Bigg[&\sum_{y=1}^{2} p_{2y}(u)\big(g(2,u,y) + \alpha J(y)\big), \\ &\sum_{y=1}^{2} p_{2y}(v)\big(g(2,v,y) + \alpha J(y)\big)\Bigg].\end{aligned} \tag{1.33}$$

Thus, $(TJ)(1)$ and $(TJ)(2)$ are concave and piecewise linear as functions of the two-dimensional vector J (with two pieces; more generally, as many linear pieces as the number of controls). This concavity property holds in general

since $(TJ)(x)$ is the minimum of a collection of linear functions of J, one for each $u \in U(x)$. Figure 1.3.1 illustrates $(T_\mu J)(1)$ for the cases where $\mu(1) = u$ and $\mu(1) = v$, $(T_\mu J)(2)$ for the cases where $\mu(2) = u$ and $\mu(2) = v$, $(TJ)(1)$, and $(TJ)(2)$, as functions of $J = \big(J(1), J(2)\big)$.

Critical properties from the DP point of view are whether T and T_μ have fixed points; equivalently, whether the Bellman equations $J = TJ$ and $J = T_\mu J$ have solutions within the class of real-valued functions, and whether the set of solutions includes J^* and J_μ, respectively. It may thus be important to verify that T or T_μ are contraction mappings. This is true for example in the benign case of discounted problems ($\alpha < 1$) with bounded cost per stage. However, for undiscounted problems, asserting the contraction property of T or T_μ may be more complicated, and even impossible. In this book we will deal extensively with such questions and related issues regarding the solution set of the Bellman equation.

Geometrical Interpretations

We will now interpret the Bellman operators geometrically, starting with T_μ, which is linear as noted earlier. Figure 1.3.2 illustrates its form. Note here that the functions J and $T_\mu J$ are multidimensional. They have as many scalar components $J(x)$ and $(T_\mu J)(x)$, respectively, as there are states x, but they can only be shown projected onto one dimension. The cost function J_μ satisfies $J_\mu = T_\mu J_\mu$, so it is obtained from the intersection of the graph of $T_\mu J$ and the 45 degree line, when J_μ is real-valued. We interpret the situation where J_μ is not real-valued with lack of system stability under μ [so μ will be viewed as unstable if we have $J_\mu(x) = \infty$ for some initial states x]. For further discussion of stability issues, see the book [Ber22].

The form of the Bellman operator T is illustrated in Fig. 1.3.3. Again the functions J, J^*, TJ, $T_\mu J$, etc, are multidimensional, but they are shown projected onto one dimension. The Bellman equation $J = TJ$ may have one or many real-valued solutions. It may also have no real-valued solution in exceptional situations, as we will discuss later. The figure assumes that the Bellman equations $J = TJ$ and $J = T_\mu J$ have a unique real-valued solution, which is true if T and T_μ are contraction mappings, as is the case for discounted problems with bounded cost per stage. Otherwise, these equations may have no solution or multiple solutions within the class of real-valued functions. The equation $J = TJ$ typically has J^* as a solution, but may have more than one solution in cases where either $\alpha = 1$ or $\alpha < 1$, and the cost per stage is unbounded.

Example 1.3.2 (A Two-State and Infinite Controls Problem)

Let us consider the mapping T for a problem that involves two states, 1 and 2, but an infinite number of controls. In particular, the control space at both

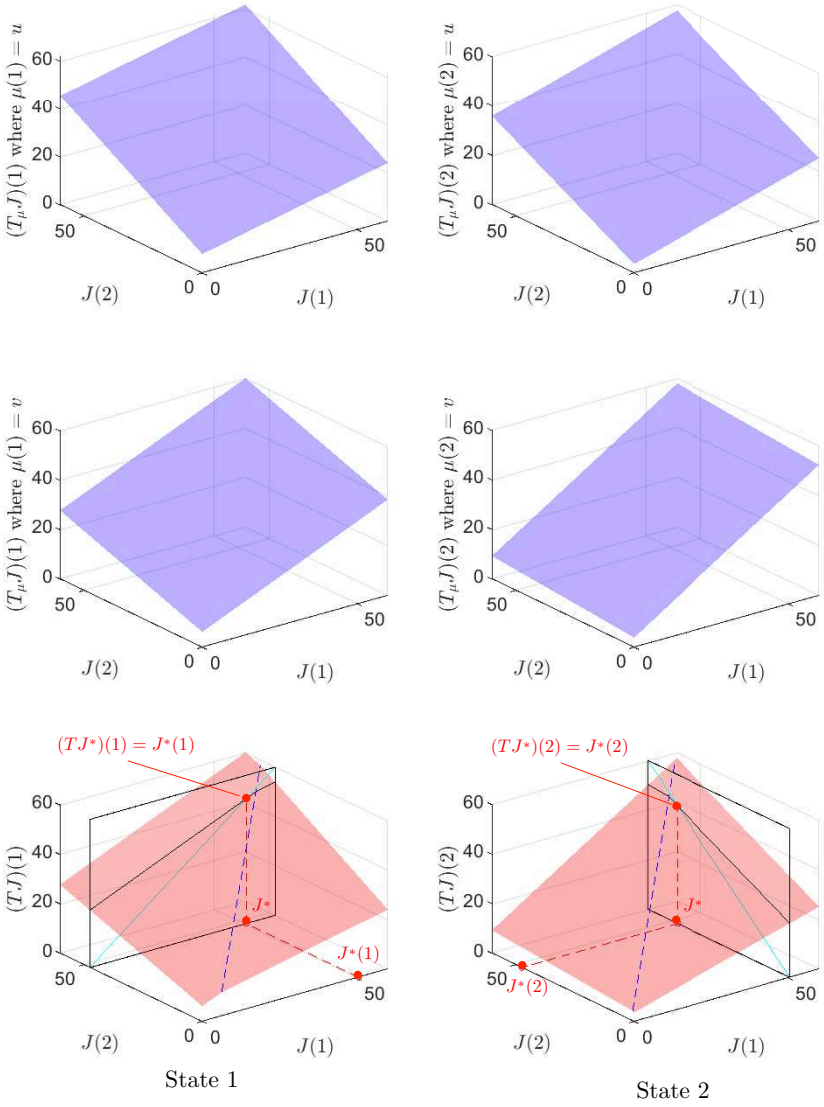


Figure 1.3.1 Geometric illustrations of the Bellman operators T_μ and T for states 1 and 2 in Example 1.3.1; cf. Eqs. (1.30)-(1.33). The problem's transition probabilities are: $p_{11}(u) = 0.3$, $p_{12}(u) = 0.7$, $p_{21}(u) = 0.4$, $p_{22}(u) = 0.6$, $p_{11}(v) = 0.6$, $p_{12}(v) = 0.4$, $p_{21}(v) = 0.9$, $p_{22}(v) = 0.1$. The stage costs are $g(1, u, 1) = 3$, $g(1, u, 2) = 10$, $g(2, u, 1) = 0$, $g(2, u, 2) = 6$, $g(1, v, 1) = 7$, $g(1, v, 2) = 5$, $g(2, v, 1) = 3$, $g(2, v, 2) = 12$. The discount factor is $\alpha = 0.9$, and the optimal costs are $J^*(1) = 50.59$ and $J^*(2) = 47.41$. The optimal policy is $\mu^*(1) = v$ and $\mu^*(2) = u$. The figure also shows the one-dimensional "slices" of T that pass through J^*.

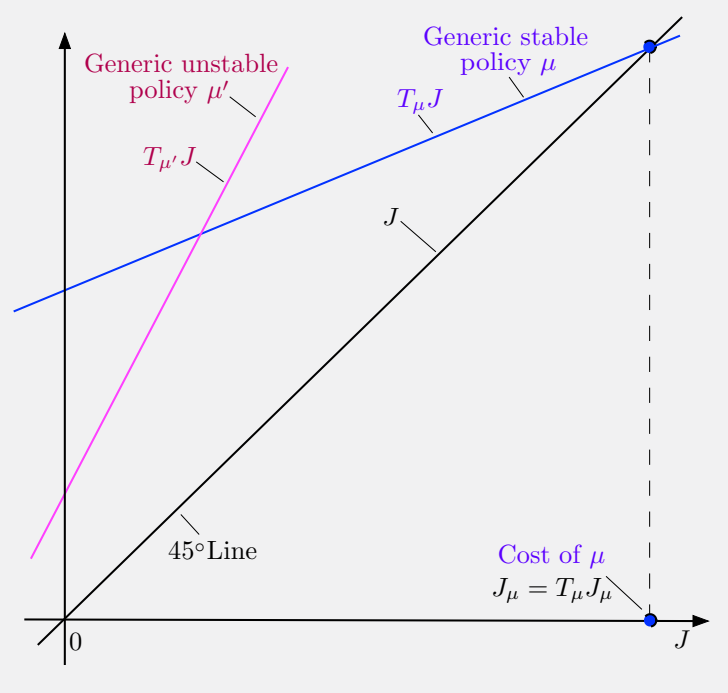


Figure 1.3.2 Geometric interpretation of the linear Bellman operator T_μ and the corresponding Bellman equation. The graph of T_μ is a plane in the space $\Re \times \Re$, and when projected on a one-dimensional plane that corresponds to a single state and passes through J_μ, it becomes a line. Then there are three cases:

(a) The line has slope less than 45 degrees, so it intersects the 45-degree line at a unique point, which is equal to J_μ, the solution of the Bellman equation $J = T_\mu J$. This is true if T_μ is a contraction mapping, as is the case for discounted problems with bounded cost per stage.

(b) The line has slope less than 45 degrees. Then it intersects the 45-degree line at a unique point, which is a solution of the Bellman equation $J = T_\mu J$, but is not equal to J_μ. Then J_μ is not real-valued; we consider such μ to be *unstable* under μ.

(c) The line has slope exactly equal to 45 degrees. This is an exceptional case where the Bellman equation $J = T_\mu J$ has an infinite number of real-valued solutions or no real-valued solution at all; we will provide examples where this occurs later.

states is the unit interval, $U(1) = U(2) = [0, 1]$. Here $(TJ)(1)$ and $(TJ)(2)$ are given by

$$(TJ)(1) = \min_{u \in [0,1]} \Big\{ g_1 + r_{11} u^2 + r_{12} (1-u)^2 + \alpha u J(1) + \alpha (1-u) J(2) \Big\},$$

$$(TJ)(2) = \min_{u \in [0,1]} \Big\{ g_2 + r_{21} u^2 + r_{22} (1-u)^2 + \alpha u J(1) + \alpha (1-u) J(2) \Big\}.$$

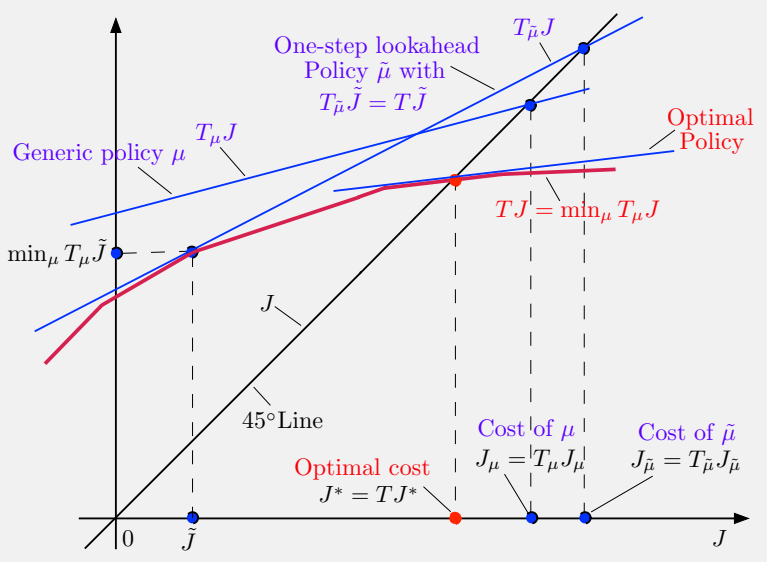


Figure 1.3.3 Geometric interpretation of the Bellman operator T, and the corresponding Bellman equation. For a fixed x, the function $(TJ)(x)$ can be written as $\min_\mu (T_\mu J)(x)$, so it is concave as a function of J. The optimal cost function J^* satisfies $J^* = TJ^*$, so it is obtained from the intersection of the graph of TJ and the 45 degree line shown, assuming J^* is real-valued.

Note that the graph of T lies below the graph of every operator T_μ, and is in fact obtained as the lower envelope of the graphs of T_μ as μ ranges over the set of policies $\mathcal{M}$. In particular, for any given function $\tilde{J}$, for every x, the value $(T\tilde{J})(x)$ is obtained by finding a support hyperplane/subgradient of the graph of the concave function $(T\,J)(x)$ at $\tilde{J}$, as shown in the figure. This support hyperplane is defined by the control $\mu(x)$ of a policy $\tilde{\mu}$ that attains the minimum of $(T_\mu \tilde{J})(x)$ over μ:

$$\tilde{\mu}(x) \in \arg\min_{\mu \in \mathcal{M}} (T_\mu \tilde{J})(x)$$

(there may be multiple policies attaining this minimum, defining multiple support hyperplanes). This construction also shows how the minimization

$$(T\tilde{J})(x) = \min_{\mu \in \mathcal{M}} (T_\mu \tilde{J})(x)$$

corresponds to a linearization of the mapping T at the point $\tilde{J}$.

The control u at each state $x = 1, 2$ has the meaning of a probability that we must select at that state. In particular, we control the probabilities u and $(1-u)$ of moving to states $y = 1$ and $y = 2$, at a control cost that is quadratic in u and $(1-u)$, respectively. For this problem $(TJ)(1)$ and $(TJ)(2)$ can be calculated in closed form, so they are easy to plot and understand. They are piecewise quadratic, unlike the corresponding plots of Fig. 1.3.1, which are piecewise linear; see Fig. 1.3.4.

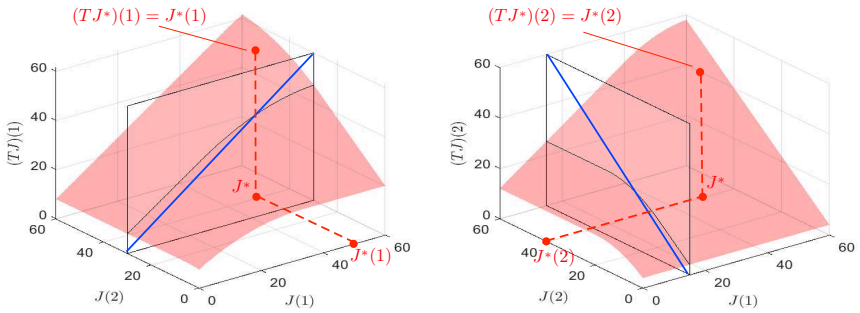


Figure 1.3.4 Illustration of the Bellman operator T for states 1 and 2 in Example 1.3.2. The parameter values are $g_1 = 5$, $g_2 = 3$, $r_{11} = 3$, $r_{12} = 15$, $r_{21} = 9$, $r_{22} = 1$, and the discount factor is $\alpha = 0.9$. The optimal costs are $J^*(1) = 49.7$ and $J^*(2) = 40.0$, and the optimal policy is $\mu^*(1) = 0.59$ and $\mu^*(2) = 0$. The figure also shows the one-dimensional slices of the operators at $J(1) = 15$ and $J(2) = 30$, together with the corresponding 45-degree lines.

Visualization of Value Iteration

The operator notation simplifies algorithmic descriptions, derivations, and proofs related to DP. For example, the value iteration (VI) algorithm can be written in the compact form

$$J_{k+1} = TJ_k, \qquad k = 0, 1, \ldots,$$

as illustrated in Fig. 1.3.5. Moreover, the VI algorithm for a given policy μ can be written as

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

and it can be similarly interpreted, except that the graph of the function $T_\mu J$ is linear. Also we will see shortly that there is a similarly compact description for the policy iteration algorithm.

1.3.2 Approximation in Value Space and Newton's Method

Let us now interpret approximation in value space in terms of abstract geometric constructions. Here we approximate J^* with some function $\tilde{J}$, and we obtain by minimization a corresponding policy, called a *one-step lookahead policy*. In particular, for a given $\tilde{J}$, a one-step lookahead policy $\tilde{\mu}$ is characterized by the equation

$$T_{\tilde{\mu}} \tilde{J} = T\tilde{J},$$

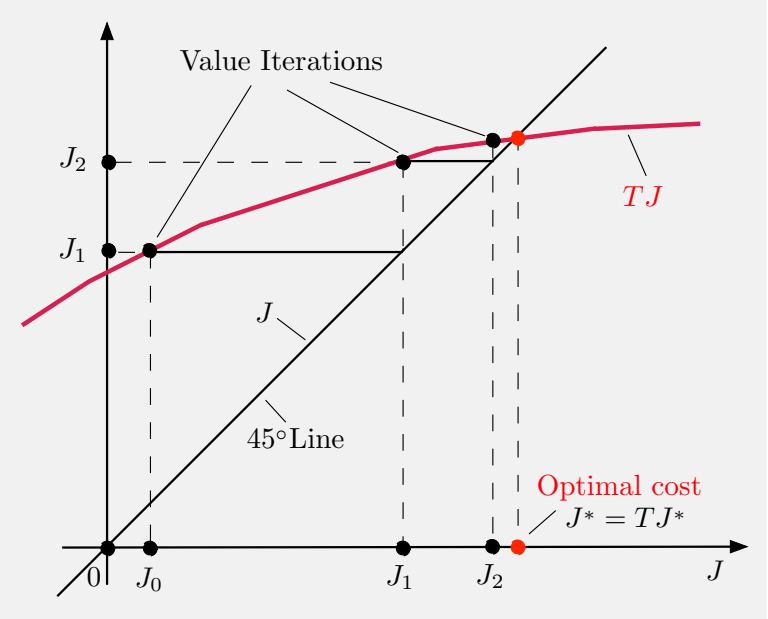


Figure 1.3.5 Geometric interpretation of the VI algorithm $J_{k+1} = TJ_k$, starting from some initial function J_0. Successive iterates are obtained through the staircase construction shown in the figure. The VI algorithm $J_{k+1} = T_\mu J_k$ for a given policy μ can be similarly interpreted, except that the graph of the function $T_\mu J$ is linear.

as in Fig. 1.3.6. This equation implies that the graph of $T_{\tilde{\mu}}J$ just touches the graph of TJ at $\tilde{J}$, as shown in the figure. Moreover, for each state $x \in X$ the hyperplane $H_{\tilde{\mu}}(x)$

$$H_{\tilde{\mu}}(x) = \Big\{\big(J(x), \xi\big) \mid (T_{\tilde{\mu}}J)(x) \geq \xi\Big\},$$

supports from above the convex set

$$\Big\{\big(J(x), \xi\big) \mid (TJ)(x) \geq \xi\Big\}$$

at the point $\big(\tilde{J}(x), (T\tilde{J})(x)\big)$ and defines a subgradient of $(TJ)(x)$ at $\tilde{J}$. Note that the one-step lookahead policy $\tilde{\mu}$ need not be unique, since T need not be differentiable.

In conclusion, the equation

$$J = T_{\tilde{\mu}}J$$

is a pointwise (for each x) linearization of the equation

$$J = TJ$$

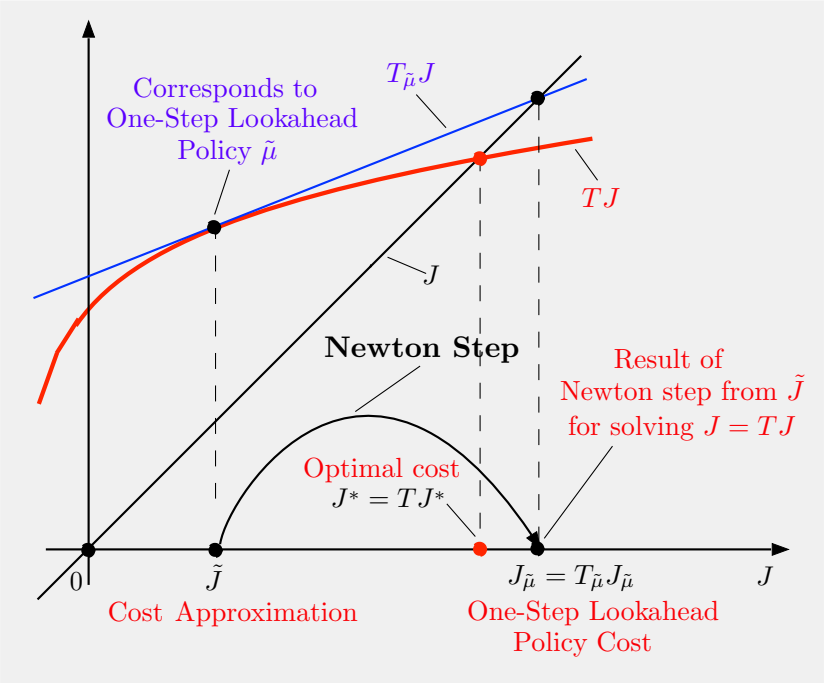


Figure 1.3.6 Geometric interpretation of approximation in value space and the one-step lookahead policy $\tilde{\mu}$ as a step of Newton's method. Given $\tilde{J}$, we find a policy $\tilde{\mu}$ that attains the minimum in the relation

$$T\tilde{J} = \min_{\mu} T_{\mu}\tilde{J}.$$

This policy satisfies $T\tilde{J} = T_{\tilde{\mu}}\tilde{J}$, so the graph of TJ and $T_{\tilde{\mu}}J$ touch at $\tilde{J}$, as shown. It may not be unique. Because TJ has concave components, the equation

$$J = T_{\tilde{\mu}}J$$

is the linearization of the equation $J = TJ$ at $\tilde{J}$. The linearized equation is solved at the typical step of Newton's method to provide the next iterate, which is just $J_{\tilde{\mu}}$.

at $\tilde{J}$, and its solution, $J_{\tilde{\mu}}$, can be viewed as the result of a Newton iteration at the point $\tilde{J}$. In summary, *the Newton iterate at $\tilde{J}$ is $J_{\tilde{\mu}}$*, the solution of the linearized equation $J = T_{\tilde{\mu}}J$.†

We may also consider approximation in value space with ℓ-step looka-

† The classical Newton's method for solving a fixed point problem of the form $y = T(y)$, where y is an n-dimensional vector, operates as follows: At the current iterate y_k, we linearize T and find the solution y_{k+1} of the corresponding linear fixed point problem. Assuming T is differentiable, the linearization is obtained

head using $\tilde{J}$. This is the same as approximation in value space with one-step lookahead using the $(\ell-1)$-fold operation of T on $\tilde{J}$, $T^{\ell-1}\tilde{J}$. Thus it can be interpreted as a Newton step starting from $T^{\ell-1}\tilde{J}$, the result of $\ell-1$ value iterations applied to $\tilde{J}$. This is illustrated in Fig. 1.3.7.†

1.3.3 Policy Iteration and Newton's Method

Another major class of infinite horizon algorithms is based on *policy iteration* (PI for short). We will discuss several abstract versions of PI in subsequent chapters, under a variety of assumptions. Generally, each iteration of the PI algorithm starts with a policy (which we call *current* or *base* policy), and generates another policy (which we call *new* or *rollout* policy, respectively). For the stochastic optimal control problem of Example 1.2.1, given the base policy μ, a policy iteration consists of two phases:

---

by using a first order Taylor expansion:

$$y_{k+1} = T(y_k) + \frac{\partial T(y_k)}{\partial y}(y_{k+1} - y_k),$$

where $\partial T(y_k)/\partial y$ is the $n \times n$ Jacobian matrix of T evaluated at the vector y_k. The most commonly given convergence rate property of Newton's method is *quadratic convergence*. It states that near the solution y^*, we have

$$\|y_{k+1} - y^*\| = O\big(\|y_k - y^*\|^2\big),$$

where $\|\cdot\|$ is the Euclidean norm, and holds assuming the Jacobian matrix exists and is Lipschitz continuous (see [Ber16], Section 1.4). There are extensions of Newton's method that are based on solving a linearized system at the current iterate, but relax the differentiability requirement to piecewise differentiability, and/or component concavity, while maintaining the superlinear convergence property of the method.

The structure of the Bellman operators (1.28) and (1.27), with their monotonicity and concavity properties, tends to enhance the convergence and rate of convergence properties of Newton's method, even in the absence of differentiability, as evidenced by the convergence analysis of PI, and the extensive favorable experience with rollout, PI, and MPC. In this connection, it is worth noting that in the case of Markov games, where the concavity property does not hold, the PI method may oscillate, as shown by Pollatschek and Avi-Itzhak [PoA69], and needs to be modified to restore its global convergence; see the author's paper [Ber21c]. We will discuss abstract versions of game and minimax contexts n Chapter 5.

† Variants of Newton's method that involve combinations of first order iterative methods, such as the Gauss-Seidel and Jacobi algorithms, and Newton's method, and they belong to the general family of *Newton-SOR methods* (SOR stands for "successive over-relaxation"); see the classic book by Ortega and Rheinboldt [OrR70] (Section 13.4).

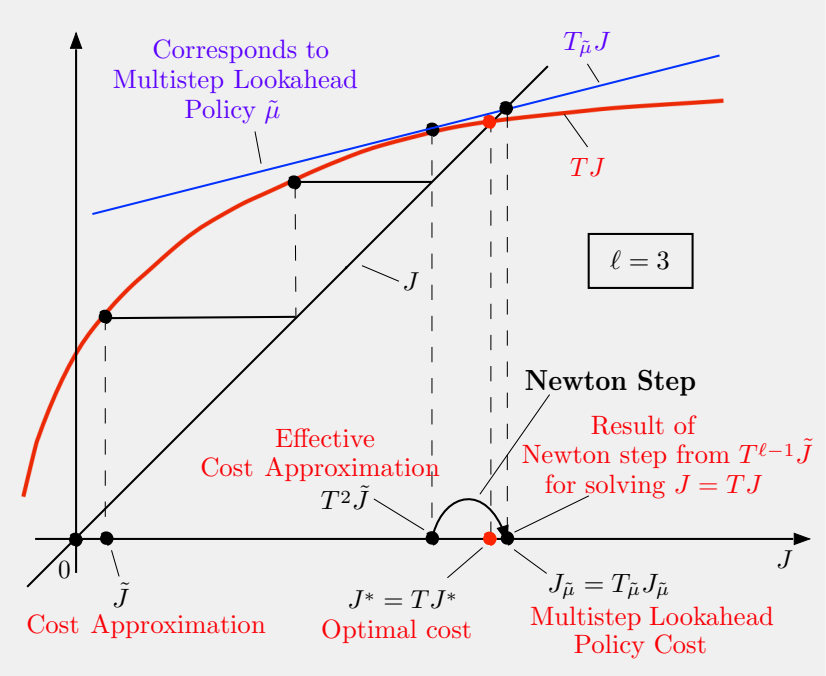


Figure 1.3.7 Geometric interpretation of approximation in value space with ℓ-step lookahead (in this figure $\ell = 3$). It is the same as approximation in value space with one-step lookahead using $T^{\ell-1}\tilde{J}$ as cost approximation. It can be viewed as a Newton step at the point $T^{\ell-1}\tilde{J}$, the result of $\ell - 1$ value iterations applied to $\tilde{J}$. Note that as ℓ increases the cost function $J_{\tilde{\mu}}$ of the ℓ-step lookahead policy $\tilde{\mu}$ approaches more closely the optimal J^*, and that $\lim_{\ell\to\infty} J_{\tilde{\mu}} = J^*$.

(a) *Policy evaluation*, which computes the cost function J_μ. One possibility is to solve the corresponding Bellman equation

$$J_\mu(x) = E\Big\{g\big(x,\mu(x),w\big) + \alpha J_\mu\big(f(x,\mu(x),w)\big)\Big\}, \qquad \text{for all } x. \tag{1.34}$$

However, the value $J_\mu(x)$ for any x can also be computed by Monte Carlo simulation, by averaging over many randomly generated trajectories the cost of the policy starting from x. Other possibilities include the use of specialized simulation-based methods, based on the projected and aggregation Bellman equations discussed in Section 1.2.4, for which there is extensive literature (see e.g., the books [BeT96], [SuB98], [Ber12a], [Ber19b]).

(b) *Policy improvement*, which computes the rollout policy $\tilde{\mu}$ using the one-step lookahead minimization

$$\tilde{\mu}(x) \in \arg\min_{u\in U(x)} E\Big\{g(x,u,w) + \alpha J_\mu\big(f(x,u,w)\big)\Big\}, \qquad \text{for all } x. \tag{1.35}$$

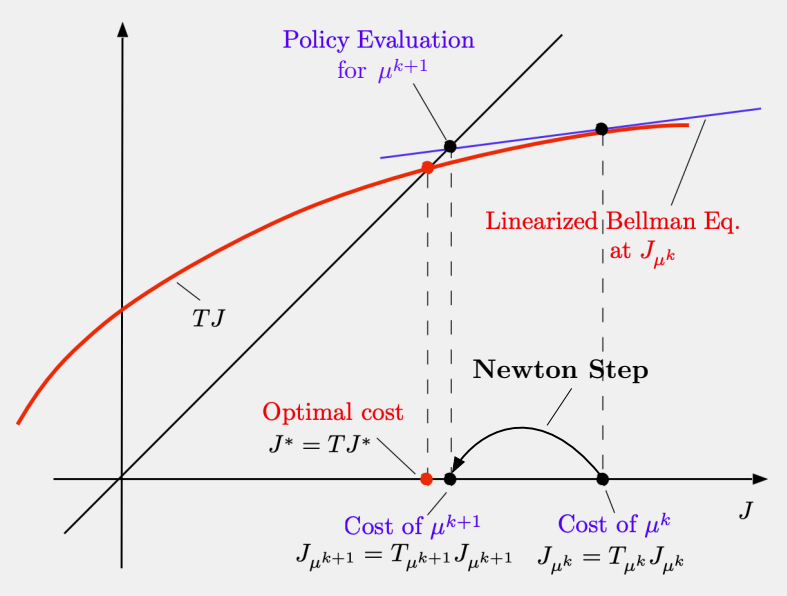


Figure 1.3.8 Geometric interpretation of a single policy iteration. Starting from the stable current policy μ^k, it evaluates the corresponding cost function J_{μ^k}, and computes the next policy μ^{k+1} according to $T_{\mu^{k+1}} J_{\mu^k} = TJ_{\mu^k}$. The corresponding cost function $J_{\mu^{k+1}}$ is obtained as the solution of the linearized equation $J = T_{\mu^{k+1}} J$, so it is the result of a Newton step for solving the Bellman equation $J = TJ$, starting from J_{μ^k}. Note than in policy iteration, the Newton step always starts at a function J_μ, which satisfies $J_\mu \geq J^*$.

It is generally expected (and can be proved under mild conditions) that the rollout policy is improved in the sense that $J_{\tilde{\mu}}(x) \leq J_\mu(x)$ for all x.

Thus the PI process generates a sequence of policies $\{\mu^k\}$, by obtaining μ^{k+1} through a policy improvement operation using J_{μ^k} in place of J_μ in Eq. (1.35), which is obtained through policy evaluation of the preceding policy μ^k using Eq. (1.34). In subsequent chapters, we will show under appropriate assumptions that general forms of PI have interesting and often solid convergence properties, which may hold even when the method is implemented (with appropriate modifications) in unconventional computing environments, involving asynchronous distributed computation.

In terms of our abstract notation, the PI algorithm can be written in a compact form. For the generated policy sequence $\{\mu^k\}$, the policy evaluation phase obtains J_{μ^k} from the equation

$$J_{\mu^k} = T_{\mu^k} J_{\mu^k}, \tag{1.36}$$

while the policy improvement phase obtains μ^{k+1} through the equation

$$T_{\mu^{k+1}} J_{\mu^k} = TJ_{\mu^k}. \tag{1.37}$$

As Fig. 1.3.8 illustrates, PI can be viewed as Newton's method for solving the Bellman equation in the function space of cost functions J. In particular, *the policy improvement Eq. (1.37) is the Newton step starting from J_{μ^k}, and yields μ^{k+1} as the corresponding one-step lookahead/rollout policy*.

The interpretation of PI as a form of Newton's method has a long history, for which we refer to the original works for linear quadratic problems by Kleinman [Klei68],† and for finite-state infinite horizon discounted and Markov game problems by Pollatschek and Avi-Itzhak [PoA69] (who also showed that the method may oscillate in the game case; see the discussion in Chapter 5).

1.3.4 Approximation in Value Space for General Abstract Dynamic Programming

Let us now consider the general case where the mapping T_μ is not assumed linear for all stationary policies $\mu \in \mathcal{M}$. In this case we still have the alternative description of T

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

but T need not be concave, i.e., for some $x \in X$, the function $(TJ)(x)$ may not be concave as a function of J. We illustrate this fact in Fig. 1.3.9.

The nonlinearity of the mapping T_μ can have profound consequences on the validity of the PI algorithm and its interpretation in terms of Newton's method. A prominent case where this is so arises in minimax problems and related two-person zero sum game settings (cf. Example 1.2.5). We will discuss this case in Chapter 5, where we will introduce modifications to the PI algorithm that restore its convergence property.

We note, however, that it is possible that the mappings T_μ are nonlinear and convex, but that T has concave and differentiable components $(TJ)(x)$, in which case the Newton step interpretation applies. This occurs in particular in the important case of zero-sum dynamic games involving a linear system and a quadratic cost function.

1.4 ORGANIZATION OF THE BOOK

The examples in the preceding sections demonstrate that while the monotonicity assumption is satisfied for most DP models, the contraction assumption may or may not hold. In particular, the contraction assumption

† This was part of Kleinman's Ph.D. thesis [Kle67] at M.I.T., supervised by M. Athans. Kleinman gives credit for the one-dimensional version of his results to Bellman and Kalaba [BeK65]. Note also that the first proposal of the PI method was given by Bellman in his classic book [Bel57], under the name "approximation in policy space."

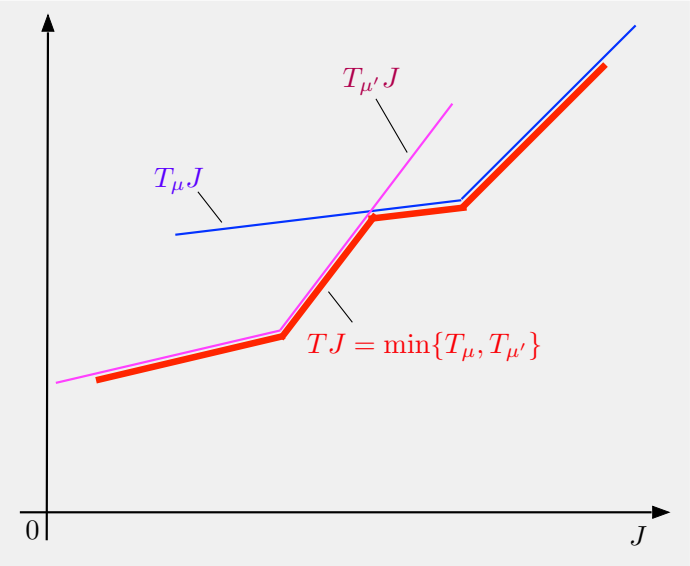


Figure 1.3.9 Geometric interpretation of the Bellman operator, in the general case where the policy mappings T_μ are not linear. The figure illustrates the case of two policies μ and μ', whose mappings T_μ and $T_{\mu'}$ are piecewise linear and convex. In this case the mapping T, given by $(TJ)(x) = \min\big\{T_\mu J(x), T_{\mu'}J(x)\big\}$, is piecewise linear, but it is neither convex nor concave, and the Newton step interpretation breaks down; see also Chapter 5.

is satisfied for the mapping H in Examples 1.2.1-1.2.5, provided there is discounting and that the cost per stage is bounded. However, it need not hold in the SSP Example 1.2.6, the multiplicative Example 1.2.8, and the affine monotonic Example 1.2.9.

The book's central theme is that the presence or absence of monotonicity and contraction fundamentally shapes the analytical and algorithmic theories for abstract DP. In our development, with few exceptions, we will assume that monotonicity holds. Consequently, the book is organized around the presence or absence of the contraction property. In the next three chapters we will discuss three types of DP models.

(a) **Contractive models:** These models, discussed in Chapter 2, have the richest and strongest algorithmic theory, and serve as a benchmark for other models. Notable examples include discounted stochastic optimal control problems (cf. Example 1.2.1), finite-state discounted MDP (cf. Example 1.2.2), and some special types of SSP problems (cf. Example 1.2.6).

(b) **Semicontractive models:** In these models, T_μ is monotone but is not a contraction for all $\mu \in \mathcal{M}$. Most practical deterministic, stochastic, and minimax-type shortest path problems fall into this

category. One challenge here is that, under certain conditions, some of the problem's cost functions may take the values $+\infty$ or $-\infty$, and the mappings T_μ and T must be able to deal with such functions.

The distinguishing feature of semicontractive models is the separation of policies into those that "behave well" within our optimization framework and those that do not. Contraction-based analysis is insufficient to deal with "ill-behaved" policies, so we introduce a notion of "regularity," which is connected to contraction, but is more general. In particular, a policy μ is considered "regular" if the dynamic system underlying T_μ has J_μ has an asymptotically stable equilibrium within a suitable domain. Our models and analysis are patterned to a large extent after the SSP problems of Example 1.2.6 (the regular μ correspond to the proper policies). We show that the (restricted) optimal cost function over just the regular policies can typically be obtained with value and policy iteration algorithms. By contrast, the optimal cost function over all policies J^* may not be obtainable by these algorithms, and indeed J^* may not even be a solution of Bellman's equation, as we will show with a simple example in Section 3.1.2.

The key idea is that under certain conditions, the restricted optimization (the one that optimizes over the regular policies only) is well behaved, both analytically and algorithmically. Under additional conditions, which directly or indirectly ensure the existence of an optimal regular policy, we obtain semicontractive models with properties nearly as robust as contractive models.

In Chapter 3, we develop the basic theory of semicontractive models for the case where the regular policies are stationary, while in Chapter 4 (Section 4.4), we extend the notion of regularity to nonstationary policies. Moreover, we illustrate the theory with a variety of interesting shortest path-type problems (stochastic, minimax, affine monotonic, and risk sensitive/exponential cost), linear-quadratic optimal control problems, and deterministic and stochastic optimal control problems.

(c) **Noncontractive models:** These models rely on just the monotonicity property of T_μ, and are more complex than the preceding ones. Like semicontractive models, the problem's cost functions may take the values of $+\infty$ or $-\infty$, and in fact the optimal cost function may take the values ∞ and $-\infty$ as a matter of course (rather than on an exceptional basis, as in semicontractive models). This complexity presents considerable challenges, as much of the contractive model theory either does not extend or does so in a weaker form only. For instance, the fixed point equation $J = TJ$ may lack a unique solution, value iteration may succeed starting with some functions but

not with others, and policy iteration may fail altogether. Some of these issues may be mitigated when additional structure is present, which we discuss in Sections 4.4-4.6, focusing on noncontractive models that also have some semicontractive structure, and corresponding favorable properties.

Examples of DP problems from each of the model categories above, primarily special cases of the specific DP models discussed in Section 1.2, are scattered throughout the book. They serve both to illustrate the theory and its exceptions, and to highlight the beneficial role of additional special structure.

We finally note some other types of models where there are restrictions to the set of policies, i.e., $\mathcal{M}$ may be a strict subset of the set of functions $\mu : X \mapsto U$ with $\mu(x) \in U(x)$ for all $x \in X$. Such restrictions may include measurability (needed to establish a mathematically rigorous probabilistic framework) or special structure that enhances the characterization of optimal policies and facilitates their computation. These models were treated in Chapter 5 of the first edition of this book, and also in Chapter 6 of [BeS78].†

Algorithms

Our discussion of algorithms centers on abstract forms of value and policy iteration, and is organized along three characteristics: *exact, approximate, and asynchronous*. The exact algorithms represent idealized versions, the approximate represent implementations that use approximations of various kinds, and the asynchronous involve irregular computation orders, where the costs and controls at different states are updated at different iterations (for example the cost of a single state being iterated at a time, as in Gauss-Seidel and other methods; see [Ber12a] for several examples of distributed asynchronous DP algorithms).

Approximate and asynchronous implementations have been the subject of intensive investigations since the 1980s, in the context of the solution of large-scale problems. Some of this methodology relies on the use of simulation, which is asynchronous by nature and is prominent in approximate DP. Generally, the monotonicity and sup-norm contraction structures of many prominent DP models favors the use of asynchronous algorithms in DP, as first shown in the author's paper [Ber82], and discussed at various points in this book: Section 2.6 for contractive models, Section 3.6 for semicontractive models, and Sections 5.3-5.4 for minimax problems and zero-sum games.

† Chapter 5 of the first edition is accessible from the author's web site and the book's web page, and uses terminology and notation that are consistent with the present edition.

1.5 NOTES, SOURCES, AND EXERCISES

This monograph is written in a mathematical style that emphasizes simplicity and abstraction. According to the relevant Wikipedia article:

"Abstraction in mathematics is the process of extracting the underlying essence of a mathematical concept, removing any dependence on real world objects with which it might originally have been connected, and generalizing it so that it has wider applications or matching among other abstract descriptions of equivalent phenomena ... The advantages of abstraction are:

(1) It reveals deep connections between different areas of mathematics.

(2) Known results in one area can suggest conjectures in a related area.

(3) Techniques and methods from one area can be applied to prove results in a related area.

One disadvantage of abstraction is that highly abstract concepts can be difficult to learn. A degree of mathematical maturity and experience may be needed for conceptual assimilation of abstractions."

Consistent with the preceding view of abstraction, our aim has been to construct a minimalist framework, where the important mathematical structures stand out, while the application context is deliberately blurred. Of course, our development has to pass the test of relevance to applications. In this connection, we note that our presentation has integrated the relation of our abstract DP models with the applications of Section 1.2, and particularly discounted stochastic optimal control models (Chapter 2), shortest path-type models (Chapters 3 and 4), undiscounted deterministic and stochastic optimal control models (Chapter 4), and minimax and zero-sum game problems (Chapter 5). We have given illustrations of the abstract mathematical theory using these models and others throughout the text. A much broader and accessible account of applications is given in the author's two-volume DP textbook.

Section 1.2: The abstract style of mathematical development has a long history in DP. In particular, the connection between DP and fixed point theory may be traced to Shapley [Sha53], who exploited contraction mapping properties in analysis of the two-player dynamic game model of Example 1.2.4. Since then, the underlying contraction properties of discounted DP problems with bounded cost per stage have been explicitly or implicitly used by most authors that have dealt with the subject. Moreover, the value of the abstract viewpoint as the basis for economical and insightful analysis has been widely recognized.

An abstract DP model, based on unweighted sup-norm contraction assumptions, was introduced in the paper by Denardo [Den67]. This model pointed to the fundamental connections between DP and fixed point theory, and provided generality and insight into the principal analytical and

algorithmic ideas underlying the discounted DP research up to that time. Abstract DP ideas were also researched earlier, notably in the paper by Mitten (Denardo's Ph.D. thesis advisor) [Mit64]; see also Denardo and Mitten [DeM67]. The properties of monotone contractions were also used in the analysis of sequential games by Zachrisson [Zac64].

Two abstract DP models that rely only on monotonicity properties were given by the author in the papers [Ber75], [Ber77]. They were patterned after the negative cost DP problem of Blackwell [Bla65] and the positive cost DP problem of Strauch [Str66] (see the monotone decreasing and monotone increasing models of Section 4.3). These two abstract DP models, together with the finite horizon models of Section 4.2, were used extensively in the book by Bertsekas and Shreve [BeS78] for the analysis of both discounted and undiscounted DP problems, ranging over MDP, minimax, multiplicative, and Borel space models.

Extensions of the monotonicity-based analysis of the author's paper [Ber77] were given by Verdu and Poor [VeP87], who introduced additional structure for developing backward and forward value iterations, and by Szepesvari [Sze98a, Sze98b], who incorporated non-Markovian policies into the abstract DP framework. The model from [Ber77] also provided a foundation for asynchronous value and policy iteration methods for abstract contractive and noncontractive DP models in Bertsekas [Ber82] and Bertsekas and Yu [BeY10]. An extended contraction framework, whereby the sup-norm contraction norm is allowed to be weighted, was given in the author's paper [Ber12b]. Another line of related research involving abstract DP mappings that are not necessarily scalar-valued was initiated by Mitten [Mit74], and was followed up by a number of authors, including Sobel [Sob75], Morin [Mor82], and Carraway and Morin [CaM88].

Section 1.3: The central role of Newton's method for understanding approximation value space, rollout, and other reinforcement learning and approximate DP methods, was articulated in the author's monograph [Ber20], and was described in more detail in the book [Ber22].

Section 1.4: Generally, noncontractive total cost DP models with some special structure beyond monotonicity, fall in three major categories: *monotone increasing models*, principally represented by positive cost DP, *monotone decreasing models*, principally represented by negative cost DP, and *transient models*, exemplified by the SSP model of Example 1.2.6, where the decision process terminates after a period that is random and subject to control. Abstract DP models patterned after the first two categories have been known since the author's papers [Ber75], [Ber77], and are further discussed in Section 4.3.

The semicontractive models, further discussed Chapter 3 and Sections 4.4-4.6, are patterned after the third category. They were introduced and analyzed in the first edition of this book, as well as the subsequent series of papers and reports, [Ber15], [Ber16a], [BeY16], [Ber17b], [Ber17c],

[Ber17d], [Ber19c]. Their analysis is based on the idea of separating policies into those that are well-behaved (these are called *regular*, and have contraction-like properties) and those that are not (these are called *irregular*). The objective of the analysis is then to explain the detrimental effects of the irregular policies, and to delineate the kind of model structure that can limit these effects. As far as the author knows, this idea is new in the context of abstract DP. One of the aims of the present monograph is to develop this idea and to show that it leads to an important and insightful paradigm for conceptualization and solution of major classes of practical DP problems.

E X E R C I S E S

1.1 (Multistep Contraction Mappings)

This exercise shows how starting with an abstract mapping, we can obtain multistep mappings with the same fixed points and a stronger contraction modulus. Consider a set of mappings $T_\mu : \mathcal{B}(X) \mapsto \mathcal{B}(X)$, $\mu \in \mathcal{M}$, satisfying the contraction Assumption 1.2.2, let m be a positive integer, and let $\mathcal{M}_m$ be the set of m-tuples $\nu = (\mu_0, \ldots, \mu_{m-1})$, where $\mu_k \in \mathcal{M}$, $k = 1, \ldots, m-1$. For each $\nu = (\mu_0, \ldots, \mu_{m-1}) \in \mathcal{M}_m$, define the mapping $\overline{T}_\nu$, by

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

Show the contraction properties

$$\|\overline{T}_\nu J - \overline{T}_\nu J'\| \leq \alpha^m \|J - J'\|, \qquad \forall\ J, J' \in \mathcal{B}(X), \tag{1.39}$$

and

$$\|\overline{T} J - \overline{T} J'\| \leq \alpha^m \|J - J'\|, \qquad \forall\ J, J' \in \mathcal{B}(X), \tag{1.40}$$

where $\overline{T}$ is defined by

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

Solution: By the contraction property of $T_{\mu_0}, \ldots, T_{\mu_{m-1}}$, we have for all $J, J' \in \mathcal{B}(X)$,

$$\begin{aligned}
\|\overline{T}_\nu J - \overline{T}_\nu J'\| &= \|T_{\mu_0} \cdots T_{\mu_{m-1}} J - T_{\mu_0} \cdots T_{\mu_{m-1}} J'\| \\
&\leq \alpha \|T_{\mu_1} \cdots T_{\mu_{m-1}} J - T_{\mu_1} \cdots T_{\mu_{m-1}} J'\| \\
&\leq \alpha^2 \|T_{\mu_2} \cdots T_{\mu_{m-1}} J - T_{\mu_2} \cdots T_{\mu_{m-1}} J'\| \\
&\vdots \\
&\leq \alpha^m \|J - J'\|,
\end{aligned}$$

thus showing Eq. (1.39).

We have from Eq. (1.39)

$$(T_{\mu_0}\cdots T_{\mu_{m-1}}J)(x) \le (T_{\mu_0}\cdots T_{\mu_{m-1}}J')(x) + \alpha^m \|J-J'\|\, v(x), \qquad \forall\ x\in X,$$

and by taking infimum of both sides over $(T_{\mu_0}\cdots T_{\mu_{m-1}}) \in \mathcal{M}_m$ and dividing by $v(x)$, we obtain

$$\frac{(\overline{T}J)(x)-(\overline{T}J')(x)}{v(x)} \le \alpha^m\|J-J'\|, \qquad \forall\ x\in X.$$

Similarly

$$\frac{(\overline{T}J')(x)-(\overline{T}J)(x)}{v(x)} \le \alpha^m\|J-J'\|, \qquad \forall\ x\in X,$$

and by combining the last two relations and taking supremum over $x\in X$, Eq. (1.40) follows.

1.2 (Relation of Temporal Difference Methods and Proximal Algorithms [Ber16b], [Ber18c])

The purpose of this exercise is establish a close connection between the mappings underlying temporal difference and proximal methods (cf. Section 1.2.5). Consider a linear mapping of the form

$$TJ = AJ + b,$$

where b is a vector in $\Re^n$, and A is an $n\times n$ matrix with eigenvalues strictly within the unit circle. Let $\lambda\in(0,1)$ and $c=\frac{\lambda}{1-\lambda}$, and consider the multistep mapping $T^{(\lambda)}$ given by

$$T^{(\lambda)}J = (1-\lambda)\sum_{\ell=0}^{\infty}\lambda^\ell T^{\ell+1}J, \qquad J\in\Re^n,$$

and the proximal mapping $P^{(c)}$ given by

$$P^{(c)}J = \left(\frac{c+1}{c}I - A\right)^{-1}\left(b+\frac{1}{c}J\right), \qquad J\in\Re^n;$$

cf. Eq. (1.23) [equivalently, for a given J, $P^{(c)}J$ is the unique vector $Y\in\Re^n$ that solves the equation

$$Y - TY = \frac{1}{c}(J-Y),$$

(cf. Fig. 1.5.1)].

(a) Show that $P^{(c)}$ is given by

$$P^{(c)} = (1-\lambda)\sum_{\ell=0}^{\infty}\lambda^\ell T^\ell,$$

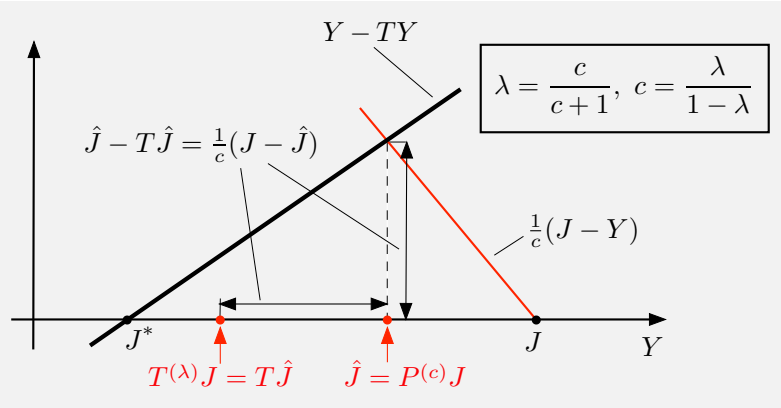


Figure 1.5.1. Illustration of the iterates $T^{(\lambda)}J$ and $P^{(c)}J$ for finding the fixed point J^* of a linear mapping T. Given J, we find the proximal iterate $\hat{J} = P^{(c)}J$ and then add the amount $\frac{1}{c}\big(\hat{J} - J\big)$ to obtain $T^{(\lambda)}J = TP^{(c)}J$. If T is a contraction mapping, $T^{(\lambda)}J$ is closer to J^* than $P^{(c)}J$.

and can be written as

$$P^{(c)}J = \overline{A}^{(\lambda)}J + \overline{b}^{(\lambda)},$$

where

$$\overline{A}^{(\lambda)} = (1-\lambda)\sum_{\ell=0}^{\infty}\lambda^\ell A^\ell, \qquad \overline{b}^{(\lambda)} = \sum_{\ell=0}^{\infty}\lambda^{\ell+1}A^\ell b.$$

(b) Verify that

$$T^{(\lambda)}J = A^{(\lambda)}J + b^{(\lambda)},$$

where

$$A^{(\lambda)} = (1-\lambda)\sum_{\ell=0}^{\infty}\lambda^\ell A^{\ell+1}, \qquad b^{(\lambda)} = \sum_{\ell=0}^{\infty}\lambda^\ell A^\ell b,$$

and show that

$$T^{(\lambda)} = TP^{(c)} = P^{(c)}T, \tag{1.41}$$

and that for all $J \in \Re^n$,

$$P^{(c)}J = J + \lambda\big(T^{(\lambda)}J - J\big), \qquad T^{(\lambda)}J = J + \frac{c+1}{c}\big(P^{(c)}J - J\big). \tag{1.42}$$

Thus $T^{(\lambda)}J$ is obtained by extrapolation along the line segment $P^{(c)}J - J$, as illustrated in Fig. 1.5.1. Note that since T is a contraction mapping, $T^{(\lambda)}J$ is closer to J^* than $P^{(c)}J$.

(c) Show that for a given $J \in \Re^n$, the multistep and proximal iterates $T^{(\lambda)}J$ and $P^{(c)}J$ are the unique fixed points of the contraction mappings W_J and $\overline{W}_J$ given by

$$W_JY = (1-\lambda)TJ + \lambda TY, \qquad \overline{W}_JY = (1-\lambda)J + \lambda TY, \qquad Y \in \Re^n,$$

respectively.

(d) Show that the fixed point property of part (c) yields the following formula for the multistep mapping $T^{(\lambda)}$:

$$T^{(\lambda)}J = (1-\lambda A)^{-1}\big(b + (1-\lambda)AJ\big). \tag{1.43}$$

(e) (*Multistep Contraction Property for Nonexpansive A [BeY09]*) Instead of assuming that A has eigenvalues strictly within the unit circle, assume that the matrix $I - A$ is invertible and A is nonexpansive [i.e., has all its eigenvalues within the unit circle (possibly on the unit circle)]. Show that $A^{(\lambda)}$ is contractive (i.e., has eigenvalues that lie strictly within the unit circle) and its eigenvalues have the form

$$\theta_i = (1-\lambda)\sum_{\ell=0}^{\infty}\lambda^\ell\zeta_i^{\ell+1} = \frac{\zeta_i(1-\lambda)}{1-\zeta_i\lambda}, \qquad i = 1,\ldots,n, \tag{1.44}$$

where ζ_i, $i = 1,\ldots,n$, are the eigenvalues of A. *Note*: For an intuitive explanation of the result, note that the eigenvalues of $A^{(\lambda)}$ can be viewed as convex combinations of complex numbers from the unit circle at least two of which are different from each other, since $\zeta_i \neq 1$ by assumption (the nonzero corresponding eigenvalues of A and A^2 are different from each other). As a result the eigenvalues of $A^{(\lambda)}$ lie strictly within the unit circle.

(f) (*Contraction Property of Projected Multistep Mappings*) Under the assumptions of part (e), show that $\lim_{\lambda\to 1} A^{(\lambda)} = 0$. Furthermore, for any $n \times n$ matrix W, the matrix $WA^{(\lambda)}$ is contractive for λ sufficiently close to 1. In particular the projected mapping $\Pi A^{(\lambda)}$ and corresponding projected proximal mapping (cf. Section 1.2.5) become contractions as $\lambda \to 1$.

Solution: (a) The inverse in the definition of $P^{(c)}$ is written as

$$\left(\frac{c+1}{c}I - A\right)^{-1} = \left(\frac{1}{\lambda}I - A\right)^{-1} = \lambda(I-\lambda A)^{-1} = \lambda\sum_{\ell=0}^{\infty}(\lambda A)^\ell.$$

Thus, using the equation $\frac{1}{c} = \frac{1-\lambda}{\lambda}$,

$$\begin{aligned} P^{(c)}J &= \left(\frac{c+1}{c}I - A\right)^{-1}\left(b + \frac{1}{c}J\right) \\ &= \lambda\sum_{\ell=0}^{\infty}(\lambda A)^\ell\left(b + \frac{1-\lambda}{\lambda}J\right) \\ &= (1-\lambda)\sum_{\ell=0}^{\infty}(\lambda A)^\ell J + \lambda\sum_{\ell=0}^{\infty}(\lambda A)^\ell b, \end{aligned}$$

which is equal to $\overline{A}^{(\lambda)}J + \overline{b}^{(\lambda)}$. The formula $P^{(c)} = (1-\lambda)\sum_{\ell=0}^{\infty}\lambda^\ell T^\ell$ follows from this expression.

(b) The formula $T^{(\lambda)}J = A^{(\lambda)}J + b^{(\lambda)}$ is verified by straightforward calculation. We have,

$$\begin{aligned} TP^{(c)}J &= A\big(\overline{A}^{(\lambda)}J + \overline{b}^{(\lambda)}\big) + b \\ &= (1-\lambda)\sum_{\ell=0}^{\infty}\lambda^{\ell}A^{\ell+1}J + \sum_{\ell=0}^{\infty}\lambda^{\ell+1}A^{\ell+1}b + b = A^{(\lambda)}J + b^{(\lambda)} \\ &= T^{(\lambda)}J, \end{aligned}$$

thus proving the left side of Eq. (1.41). The right side is proved similarly. The interpolation/extrapolation formula (1.42) follows by a straightforward calculation from the definition of $T^{(\lambda)}$. As an example, to show the left side of Eq. (1.42), we write

$$\begin{aligned} J + \lambda\big(T^{(\lambda)}J - J\big) &= (1-\lambda)J + \lambda T^{(\lambda)}J \\ &= (1-\lambda)J + \lambda\left((1-\lambda)\sum_{\ell=0}^{\infty}\lambda^{\ell}A^{\ell+1}J + \sum_{\ell=0}^{\infty}\lambda^{\ell}A^{\ell}b\right) \\ &= (1-\lambda)\left(J + \sum_{\ell=1}^{\infty}\lambda^{\ell}A^{\ell}J\right) + \sum_{\ell=0}^{\infty}\lambda^{\ell+1}A^{\ell}b \\ &= \overline{A}^{(\lambda)}J + \overline{b}^{(\lambda)} \\ &= P^{(c)}J. \end{aligned}$$

(c) To show that $T^{(\lambda)}J$ is the fixed point of W_J, we must verify that

$$T^{(\lambda)}J = W_J\big(T^{(\lambda)}J\big),$$

or equivalently that

$$T^{(\lambda)}J = (1-\lambda)TJ + \lambda T\big(T^{(\lambda)}J\big) = (1-\lambda)TJ + \lambda T^{(\lambda)}(TJ).$$

The right-hand side, in view of the interpolation formula

$$(1-\lambda)J + \lambda T^{(\lambda)}J = P^{(c)}J, \qquad \forall\, x \in \Re^n,$$

is equal to $P^{(c)}(TJ)$, which from the formula $T^{(\lambda)} = P^{(c)}T$ [cf. part (b)], is equal to $T^{(\lambda)}J$. The proof is similar for $\overline{W}_J$.

(d) The fixed point property of part (c) states that $T^{(\lambda)}J$ is the unique solution of the following equation in Y:

$$Y = (1-\lambda)TJ + \lambda TY = (1-\lambda)(AJ+b) + \lambda(AY+b),$$

from which the desired relation follows.

(e), (f) The formula (1.44) follows from the expression for $A^{(\lambda)}$ given in part (b). This formula can be used to show that the eigenvalues of $A^{(\lambda)}$ lie strictly within the unit circle, using also the fact that the matrices A^m, $m \geq 1$, and $A^{(\lambda)}$ have the same eigenvectors (see [BeY09] for details). Moreover, the eigenvalue formula shows that all eigenvalues of $A^{(\lambda)}$ converge to 0 as $\lambda \to 1$, so that $\lim_{\lambda\to 1} A^{(\lambda)} = 0$. This also implies that $WA^{(\lambda)}$ is contractive for λ sufficiently close to 1.

1.3 (State-Dependent Weighted Multistep Mappings [YuB12])

Consider a set of mappings $T_\mu : \mathcal{B}(X) \mapsto \mathcal{B}(X)$, $\mu \in \mathcal{M}$, satisfying the contraction Assumption 1.2.2. Consider also the mappings $T_\mu^{(w)} : \mathcal{B}(X) \mapsto \mathcal{B}(X)$ defined by

$$(T_\mu^{(w)} J)(x) = \sum_{\ell=1}^{\infty} w_\ell(x) \big(T_\mu^\ell J\big)(x), \qquad x \in X,\ J \in \mathcal{B}(X),$$

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

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

Show that

$$\frac{\big|(T_\mu^{(w)} J)(x) - (T_\mu^{(w)} J')(x)\big|}{v(x)} \leq \sum_{\ell=1}^{\infty} w_\ell(x)\, \alpha^\ell \|J - J'\|, \qquad \forall\ x \in X,$$

so that $T_\mu^{(w)}$ is a contraction with modulus

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

Moreover, for all $\mu \in \mathcal{M}$, the mappings T_μ and $T_\mu^{(w)}$ have the same fixed point.

Solution: By the contraction property of T_μ, we have for all $J, J' \in \mathcal{B}(X)$ and $x \in X$,

$$\begin{aligned}
\frac{\big|(T_\mu^{(w)} J)(x) - (T_\mu^{(w)} J')(x)\big|}{v(x)} &= \frac{\big|\sum_{\ell=1}^{\infty} w_\ell(x)(T_\mu^\ell J)(x) - \sum_{\ell=1}^{\infty} w_\ell(x)(T_\mu^\ell J')(x)\big|}{v(x)} \\
&\leq \sum_{\ell=1}^{\infty} w_\ell(x) \|T_\mu^\ell J - T_\mu^\ell J'\| \\
&\leq \left(\sum_{\ell=1}^{\infty} w_\ell(x) \alpha^\ell\right) \|J - J'\|,
\end{aligned}$$

showing the contraction property of $T_\mu^{(w)}$.

Let J_μ be the fixed point of T_μ. By using the relation $(T_\mu^\ell J_\mu)(x) = J_\mu(x)$, we have for all $x \in X$,

$$(T_\mu^{(w)} J_\mu)(x) = \sum_{\ell=1}^{\infty} w_\ell(x) \big(T_\mu^\ell J_\mu\big)(x) = \left(\sum_{\ell=1}^{\infty} w_\ell(x)\right) J_\mu(x) = J_\mu(x),$$

so J_μ is the fixed point of $T_\mu^{(w)}$ [which is unique since $T_\mu^{(w)}$ is a contraction].

2
Contractive Models

Contents

In this chapter we consider the abstract DP model of Section 1.2 under the most favorable assumptions: monotonicity and weighted sup-norm contraction. Important special cases of this model are the discounted problems with bounded cost per stage (Example 1.2.1-1.2.5), the stochastic shortest path problem of Example 1.2.6 in the case where all policies are proper, as well as other problems involving special structures.

We first provide some basic analytical results and then focus on two types of algorithms: *value iteration* and *policy iteration*. In addition to exact forms of these algorithms, we discuss combinations and approximate versions, as well as asynchronous distributed versions.

2.1 BELLMAN'S EQUATION AND OPTIMALITY CONDITIONS

In this section we recall the abstract DP model of Section 1.2, and derive some of its basic properties under the monotonicity and contraction assumptions of Section 1.3. We consider a set X of states and a set U of controls, and for each $x \in X$, a nonempty control constraint set $U(x) \subset U$. We denote by $\mathcal{M}$ the set of all functions $\mu : X \mapsto U$ with $\mu(x) \in U(x)$ for all $x \in X$, which we refer to as *policies* (or "stationary policies," when we want to emphasize the distinction from nonstationary policies, to be discussed later).

We denote by $\mathcal{R}(X)$ the set of real-valued functions $J : X \mapsto \Re$. We have a mapping $H : X \times U \times \mathcal{R}(X) \mapsto \Re$ and for each policy $\mu \in \mathcal{M}$, we consider the mapping $T_\mu : \mathcal{R}(X) \mapsto \mathcal{R}(X)$ defined by

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

We also consider the mapping T defined by

$$(TJ)(x) = \inf_{u \in U(x)} H(x, u, J) = \inf_{\mu \in \mathcal{M}} (T_\mu J)(x), \qquad \forall\ x \in X.$$

[We will use frequently the second equality above, which holds because $\mathcal{M}$ can be viewed as the Cartesian product $\Pi_{x \in X} U(x)$.] We want to find a function $J^* \in \mathcal{R}(X)$ such that

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

i.e., to find a fixed point of T within $\mathcal{R}(X)$. We also want to obtain a policy $\mu^* \in \mathcal{M}$ such that $T_{\mu^*} J^* = TJ^*$.

Let us restate for convenience the contraction and monotonicity assumptions of Section 1.2.2.

Assumption 2.1.1: (Monotonicity) If $J, J' \in \mathcal{R}(X)$ and $J \leq J'$, then

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

Note that the monotonicity assumption implies the following properties, for all $J, J' \in \mathcal{R}(X)$ and $k = 0, 1, \ldots$, which we will use extensively:

$$J \leq J' \qquad \Rightarrow \qquad T^k J \leq T^k J', \qquad T_\mu^k J \leq T_\mu^k J', \quad \forall\ \mu \in \mathcal{M},$$

$$J \leq TJ \qquad \Rightarrow \qquad T^k J \leq T^{k+1} J, \qquad T_\mu^k J \leq T_\mu^{k+1} J, \quad \forall\ \mu \in \mathcal{M}.$$

Here T^k and T_μ^k denotes the k-fold composition of T and T_μ, respectively.

For the contraction assumption, we introduce a function $v : X \mapsto \Re$ with

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

We consider the weighted sup-norm

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

on $\mathcal{B}(X)$, the space of real-valued functions J on X such that $J(x)/v(x)$ is bounded over $x \in X$ (see Appendix B for a discussion of the properties of this space).

Assumption 2.1.2: (Contraction) For all $J \in \mathcal{B}(X)$ and $\mu \in \mathcal{M}$, the functions $T_\mu J$ and TJ belong to $\mathcal{B}(X)$. Furthermore, for some $\alpha \in (0, 1)$, we have

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

The classical DP models where both the monotonicity and contraction assumptions are satisfied are the discounted finite-state Markovian decision problem of Example 1.2.2, and the stochastic shortest path problem of Example 1.2.6 in the special case where all policies are proper; see the textbook [Ber12a] for an extensive discussion. In the context of these problems, the fixed point equation $J = TJ$ is called *Bellman's equation*, a term that we will use more generally in this book as well. The following proposition summarizes some of the basic consequences of the contraction assumption.

Proposition 2.1.1: Let the contraction Assumption 2.1.2 hold. Then:

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

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

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

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

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

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

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

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

Proof: We showed in Section 1.2.2 that T is a contraction with modulus α over $\mathcal{B}(X)$. Parts (a) and (b) follow from Prop. B.1 of Appendix B.

To show part (c), note that if $T_\mu J^* = TJ^*$, then in view of $TJ^* = J^*$, we have $T_\mu J^* = J^*$, which implies that $J^* = J_\mu$, since J_μ is the unique fixed point of T_μ. Conversely, if $J^* = J_\mu$, we have $T_\mu J^* = T_\mu J_\mu = J_\mu = J^* = TJ^*$.

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

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

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

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

The preceding proposition and some of the subsequent results may also be proved if $\mathcal{B}(X)$ is replaced by a closed subset $\overline{\mathcal{B}}(X) \subset \mathcal{B}(X)$. This is because the contraction mapping fixed point theorem (Prop. B.1) applies to closed subsets of complete spaces. For simplicity, however, we will disregard this possibility in the present chapter.

An important consequence of monotonicity of H, when it holds in addition to contraction, is that it implies that J^*, the unique fixed point of T, is the infimum over $\mu \in \mathcal{M}$ of J_μ, the unique fixed point of T_μ.

Proposition 2.1.2: Let the monotonicity and contraction Assumptions 2.1.1 and 2.1.2 hold. Then

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

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

$$J^*(x) \leq J_{\mu_\epsilon}(x) \leq J^*(x) + \epsilon, \qquad \forall\ x \in X. \tag{2.1}$$

Proof: We note that the right-hand side of Eq. (2.1) holds by Prop. 2.1.1(e) (see the remark following its proof). Thus $\inf_{\mu \in \mathcal{M}} J_\mu(x) \leq J^*(x)$ for all $x \in X$. To show the reverse inequality as well as the left-hand side of Eq. (2.1), we note that for all $\mu \in \mathcal{M}$, we have $TJ^* \leq T_\mu J^*$, and since $J^* = TJ^*$, it follows that $J^* \leq T_\mu J^*$. By applying repeatedly T_μ to both sides of this inequality and by using the monotonicity Assumption 2.1.1, we obtain $J^* \leq T_\mu^k J^*$ for all $k > 0$. Taking the limit as $k \to \infty$, we see that $J^* \leq J_\mu$ for all $\mu \in \mathcal{M}$, so that $J^*(x) \leq \inf_{\mu \in \mathcal{M}} J_\mu(x)$ for all $x \in X$. **Q.E.D.**

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

Example 2.1.1 (Counterexample Without Monotonicity)

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

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

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

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

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

Optimality over Nonstationary Policies

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

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

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

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

Note that under the contraction Assumption 2.1.2, *the choice of $\bar{J}$ in the definition of J_π does not matter*, since for any two $J, J' \in \mathcal{B}(X)$, we have

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

so the value of $J_\pi(x)$ is independent of $\bar{J}$. Since by Prop. 2.1.1(b), $J_\mu(x) = \lim_{k\to\infty} (T_\mu^k J)(x)$ for all $\mu \in \mathcal{M}$, $J \in \mathcal{B}(X)$, and $x \in X$, in the DP context we recognize J_μ as the cost function of the stationary policy $\{\mu, \mu, \ldots\}$.

We now claim that under the monotonicity and contraction Assumptions 2.1.1 and 2.1.2, *J^*, which was defined as the unique fixed point of T, is equal to the optimal value of J_π*, i.e.,

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

Indeed, since $\mathcal{M}$ defines a subset of Π, we have from Prop. 2.1.2,

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

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

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

[the monotonicity Assumption 2.1.1 can be used to show that

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

and the last equality holds by Prop. 2.1.1(b)]. Combining the preceding relations, we obtain $J^*(x) = \inf_{\pi\in\Pi} J_\pi(x)$.

Thus, in DP terms, *we may view J^* as an optimal cost function over all policies*, including nonstationary ones. At the same time, Prop. 2.1.2 states that stationary policies are sufficient in the sense that *the optimal cost can be attained to within arbitrary accuracy with a stationary policy* [uniformly for all $x \in X$, as Eq. (2.1) shows].

Error Bounds and Other Inequalities

The analysis of abstract DP algorithms and related approximations requires the use of some basic inequalities that follow from the assumptions of contraction and monotonicity. We have obtained two such results in Prop. 2.1.1(d),(e), which assume only the contraction assumption. These results can be strengthened if in addition to contraction, we have monotonicity. To this end we first show the following useful characterization.

Proposition 2.1.3: The monotonicity and contraction Assumptions 2.1.1 and 2.1.2 hold if and only if for all $J, J' \in \mathcal{B}(X)$, $\mu \in \mathcal{M}$, and scalar $c \geq 0$, we have

$$J \leq J' + c\,v \quad \Rightarrow \quad T_\mu J \leq T_\mu J' + \alpha c\,v, \tag{2.2}$$

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

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

$$H(x,u,J) \leq H(x,u,J'+c\,v) \leq H(x,u,J') + \alpha c\,v(x), \quad \forall\, x \in X,\ u \in U(x), \tag{2.3}$$

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

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

The condition (2.3) implies the desired condition (2.2). Conversely, condition (2.2) for $c = 0$ yields the monotonicity assumption, while for $c = \|J' - J\|$ it yields the contraction assumption. **Q.E.D.**

We can now derive the following useful variant of Prop. 2.1.1(d),(e), which involves one-sided inequalities. This variant will be used in the derivation of error bounds for various computational methods.

Proposition 2.1.4: (Error Bounds Under Contraction and Monotonicity) Let the monotonicity and contraction Assumptions 2.1.1 and 2.1.2 hold. Then:

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

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

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

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

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

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

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

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

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

Proof: (a) We show the first relation. Applying Eq. (2.2) with J' and J replaced by J and TJ, respectively, and taking infimum over $\mu \in \mathcal{M}$, we see that if $TJ \leq J + c\,v$, then $T^2 J \leq TJ + \alpha c\,v$. Proceeding similarly, it follows that

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

We now write for every k,

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

from which, by taking the limit as $k \to \infty$, we obtain

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

The second relation follows similarly.

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

(c) Similar to the proof of part (a), the inequality

$$TJ \leq J + c\,v$$

implies that for all k we have

$$T^{k+1}J \leq T^kJ + \alpha^k c\, v.$$

Applying part (a) with J and c replaced by T^kJ and $\alpha^k c$, respectively, we obtain the first desired relation. The second relation follows similarly. **Q.E.D.**

2.2 LIMITED LOOKAHEAD POLICIES

In this section, we discuss a basic building block in the algorithmic methodology of abstract DP. Given some function $\tilde{J}$ that approximates J^*, we obtain a policy by solving a finite-horizon problem where $\tilde{J}$ is the terminal cost function. The simplest possibility is a *one-step lookahead policy* $\overline{\mu}$ defined by

$$\overline{\mu}(x) \in \arg\min_{u\in U(x)} H(x,u,\tilde{J}), \qquad x \in X. \tag{2.4}$$

Its cost function $J_{\overline{\mu}}$ was interpreted in Section 1.3.1 as the result of a Newton iteration that starts from $\tilde{J}$ and aims to solve the Bellman equation $J = TJ$. The following proposition gives some bounds for its performance.

Proposition 2.2.1: (One-Step Lookahead Error Bounds) Let the contraction Assumption 2.1.2 hold, and let $\overline{\mu}$ be a one-step lookahead policy obtained by minimization in Eq. (2.4), i.e., satisfying $T_{\overline{\mu}}\tilde{J} = T\tilde{J}$. Then

$$\|J_{\overline{\mu}} - T\tilde{J}\| \leq \frac{\alpha}{1-\alpha}\|T\tilde{J} - \tilde{J}\|, \tag{2.5}$$

where $\|\cdot\|$ denotes the weighted sup-norm. Moreover

$$\|J_{\overline{\mu}} - J^*\| \leq \frac{2\alpha}{1-\alpha}\|\tilde{J} - J^*\|, \tag{2.6}$$

and

$$\|J_{\overline{\mu}} - J^*\| \leq \frac{2}{1-\alpha}\|T\tilde{J} - \tilde{J}\|. \tag{2.7}$$

Proof: Equation (2.5) follows from the second relation of Prop. 2.1.1(e) with $J = \tilde{J}$. Also from the first relation of Prop. 2.1.1(e) with $J = J^*$, we have

$$\|J_{\overline{\mu}} - J^*\| \leq \frac{1}{1-\alpha}\|T_{\overline{\mu}}J^* - J^*\|.$$

By using the triangle inequality, and the relations $T_{\overline{\mu}}\tilde{J} = T\tilde{J}$ and $J^* = TJ^*$, we obtain

$$\begin{aligned}\|T_{\overline{\mu}}J^* - J^*\| &\leq \|T_{\overline{\mu}}J^* - T_{\overline{\mu}}\tilde{J}\| + \|T_{\overline{\mu}}\tilde{J} - T\tilde{J}\| + \|T\tilde{J} - J^*\| \\ &= \|T_{\overline{\mu}}J^* - T_{\overline{\mu}}\tilde{J}\| + \|T\tilde{J} - TJ^*\| \\ &\leq \alpha\|J^* - \tilde{J}\| + \alpha\|\tilde{J} - J^*\| \\ &= 2\alpha\,\|\tilde{J} - J^*\|,\end{aligned}$$

and Eq. (2.6) follows by combining the preceding two relations.

Also, from the first relation of Prop. 2.1.1(d) with $J = \tilde{J}$,

$$\|J^* - \tilde{J}\| \leq \frac{1}{1-\alpha}\|T\tilde{J} - \tilde{J}\|. \tag{2.8}$$

Thus

$$\begin{aligned}\|J_{\overline{\mu}} - J^*\| &\leq \|J_{\overline{\mu}} - T\tilde{J}\| + \|T\tilde{J} - \tilde{J}\| + \|\tilde{J} - J^*\| \\ &\leq \frac{\alpha}{1-\alpha}\|T\tilde{J} - \tilde{J}\| + \|T\tilde{J} - \tilde{J}\| + \frac{1}{1-\alpha}\|T\tilde{J} - \tilde{J}\| \\ &= \frac{2}{1-\alpha}\|T\tilde{J} - \tilde{J}\|,\end{aligned}$$

where the second inequality follows from Eqs. (2.5) and (2.8). This proves Eq. (2.7). **Q.E.D.**

Equation (2.5) provides a computable bound on the cost function $J_{\overline{\mu}}$ of the one-step lookahead policy. The bound (2.6) says that if the one-step lookahead approximation $\tilde{J}$ is within ϵ of the optimal, the performance of the one-step lookahead policy is within

$$\frac{2\alpha\epsilon}{1-\alpha}$$

of the optimal. Unfortunately, this is not very reassuring when α is close to 1, in which case the error bound is large relative to ϵ. Nonetheless, the following example from [BeT96], Section 6.1.1, shows that this bound is tight, i.e., for any $\alpha < 1$, there is a problem with just two states where the error bound is satisfied with equality. What is happening is that an $O(\epsilon)$ difference in single stage cost between two controls can generate an $O\big(\epsilon/(1-\alpha)\big)$ difference in policy costs, yet it can be "nullified" in the fixed point equation $J^* = TJ^*$ by an $O(\epsilon)$ difference between J^* and $\tilde{J}$.

Example 2.2.1

Consider a discounted optimal control problem with two states, 1 and 2, and deterministic transitions. State 2 is absorbing, but at state 1 there are two possible decisions: move to state 2 (policy μ^*) or stay at state 1 (policy μ).

The cost of each transition is 0 except for the transition from 1 to itself under policy μ, which has cost $2\alpha\epsilon$, where ϵ is a positive scalar and $\alpha \in [0, 1)$ is the discount factor. The optimal policy μ^* is to move from state 1 to state 2, and the optimal cost-to-go function is $J^*(1) = J^*(2) = 0$. Consider the vector $\tilde{J}$ with $\tilde{J}(1) = -\epsilon$ and $\tilde{J}(2) = \epsilon$, so that

$$\|\tilde{J} - J^*\| = \epsilon,$$

as assumed in Eq. (2.6) (cf. Prop. 2.2.1). The policy μ that decides to stay at state 1 is a one-step lookahead policy based on $\tilde{J}$, because

$$2\alpha\epsilon + \alpha\tilde{J}(1) = \alpha\epsilon = 0 + \alpha\tilde{J}(2).$$

We have

$$J_\mu(1) = \frac{2\alpha\epsilon}{1-\alpha} = \frac{2\alpha}{1-\alpha}\|\tilde{J} - J^*\|,$$

so the bound of Eq. (2.6) holds with equality.

Multistep Lookahead Policies with Approximations

Let us now consider a more general form of lookahead involving multiple stages as well as other approximations of the type that we will consider later in the implementation of various approximate value and policy iteration algorithms. In particular, we will assume that given any $J \in \mathcal{B}(X)$, we cannot compute exactly TJ, but instead we can compute $\tilde{J} \in \mathcal{B}(X)$ and $\mu \in \mathcal{M}$ such that

$$\|\tilde{J} - TJ\| \leq \delta, \qquad \|T_\mu J - TJ\| \leq \epsilon, \tag{2.9}$$

where δ and ϵ are nonnegative scalars. These scalars are usually unknown, so the resulting analysis will have a mostly qualitative character.

The case $\delta > 0$ arises when the state space is either infinite or it is finite but 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. Alternatively, one may use simulation data [e.g., noisy values of $(TJ)(x)$ for some or all x] and some kind of least-squares error fit of $(TJ)(x)$ with a function from a suitable parametric class. The function $\tilde{J}$ thus obtained will satisfy $\|\tilde{J} - TJ\| \leq \delta$ with $\delta > 0$. Note that δ may not be small in this context, and the resulting performance degradation may be a primary concern.

Cases where $\epsilon > 0$ may arise when the control space is infinite or finite but large, and the minimization involved in the calculation of $(TJ)(x)$ cannot be done exactly. Note, however, that it is possible that

$$\delta > 0, \qquad \epsilon = 0,$$

and in fact this occurs often in practice. In an alternative scenario, we may first obtain the policy μ subject to a restriction that it belongs to a certain subset of structured policies, so it satisfies

$$\|T_\mu J - TJ\| \leq \epsilon$$

for some $\epsilon > 0$, and then we may set $\tilde{J} = T_\mu J$. In this case we have $\epsilon = \delta$ in Eq. (2.9).

In a multistep method with approximations, we are given a positive integer m and a lookahead function J_m, and we successively compute (backwards in time) $J_{m-1}, \ldots, J_0$ and policies $\mu_{m-1}, \ldots, \mu_0$ satisfying

$$\|J_k - TJ_{k+1}\| \leq \delta, \quad \|T_{\mu_k} J_{k+1} - TJ_{k+1}\| \leq \epsilon, \qquad k = 0, \ldots, m-1. \tag{2.10}$$

Note that in the context of MDP, J_k can be viewed as an approximation to the optimal cost function of an $(m-k)$-stage problem with terminal cost function J_m. We have the following proposition.

Proposition 2.2.2: (Multistep Lookahead Error Bound) Let the contraction Assumption 2.1.2 hold. The periodic policy

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

generated by the method of Eq. (2.10) satisfies

$$\|J_\pi - J^*\| \leq \frac{2\alpha^m}{1-\alpha^m}\|J_m - J^*\| + \frac{\epsilon}{1-\alpha^m} + \frac{\alpha(\epsilon + 2\delta)(1-\alpha^{m-1})}{(1-\alpha)(1-\alpha^m)}. \tag{2.11}$$

Proof: Using the triangle inequality, Eq. (2.10), and the contraction property of T, we have for all k

$$\begin{aligned} \|J_{m-k} - T^k J_m\| &\leq \|J_{m-k} - TJ_{m-k+1}\| + \|TJ_{m-k+1} - T^2 J_{m-k+2}\| \\ &\quad + \cdots + \|T^{k-1} J_{m-1} - T^k J_m\| \\ &\leq \delta + \alpha\delta + \cdots + \alpha^{k-1}\delta, \end{aligned} \tag{2.12}$$

showing that

$$\|J_{m-k} - T^k J_m\| \leq \frac{\delta(1-\alpha^k)}{1-\alpha}, \qquad k = 1, \ldots, m. \tag{2.13}$$

From Eq. (2.10), we have $\|J_k - T_{\mu_k} J_{k+1}\| \leq \delta + \epsilon$, so for all k

$$\begin{aligned}\|J_{m-k} - T_{\mu_{m-k}} \cdots T_{\mu_{m-1}} J_m\| &\leq \|J_{m-k} - T_{\mu_{m-k}} J_{m-k+1}\| \\ &\quad + \|T_{\mu_{m-k}} J_{m-k+1} - T_{\mu_{m-k}} T_{\mu_{m-k+1}} J_{m-k+2}\| \\ &\quad + \cdots \\ &\quad + \|T_{\mu_{m-k}} \cdots T_{\mu_{m-2}} J_{m-1} - T_{\mu_{m-k}} \cdots T_{\mu_{m-1}} J_m\| \\ &\leq (\delta + \epsilon) + \alpha(\delta + \epsilon) + \cdots + \alpha^{k-1}(\delta + \epsilon),\end{aligned}$$

showing that

$$\|J_{m-k} - T_{\mu_{m-k}} \cdots T_{\mu_{m-1}} J_m\| \leq \frac{(\delta + \epsilon)(1 - \alpha^k)}{1 - \alpha}, \qquad k = 1, \ldots, m. \tag{2.14}$$

Using the fact $\|T_{\mu_0} J_1 - T J_1\| \leq \epsilon$ [cf. Eq. (2.10)], we obtain

$$\begin{aligned}\|T_{\mu_0} \cdots T_{\mu_{m-1}} J_m - T^m J_m\| &\leq \|T_{\mu_0} \cdots T_{\mu_{m-1}} J_m - T_{\mu_0} J_1\| \\ &\quad + \|T_{\mu_0} J_1 - T J_1\| + \|T J_1 - T^m J_m\| \\ &\leq \alpha \|T_{\mu_1} \cdots T_{\mu_{m-1}} J_m - J_1\| + \epsilon + \alpha \|J_1 - T^{m-1} J_m\| \\ &\leq \epsilon + \frac{\alpha(\epsilon + 2\delta)(1 - \alpha^{m-1})}{1 - \alpha},\end{aligned}$$

where the last inequality follows from Eqs. (2.13) and (2.14) for $k = m - 1$.

From this relation and the fact that $T_{\mu_0} \cdots T_{\mu_{m-1}}$ and T^m are contractions with modulus α^m, we obtain

$$\begin{aligned}\|T_{\mu_0} \cdots T_{\mu_{m-1}} J^* - J^*\| &\leq \|T_{\mu_0} \cdots T_{\mu_{m-1}} J^* - T_{\mu_0} \cdots T_{\mu_{m-1}} J_m\| \\ &\quad + \|T_{\mu_0} \cdots T_{\mu_{m-1}} J_m - T^m J_m\| + \|T^m J_m - J^*\| \\ &\leq 2\alpha^m \|J^* - J_m\| + \epsilon + \frac{\alpha(\epsilon + 2\delta)(1 - \alpha^{m-1})}{1 - \alpha}.\end{aligned}$$

We also have using Prop. 2.1.1(e), applied in the context of the multistep mapping of Example 1.3.1,

$$\|J_\pi - J^*\| \leq \frac{1}{1 - \alpha^m} \|T_{\mu_0} \cdots T_{\mu_{m-1}} J^* - J^*\|.$$

Combining the last two relations, we obtain the desired result. **Q.E.D.**

Note that for $m = 1$ and $\delta = \epsilon = 0$, i.e., the case of one-step lookahead policy $\overline{\mu}$ with lookahead function J_1 and no approximation error in the minimization involved in TJ_1, Eq. (2.11) yields the bound

$$\|J_{\overline{\mu}} - J^*\| \leq \frac{2\alpha}{1 - \alpha} \|J_1 - J^*\|,$$

which coincides with the bound (2.6) derived earlier.

Also, in the special case where $\epsilon = \delta$ and $J_k = T_{\mu_k} J_{k+1}$ (cf. the discussion preceding Prop. 2.2.2), the bound (2.11) can be strengthened somewhat. In particular, we have for all k, $J_{m-k} = T_{\mu_{m-k}} \cdots T_{\mu_{m-1}} J_m$, so the right-hand side of Eq. (2.14) becomes 0 and the preceding proof yields, with some calculation,

$$\begin{aligned} \|J_\pi - J^*\| &\leq \frac{2\alpha^m}{1-\alpha^m}\|J_m - J^*\| + \frac{\delta}{1-\alpha^m} + \frac{\alpha\delta(1-\alpha^{m-1})}{(1-\alpha)(1-\alpha^m)} \\ &= \frac{2\alpha^m}{1-\alpha^m}\|J_m - J^*\| + \frac{\delta}{1-\alpha}. \end{aligned}$$

We finally note that Prop. 2.2.2 shows that as the lookahead size m increases, the corresponding bound for $\|J_\pi - J^*\|$ tends to $\epsilon + \alpha(\epsilon + 2\delta)/(1-\alpha)$, or

$$\limsup_{m\to\infty} \|J_\pi - J^*\| \leq \frac{\epsilon + 2\alpha\delta}{1-\alpha}.$$

We will see that this error bound is superior to corresponding error bounds for approximate versions of value and policy iteration by essentially a factor $1/(1-\alpha)$. In practice, however, periodic suboptimal policies, as required by Prop. 2.2.2, are typically not used.

There is an alternative and often used form of *on-line* multistep lookahead, whereby at the current state x we compute a multistep policy $\{\mu_0, \ldots, \mu_{m-1}\}$, we apply the first component $\mu_0(x)$ of that policy at state x, then at the next state $\bar{x}$ we recompute a new multistep policy $\{\bar{\mu}_0, \ldots, \bar{\mu}_{m-1}\}$, apply $\bar{\mu}_0(\bar{x})$, etc. However, no error bound similar to the one of Prop. 2.2.2 is currently known for this type of lookahead.

2.3 VALUE ITERATION

In this section, we discuss value iteration (VI for short), the algorithm that starts with some $J \in \mathcal{B}(X)$, and generates $TJ, T^2J, \ldots$. Since T is a weighted sup-norm contraction under Assumption 2.1.2, the algorithm converges to J^*, and the rate of convergence is governed by

$$\|T^k J - J^*\| \leq \alpha^k \|J - J^*\|, \qquad k = 0, 1, \ldots.$$

Similarly, for a given policy $\mu \in \mathcal{M}$, we have

$$\|T_\mu^k J - J_\mu\| \leq \alpha^k \|J - J_\mu\|, \qquad k = 0, 1, \ldots.$$

From Prop. 2.1.1(d), we also have the error bound

$$\|T^{k+1} J - J^*\| \leq \frac{\alpha}{1-\alpha}\|T^{k+1} J - T^k J\|, \qquad k = 0, 1, \ldots.$$

This bound does not rely on the monotonicity Assumption 2.1.1.

The VI algorithm is often used to compute an approximation $\tilde{J}$ to J^*, and then to obtain a policy $\overline{\mu}$ by minimizing $H(x,u,\tilde{J})$ over $u \in U(x)$ for each $x \in X$. In other words $\tilde{J}$ and $\overline{\mu}$ satisfy

$$\|\tilde{J} - J^*\| \leq \gamma, \qquad T_{\overline{\mu}}\tilde{J} = T\tilde{J},$$

where γ is some positive scalar. Then by using Eq. (2.6), we have

$$\|J_{\overline{\mu}} - J^*\| \leq \frac{2\alpha\gamma}{1-\alpha}. \tag{2.15}$$

If the set of policies is finite, this procedure can be used to compute an optimal policy with a finite but sufficiently large number of exact VI, as shown in the following proposition.

Proposition 2.3.1: Let the contraction Assumption 2.1.2 hold and let $J \in \mathcal{B}(X)$. If the set of policies $\mathcal{M}$ is finite, there exists an integer $\overline{k} \geq 0$ such that $J_{\mu^*} = J^*$ for all μ^* and $k \geq \overline{k}$ with $T_{\mu^*}T^kJ = T^{k+1}J$.

Proof: Let $\tilde{\mathcal{M}}$ be the set of policies such that $J_\mu \neq J^*$. Since $\tilde{\mathcal{M}}$ is finite, we have

$$\inf_{\mu \in \tilde{\mathcal{M}}} \|J_\mu - J^*\| > 0,$$

so by Eq. (2.15), there exists sufficiently small $\beta > 0$ such that

$$\|\tilde{J} - J^*\| \leq \beta \ \text{ and } \ T_\mu\tilde{J} = T\tilde{J} \quad \Rightarrow \quad \|J_\mu - J^*\| = 0 \quad \Rightarrow \quad \mu \notin \tilde{\mathcal{M}}. \tag{2.16}$$

It follows that if k is sufficiently large so that $\|T^kJ - J^*\| \leq \beta$, then $T_{\mu^*}T^kJ = T^{k+1}J$ implies that $\mu^* \notin \tilde{\mathcal{M}}$ so $J_{\mu^*} = J^*$. **Q.E.D.**

2.3.1 Approximate Value Iteration

We will now consider situations where the VI method may be implementable only through approximations. In particular, given a function J, assume that we may only be able to calculate an approximation $\tilde{J}$ to TJ such that

$$\big\|\tilde{J} - TJ\big\| \leq \delta,$$

where δ is a given positive scalar. In the corresponding approximate VI method, we start from an arbitrary bounded function J_0, and we generate a sequence $\{J_k\}$ satisfying

$$\|J_{k+1} - TJ_k\| \leq \delta, \qquad k = 0, 1, \ldots. \tag{2.17}$$

This approximation may be the result of representing J_{k+1} compactly, as a linear combination of basis functions, through a projection or aggregation process, as is common in approximate DP (cf. the discussion of Section 1.2.4).

We may also simultaneously generate a sequence of policies $\{\mu^k\}$ such that

$$\|T_{\mu^k} J_k - T J_k\| \leq \epsilon, \qquad k = 0, 1, \ldots, \tag{2.18}$$

where ϵ is some scalar [which could be equal to 0, as in case of Eq. (2.10), considered earlier]. The following proposition shows that the corresponding cost functions J_{μ^k} "converge" to J^* to within an error of order $O\big(\delta/(1-\alpha)^2\big)$ [plus a less significant error of order $O\big(\epsilon/(1-\alpha)\big)$].

Proposition 2.3.2: (Error Bounds for Approximate VI) Let the contraction Assumption 2.1.2 hold. A sequence $\{J_k\}$ generated by the approximate VI method (2.17)-(2.18) satisfies

$$\limsup_{k\to\infty} \|J_k - J^*\| \leq \frac{\delta}{1-\alpha}, \tag{2.19}$$

while the corresponding sequence of policies $\{\mu^k\}$ satisfies

$$\limsup_{k\to\infty} \|J_{\mu^k} - J^*\| \leq \frac{\epsilon}{1-\alpha} + \frac{2\alpha\delta}{(1-\alpha)^2}. \tag{2.20}$$

Proof: Using the triangle inequality, Eq. (2.17), and the contraction property of T, we have

$$\begin{aligned} \|J_k - T^k J_0\| &\leq \|J_k - T J_{k-1}\| \\ &\quad + \|T J_{k-1} - T^2 J_{k-2}\| + \cdots + \|T^{k-1} J_1 - T^k J_0\| \\ &\leq \delta + \alpha\delta + \cdots + \alpha^{k-1}\delta, \end{aligned}$$

and finally

$$\|J_k - T^k J_0\| \leq \frac{(1-\alpha^k)\delta}{1-\alpha}, \qquad k = 0, 1, \ldots. \tag{2.21}$$

By taking limit as $k \to \infty$ and by using the fact $\lim_{k\to\infty} T^k J_0 = J^*$, we obtain Eq. (2.19).

We also have using the triangle inequality and the contraction property of T_{μ^k} and T,

$$\begin{aligned} \|T_{\mu^k} J^* - J^*\| &\leq \|T_{\mu^k} J^* - T_{\mu^k} J_k\| + \|T_{\mu^k} J_k - T J_k\| + \|T J_k - J^*\| \\ &\leq \alpha\|J^* - J_k\| + \epsilon + \alpha\|J_k - J^*\|, \end{aligned}$$

while by using also Prop. 2.1.1(e), we obtain

$$\|J_{\mu^k} - J^*\| \leq \frac{1}{1-\alpha}\|T_{\mu^k}J^* - J^*\| \leq \frac{\epsilon}{1-\alpha} + \frac{2\alpha}{1-\alpha}\|J_k - J^*\|.$$

By combining this relation with Eq. (2.19), we obtain Eq. (2.20). **Q.E.D.**

The error bound (2.20) relates to stationary policies obtained from the functions J_k by one-step lookahead. We may also obtain an m-step periodic policy π from J_k by using m-step lookahead. Then Prop. 2.2.2 shows that the corresponding bound for $\|J_\pi - J^*\|$ tends to $(\epsilon+2\alpha\delta)/(1-\alpha)$ as $m \to \infty$, which improves on the error bound (2.20) by a factor $1/(1-\alpha)$.

Finally, let us note that the error bound of Prop. 2.3.2 is predicated upon generating a sequence $\{J_k\}$ satisfying $\|J_{k+1} - TJ_k\| \leq \delta$ for all k [cf. Eq. (2.17)]. Unfortunately, some practical approximation schemes guarantee the existence of such a δ only if $\{J_k\}$ is a bounded sequence. The following example from [BeT96], Section 6.5.3, shows that boundedness of the iterates is not automatically guaranteed, and is a serious issue that should be addressed in approximate VI schemes.

Example 2.3.1 (Error Amplification in Approximate Value Iteration)

Consider a two-state α-discounted MDP with states 1 and 2, and a single policy. The transitions are deterministic: from state 1 to state 2, and from state 2 to state 2. These transitions are also cost-free. Thus we have $(TJ(1) = (TJ)(2) = \alpha J(2)$, and $J^*(1) = J^*(2) = 0$.

We consider a VI scheme that approximates cost functions within the one-dimensional subspace of linear functions $S = \big\{(r, 2r) \mid r \in \Re\big\}$ by using a weighted least squares minimization; i.e., we approximate a vector J by its weighted Euclidean projection onto S. In particular, given $J_k = (r_k, 2r_k)$, we find $J_{k+1} = (r_{k+1}, 2r_{k+1})$, where for weights $\xi_1, \xi_2 > 0$, r_{k+1} is obtained as

$$r_{k+1} \in \arg\min_r \Big[\xi_1\big(r - (TJ_k)(1)\big)^2 + \xi_2\big(2r - (TJ_k)(2)\big)^2\Big].$$

Since for a zero cost per stage and the given deterministic transitions, we have $TJ_k = (2\alpha r_k, 2\alpha r_k)$, the preceding minimization is written as

$$r_{k+1} \in \arg\min_r \big[\xi_1(r - 2\alpha r_k)^2 + \xi_2(2r - 2\alpha r_k)^2\big],$$

which by writing the corresponding optimality condition yields $r_{k+1} = \alpha\beta r_k$, where $\beta = 2(\xi_1 + 2\xi_2)(\xi_1 + 4\xi_2) > 1$. Thus if $\alpha > 1/\beta$, the sequence $\{r_k\}$ diverges and so does $\{J_k\}$. Note that in this example the optimal cost function $J^* = (0, 0)$ belongs to the subspace S. The difficulty here is that the approximate VI mapping that generates J_{k+1} as the weighted Euclidean projection of TJ_k is not a contraction (this is a manifestation of an important issue in approximate DP and projected equation approximation, namely that the projected mapping ΠT need not be a contraction even if T is a sup-norm contraction; see [DFV00], [Ber12b] for examples and related discussions). At the same time there is no δ such that $\|J_{k+1} - TJ_k\| \leq \delta$ for all k, because of error amplification in each approximate VI.

2.4 POLICY ITERATION

In this section, we discuss policy iteration (PI for short), an algorithm whereby we maintain and update a policy μ^k, starting from some initial policy μ^0. The typical iteration has the following form (see Fig. 2.4.1 for a one-dimensional illustration).

Policy iteration given the current policy μ^k:

Policy evaluation: We compute J_{μ^k} as the unique solution of the equation

$$J_{\mu^k} = T_{\mu^k} J_{\mu^k}.$$

Policy improvement: We obtain a policy μ^{k+1} that satisfies

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

We assume that the minimum of $H(x, u, J_{\mu^k})$ over $u \in U(x)$ is attained for all $x \in X$, so that the improved policy μ^{k+1} is defined (we use this assumption for all the PI algorithms of the book). The following proposition establishes a basic cost improvement property, as well as finite convergence for the case where the set of policies is finite.

Proposition 2.4.1: (Convergence of PI) Let the monotonicity and contraction Assumptions 2.1.1 and 2.1.2 hold, and let $\{\mu^k\}$ be a sequence generated by the PI algorithm. Then for all k, we have $J_{\mu^{k+1}} \leq J_{\mu^k}$, with equality if and only if $J_{\mu^k} = J^*$. Moreover,

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

and if the set of policies is finite, we have $J_{\mu^k} = J^*$ for some k.

Proof: We have

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

Applying $T_{\mu^{k+1}}$ to this inequality while using the monotonicity Assumption 2.1.1, we obtain

$$T^2_{\mu^{k+1}} J_{\mu^k} \leq T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k} \leq T_{\mu^k} J_{\mu^k} = J_{\mu^k}.$$

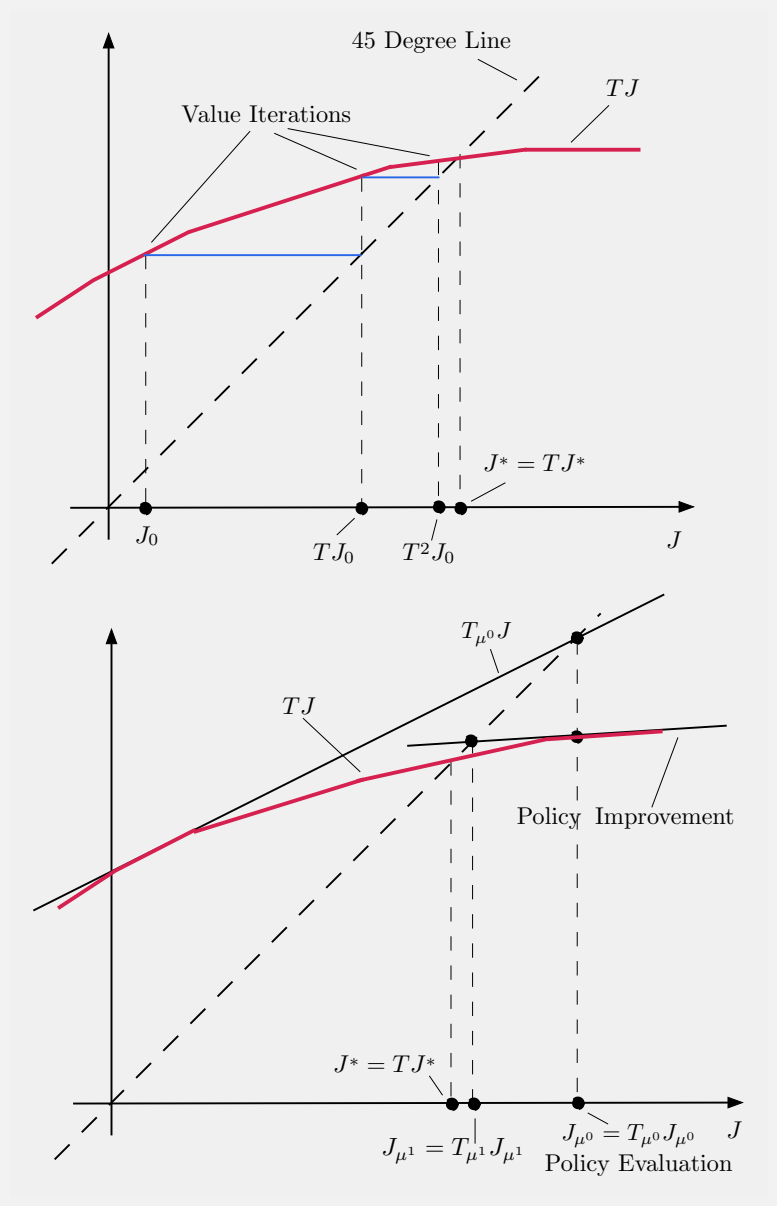


Figure 2.4.1 Geometric interpretation of PI and VI in one dimension (a single state). Each policy μ defines the mapping T_μ, and TJ is the function $\min_\mu T_\mu J$. When the number of policies is finite, TJ is a piecewise linear concave function, with each piece being a linear function $T_\mu J$ that corresponds to a policy μ. The optimal cost function J^* satisfies $J^* = TJ^*$, so it is obtained from the intersection of the graph of TJ and the 45 degree line shown. Similarly J_μ is the intersection of the graph of $T_\mu J$ and the 45 degree line. The VI sequence is indicated in the top figure by the staircase construction, which asymptotically leads to J^*. A single policy iteration is illustrated in the bottom figure, and illustrates the connection of PI with Newton's method that was discussed in Section 1.3.2.

Similarly, we have for all $m > 0$,

$$T^m_{\mu^{k+1}} J_{\mu^k} \leq T J_{\mu^k} \leq J_{\mu^k},$$

and by taking the limit as $m \to \infty$, we obtain

$$J_{\mu^{k+1}} \leq T J_{\mu^k} \leq J_{\mu^k}, \qquad k = 0, 1, \ldots. \tag{2.22}$$

If $J_{\mu^{k+1}} = J_{\mu^k}$, it follows that $T J_{\mu^k} = J_{\mu^k}$, so J_{μ^k} is a fixed point of T and must be equal to J^*. Moreover by using induction, Eq. (2.22) implies that

$$J_{\mu^k} \leq T^k J_{\mu^0}, \qquad k = 0, 1, \ldots.$$

Since

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

it follows that $\lim_{k\to\infty} \|J_{\mu^k} - J^*\| = 0$.

Finally, if the number of policies is finite, Eq. (2.22) implies that there can be only a finite number of iterations for which $J_{\mu^{k+1}}(x) < J_{\mu^k}(x)$ for some x. Thus we must have $J_{\mu^{k+1}} = J_{\mu^k}$ for some k, at which time $J_{\mu^k} = J^*$ as shown earlier [cf. Eq. (2.22)]. **Q.E.D.**

In the case where the set of policies is infinite, we may assert the convergence of the sequence of generated policies under some compactness and continuity conditions. In particular, we will assume that the state space is finite, $X = \{1, \ldots, n\}$, and that each control constraint set $U(x)$ is a compact subset of $\Re^m$. We will view a cost function J as an element of $\Re^n$, and a policy μ as an element of the set $U(1) \times \cdots \times U(n) \subset \Re^{mn}$, which is compact. Then $\{\mu^k\}$ has at least one limit point $\overline{\mu}$, which must be an admissible policy. The following proposition guarantees, under an additional continuity assumption for $H(x, \cdot, \cdot)$, that every limit point $\overline{\mu}$ is optimal.

Assumption 2.4.1: (Compactness and Continuity)

(a) The state space is finite, $X = \{1, \ldots, n\}$.

(b) Each control constraint set $U(x)$, $x = 1, \ldots, n$, is a compact subset of $\Re^m$.

(c) Each function $H(x, \cdot, \cdot)$, $x = 1, \ldots, n$, is continuous over $U(x) \times \Re^n$.

Proposition 2.4.2: Let the monotonicity and contraction Assumptions 2.1.1 and 2.1.2 hold, together with Assumption 2.4.1, and let $\{\mu^k\}$ be a sequence generated by the PI algorithm. Then for every limit point $\overline{\mu}$ of $\{\mu^k\}$, we have $J_{\overline{\mu}} = J^*$.

Proof: We have $J_{\mu^k} \to J^*$ by Prop. 2.4.1. Let $\overline{\mu}$ be the limit of a subsequence $\{\mu^k\}_{k\in\mathcal{K}}$. We will show that $T_{\overline{\mu}}J^* = TJ^*$, from which it follows that $J_{\overline{\mu}} = J^*$ [cf. Prop. 2.1.1(c)]. Indeed, we have $T_{\overline{\mu}}J^* \geq TJ^*$, so we focus on showing the reverse inequality. From the equation

$$T_{\mu^k}J_{\mu^{k-1}} = TJ_{\mu^{k-1}},$$

we have

$$H\big(x,\mu^k(x),J_{\mu^{k-1}}\big) \leq H(x,u,J_{\mu^{k-1}}), \qquad x=1,\ldots,n,\ u\in U(x).$$

By taking limit in this relation as $k \to \infty$, $k \in \mathcal{K}$, and by using the continuity of $H(x,\cdot,\cdot)$ [cf. Assumption 2.4.1(c)], we obtain

$$H\big(x,\overline{\mu}(x),J^*\big) \leq H(x,u,J^*), \qquad x=1,\ldots,n,\ u\in U(x).$$

By taking the minimum of the right-hand side over $u \in U(x)$, we obtain $T_{\overline{\mu}}J^* \leq TJ^*$. **Q.E.D.**

2.4.1 Approximate Policy Iteration

We now consider the PI method where the policy evaluation step and/or the policy improvement step of the method are implemented through approximations. This method generates a sequence of policies $\{\mu^k\}$ and a corresponding sequence of approximate cost functions $\{J_k\}$ satisfying

$$\|J_k - J_{\mu^k}\| \leq \delta, \qquad \|T_{\mu^{k+1}}J_k - TJ_k\| \leq \epsilon, \qquad k=0,1,\ldots, \tag{2.23}$$

where δ and ϵ are some scalars, and $\|\cdot\|$ denotes the weighted sup-norm (the one used in the contraction Assumption 2.1.2). The following proposition provides an error bound for this algorithm.

Proposition 2.4.3: (Error Bound for Approximate PI) Let the monotonicity and contraction Assumptions 2.1.1 and 2.1.2 hold. The sequence $\{\mu^k\}$ generated by the approximate PI algorithm (2.23) satisfies

$$\limsup_{k\to\infty} \|J_{\mu^k} - J^*\| \leq \frac{\epsilon + 2\alpha\delta}{(1-\alpha)^2}. \tag{2.24}$$

The essence of the proof is contained in the following proposition, which quantifies the amount of approximate policy improvement at each iteration.

Proposition 2.4.4: Let the monotonicity and contraction Assumptions 2.1.1 and 2.1.2 hold. Let J, $\overline{\mu}$, and μ satisfy

$$\|J - J_\mu\| \leq \delta, \qquad \|T_{\overline{\mu}} J - TJ\| \leq \epsilon,$$

where δ and ϵ are some scalars. Then

$$\|J_{\overline{\mu}} - J^*\| \leq \alpha \|J_\mu - J^*\| + \frac{\epsilon + 2\alpha\delta}{1-\alpha}. \tag{2.25}$$

Proof: We denote by v the weight function corresponding to the weighted sup-norm. Using the contraction property of T and $T_{\overline{\mu}}$, which implies that $\|T_{\overline{\mu}} J_\mu - T_{\overline{\mu}} J\| \leq \alpha\delta$ and $\|TJ - TJ_\mu\| \leq \alpha\delta$, and hence $T_{\overline{\mu}} J_\mu \leq T_{\overline{\mu}} J + \alpha\delta\, v$ and $TJ \leq TJ_\mu + \alpha\delta\, v$, we have

$$T_{\overline{\mu}} J_\mu \leq T_{\overline{\mu}} J + \alpha\delta\, v \leq TJ + (\epsilon + \alpha\delta)\, v \leq TJ_\mu + (\epsilon + 2\alpha\delta)\, v. \tag{2.26}$$

Since $TJ_\mu \leq T_\mu J_\mu = J_\mu$, this relation yields

$$T_{\overline{\mu}} J_\mu \leq J_\mu + (\epsilon + 2\alpha\delta)\, v,$$

and applying Prop. 2.1.4(b) with $\mu = \overline{\mu}$, $J = J_\mu$, and $c = \epsilon + 2\alpha\delta$, we obtain

$$J_{\overline{\mu}} \leq J_\mu + \frac{\epsilon + 2\alpha\delta}{1-\alpha}\, v. \tag{2.27}$$

Using this relation, we have

$$J_{\overline{\mu}} = T_{\overline{\mu}} J_{\overline{\mu}} = T_{\overline{\mu}} J_\mu + (T_{\overline{\mu}} J_{\overline{\mu}} - T_{\overline{\mu}} J_\mu) \leq T_{\overline{\mu}} J_\mu + \frac{\alpha(\epsilon + 2\alpha\delta)}{1-\alpha}\, v,$$

where the inequality follows by using Prop. 2.1.3 and Eq. (2.27). Subtracting J^* from both sides, we have

$$J_{\overline{\mu}} - J^* \leq T_{\overline{\mu}} J_\mu - J^* + \frac{\alpha(\epsilon + 2\alpha\delta)}{1-\alpha}\, v. \tag{2.28}$$

Also by subtracting J^* from both sides of Eq. (2.26), and using the contraction property

$$TJ_\mu - J^* = TJ_\mu - TJ^* \leq \alpha \|J_\mu - J^*\|\, v,$$

we obtain

$$T_{\overline{\mu}} J_\mu - J^* \leq TJ_\mu - J^* + (\epsilon + 2\alpha\delta)\, v \leq \alpha \|J_\mu - J^*\|\, v + (\epsilon + 2\alpha\delta)\, v.$$

Combining this relation with Eq. (2.28), yields

$$J_{\overline{\mu}} - J^* \leq \alpha \|J_\mu - J^*\| \, v + \frac{\alpha(\epsilon + 2\alpha\delta)}{1-\alpha} \, v + (\epsilon + \alpha\delta) e = \alpha \|J_\mu - J^*\| \, v + \frac{\epsilon + 2\alpha\delta}{1-\alpha} \, v,$$

which is equivalent to the desired relation (2.25). **Q.E.D.**

Proof of Prop. 2.4.3: Applying Prop. 2.4.4, we have

$$\|J_{\mu^{k+1}} - J^*\| \leq \alpha \|J_{\mu^k} - J^*\| + \frac{\epsilon + 2\alpha\delta}{1-\alpha},$$

which by taking the lim sup of both sides as $k \to \infty$ yields the desired result. **Q.E.D.**

We note that the error bound of Prop. 2.4.3 is tight, as can be shown with an example from [BeT96], Section 6.2.3. The error bound is comparable to the one for approximate VI, derived earlier in Prop. 2.3.2. In particular, the error $\|J_{\mu^k} - J^*\|$ is asymptotically proportional to $1/(1-\alpha)^2$ and to the approximation error in policy evaluation or value iteration, respectively. This is noteworthy, as it indicates that contrary to the case of exact implementation, approximate PI need not hold a convergence rate advantage over approximate VI, despite its greater overhead per iteration.

Note that when $\delta = \epsilon = 0$, Eq. (2.25) yields

$$\|J_{\mu^{k+1}} - J^*\| \leq \alpha \|J_{\mu^k} - J^*\|.$$

Thus in the case of an infinite state space and/or control space, exact PI converges at a geometric rate under the contraction and monotonicity assumptions of this section. This rate is the same as the rate of convergence of exact VI. It follows that judging solely from the point of view of rate of convergence estimates, exact PI holds an advantage over exact VI only when the number of states is finite. This raises the question what happens when the number of states is finite but very large. However, this question is not very interesting from a practical point of view, since for a very large number of states, neither VI or PI can be implemented in practice without approximations (see the discussion of Section 1.2.4).

2.4.2 Approximate Policy Iteration Where Policies Converge

Generally, the policy sequence $\{\mu^k\}$ generated by approximate PI may oscillate between several policies. However, under some circumstances this sequence may be guaranteed to converge to some $\overline{\mu}$, in the sense that

$$\mu^{\overline{k}+1} = \mu^{\overline{k}} = \overline{\mu} \qquad \text{for some } \overline{k}. \tag{2.29}$$

An example arises when the policy sequence $\{\mu^k\}$ is generated by *exact PI* applied with a *different* mapping $\tilde{H}$ in place of H, but the policy evaluation and policy improvement error bounds of Eq. (2.23) are satisfied. The mapping $\tilde{H}$ may for example correspond to an approximation of the original problem (as in the aggregation methods of Example 1.2.10; see [Ber11c] and [Ber12a] for further discussion). In this case we can show the following bound, which is much more favorable than the one of Prop. 2.4.3.

Proposition 2.4.5: (Error Bound for Approximate PI when Policies Converge) Let the monotonicity and contraction Assumptions 2.1.1 and 2.1.2 hold, and assume that the approximate PI algorithm (2.23) terminates with a policy $\overline{\mu}$ that satisfies condition (2.29). Then we have

$$\|J_{\overline{\mu}} - J^*\| \leq \frac{\epsilon + 2\alpha\delta}{1-\alpha}. \tag{2.30}$$

Proof: Let $\tilde{J}$ be the cost function obtained by approximate policy evaluation of $\overline{\mu}$ [i.e., $\tilde{J} = J_{\overline{k}}$, where $\overline{k}$ satisfies the condition (2.29)]. Then we have

$$\|\tilde{J} - J_{\overline{\mu}}\| \leq \delta, \qquad \|T_{\overline{\mu}}\tilde{J} - T\tilde{J}\| \leq \epsilon, \tag{2.31}$$

where the latter inequality holds since we have

$$\|T_{\overline{\mu}}\tilde{J} - T\tilde{J}\| = \|T_{\mu^{\overline{k}+1}}J_{\overline{k}} - TJ_{\overline{k}}\| \leq \epsilon,$$

cf. Eq. (2.23). Using Eq. (2.31) and the fact $J_{\overline{\mu}} = T_{\overline{\mu}}J_{\overline{\mu}}$, we have

$$\begin{aligned} \|TJ_{\overline{\mu}} - J_{\overline{\mu}}\| &\leq \|TJ_{\overline{\mu}} - T\tilde{J}\| + \|T\tilde{J} - T_{\overline{\mu}}\tilde{J}\| + \|T_{\overline{\mu}}\tilde{J} - J_{\overline{\mu}}\| \\ &= \|TJ_{\overline{\mu}} - T\tilde{J}\| + \|T\tilde{J} - T_{\overline{\mu}}\tilde{J}\| + \|T_{\overline{\mu}}\tilde{J} - T_{\overline{\mu}}J_{\overline{\mu}}\| \\ &\leq \alpha\|J_{\overline{\mu}} - \tilde{J}\| + \epsilon + \alpha\|\tilde{J} - J_{\overline{\mu}}\| \\ &\leq \epsilon + 2\alpha\delta. \end{aligned} \tag{2.32}$$

Using Prop. 2.1.1(d) with $J = J_{\overline{\mu}}$, we obtain the error bound (2.30). **Q.E.D.**

The preceding error bound can be extended to the case where two successive policies generated by the approximate PI algorithm are "not too different" rather than being identical. In particular, suppose that μ and $\overline{\mu}$ are successive policies, which in addition to

$$\|\tilde{J} - J_{\mu}\| \leq \delta, \qquad \|T_{\overline{\mu}}\tilde{J} - T\tilde{J}\| \leq \epsilon,$$

[cf. Eq. (2.23)], also satisfy

$$\|T_{\mu}\tilde{J} - T_{\overline{\mu}}\tilde{J}\| \leq \zeta,$$

where ζ is some scalar (instead of $\mu = \overline{\mu}$, which is the case where policies converge exactly). Then we also have

$$\|T\tilde{J} - T_\mu \tilde{J}\| \leq \|T\tilde{J} - T_{\overline{\mu}}\tilde{J}\| + \|T_{\overline{\mu}}\tilde{J} - T_\mu \tilde{J}\| \leq \epsilon + \zeta,$$

and by replacing ϵ with $\epsilon + \zeta$ and $\overline{\mu}$ with μ in Eq. (2.32), we obtain

$$\|J_\mu - J^*\| \leq \frac{\epsilon + \zeta + 2\alpha\delta}{1-\alpha}.$$

When ζ is small enough to be of the order of $\max\{\delta, \epsilon\}$, this error bound is comparable to the one for the case where policies converge.

2.5 OPTIMISTIC POLICY ITERATION AND λ-POLICY ITERATION

In this section, we discuss some variants of the PI algorithm of the preceding section, where the policy evaluation

$$J_{\mu^k} = T_{\mu^k} J_{\mu^k}$$

is approximated by using VI. The most straightforward of these methods is *optimistic PI* (also called "modified" PI, see e.g., [Put94]), where a policy μ^k is evaluated approximately, using a finite number of VI. Thus, starting with a function $J_0 \in \mathcal{B}(X)$, we generate sequences $\{J_k\}$ and $\{\mu^k\}$ with the algorithm

$$T_{\mu^k} J_k = TJ_k, \qquad J_{k+1} = T_{\mu^k}^{m_k} J_k, \qquad k = 0, 1, \ldots, \tag{2.33}$$

where $\{m_k\}$ is a sequence of positive integers (see Fig. 2.5.1, which shows one iteration of the method where $m_k = 3$). There is no systematic guideline for selecting the integers m_k. Usually their best values are chosen empirically, and tend to be considerably larger than 1 (in the case where $m_k \equiv 1$ the optimistic PI method coincides with the VI method). The convergence of this method is discussed in Section 2.5.1.

Variants of optimistic PI include methods with approximations in the policy evaluation and policy improvement phases (Section 2.5.2), and methods where the number m_k is randomized (Section 2.5.3). An interesting advantage of the latter methods is that *they do not require the monotonicity Assumption 2.1.1* for convergence in problems with a finite number of policies.

A method that is conceptually similar to the optimistic PI method is the λ-PI method defined by

$$T_{\mu^k} J_k = TJ_k, \qquad J_{k+1} = T_{\mu^k}^{(\lambda)} J_k, \qquad k = 0, 1, \ldots, \tag{2.34}$$

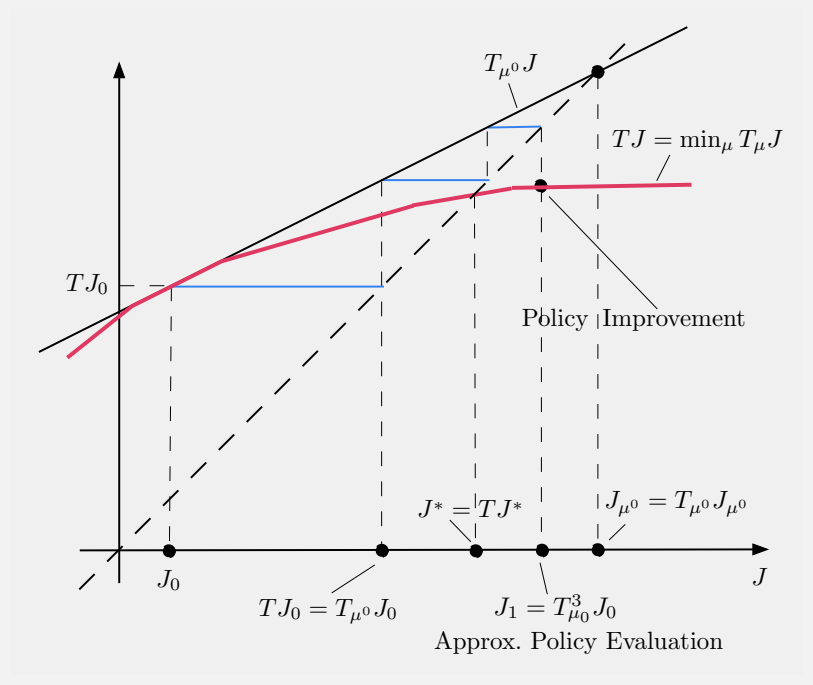


Figure 2.5.1 Illustration of optimistic PI in one dimension. In this example, the policy μ^0 is evaluated approximately with just three applications of T_{μ^0} to yield $J_1 = T^3_{\mu^0}J_0$.

where J_0 is an initial function in $\mathcal{B}(X)$, and for any policy μ and scalar $\lambda \in (0,1)$, $T_\mu^{(\lambda)}$ is the multistep mapping defined by

$$T_\mu^{(\lambda)}J = (1-\lambda)\sum_{\ell=0}^{\infty}\lambda^\ell T_\mu^{\ell+1}J, \qquad J \in \mathcal{B}(X),$$

(cf. Section 1.2.5). To compare optimistic PI and λ-PI, note that they both involve multiple applications of the VI mapping T_{μ^k}: a fixed number m_k in the former case, and a geometrically weighted number in the latter case. In fact, we may view the λ-PI iterate $T_{\mu^k}^{(\lambda)}J_k$ as the expected value of the optimistic PI iterate $T_{\mu^k}^{m_k}J_{\mu^k}$ when m_k is chosen by a geometric probability distribution with parameter λ.

One of the reasons that make λ-PI interesting is its relation with TD(λ) and other temporal difference methods on one hand, and the proximal algorithm on the other. In particular, in λ-PI a policy evaluation is performed with a single iteration of an extrapolated proximal algorithm; cf. the discussion of Section 1.2.5 and Exercise 1.2. Thus implementation

of λ-PI can benefit from the rich methodology that has developed around temporal difference and proximal methods.

Generally the optimistic and λ-PI methods have similar convergence properties. In this section, we focus primarily on optimistic PI, and we discuss briefly λ-PI in Section 2.5.3, where we will prove convergence for a randomized version. For a convergence proof of λ-PI without randomization in discounted stochastic optimal control and stochastic shortest path problems, see the paper [BeI96] and the book [BeT96] (Section 2.3.1).

2.5.1 Convergence of Optimistic Policy Iteration

We will now focus on the optimistic PI algorithm (2.33). The following two propositions provide its convergence properties.

Proposition 2.5.1: (Convergence of Optimistic PI) Let the monotonicity and contraction Assumptions 2.1.1 and 2.1.2 hold, and let $\{(J_k, \mu^k)\}$ be a sequence generated by the optimistic PI algorithm (2.33). Then

$$\lim_{k\to\infty} \|J_k - J^*\| = 0,$$

and if the number of policies is finite, we have $J_{\mu^k} = J^*$ for all k greater than some index $\bar{k}$.

Proposition 2.5.2: Let the monotonicity and contraction Assumptions 2.1.1 and 2.1.2 hold, together with Assumption 2.4.1, and let $\{(J_k, \mu^k)\}$ be a sequence generated by the optimistic PI algorithm (2.33). Then for every limit point $\overline{\mu}$ of $\{\mu^k\}$, we have $J_{\overline{\mu}} = J^*$.

We develop the proofs of the propositions through four lemmas. The first lemma collects some properties of monotone weighted sup-norm contractions, variants of which we noted earlier and we restate for convenience.

Lemma 2.5.1: Let $W : \mathcal{B}(X) \mapsto \mathcal{B}(X)$ be a mapping that satisfies the monotonicity assumption

$$J \leq J' \quad \Rightarrow \quad WJ \leq WJ', \qquad \forall\ J, J' \in \mathcal{B}(X),$$

and the contraction assumption

$$\|WJ - WJ'\| \leq \alpha \|J - J'\|, \qquad \forall\ J, J' \in \mathcal{B}(X),$$

for some $\alpha \in (0, 1)$.

(a) For all $J, J' \in \mathcal{B}(X)$ and scalar $c \geq 0$, we have

$$J \geq J' - c\,v \qquad \Rightarrow \qquad WJ \geq WJ' - \alpha c\,v. \tag{2.35}$$

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

$$J \geq WJ - c\,v \qquad \Rightarrow \qquad W^k J \geq J^* - \frac{\alpha^k}{1-\alpha} c\,v, \tag{2.36}$$

$$WJ \geq J - c\,v \qquad \Rightarrow \qquad J^* \geq W^k J - \frac{\alpha^k}{1-\alpha} c\,v, \tag{2.37}$$

where J^* is the fixed point of W.

Proof: The proof of part (a) follows the one of Prop. 2.1.4(b), while the proof of part (b) follows the one of Prop. 2.1.4(c). **Q.E.D.**

Lemma 2.5.2: Let the monotonicity and contraction Assumptions 2.1.1 and 2.1.2 hold, let $J \in \mathcal{B}(X)$ and $c \geq 0$ satisfy

$$J \geq TJ - c\,v,$$

and let $\mu \in \mathcal{M}$ be such that $T_\mu J = TJ$. Then for all $k > 0$, we have

$$TJ \geq T_\mu^k J - \frac{\alpha}{1-\alpha} c\,v, \tag{2.38}$$

and

$$T_\mu^k J \geq T(T_\mu^k J) - \alpha^k c\,v. \tag{2.39}$$

Proof: Since $J \geq TJ - c\,v = T_\mu J - c\,v$, by using Lemma 2.5.1(a) with $W = T_\mu^j$ and $J' = T_\mu J$, we have for all $j \geq 1$,

$$T_\mu^j J \geq T_\mu^{j+1} J - \alpha^j c\,v. \tag{2.40}$$

By adding this relation over $j = 1, \ldots, k-1$, we have

$$TJ = T_\mu J \geq T_\mu^k J - \sum_{j=1}^{k-1} \alpha^j c\,v = T_\mu^k J - \frac{\alpha - \alpha^k}{1-\alpha} c\,v \geq T_\mu^k J - \frac{\alpha}{1-\alpha} c\,v,$$

showing Eq. (2.38). From Eq. (2.40) for $j = k$, we obtain

$$T_\mu^k J \geq T_\mu^{k+1} J - \alpha^k c\, v = T_\mu(T_\mu^k J) - \alpha^k c\, v \geq T(T_\mu^k J) - \alpha^k c\, v,$$

showing Eq. (2.39). **Q.E.D.**

The next lemma applies to the optimistic PI algorithm (2.33) and proves a preliminary bound.

Lemma 2.5.3: Let the monotonicity and contraction Assumptions 2.1.1 and 2.1.2 hold, let $\{(J_k, \mu^k)\}$ be a sequence generated by the optimistic PI algorithm (2.33), and assume that for some $c \geq 0$ we have

$$J_0 \geq TJ_0 - c\, v.$$

Then for all $k \geq 0$,

$$TJ_k + \frac{\alpha}{1-\alpha}\beta_k c\, v \geq J_{k+1} \geq TJ_{k+1} - \beta_{k+1} c\, v, \tag{2.41}$$

where β_k is the scalar given by

$$\beta_k = \begin{cases} 1 & \text{if } k = 0, \\ \alpha^{m_0 + \cdots + m_{k-1}} & \text{if } k > 0, \end{cases} \tag{2.42}$$

with m_j, $j = 0, 1, \ldots$, being the integers used in the algorithm (2.33).

Proof: We prove Eq. (2.41) by induction on k, using Lemma 2.5.2. For $k = 0$, using Eq. (2.38) with $J = J_0$, $\mu = \mu^0$, and $k = m_0$, we have

$$TJ_0 \geq J_1 - \frac{\alpha}{1-\alpha} c\, v = J_1 - \frac{\alpha}{1-\alpha}\beta_0 c\, v,$$

showing the left-hand side of Eq. (2.41) for $k = 0$. Also by Eq. (2.39) with $\mu = \mu^0$ and $k = m_0$, we have

$$J_1 \geq TJ_1 - \alpha^{m_0} c\, v = TJ_1 - \beta_1 c\, v.$$

showing the right-hand side of Eq. (2.41) for $k = 0$.

Assuming that Eq. (2.41) holds for $k - 1 \geq 0$, we will show that it holds for k. Indeed, the right-hand side of the induction hypothesis yields

$$J_k \geq TJ_k - \beta_k c\, v.$$

Using Eqs. (2.38) and (2.39) with $J = J_k$, $\mu = \mu^k$, and $k = m_k$, we obtain

$$TJ_k \geq J_{k+1} - \frac{\alpha}{1-\alpha}\beta_k c\, v,$$

and

$$J_{k+1} \geq TJ_{k+1} - \alpha^{m_k}\beta_k c\, v = TJ_{k+1} - \beta_{k+1} c\, v,$$

respectively. This completes the induction. **Q.E.D.**

The next lemma essentially proves the convergence of the optimistic PI (Prop. 2.5.1) and provides associated error bounds.

Lemma 2.5.4: Let the monotonicity and contraction Assumptions 2.1.1 and 2.1.2 hold, let $\{(J_k, \mu^k)\}$ be a sequence generated by the optimistic PI algorithm (2.33), and let $c \geq 0$ be a scalar such that

$$\|J_0 - TJ_0\| \leq c. \tag{2.43}$$

Then for all $k \geq 0$,

$$J_k + \frac{\alpha^k}{1-\alpha} c\, v \geq J_k + \frac{\beta_k}{1-\alpha} c\, v \geq J^* \geq J_k - \frac{(k+1)\alpha^k}{1-\alpha} c\, v, \tag{2.44}$$

where β_k is defined by Eq. (2.42).

Proof: Using the relation $J_0 \geq TJ_0 - c\, v$ [cf. Eq. (2.43)] and Lemma 2.5.3, we have

$$J_k \geq TJ_k - \beta_k c\, v, \quad k = 0, 1, \ldots.$$

Using this relation in Lemma 2.5.1(b) with $W = T$ and $k = 0$, we obtain

$$J_k \geq J^* - \frac{\beta_k}{1-\alpha} c\, v,$$

which together with the fact $\alpha^k \geq \beta_k$, shows the left-hand side of Eq. (2.44).

Using the relation $TJ_0 \geq J_0 - c\, v$ [cf. Eq. (2.43)] and Lemma 2.5.1(b) with $W = T$, we have

$$J^* \geq T^k J_0 - \frac{\alpha^k}{1-\alpha} c\, v, \qquad k = 0, 1, \ldots. \tag{2.45}$$

Using again the relation $J_0 \geq TJ_0 - c\, v$ in conjunction with Lemma 2.5.3, we also have

$$TJ_j \geq J_{j+1} - \frac{\alpha}{1-\alpha}\beta_j c\, v, \qquad j = 0, \ldots, k-1.$$

Applying T^{k-j-1} to both sides of this inequality and using the monotonicity and contraction properties of T^{k-j-1}, we obtain

$$T^{k-j}J_j \geq T^{k-j-1}J_{j+1} - \frac{\alpha^{k-j}}{1-\alpha}\beta_j c\, v, \qquad j = 0, \ldots, k-1,$$

cf. Lemma 2.5.1(a). By adding this relation over $j = 0, \ldots, k-1$, and using the fact $\beta_j \leq \alpha^j$, it follows that

$$T^k J_0 \geq J_k - \sum_{j=0}^{k-1} \frac{\alpha^{k-j}}{1-\alpha} \alpha^j c\, v = J_k - \frac{k\alpha^k}{1-\alpha} c\, v. \tag{2.46}$$

Finally, by combining Eqs. (2.45) and (2.46), we obtain the right-hand side of Eq. (2.44). **Q.E.D.**

Proof of Props. 2.5.1 and 2.5.2: Let c be a scalar satisfying Eq. (2.43). Then the error bounds (2.44) show that $\lim_{k\to\infty} \|J_k - J^*\| = 0$, i.e., the first part of Prop. 2.5.1. To show the second part (finite termination when the number of policies is finite), let $\widehat{\mathcal{M}}$ be the finite set of nonoptimal policies. Then there exists $\epsilon > 0$ such that $\|T_{\hat{\mu}} J^* - TJ^*\| > \epsilon$ for all $\hat{\mu} \in \widehat{\mathcal{M}}$, which implies that $\|T_{\hat{\mu}} J_k - TJ_k\| > \epsilon$ for all $\hat{\mu} \in \widehat{\mathcal{M}}$ and k sufficiently large. This implies that $\mu^k \notin \widehat{\mathcal{M}}$ for all k sufficiently large. The proof of Prop. 2.5.2 follows using the compactness and continuity Assumption 2.4.1, and the convergence argument of Prop. 2.4.2. **Q.E.D.**

Convergence Rate Issues

Let us consider the convergence rate bounds of Lemma 2.5.4 for optimistic PI, and write them in the form

$$\|J_0 - TJ_0\| \leq c \quad \Rightarrow \quad J_k - \frac{(k+1)\alpha^k}{1-\alpha} c\, v \leq J^* \leq J_k + \frac{\alpha^{m_0+\cdots+m_{k-1}}}{1-\alpha} c\, v. \tag{2.47}$$

We may contrast these bounds with the ones for VI, where

$$\|J_0 - TJ_0\| \leq c \quad \Rightarrow \quad T^k J_0 - \frac{\alpha^k}{1-\alpha} c\, v \leq J^* \leq T^k J_0 + \frac{\alpha^k}{1-\alpha} c\, v \tag{2.48}$$

[cf. Prop. 2.1.4(c)].

In comparing the bounds (2.47) and (2.48), we should also take into account the associated overhead for a single iteration of each method: optimistic PI requires at iteration k a single application of T and $m_k - 1$ applications of T_{μ^k} (each being less time-consuming than an application of T), while VI requires a single application of T. It can then be seen that the upper bound for optimistic PI is better than the one for VI (same bound for less overhead), while the lower bound for optimistic PI is worse than the one for VI (worse bound for more overhead). This suggests that the choice of the initial condition J_0 is important in optimistic PI, and in particular it is preferable to have $J_0 \geq TJ_0$ (implying convergence to J^* from above) rather than $J_0 \leq TJ_0$ (implying convergence to J^* from below). This is consistent with the results of other works, which indicate that the convergence properties of the method are fragile when the condition $J_0 \geq TJ_0$ does not hold (see [WiB93], [BeT96], [BeY10], [BeY12], [YuB13a]).

2.5.2 Approximate Optimistic Policy Iteration

We will now derive error bounds for the case where the policy evaluation and policy improvement operations are approximate, similar to the nonoptimistic PI case of Section 2.4.1. In particular, we consider a method that generates a sequence of policies $\{\mu^k\}$ and a corresponding sequence of approximate cost functions $\{J_k\}$ satisfying

$$\|J_k - T_{\mu^k}^{m_k} J_{k-1}\| \leq \delta, \qquad \|T_{\mu^{k+1}} J_k - T J_k\| \leq \epsilon, \qquad k = 0, 1, \ldots, \quad (2.49)$$

[cf. Eq. (2.23)]. For example, we may compute (perhaps approximately, by simulation) the values $(T_{\mu^k}^{m_k} J_{k-1})(x)$ for a subset of states x, and use a least squares fit of these values to select J_k from some parametric class of functions.

We will prove the same error bound as for the nonoptimistic case, cf. Eq. (2.24). However, for this we will need the following condition, which is stronger than the contraction and monotonicity conditions that we have been using so far.

Assumption 2.5.1: (Semilinear Monotonic Contraction) For all $J \in \mathcal{B}(X)$ and $\mu \in \mathcal{M}$, the functions $T_\mu J$ and TJ belong to $\mathcal{B}(X)$. Furthermore, for some $\alpha \in (0,1)$, we have for all $J, J' \in \mathcal{B}(X)$, $\mu \in \mathcal{M}$, and $x \in X$,

$$\frac{(T_\mu J')(x) - (T_\mu J)(x)}{v(x)} \leq \alpha \sup_{y \in X} \frac{J'(y) - J(y)}{v(y)}. \quad (2.50)$$

This assumption implies both the monotonicity and contraction Assumptions 2.1.1 and 2.1.2, as can be easily verified. Moreover the assumption is satisfied in the discounted DP examples of Section 1.2, as well as the stochastic shortest path problem of Example 1.2.6. It holds if T_μ is a linear mapping involving a matrix with nonnegative components that has spectral radius less than 1 (or more generally if T_μ is the minimum or the maximum of a finite number of such linear mappings).

For any function $y \in \mathcal{B}(X)$, let us use the notation

$$M(y) = \sup_{x \in X} \frac{y(x)}{v(x)}.$$

Then the condition (2.50) can be written for all $J, J' \in \mathcal{B}(X)$, and $\mu \in \mathcal{M}$ as

$$M(T_\mu J - T_\mu J') \leq \alpha M(J - J'), \quad (2.51)$$

and also implies the following multistep versions, for $\ell \geq 1$,

$$T_\mu^\ell J - T_\mu^\ell J' \leq \alpha^\ell M(J - J')v, \quad M(T_\mu^\ell J - T_\mu^\ell J') \leq \alpha^\ell M(J - J'), \quad (2.52)$$

which can be proved by induction using Eq. (2.51). We have the following proposition.

Proposition 2.5.3: (Error Bound for Optimistic Approximate PI) Let Assumption 2.5.1 hold, in addition to the monotonicity and contraction Assumptions 2.1.1 and 2.1.2. Then the sequence $\{\mu^k\}$ generated by the optimistic approximate PI algorithm (2.49) satisfies

$$\limsup_{k\to\infty} \|J_{\mu^k} - J^*\| \leq \frac{\epsilon + 2\alpha\delta}{(1-\alpha)^2}.$$

Proof: Let us fix $k \geq 1$, and for simplicity let us assume that $m_k \equiv m$ for some m, and denote

$$\underline{J} = J_{k-1}, \qquad J = J_k, \qquad \mu = \mu^k, \qquad \bar{\mu} = \mu^{k+1},$$

$$s = J_\mu - T_\mu^m \underline{J}, \qquad \bar{s} = J_{\bar{\mu}} - T_{\bar{\mu}}^m J, \qquad t = T_\mu^m \underline{J} - J^*, \qquad \bar{t} = T_{\bar{\mu}}^m J - J^*.$$

We have

$$J_\mu - J^* = J_\mu - T_\mu^m \underline{J} + T_\mu^m \underline{J} - J^* = s + t. \quad (2.53)$$

We will derive recursive relations for s and t, which will also involve the residual functions

$$r = T_\mu \underline{J} - \underline{J}, \qquad \bar{r} = T_{\bar{\mu}} J - J.$$

We first obtain a relation between r and $\bar{r}$. We have

$$\begin{aligned}
\bar{r} &= T_{\bar{\mu}} J - J \\
&= (T_{\bar{\mu}} J - T_\mu J) + (T_\mu J - J) \\
&\leq (T_{\bar{\mu}} J - TJ) + \big(T_\mu J - T_\mu(T_\mu^m \underline{J})\big) + (T_\mu^m \underline{J} - J) + \big(T_\mu^m(T_\mu \underline{J}) - T_\mu^m \underline{J}\big) \\
&\leq \epsilon v + \alpha M(J - T_\mu^m \underline{J})v + \delta v + \alpha^m M(T_\mu \underline{J} - \underline{J})v \\
&\leq (\epsilon + \delta)v + \alpha\delta v + \alpha^m M(r)v,
\end{aligned}$$

where the first inequality follows from $T_{\bar{\mu}} J \geq TJ$, and the second and third inequalities follow from Eqs. (2.49) and (2.52). From this relation we have

$$M(\bar{r}) \leq \big(\epsilon + (1+\alpha)\delta\big) + \beta M(r),$$

where $\beta = \alpha^m$. Taking lim sup as $k \to \infty$ in this relation, we obtain

$$\limsup_{k\to\infty} M(r) \leq \frac{\epsilon + (1+\alpha)\delta}{1-\beta}. \tag{2.54}$$

Next we derive a relation between s and r. We have

$$\begin{aligned} s &= J_\mu - T_\mu^m \underline{J} \\ &= T_\mu^m J_\mu - T_\mu^m \underline{J} \\ &\leq \alpha^m M(J_\mu - \underline{J})v \\ &\leq \frac{\alpha^m}{1-\alpha} M(T_\mu \underline{J} - \underline{J})v \\ &= \frac{\alpha^m}{1-\alpha} M(r)v, \end{aligned}$$

where the first inequality follows from Eq. (2.52) and the second inequality follows by using Prop. 2.1.4(b). Thus we have $M(s) \leq \frac{\alpha^m}{1-\alpha} M(r)$, from which by taking lim sup of both sides and using Eq. (2.54), we obtain

$$\limsup_{k\to\infty} M(s) \leq \frac{\beta\big(\epsilon + (1+\alpha)\delta\big)}{(1-\alpha)(1-\beta)}. \tag{2.55}$$

Finally we derive a relation between t, $\bar{t}$, and r. We first note that

$$\begin{aligned} TJ - TJ^* &\leq \alpha M(J - J^*)v \\ &= \alpha M(J - T_\mu^m \underline{J} + T_\mu^m \underline{J} - J^*)v \\ &\leq \alpha M(J - T_\mu^m \underline{J})v + \alpha M(T_\mu^m \underline{J} - J^*)v \\ &\leq \alpha\delta v + \alpha M(t)v. \end{aligned}$$

Using this relation, and Eqs. (2.49) and (2.52), we have

$$\begin{aligned} \bar{t} &= T_{\bar{\mu}}^m J - J^* \\ &= (T_{\bar{\mu}}^m J - T_{\bar{\mu}}^{m-1} J) + \cdots + (T_{\bar{\mu}}^2 J - T_{\bar{\mu}} J) + (T_{\bar{\mu}} J - TJ) + (TJ - TJ^*) \\ &\leq (\alpha^{m-1} + \cdots + \alpha) M(T_{\bar{\mu}} J - J)v + \epsilon v + \alpha\delta v + \alpha M(t)v, \end{aligned}$$

so finally

$$M(\bar{t}) \leq \frac{\alpha - \alpha^m}{1-\alpha} M(\bar{r}) + (\epsilon + \alpha\delta) + \alpha M(t).$$

By taking lim sup of both sides and using Eq. (2.54), it follows that

$$\limsup_{k\to\infty} M(t) \leq \frac{(\alpha-\beta)\big(\epsilon + (1+\alpha)\delta\big)}{(1-\alpha)^2(1-\beta)} + \frac{\epsilon + \alpha\delta}{1-\alpha}. \tag{2.56}$$

We now combine Eqs. (2.53), (2.55), and (2.56). We obtain

$$\begin{aligned}\limsup_{k\to\infty} M(J_{\mu^k} - J^*) &\le \limsup_{k\to\infty} M(s) + \limsup_{k\to\infty} M(t)\\ &\le \frac{\beta\big(\epsilon + (1+\alpha)\delta\big)}{(1-\alpha)(1-\beta)} + \frac{(\alpha-\beta)\big(\epsilon + (1+\alpha)\delta\big)}{(1-\alpha)^2(1-\beta)} + \frac{\epsilon+\alpha\delta}{1-\alpha}\\ &= \frac{\big(\beta(1-\alpha) + (\alpha-\beta)\big)\big(\epsilon + (1+\alpha)\delta\big)}{(1-\alpha)^2(1-\beta)} + \frac{\epsilon+\alpha\delta}{1-\alpha}\\ &= \frac{\alpha\big(\epsilon + (1+\alpha)\delta\big)}{(1-\alpha)^2} + \frac{\epsilon+\alpha\delta}{1-\alpha}\\ &= \frac{\epsilon + 2\alpha\delta}{(1-\alpha)^2}.\end{aligned}$$

This proves the result, since in view of $J_{\mu^k} \ge J^*$, we have $M(J_{\mu^k} - J^*) = \|J_{\mu^k} - J^*\|$. **Q.E.D.**

A remarkable fact is that approximate VI, approximate PI, and approximate optimistic PI have very similar error bounds (cf. Props. 2.3.2, 2.4.3, and 2.5.3). Approximate VI has a slightly better bound, but insignificantly so in practical terms. When approximate PI produces a convergent sequence of policies, the associated error bound is much better (cf. Prop. 2.4.5). However, special conditions are needed for convergence of policies in approximate PI. These conditions are fulfilled in some cases, notably including schemes where aggregation is used for policy evaluation (cf. Section 1.2.4). In other cases, including some where the projected equation is used for policy evaluation, approximate PI (both optimistic and nonoptimistic) will typically generate a cycle of policies satisfying the bound of Prop. 2.4.3; see Section 3.6 of the PI survey paper [Ber11c], or Chapter 6 of the book [Ber12a].

2.5.3 Randomized Optimistic Policy Iteration

We will now consider a randomized version of the optimistic PI algorithm where the number m_k of VI iterations in the kth policy evaluation is random, while the monotonicity assumption need not hold. We assume, however, that each policy mapping is a contraction in a suitable space, that the number of policies is finite, and that $m_k = 1$ with positive probability (these assumptions can be modified and/or generalized in ways suggested by the subsequent line of proof). In particular, for each positive integer j, we have a probability $p(j) \ge 0$, where

$$p(1) > 0, \qquad \sum_{j=1}^{\infty} p(j) = 1.$$

We consider the algorithm

$$T_{\mu^k} J_k = T J_k, \qquad J_{k+1} = T_{\mu^k}^{m_k} J_k, \qquad k = 0, 1, \ldots, \tag{2.57}$$

where m_k is chosen randomly according to the distribution $p(j)$,

$$P(m_k = j) = p(j), \qquad j = 1, 2, \ldots. \tag{2.58}$$

The selection of m_k is independent of previous selections. We will assume the following.

Assumption 2.5.2: Let $\|\cdot\|$ be a norm on some complete space of real-valued functions over X, denoted $\mathcal{F}(X)$, and assume the following.

(a) The set of policies $\mathcal{M}$ is finite.

(b) The mappings T_μ, $\mu \in \mathcal{M}$, and T are contraction mappings from $\mathcal{F}(X)$ into $\mathcal{F}(X)$.

The preceding assumption requires that the number of policies is finite, but does not require any monotonicity condition (cf. Assumption 2.1.1), while its contraction condition (b) is weaker than the contraction Assumption 2.1.2 since $\mathcal{F}(X)$ is a general complete normed space, not necessarily $\mathcal{B}(X)$. This flexibility may be useful in algorithms that involve cost function approximation within a subspace of basis functions. For such algorithms, however, T does not necessarily have a unique fixed point, as discussed in Section 1.2.4. By contrast since $\mathcal{F}(X)$ is assumed complete, Assumption 2.5.2 implies that T_μ and T have unique fixed points, which we denote by J_μ and J^*, respectively.

An important preliminary fact (which relies on the finiteness of $\mathcal{M}$) is given in the following proposition. The proposition implies that near J^* the generated policies μ^k are "optimal" in the sense that $J_{\mu^k} = J^*$, so the algorithm does not tend to cycle. †

Proposition 2.5.4: Let Assumption 2.5.2 hold, and let $\mathcal{M}^*$ be the subset of all $\mu \in \mathcal{M}$ such that $T_\mu J^* = T J^*$. Then for all $\mu \in \mathcal{M}^*$, we have $J_\mu = J^*$. Moreover, there exists an $\epsilon > 0$ such that for all J with $\|J - J^*\| < \epsilon$ we have $T_\mu J = T J$ only if $\mu \in \mathcal{M}^*$.

Proof: If $\mu \in \mathcal{M}^*$, we have $T_\mu J^* = T J^* = J^*$. Thus J^* is the unique fixed point J_μ of T_μ, and we have $J_\mu = J^*$.

† Note that without monotonicity, J^* need not have any formal optimality properties (cf. the discussion of Section 2.1 and Example 2.1.1).

To prove the second assertion, we argue by contradiction, so we assume that there exist a sequence of scalars $\{\epsilon_k\}$ and a sequence of policies $\{\mu^k\}$ such that $\epsilon_k \downarrow 0$ and

$$\mu^k \notin \mathcal{M}^*, \qquad T_{\mu^k} J_k = TJ_k, \qquad \|J_k - J^*\| < \epsilon_k, \qquad \forall\ k = 0, 1, \ldots.$$

Since $\mathcal{M}$ is finite, we may assume without loss of generality that for some $\overline{\mu} \notin \mathcal{M}^*$, we have $\mu^k = \overline{\mu}$ for all k, so from the preceding relation we have

$$T_{\overline{\mu}} J_k = TJ_k, \qquad \|J_k - J^*\| < \epsilon_k, \qquad \forall\ k = 0, 1, \ldots.$$

Thus $\|J_k - J^*\| \to 0$, and by the contraction Assumption 2.5.2(b), we have

$$\|T_{\overline{\mu}} J_k - T_{\overline{\mu}} J^*\| \to 0, \qquad \|TJ_k - TJ^*\| \to 0.$$

Since $T_{\overline{\mu}} J_k = TJ_k$, the limits of $\{T_{\overline{\mu}} J_k\}$ and $\{TJ_k\}$ are equal, i.e., $T_{\overline{\mu}} J^* = TJ^* = J^*$. Since $J_{\overline{\mu}}$ is the unique fixed point of $T_{\overline{\mu}}$ over $\mathcal{F}(X)$, it follows that $J_{\overline{\mu}} = J^*$, contradicting the earlier hypothesis that $\overline{\mu} \notin \mathcal{M}^*$. **Q.E.D.**

The preceding proof illustrates the key idea of the randomized optimistic PI algorithm, which is that for $\mu \in \mathcal{M}^*$, the mappings $T_\mu^{m_k}$ have a common fixed point that is equal to J^*, the fixed point of T. Thus within a distance ϵ from J^*, the iterates (2.57) aim consistently at J^*. Moreover, because the probability of a VI (an iteration with $m_k = 1$) is positive, the algorithm is guaranteed to eventually come within ϵ from J^* through a sufficiently long sequence of contiguous VI iterations. For this we need the sequence $\{J_k\}$ to be bounded, which will be shown as part of the proof of the following proposition.

Proposition 2.5.5: Let Assumption 2.5.2 hold. Then for any starting point $J_0 \in \mathcal{F}(X)$, a sequence $\{J_k\}$ generated by the randomized optimistic PI algorithm (2.57)-(2.58) belongs to $\mathcal{F}(X)$ and converges to J^* with probability one.

Proof: We will show that $\{J_k\}$ is bounded by showing that for all k, we have

$$\max_{\mu \in \mathcal{M}} \|J_k - J_\mu\| \leq \rho^k \max_{\mu \in \mathcal{M}} \|J_0 - J_\mu\| + \frac{1}{1-\rho} \max_{\mu, \mu' \in \mathcal{M}} \|J_\mu - J_{\mu'}\|, \quad (2.59)$$

where ρ is a common contraction modulus of T_μ, $\mu \in \mathcal{M}$, and T. Indeed, we have for all $\mu \in \mathcal{M}$

$$\begin{aligned}
\|J_k - J_\mu\| &\leq \|J_k - J_{\mu^{k-1}}\| + \|J_{\mu^{k-1}} - J_\mu\| \\
&= \|T_{\mu^{k-1}}^{m_{k-1}} J_{k-1} - J_{\mu^{k-1}}\| + \|J_{\mu^{k-1}} - J_\mu\| \\
&\leq \rho^{m_{k-1}} \|J_{k-1} - J_{\mu^{k-1}}\| + \|J_{\mu^{k-1}} - J_\mu\| \\
&\leq \rho^{m_{k-1}} \max_{\mu \in \mathcal{M}} \|J_{k-1} - J_\mu\| + \max_{\mu, \mu' \in \mathcal{M}} \|J_\mu - J_{\mu'}\| \\
&\leq \rho \max_{\mu \in \mathcal{M}} \|J_{k-1} - J_\mu\| + \max_{\mu, \mu' \in \mathcal{M}} \|J_\mu - J_{\mu'}\|,
\end{aligned}$$

and finally, for all k,

$$\max_{\mu\in\mathcal{M}} \|J_k - J_\mu\| \leq \rho \max_{\mu\in\mathcal{M}} \|J_{k-1} - J_\mu\| + \max_{\mu,\mu'\in\mathcal{M}} \|J_\mu - J_{\mu'}\|.$$

From this relation, we obtain Eq. (2.59) by induction.

Thus in conclusion, we have $\{J_k\} \subset D$, where D is the bounded set

$$D = \left\{ J \;\middle|\; \max_{\mu\in\mathcal{M}} \|J - J_\mu\| \leq \max_{\mu\in\mathcal{M}} \|J_0 - J_\mu\| + \frac{1}{1-\rho} \max_{\mu,\mu'\in\mathcal{M}} \|J_\mu - J_{\mu'}\| \right\}.$$

We use this fact to argue that with enough contiguous value iterations, i.e., iterations where $m_k = 1$, J_k can be brought arbitrarily close to J^*, and once this happens, the algorithm operates like the ordinary VI algorithm.

Indeed, each time the iteration $J_{k+1} = TJ_k$ is performed (i.e., when $m_k = 1$), the distance of the iterate J_k from J^* is reduced by a factor ρ, i.e., $\|J_{k+1} - J^*\| \leq \rho\|J_k - J^*\|$. Since $\{J_k\}$ belongs to the bounded set D, and our randomization scheme includes the condition $p(1) > 0$, the algorithm is guaranteed (with probability one) to eventually execute a sufficient number of contiguous iterations $J_{k+1} = TJ_k$ to enter a sphere

$$S_\epsilon = \big\{ J \in \mathcal{F}(X) \mid \|J - J^*\| < \epsilon \big\}$$

of small enough radius ϵ to guarantee that the generated policy μ^k belongs to $\mathcal{M}^*$, as per Prop. 2.5.4. Once this happens, all subsequent iterations reduce the distance $\|J_k - J^*\|$ by a factor ρ at every iteration, since

$$\big\|T_\mu^m J - J^*\| \leq \rho\|T_\mu^{m-1} J - J^*\| \leq \rho\|J - J^*\|, \quad \forall\ \mu \in \mathcal{M}^*,\ m \geq 1,\ J \in S_\epsilon.$$

Thus once $\{J_k\}$ enters S_ϵ, it stays within S_ϵ and converges to J^*. **Q.E.D.**

A Randomized Version of λ-Policy Iteration

We now turn to the λ-PI algorithm. Instead of the nonrandomized version

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

cf. Eq. (2.34), we consider a randomized version that involves a fixed probability $p \in (0, 1)$. It has the form

$$T_{\mu^k} J_k = TJ_k, \qquad J_{k+1} = \begin{cases} TJ_k & \text{with probability } p, \\ T_{\mu^k}^{(\lambda)} J_k & \text{with probability } 1 - p. \end{cases} \tag{2.60}$$

The idea of the algorithm is similar to the one of the randomized optimistic PI algorithm (2.57)-(2.58). Under the assumptions of Prop.

2.5.5, the sequence $\{J_k\}$ generated by the randomized λ-PI algorithm (2.60) belongs to $\mathcal{F}(X)$ and converges to J^* with probability one. The reason is that the contraction property of T_μ over $\mathcal{F}(X)$ with respect to the norm $\|\cdot\|$ implies that $T_\mu^{(\lambda)}$ is well-defined, and also implies that $T_\mu^{(\lambda)}$ is a contraction over $\mathcal{F}(X)$. The latter assertion follows from the calculation

$$\begin{aligned}\left\|T_\mu^{(\lambda)}J - J_\mu\right\| &= \left\|(1-\lambda)\sum_{\ell=0}^{\infty}\lambda^\ell T_\mu^{\ell+1}J - (1-\lambda)\sum_{\ell=0}^{\infty}\lambda^\ell J_\mu\right\| \\ &\le (1-\lambda)\sum_{\ell=0}^{\infty}\lambda^\ell\left\|T_\mu^{\ell+1}J - J_\mu\right\| \\ &\le (1-\lambda)\sum_{\ell=0}^{\infty}\lambda^\ell\rho^{\ell+1}\|J - J_\mu\| \\ &= \rho\,\|J - J_\mu\|,\end{aligned}$$

where the first inequality follows from the triangle inequality, and the second inequality follows from the contraction property of T_μ. Given that $T_\mu^{(\lambda)}$ is a contraction, the proof of Prop. 2.5.5 goes through with minimal changes. The idea again is that $\{J_k\}$ remains bounded, and through a sufficiently long sequence of contiguous iterations where the iteration $x_{k+1} = TJ_k$ is performed, it enters the sphere S_ϵ, and subsequently stays within S_ϵ and converges to J^*.

The convergence argument just given suggests that the choice of the randomization probability p is important. If p is too small, convergence may be slow because oscillatory behavior may go unchecked for a long time. On the other hand if p is large, a correspondingly large number of fixed point iterations $x_{k+1} = TJ_k$ may be performed, and the hoped for benefits of the use of the proximal iterations $x_{k+1} = T_{\mu^k}^{(\lambda)}J_k$ may be lost. Adaptive schemes that adjust p based on algorithmic progress may address this issue. Similarly, the choice of the probability $p(1)$ is significant in the randomized optimistic PI algorithm (2.57)-(2.58).

2.6 ASYNCHRONOUS ALGORITHMS

In this section, we extend further the computational methods of VI and PI for abstract DP models, by embedding them within an asynchronous computation framework.

2.6.1 Asynchronous Value Iteration

Each VI of the form given in Section 2.3 applies the mapping T defined by

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

for all states simultaneously, thereby producing the sequence $TJ, T^2J, \ldots$ starting with some $J \in \mathcal{B}(X)$. In a more general form of VI, at any one iteration, $J(x)$ may be updated and replaced by $(TJ)(x)$ only for a subset of states. An example is the Gauss-Seidel method for the finite-state case, where at each iteration, $J(x)$ is updated only for a single selected state $\overline{x}$ and $J(x)$ is left unchanged for all other states $x \neq \overline{x}$ (see [Ber12a]). In that method the states are taken up for iteration in a cyclic order, but more complex iteration orders are possible, deterministic as well as randomized.

Methods of the type just described are called *asynchronous VI methods* and may be motivated by several considerations such as:

(a) *Faster convergence*. Generally, computational experience with DP as well as analysis, have shown that convergence is accelerated by incorporating the results of VI updates for some states as early as possible into subsequent VI updates for other states. This is known as the *Gauss-Seidel effect*, which is discussed in some detail in the book [BeT89].

(b) *Parallel and distributed asynchronous computation*. In this context, we have several processors, each applying VI for a subset of states, and communicating the results to other processors (perhaps with some delay). One objective here may be faster computation by taking advantage of parallelism. Another objective may be computational convenience in problems where useful information is generated and processed locally at geographically dispersed points. An example is data or sensor network computations, where nodes, gateways, sensors, and data collection centers collaborate to route and control the flow of data, using DP or shortest path-type computations.

(c) *Simulation-based implementations*. In simulation-based versions of VI, iterations at various states are often performed in the order that the states are generated by some form of simulation.

With these contexts in mind, we introduce a model of asynchronous distributed solution of abstract fixed point problems of the form $J = TJ$. Let $\mathcal{R}(X)$ be the set of real-valued functions defined on some given set X and let T map $\mathcal{R}(X)$ into $\mathcal{R}(X)$. We consider a partition of X into disjoint nonempty subsets $X_1, \ldots, X_m$, and a corresponding partition of J as $J = (J_1, \ldots, J_m)$, where J_ℓ is the restriction of J on the set X_ℓ. Our computation framework involves a network of m processors, each updating corresponding components of J. In a (synchronous) distributed VI algorithm, processor ℓ updates J_ℓ at iteration t according to

$$J_\ell^{t+1}(x) = T(J_1^t, \ldots, J_m^t)(x), \qquad \forall\, x \in X_\ell,\ \ell = 1, \ldots, m.$$

Here to accommodate the distributed algorithmic framework and its overloaded notation, we will use superscript t to denote iterations/times where

some (but not all) processors update their corresponding components, reserving the index k for computation stages involving all processors, and also reserving subscript ℓ to denote component/processor index.

In an asynchronous VI algorithm, processor ℓ updates J_ℓ only for t in a selected subset $\mathcal{R}_\ell$ of iterations, and with components J_j, $j \neq \ell$, supplied by other processors with communication "delays" $t - \tau_{\ell j}(t)$,

$$J_\ell^{t+1}(x) = \begin{cases} T\left(J_1^{\tau_{\ell 1}(t)}, \ldots, J_m^{\tau_{\ell m}(t)}\right)(x) & \text{if } t \in \mathcal{R}_\ell,\ x \in X_\ell, \\ J_\ell^t(x) & \text{if } t \notin \mathcal{R}_\ell,\ x \in X_\ell. \end{cases} \tag{2.61}$$

Communication delays arise naturally in the context of asynchronous distributed computing systems of the type described in many sources (an extensive reference is the book [BeT89]). Such systems are interesting for solution of large DP problems, particularly for methods that are based on simulation, which is naturally well-suited for distributed computation. On the other hand, if the entire algorithm is centralized at a single physical processor, the algorithm (2.61) ordinarily will not involve communication delays, i.e., $\tau_{\ell j}(t) = t$ for all ℓ, j, and t.

The simpler case where X is a finite set and each subset X_ℓ consists of a single element ℓ arises often, particularly in the context of simulation. In this case we may simplify the notation of iteration (2.61) by writing J_ℓ^t in place of the scalar component $J_\ell^t(\ell)$, as we do in the following example.

Example 2.6.1 (One-State-at-a-Time Iterations)

Assuming $X = \{1, \ldots, n\}$, let us view each state as a processor by itself, so that $X_\ell = \{\ell\}$, $\ell = 1, \ldots, n$. Consider a VI algorithm that executes one-state-at-a-time, according to some state sequence $\{x^0, x^1, \ldots\}$, which is generated in some way, possibly by simulation. Thus, starting from some initial vector J^0, we generate a sequence $\{J^t\}$, with $J^t = (J_1^t, \ldots, J_n^t)$, as follows:

$$J_\ell^{t+1} = \begin{cases} T(J_1^t, \ldots, J_n^t)(\ell) & \text{if } \ell = x^t, \\ J_\ell^t & \text{if } \ell \neq x^t, \end{cases}$$

where $T(J_1^t, \ldots, J_n^t)(\ell)$ denotes the ℓ-th component of the vector

$$T(J_1^t, \ldots, J_n^t) = TJ^t,$$

and for simplicity we write J_ℓ^t instead of $J_\ell^t(\ell)$. This algorithm is a special case of iteration (2.61) where the set of times at which J_ℓ is updated is

$$\mathcal{R}_\ell = \{t \mid x^t = \ell\},$$

and there are no communication delays (as in the case where the entire algorithm is centralized at a single physical processor).

Note also that if X is finite, we can assume without loss of generality that each state is assigned to a separate processor. The reason is that a physical processor that updates a group of states may be replaced by a group of fictitious processors, each assigned to a single state, and updating their corresponding components of J simultaneously.

We will now discuss the convergence of the asynchronous algorithm (2.61). To this end we introduce the following assumption.

Assumption 2.6.1: (Continuous Updating and Information Renewal)

(1) The set of times $\mathcal{R}_\ell$ at which processor ℓ updates J_ℓ is infinite, for each $\ell = 1, \ldots, m$.

(2) $\lim_{t\to\infty} \tau_{\ell j}(t) = \infty$ for all $\ell, j = 1, \ldots, m$.

Assumption 2.6.1 is natural, and is essential for any kind of convergence result about the algorithm. † In particular, the condition $\tau_{\ell j}(t) \to \infty$ guarantees that outdated information about the processor updates will eventually be purged from the computation. It is also natural to assume that $\tau_{\ell j}(t)$ is monotonically increasing with t, but this assumption is not necessary for the subsequent analysis.

We wish to show that $J_\ell^t \to J_\ell^*$ for all ℓ, and to this end we employ the following convergence theorem for totally asynchronous iterations from the author's paper [Ber83], which has served as the basis for the treatment of totally asynchronous iterations in the book [BeT89] (Chapter 6), and their application to DP (i.e., VI and PI), and asynchronous gradient-based optimization. For the statement of the theorem, we say that a sequence $\{J^k\} \subset \mathcal{R}(X)$ *converges pointwise* to $J \in \mathcal{R}(X)$ if

$$\lim_{k\to\infty} J^k(x) = J(x)$$

for all $x \in X$.

Proposition 2.6.1 (Asynchronous Convergence Theorem): Let T have a unique fixed point J^*, let Assumption 2.6.1 hold, and assume that there is a sequence of nonempty subsets $\big\{S(k)\big\} \subset \mathcal{R}(X)$ with

† Generally, convergent distributed iterative asynchronous algorithms are classified in *totally and partially asynchronous* [cf. the book [BeT89] (Chapters 6 and 7), or the more recent survey in the book [Ber16c] (Section 2.5)]. In the former, there is no bound on the communication delays, while in the latter there must be a bound (which may be unknown). The algorithms of the present section are totally asynchronous, as reflected by Assumption 2.6.1.

$$S(k+1) \subset S(k), \quad k = 0, 1, \ldots,$$

and is such that if $\{V^k\}$ is any sequence with $V^k \in S(k)$, for all $k \geq 0$, then $\{V^k\}$ converges pointwise to J^*. Assume further the following:

(1) *Synchronous Convergence Condition:* We have

$$TJ \in S(k+1), \qquad \forall\ J \in S(k),\ k = 0, 1, \ldots.$$

(2) *Box Condition:* For all k, $S(k)$ is a Cartesian product of the form

$$S(k) = S_1(k) \times \cdots \times S_m(k),$$

where $S_\ell(k)$ is a set of real-valued functions on X_ℓ, $\ell = 1, \ldots, m$.

Then for every $J^0 \in S(0)$, the sequence $\{J^t\}$ generated by the asynchronous algorithm (2.61) converges pointwise to J^*.

Proof: To explain the idea of the proof, let us note that the given conditions imply that updating any component J_ℓ, by applying T to a function $J \in S(k)$, while leaving all other components unchanged, yields a function in $S(k)$. Thus, once enough time passes so that the delays become "irrelevant," then after J enters $S(k)$, it stays within $S(k)$. Moreover, once a component J_ℓ enters the subset $S_\ell(k)$ and the delays become "irrelevant," J_ℓ gets permanently within the smaller subset $S_\ell(k+1)$ at the first time that J_ℓ is iterated on with $J \in S(k)$. Once each component J_ℓ, $\ell = 1, \ldots, m$, gets within $S_\ell(k+1)$, the entire function J is within $S(k+1)$ by the Box Condition. Thus the iterates from $S(k)$ eventually get into $S(k+1)$ and so on, and converge pointwise to J^* in view of the assumed properties of $\{S(k)\}$.

With this idea in mind, we show by induction that for each $k \geq 0$, there is a time t_k such that:

(1) $J^t \in S(k)$ for all $t \geq t_k$.

(2) For all ℓ and $t \in \mathcal{R}_\ell$ with $t \geq t_k$, we have

$$\left(J_1^{\tau_{\ell 1}(t)}, \ldots, J_m^{\tau_{\ell m}(t)}\right) \in S(k).$$

[In words, after some time, all fixed point estimates will be in $S(k)$ and all estimates used in iteration (2.61) will come from $S(k)$.]

The induction hypothesis is true for $k = 0$ since $J^0 \in S(0)$. Assuming it is true for a given k, we will show that there exists a time t_{k+1} with the required properties. For each $\ell = 1, \ldots, m$, let $t(\ell)$ be the first element of $\mathcal{R}_\ell$ such that $t(\ell) \geq t_k$. Then by the Synchronous Convergence Condition,

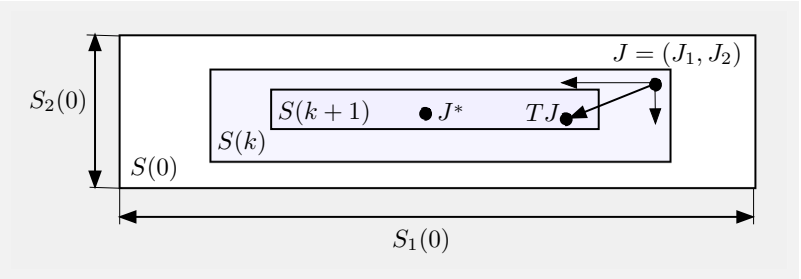


Figure 2.6.1 Geometric interpretation of the conditions of asynchronous convergence theorem. We have a nested sequence of boxes $\{S(k)\}$ such that $TJ \in S(k+1)$ for all $J \in S(k)$.

we have $TJ^{t(\ell)} \in S(k+1)$, implying (in view of the Box Condition) that

$$J_\ell^{t(\ell)+1} \in S_\ell(k+1).$$

Similarly, for every $t \in \mathcal{R}_\ell$, $t \geq t(\ell)$, we have $J_\ell^{t+1} \in S_\ell(k+1)$. Between elements of $\mathcal{R}_\ell$, J_ℓ^t does not change. Thus,

$$J_\ell^t \in S_\ell(k+1), \qquad \forall\, t \geq t(\ell)+1.$$

Let $t'_k = \max_\ell \big\{t(\ell)\big\} + 1$. Then, using the Box Condition we have

$$J^t \in S(k+1), \qquad \forall\, t \geq t'_k.$$

Finally, since by Assumption 2.6.1, we have $\tau_{\ell j}(t) \to \infty$ as $t \to \infty$, $t \in \mathcal{R}_\ell$, we can choose a time $t_{k+1} \geq t'_k$ that is sufficiently large so that $\tau_{\ell j}(t) \geq t'_k$ for all ℓ, j, and $t \in \mathcal{R}_\ell$ with $t \geq t_{k+1}$. We then have, for all $t \in \mathcal{R}_\ell$ with $t \geq t_{k+1}$ and $j = 1, \ldots, m$, $J_j^{\tau_{\ell j}(t)} \in S_j(k+1)$, which (by the Box Condition) implies that

$$\Big(J_1^{\tau_{\ell 1}(t)}, \ldots, J_m^{\tau_{\ell m}(t)}\Big) \in S(k+1).$$

The induction is complete. **Q.E.D.**

Figure 2.6.1 illustrates the assumptions of the preceding convergence theorem. The challenge in applying the theorem is to identify the set sequence $\big\{S(k)\big\}$ and to verify the assumptions of Prop. 2.6.1. In abstract DP, these assumptions are satisfied in two primary contexts of interest. The first is when $S(k)$ are weighted sup-norm spheres centered at J^*, and can be used in conjunction with the contraction framework of the preceding section (see the following proposition). The second context is based on monotonicity conditions. It will be used in Section 3.6 in conjunction

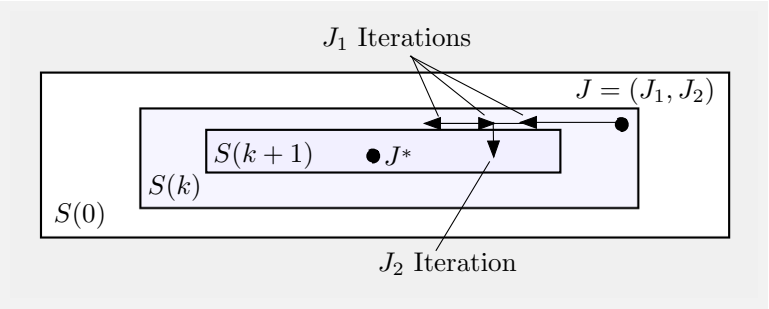


Figure 2.6.2 Geometric interpretation of the mechanism for asynchronous convergence. Iteration on a single component of a function $J \in S(k)$, say J_ℓ, keeps J in $S(k)$, while it moves J_ℓ into the corresponding component $S_\ell(k+1)$ of $S(k+1)$, where it remains throughout the subsequent iterations. Once all components J_ℓ have been iterated on at least once, the iterate is guaranteed to be in $S(k+1)$.

with semicontractive models for which there is no underlying sup-norm contraction. It is also relevant to the noncontractive models of Section 4.3 where again there is no underlying contraction. Figure 2.6.2 illustrates the mechanism by which asynchronous convergence is achieved.

We note a few extensions of the theorem. It is possible to allow T to be time-varying, so in place of T we operate with a sequence of mappings T_k, $k = 0, 1, \ldots$. Then if all T_k have a common fixed point J^*, the conclusion of the theorem holds (see Exercise 2.2 for a more precise statement). This extension is useful in some of the algorithms to be discussed later. Another extension is to allow T to have multiple fixed points and introduce an assumption that roughly says that $\cap_{k=0}^{\infty} S(k)$ is the set of fixed points. Then the conclusion is that any limit point (in an appropriate sense) of $\{J^t\}$ is a fixed point.

We now apply the preceding convergence theorem to the totally asynchronous VI algorithm under the contraction assumption. Note that the monotonicity Assumption 2.1.1 is not necessary (just like it is not needed for the synchronous convergence of $\{T^k J\}$ to J^*).

Proposition 2.6.2: Let the contraction Assumption 2.1.2 hold, together with Assumption 2.6.1. Then if $J^0 \in \mathcal{B}(X)$, a sequence $\{J^t\}$ generated by the asynchronous VI algorithm (2.61) converges to J^*.

Proof: We apply Prop. 2.6.1 with

$$S(k) = \big\{J \in \mathcal{B}(X) \mid \|J^k - J^*\| \leq \alpha^k \|J^0 - J^*\|\big\}, \qquad k = 0, 1, \ldots.$$

Since T is a contraction with modulus α, the synchronous convergence

condition is satisfied. Since T is a weighted sup-norm contraction, the box condition is also satisfied, and the result follows. **Q.E.D.**

2.6.2 Asynchronous Policy Iteration

We will now develop asynchronous PI algorithms that have comparable properties to the asynchronous VI algorithm of the preceding subsection. The processors collectively maintain and update an estimate J^t of the optimal cost function, and an estimate μ^t of an optimal policy. The local portions of J^t and μ^t of processor ℓ are denoted J_ℓ^t and μ_ℓ^t, respectively, i.e., $J_\ell^t(x) = J^t(x)$ and $\mu_\ell^t(x) = \mu^t(x)$ for all $x \in X_\ell$.

For each processor ℓ, there are two disjoint subsets of times $\mathcal{R}_\ell, \overline{\mathcal{R}}_\ell \subset \{0, 1, \ldots\}$, corresponding to policy improvement and policy evaluation iterations, respectively. At the times $t \in \mathcal{R}_\ell \cup \overline{\mathcal{R}}_\ell$, the local cost function J_ℓ^t of processor ℓ is updated using "delayed" local costs $J_j^{\tau_{\ell j}(t)}$ of other processors $j \neq \ell$, where $0 \leq \tau_{\ell j}(t) \leq t$. At the times $t \in \mathcal{R}_\ell$ (the local policy improvement times), the local policy μ_ℓ^t is also updated. For various choices of $\mathcal{R}_\ell$ and $\overline{\mathcal{R}}_\ell$, the algorithm takes the character of VI (when $\mathcal{R}_\ell = \{0, 1, \ldots\}$), and PI (when $\overline{\mathcal{R}}_\ell$ contains a large number of time indices between successive elements of $\mathcal{R}_\ell$). As before, we view $t - \tau_{\ell j}(t)$ as a "communication delay," and we require Assumption 2.6.1. †

In a natural asynchronous version of optimistic PI, at each time t, each processor ℓ does one of the following:

(a) *Local policy improvement*: If $t \in \mathcal{R}_\ell$, processor ℓ sets for all $x \in X_\ell$,

$$J_\ell^{t+1}(x) = \min_{u \in U(x)} H\big(x, u, J_1^{\tau_{\ell 1}(t)}, \ldots, J_m^{\tau_{\ell m}(t)}\big), \tag{2.62}$$

$$\mu_\ell^{t+1}(x) \in \arg\min_{u \in U(x)} H\big(x, u, J_1^{\tau_{\ell 1}(t)}, \ldots, J_m^{\tau_{\ell m}(t)}\big). \tag{2.63}$$

(b) *Local policy evaluation*: If $t \in \overline{\mathcal{R}}_\ell$, processor ℓ sets for all $x \in X_\ell$,

$$J_\ell^{t+1}(x) = H\big(x, \mu^t(x), J_1^{\tau_{\ell 1}(t)}, \ldots, J_m^{\tau_{\ell m}(t)}\big), \tag{2.64}$$

and leaves μ_ℓ unchanged, i.e., $\mu_\ell^{t+1}(x) = \mu_\ell^t(x)$ for all $x \in X_\ell$.

(c) *No local change*: If $t \notin \mathcal{R}_\ell \cup \overline{\mathcal{R}}_\ell$, processor ℓ leaves J_ℓ and μ_ℓ unchanged, i.e., $J_\ell^{t+1}(x) = J_\ell^t(x)$ and $\mu_\ell^{t+1}(x) = \mu_\ell^t(x)$ for all $x \in X_\ell$.

Unfortunately, even when implemented without the delays $\tau_{\ell j}(t)$, the preceding PI algorithm is unreliable. The difficulty is that the algorithm

† As earlier in all PI algorithms we assume that the infimum over $u \in U(x)$ in the policy improvement operation is attained, and we write min in place of inf.

involves a mix of applications of T and various mappings T_μ that have different fixed points, so in the absence of some systematic tendency towards J^* there is the possibility of oscillation (see Fig. 2.6.3). While this does not happen in synchronous versions (cf. Prop. 2.5.1), asynchronous versions of the algorithm (2.33) may oscillate unless J^0 satisfies some special condition (examples of this type of oscillation have been constructed in the paper [WiB93]; see also [Ber10], which translates an example from [WiB93] to the notation of the present book).

In this subsection and the next we will develop two distributed asynchronous PI algorithms, each embodying a distinct mechanism that precludes the oscillatory behavior just described. In the first algorithm, there is a simple randomization scheme, according to which a policy evaluation of the form (2.64) is replaced by a policy improvement (2.62)-(2.63) with some positive probability. In the second algorithm, given in Section 2.6.3, we introduce a mapping F_μ, which has a common fixed point property: its fixed point is related to J^* and is the same for all μ, so the anomaly illustrated in Fig. 2.6.3 cannot occur. The first algorithm is simple but requires some restrictions, including that the set of policies is finite. The second algorithm is more sophisticated and does not require this restriction. Both of these algorithms *do not require the monotonicity assumption*.

An Optimistic Asynchronous PI Algorithm with Randomization

We introduce a randomization scheme for avoiding oscillatory behavior. It is defined by a small probability $p > 0$, according to which a policy evaluation iteration is replaced by a policy improvement iteration with probability p, independently of the results of past iterations. We model this randomization by assuming that before the algorithm is started, we restructure the sets $\mathcal{R}_\ell$ and $\overline{\mathcal{R}}_\ell$ as follows: we take each element of each set $\overline{\mathcal{R}}_\ell$, and with probability p, remove it from $\overline{\mathcal{R}}_\ell$, and add it to $\mathcal{R}_\ell$ (independently of other elements). We will assume the following:

Assumption 2.6.2:

(a) The set of policies $\mathcal{M}$ is finite.

(b) There exists an integer $B \geq 0$ such that

$$(\mathcal{R}_\ell \cup \overline{\mathcal{R}}_\ell) \cap \{\tau \mid t < \tau \leq t + B\} \neq \emptyset, \qquad \forall\ t,\ \ell.$$

(c) There exists an integer $B' \geq 0$ such that

$$0 \leq t - \tau_{\ell j}(t) \leq B', \qquad \forall\ t,\ \ell,\ j.$$

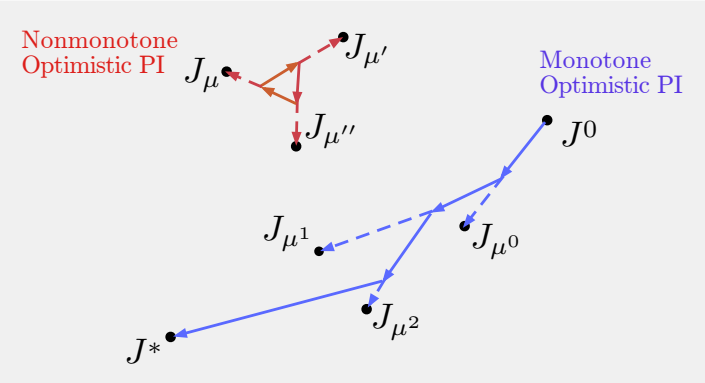


Figure 2.6.3 Illustration of optimistic asynchronous PI under the monotonicity and the contraction assumptions. When started with J^0 and μ^0 satisfying

$$J^0 \geq TJ^0 = T_{\mu^0} J^0,$$

the algorithm converges monotonically to J^* (see the trajectory on the right). However, for other initial conditions, there is a possibility for oscillations, since with changing values of μ, the mappings T_μ have different fixed points and "aim at different targets" (see the trajectory on the left, which illustrates a cycle between three policies μ, μ', μ''). It turns out that such oscillations are not possible when the algorithm is implemented synchronously (cf. Prop. 2.5.1), but may occur in asynchronous implementations.

Assumption 2.6.2 guarantees that each processor ℓ will execute at least one policy evaluation or policy improvement iteration within every block of B consecutive iterations, and places a bound B' on the communication delays. The convergence of the algorithm is shown in the following proposition.

Proposition 2.6.3: Under the contraction Assumption 2.1.2, and Assumptions 2.6.1, and 2.6.2, for the preceding algorithm with randomization, we have

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

with probability one.

Proof: Let J^* and J_μ be the fixed points of T and T_μ, respectively, and denote by $\mathcal{M}^*$ the set of optimal policies:

$$\mathcal{M}^* = \{\mu \in \mathcal{M} \mid J_\mu = J^*\} = \{\mu \in \mathcal{M} \mid T_\mu J^* = TJ^*\}.$$

We will show that the algorithm eventually (with probability one) enters a small neighborhood of J^* within which it remains, generates policies in $\mathcal{M}^*$, becomes equivalent to asynchronous VI, and therefore converges to J^* by Prop. 2.6.2. The idea of the proof is twofold; cf. Props. 2.5.4 and 2.5.5.

(1) There exists a small enough weighted sup-norm sphere centered at J^*, call it $\mathcal{S}^*$, within which policy improvement generates only policies in $\mathcal{M}^*$, so policy evaluation with such policies as well as policy improvement keep the algorithm within $\mathcal{S}^*$ if started there, and reduce the weighted sup-norm distance to J^*, in view of the contraction and common fixed point property of T and T_μ, $\mu \in \mathcal{M}^*$. This is a consequence of Prop. 2.3.1 [cf. Eq. (2.16)].

(2) With probability one, thanks to the randomization device, the algorithm will eventually enter permanently $\mathcal{S}^*$ with a policy in $\mathcal{M}^*$.

We now establish (1) and (2) in suitably refined form to account for the presence of delays and asynchronism. As in the proof of Prop. 2.5.5, we can prove that given J^0, we have that $\{J^t\} \subset D$, where D is a bounded set that depends on J^0. We define

$$S(k) = \big\{J \mid \|J - J^*\| \leq \alpha^k c\big\},$$

where c is sufficiently large so that $D \subset S(0)$. Then $J^t \in D$ and hence $J^t \in S(0)$ for all t.

Let k^* be such that

$$J \in S(k^*) \ \text{ and } \ T_\mu J = TJ \qquad \Rightarrow \qquad \mu \in \mathcal{M}^*. \tag{2.65}$$

Such a k^* exists in view of the finiteness of $\mathcal{M}$ and Prop. 2.3.1 [cf. Eq. (2.16)].

We now claim that with probability one, for any given $k \geq 1$, J^t will eventually enter $S(k)$ and stay within $S(k)$ for at least B' additional consecutive iterations. This is because our randomization scheme is such that for any t and k, with probability at least $p^{k(B+B')}$ the next $k(B+B')$ iterations are policy improvements, so that

$$J^{t+k(B+B')-\xi} \in S(k)$$

for all ξ with $0 \leq \xi < B'$ [if $t \geq B' - 1$, we have $J^{t-\xi} \in S(0)$ for all ξ with $0 \leq \xi < B'$, so $J^{t+B+B'-\xi} \in S(1)$ for $0 \leq \xi < B'$, which implies that $J^{t+2(B+B')-\xi} \in S(2)$ for $0 \leq \xi < B'$, etc].

It follows that with probability one, for some $\overline{t}$ we will have $J^\tau \in S(k^*)$ for all τ with $\overline{t} - B' \leq \tau \leq \overline{t}$, as well as $\mu^{\overline{t}} \in \mathcal{M}^*$ [cf. Eq. (2.65)]. Based on property (2.65) and the definition (2.63)-(2.64) of the algorithm, we see that at the next iteration, we have $\mu^{\overline{t}+1} \in \mathcal{M}^*$ and

$$\|J^{\overline{t}+1} - J^*\| \leq \|J^{\overline{t}} - J^*\| \leq \alpha^{k^*} c,$$

so $J^{\overline{t}+1} \in S(k^*)$; this is because in view of $J_{\mu^{\overline{t}}} = J^*$, and the contraction property of T and $T_{\mu^{\overline{t}}}$, we have

$$\frac{\left|J_\ell^{\overline{t}+1}(x) - J_\ell^*(x)\right|}{v(x)} \leq \alpha \|J^{\overline{t}} - J^*\| \leq \alpha^{k^*+1} c, \tag{2.66}$$

for all $x \in X_\ell$ and ℓ such that $\overline{t} \in \overline{\mathcal{R}}_\ell \cup \mathcal{R}_\ell$, while

$$J^{\overline{t}+1}(x) = J^{\overline{t}}(x)$$

for all other x. Proceeding similarly, it follows that for all $t > \overline{t}$ we will have

$$J^\tau \in S(k^*), \qquad \forall\ t \text{ with } t - B' \leq \tau \leq t,$$

as well as $\mu^t \in \mathcal{M}^*$. Thus, after at most B iterations following $\overline{t}$ [after all components J_ℓ are updated through policy evaluation or policy improvement at least once so that

$$\frac{\left|J_\ell^{t+1}(x) - J_\ell^*(x)\right|}{v(x)} \leq \alpha \|J^{\overline{t}} - J^*\| \leq \alpha^{k^*+1} c,$$

for every i, $x \in X_\ell$, and some t with $\overline{t} \leq t < \overline{t} + B$, cf. Eq. (2.66)], J^t will enter $S(k^*+1)$ permanently, with $\mu^t \in \mathcal{M}^*$ (since $\mu^t \in \mathcal{M}^*$ for all $t \geq \overline{t}$ as shown earlier). Then, with the same reasoning, after at most another $B' + B$ iterations, J^t will enter $S(k^*+2)$ permanently, with $\mu^t \in \mathcal{M}^*$, etc. Thus J^t will converge to J^* with probability one. **Q.E.D.**

The proof of Prop. 2.6.3 shows that eventually (with probability one after some iteration) the algorithm will become equivalent to asynchronous VI (each policy evaluation will produce the same results as a policy improvement), while generating optimal policies exclusively. However, the expected number of iterations for this to happen can be very large. Moreover the proof depends on the set of policies being finite. These observations raise questions regarding the practical effectiveness of the algorithm. However, it appears that for many problems the algorithm works well, particularly when oscillatory behavior is a rare occurrence.

A potentially important issue is the choice of the randomization probability p. If p is too small, convergence may be slow because oscillatory behavior may go unchecked for a long time. On the other hand if p is large, a correspondingly large number of policy improvement iterations may be performed, and the hoped for benefits of optimistic PI may be lost. Adaptive schemes which adjust p based on algorithmic progress may be an interesting possibility for addressing this issue.

2.6.3 Optimistic Asynchronous Policy Iteration with a Uniform Fixed Point

We will now discuss another approach to address the convergence difficulties of the "natural" asynchronous PI algorithm (2.62)-(2.64). As illustrated in Fig. 2.6.3 in connection with optimistic PI, the mappings T and T_μ have different fixed points. As a result, optimistic and distributed PI, which involve an irregular mixture of applications of T_μ and T, do not have a "consistent target" at which to aim.

With this in mind, we introduce a new mapping that is parametrized by μ and has *a common fixed point for all* μ, which in turn yields J^*. This mapping is a weighted sup-norm contraction with modulus α, so it may be used in conjunction with asynchronous VI and PI. An additional benefit is that the monotonicity Assumption 2.1.1 is not needed to prove convergence in the analysis that follows; the contraction Assumption 2.1.2 is sufficient (see Exercise 2.3 for an application).

The mapping operates on a pair (V, Q) where:

- V is a function with a component $V(x)$ for each x (in the DP context it may be viewed as a cost function).
- Q is a function with a component $Q(x, u)$ for each pair (x, u) [in the DP context $Q(x, u)$, is known as a *Q-factor*].

The mapping produces a pair

$$\big(MF_\mu(V, Q), F_\mu(V, Q)\big),$$

where

- $F_\mu(V, Q)$ is a function with a component $F_\mu(V, Q)(x, u)$ for each (x, u), defined by
$$F_\mu(V, Q)(x, u) = H\big(x, u, \min\{V, Q_\mu\}\big), \tag{2.67}$$
where for any Q and μ, we denote by Q_μ the function of x defined by
$$Q_\mu(x) = Q\big(x, \mu(x)\big), \qquad x \in X, \tag{2.68}$$
and for any two functions V_1 and V_2 of x, we denote by $\min\{V_1, V_2\}$ the function of x given by
$$\min\{V_1, V_2\}(x) = \min\big\{V_1(x), V_2(x)\big\}, \qquad x \in X.$$

- $MF_\mu(V, Q)$ is a function with a component $\big(MF_\mu(V, Q)\big)(x)$ for each x, where M denotes minimization over u, so that
$$\big(MF_\mu(V, Q)\big)(x) = \min_{u \in U(x)} F_\mu(V, Q)(x, u). \tag{2.69}$$

Example 2.6.2 (Asynchronous Optimistic Policy Iteration for Discounted Finite-State MDP)

Consider the special case of the finite-state discounted MDP of Example 1.2.2. We have

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

and

$$\begin{aligned} F_\mu(V,Q)(x,u) &= H\big(x,u,\min\{V,Q_\mu\}\big) \\ &= \sum_{y=1}^{n} p_{xy}(u)\Big(g(x,u,y) + \alpha \min\big\{V(y), Q(y,\mu(y))\big\}\Big), \end{aligned}$$

$$\big(MF_\mu(V,Q)\big)(x) = \min_{u\in U(x)} \sum_{y=1}^{n} p_{xy}(u)\Big(g(x,u,y) + \alpha \min\big\{V(y), Q(y,\mu(y))\big\}\Big),$$

[cf. Eqs. (2.67)-(2.69)]. Note that $F_\mu(V,Q)$ is the mapping that defines Bellman's equation for the Q-factors of a policy μ in an optimal stopping problem where the stopping cost at state y is equal to $V(y)$.

We now consider the mapping G_μ given by

$$G_\mu(V,Q) = \big(MF_\mu(V,Q), F_\mu(V,Q)\big), \tag{2.70}$$

and show that it has a uniform contraction property and a corresponding uniform fixed point. To this end, we introduce the norm

$$\big\|(V,Q)\big\| = \max\big\{\|V\|, \|Q\|\big\}$$

in the space of (V,Q), where $\|V\|$ is the weighted sup-norm of V, and $\|Q\|$ is defined by

$$\|Q\| = \sup_{x\in X,\, u\in U(x)} \frac{\big|Q(x,u)\big|}{v(x)}.$$

We have the following proposition.

Proposition 2.6.4: Let the contraction Assumption 2.1.2 hold. Consider the mapping G_μ defined by Eqs. (2.67)-(2.70). Then for all μ:

(a) (J^*, Q^*) is the unique fixed point of G_μ, where Q^* is defined by

$$Q^*(x,u) = H(x,u,J^*), \qquad x \in X,\ u \in U(x). \tag{2.71}$$

(b) The following uniform contraction property holds for all (V,Q) and $(\tilde V, \tilde Q)$:

$$\big\|G_\mu(V,Q) - G_\mu(\tilde V, \tilde Q)\big\| \le \alpha \big\|(V,Q) - (\tilde V, \tilde Q)\big\|.$$

Proof: (a) Using the definition (2.71) of Q^*, we have

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

so that

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

Using the definition (2.67) of F_μ, it follows that $F_\mu(J^*,Q^*) = Q^*$ and also that $MF_\mu(J^*,Q^*) = J^*$, so (J^*,Q^*) is a fixed point of G_μ for all μ. The uniqueness of this fixed point will follow from the contraction property of part (b).

(b) We first show that for all (V,Q) and $(\tilde V,\tilde Q)$, we have

$$\begin{aligned}\big\|F_\mu(V,Q) - F_\mu(\tilde V,\tilde Q)\big\| &\le \alpha\big\|\min\{V,Q_\mu\} - \min\{\tilde V,\tilde Q_\mu\}\big\| \\ &\le \alpha \max\big\{\|V-\tilde V\|,\, \|Q-\tilde Q\|\big\}.\end{aligned} \tag{2.72}$$

Indeed, the first inequality follows from the definition (2.67) of F_μ and the contraction Assumption 2.1.2. The second inequality follows from a nonexpansiveness property of the minimization map: for any J_1, J_2, $\tilde J_1$, $\tilde J_2$, we have

$$\big\|\min\{J_1,J_2\} - \min\{\tilde J_1,\tilde J_2\}\big\| \le \max\big\{\|J_1-\tilde J_1\|, \|J_2-\tilde J_2\|\big\}; \tag{2.73}$$

[to see this, write for every x,

$$\frac{J_m(x)}{v(x)} \le \max\big\{\|J_1-\tilde J_1\|, \|J_2-\tilde J_2\|\big\} + \frac{\tilde J_m(x)}{v(x)}, \qquad m = 1,2,$$

take the minimum of both sides over m, exchange the roles of J_m and $\tilde J_m$, and take supremum over x]. Here we use the relation (2.73) for $J_1 = V$, $\tilde J_1 = \tilde V$, and $J_2(x) = Q\big(x,\mu(x)\big)$, $\tilde J_2(x) = \tilde Q\big(x,\mu(x)\big)$, for all $x \in X$.

We next note that for all $Q, \tilde Q$, †

$$\|MQ - M\tilde Q\| \le \|Q - \tilde Q\|,$$

† For a proof, we write

$$\frac{Q(x,u)}{v(x)} \le \|Q-\tilde Q\| + \frac{\tilde Q(x,u)}{v(x)}, \qquad \forall\ u \in U(x),\ x \in X,$$

take infimum of both sides over $u \in U(x)$, exchange the roles of Q and $\tilde Q$, and take supremum over $x \in X$.

which together with Eq. (2.72) yields

$$\max\big\{\big\|MF_\mu(V,Q)-MF_\mu(\tilde V,\tilde Q)\big\|,\big\|F_\mu(V,Q)-F_\mu(\tilde V,\tilde Q)\big\|\big\} \le \alpha\max\big\{\|V-\tilde V\|,\,\|Q-\tilde Q\|\big\},$$

or equivalently $\big\|G_\mu(V,Q)-G_\mu(\tilde V,\tilde Q)\big\| \le \alpha\big\|(V,Q)-(\tilde V,\tilde Q)\big\|$. **Q.E.D.**

Because of the uniform contraction property of Prop. 2.6.4(b), a distributed fixed point iteration, like the VI algorithm of Eq. (2.61), can be used in conjunction with the mapping (2.70) to generate asynchronously a sequence $\big\{(V^t,Q^t)\big\}$ that is guaranteed to converge to (J^*,Q^*) for any sequence $\{\mu^t\}$. This can be verified using the proof of Prop. 2.6.2 (more precisely, a proof that closely parallels the one of that proposition); the mapping (2.70) plays the role of T in Eq. (2.61). †

Asynchronous PI Algorithm

We now describe a PI algorithm, which applies asynchronously the components $MF_\mu(V,Q)$ and $F_\mu(V,Q)$ of the mapping $G_\mu(V,Q)$ of Eq. (2.70). The first component is used for local policy improvement and makes a local update to V and μ, while the second component is used for local policy evaluation and makes a local update to Q. The algorithm draws its validity from the weighted sup-norm contraction property of Prop. 2.6.4(b) and the asynchronous convergence theory (Prop. 2.6.2 and Exercise 2.2).

The algorithm is a modification of the "natural" asynchronous PI algorithm (2.63)-(2.64) [without the "communication delays" $t-\tau_{\ell j}(t)$]. It generates sequences $\{V^t,Q^t,\mu^t\}$, which will be shown to converge, in the sense that $V^t\to J^*$, $Q^t\to Q^*$. Note that this is not the only distributed iterative algorithm that can be constructed using the contraction property of Prop. 2.6.4, because this proposition allows a lot of freedom of choice for the policy μ. The paper by Bertsekas and Yu [BeY12] provides an extensive discussion of alternative possibilities, including stochastic simulation-based iterative algorithms, and algorithms that involve function approximation.

To define the asynchronous computation framework, we consider again m processors, a partition of X into sets $X_1,\ldots,X_m$, and assignment of each subset X_ℓ to a processor $\ell\in\{1,\ldots,m\}$. For each ℓ, there are two infinite disjoint subsets of times $\mathcal{R}_\ell,\overline{\mathcal{R}}_\ell\subset\{0,1,\ldots\}$, corresponding to policy improvement and policy evaluation iterations, respectively. Each processor ℓ operates on $V^t(x)$, $Q^t(x,u)$, and $\mu^t(x)$, only for the states x within its "local" state space X_ℓ. Moreover, to execute the steps (a) and (b) of the algorithm, processor ℓ needs only the values $Q^t\big(x,\mu^t(x)\big)$ of Q^t [which are

† Because F_μ and G_μ depend on μ, which changes as the algorithm progresses, it is necessary to use a minor extension of the asynchronous convergence theorem, given in Exercise 2.2, for the convergence proof.

equal to $Q^t_{\mu^t}(x)$; cf. Eq. (2.68)]. In particular, at each time t, each processor ℓ does one of the following:

(a) *Local policy improvement*: If $t \in \mathcal{R}_\ell$, processor ℓ sets for all $x \in X_\ell$, †

$$V^{t+1}(x) = \min_{u \in U(x)} H\big(x, u, \min\{V^t, Q^t_{\mu^t}\}\big) = \big(MF_{\mu^t}(V^t, Q^t)\big)(x),$$

sets $\mu^{t+1}(x)$ to a u that attains the minimum, and leaves Q unchanged, i.e., $Q^{t+1}(x,u) = Q^t(x,u)$ for all $x \in X_\ell$ and $u \in U(x)$.

(b) *Local policy evaluation*: If $t \in \overline{\mathcal{R}}_\ell$, processor ℓ sets for all $x \in X_\ell$ and $u \in U(x)$,

$$Q^{t+1}(x,u) = H\big(x, u, \min\{V^t, Q^t_{\mu^t}\}\big) = F_{\mu^t}(V^t, Q^t)(x,u),$$

and leaves V and μ unchanged, i.e., $V^{t+1}(x) = V^t(x)$ and $\mu^{t+1}(x) = \mu^t(x)$ for all $x \in X_\ell$.

(c) *No local change*: If $t \notin \mathcal{R}_\ell \cup \overline{\mathcal{R}}_\ell$, processor ℓ leaves Q, V, and μ unchanged, i.e., $Q^{t+1}(x,u) = Q^t(x,u)$ for all $x \in X_\ell$ and $u \in U(x)$, $V^{t+1}(x) = V^t(x)$, and $\mu^{t+1}(x) = \mu^t(x)$ for all $x \in X_\ell$.

Note that while this algorithm does not involve the "communication delays" $t - \tau_{\ell j}(t)$, it can clearly be extended to include them. The reason is that our asynchronous convergence analysis framework in combination with the uniform weighted sup-norm contraction property of Prop. 2.6.4 can tolerate the presence of such delays.

Reduced Space Implementation

The preceding PI algorithm may be used for the calculation of both J^* and Q^*. However, if the objective is just to calculate J^*, a simpler and more efficient algorithm is possible. To this end, we observe that the preceding algorithm can be operated so that it does not require the maintenance of the entire function Q. The reason is that the values $Q^t(x,u)$ with $u \neq \mu^t(x)$ do not appear in the calculations, and hence we need only the values $Q^t_{\mu^t}(x) = Q\big(x, \mu^t(x)\big)$, which we store in a function J^t:

$$J^t(x) = Q\big(x, \mu^t(x)\big).$$

This observation is the basis for the following algorithm.

At each time t and for each processor ℓ:

(a) *Local policy improvement*: If $t \in \mathcal{R}_\ell$, processor ℓ sets for all $x \in X_\ell$,

$$J^{t+1}(x) = V^{t+1}(x) = \min_{u \in U(x)} H\big(x, u, \min\{V^t, J^t\}\big), \tag{2.74}$$

† As earlier we assume that the infimum over $u \in U(x)$ in the policy improvement operation is attained, and we write min in place of inf.

and sets $\mu^{t+1}(x)$ to a u that attains the minimum.

(b) *Local policy evaluation*: If $t \in \overline{\mathcal{R}}_\ell$, processor ℓ sets for all $x \in X_\ell$,

$$J^{t+1}(x) = H\big(x, \mu^t(x), \min\{V^t, J^t\}\big), \tag{2.75}$$

and leaves V and μ unchanged, i.e., for all $x \in X_\ell$,

$$V^{t+1}(x) = V^t(x), \qquad \mu^{t+1}(x) = \mu^t(x).$$

(c) *No local change*: If $t \notin \mathcal{R}_\ell \cup \overline{\mathcal{R}}_\ell$, processor ℓ leaves J, V, and μ unchanged, i.e., for all $x \in X_\ell$,

$$J^{t+1}(x) = J^t(x), \quad V^{t+1}(x) = V^t(x), \quad \mu^{t+1}(x) = \mu^t(x).$$

Example 2.6.3 (Asynchronous Optimistic Policy Iteration for Discounted Finite-State MDP - Continued)

As an illustration of the preceding reduced space implementation, consider the special case of the finite-state discounted MDP of Example 2.6.2. Here

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

and the mapping $F_\mu(V,Q)$ given by

$$F_\mu(V,Q)(x,u) = \sum_{y=1}^{n} p_{xy}(u)\Big(g(x,u,y) + \alpha \min\big\{V(y), Q(y,\mu(y))\big\}\Big),$$

defines the Q-factors of μ in a corresponding stopping problem. In the PI algorithm (2.74)-(2.75), policy evaluation of μ aims to solve this stopping problem, rather than solve a linear system of equations, as in classical PI. In particular, the policy evaluation iteration (2.75) is

$$J^{t+1}(x) = \sum_{y=1}^{n} p_{xy}\big(\mu^t(x)\big)\Big(g\big(x,\mu^t(x),y\big) + \alpha \min\big\{V^t(y), J^t(y)\big\}\Big),$$

for all $x \in X_\ell$. The policy improvement iteration (2.74) is a VI for the stopping problem:

$$J^{t+1}(x) = V^{t+1}(x) = \min_{u \in U(x)} \sum_{y=1}^{n} p_{xy}(u)\Big(g(x,u,y) + \alpha \min\big\{V^t(y), J^t(y)\big\}\Big),$$

for all $x \in X_\ell$, while the current policy is locally updated by

$$\mu^{t+1}(x) \in \arg\min_{u \in U(x)} \sum_{y=1}^{n} p_{xy}(u)\Big(g(x,u,y) + \alpha \min\big\{V^t(y), J^t(y)\big\}\Big),$$

for all $x \in X_\ell$. The "stopping cost" $V^t(y)$ is the most recent cost value, obtained by local policy improvement at y.

Example 2.6.4 (Asynchronous Optimistic Policy Iteration for Minimax Problems and Dynamic Games)

Consider the optimistic PI algorithm (2.74)-(2.75) for the case of the minimax problem of Example 1.2.5 of Chapter 1, where

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

Then the local policy evaluation step [cf. Eq. (2.75)] is written as

$$J^{t+1}(x)=\sup_{w\in W(x,\mu^t(x))}\Big[g\big(x,\mu^t(x),w\big)+\alpha\min\Big\{V^t\big(f(x,\mu^t(x),w)\big),J^t\big(f(x,\mu^t(x),w)\big)\Big\}\Big].$$

The local policy improvement step [cf. Eq. (2.74)] takes the form

$$J^{t+1}(x)=V^{t+1}(x)=\min_{u\in U(x)}\sup_{w\in W(x,u)}\Big[g(x,u,w)+\alpha\min\Big\{V^t\big(f(x,u,w)\big),J^t\big(f(x,u,w)\big)\Big\}\Big],$$

and sets $\mu^{t+1}(x)$ to a u that attains the minimum.

Similarly for the discounted dynamic game problem of Example 1.2.4 of Chapter 1, a local policy evaluation step [cf. Eq. (2.75)] consists of a local VI for the maximizer's DP problem assuming a fixed policy for the minimizer, and a stopping cost V^t as per Eq. (2.75). A local policy improvement step [cf. Eq. (2.74)] at state x consists of the solution of a static game with a payoff matrix that also involves $\min\{V^t,J^t\}$ in place of J^t, as per Eq. (2.74).

A Variant with Interpolation

While the use of $\min\{V^t,J^t\}$ (rather than J^t) in Eq. (2.75) provides a convergence enforcement mechanism for the algorithm, it may also become a source of inefficiency, particularly when $V^t(x)$ approaches its limit $J^*(x)$ from lower values for many x. Then $J^{t+1}(x)$ is set to a lower value than the iterate

$$\hat{J}^{t+1}(x)=H\big(x,\mu^t(x),J^t\big),\tag{2.76}$$

given by the "standard" policy evaluation iteration, and in some cases this may slow down the algorithm.

A possible way to address this is to use an algorithmic variation that modifies appropriately Eq. (2.75), using interpolation with a parameter $\gamma_t\in(0,1]$, with $\gamma_t\to 0$. In particular, for $t\in\overline{\mathcal{R}}_\ell$ and $x\in X_\ell$, we calculate the values $J^{t+1}(x)$ and $\hat{J}^{t+1}(x)$ given by Eqs. (2.75) and (2.76), and if

$$J^{t+1}(x)<\hat{J}^{t+1}(x),\tag{2.77}$$

we reset $J^{t+1}(x)$ to

$$(1-\gamma_t)J^{t+1}(x)+\gamma_t\hat{J}^{t+1}(x). \tag{2.78}$$

The idea of the algorithm is to aim for a larger value of $J^{t+1}(x)$ when the condition (2.77) holds. Asymptotically, as $\gamma_t \to 0$, the iteration (2.77)-(2.78) becomes identical to the convergent update (2.75). For a detailed analysis we refer to the paper by Bertsekas and Yu [BeY10].

2.7 NOTES, SOURCES, AND EXERCISES

Section 2.1: The contractive DP model of this section was first studied systematically by Denardo [Den67], who assumed an unweighted sup-norm, proved the basic results of Section 2.1, and described some of their applications. In this section, we have extended the analysis of [Den67] to the case of weighted sup-norm contractions.

Section 2.2: The abstraction of the computational methodology for finite-state discounted MDP within the broader framework of weighted sup-norm contractions and an infinite state space (Sections 2.2-2.6) follows the author's survey [Ber12b], and relies on several earlier analyses that use more specialized assumptions.

Section 2.3: The multistep error bound of Prop. 2.2.2 is based on Scherrer [Sch12], which explores periodic policies in approximate VI and PI in finite-state discounted MDP (see also Scherrer and Lesner [ShL12], who give an example showing that the bound for approximate VI of Prop. 2.3.2 is essentially sharp for discounted finite-state MDP). For a related discussion of approximate VI, including the error amplification phenomenon of Example 2.3.1, and associated error bounds, see de Farias and Van Roy [DFV00].

Section 2.4: The error bound of Prop. 2.4.3 extends a standard bound for finite-state discounted MDP, derived by Bertsekas and Tsitsiklis [BeT96] (Section 6.2.2), and shown to be tight by an example.

Section 2.5: Optimistic PI has received a lot of attention in the literature, particularly for finite-state discounted MDP, and it is generally thought to be computationally more efficient in practice than ordinary PI (see e.g., Puterman [Put94], who refers to the method as "modified PI"). The convergence analysis of the synchronous optimistic PI (Section 2.5.1) follows Rothblum [Rot79], who considered the case of an unweighted sup-norm ($v = e$); see also Canbolat and Rothblum [CaR13]. The error bound for optimistic PI (Section 2.5.2) is due to Thierry and Scherrer [ThS10b], which was given for the case of a finite-state discounted MDP. We follow closely their line of proof. Related error bounds and analysis are given by Scherrer [Sch11].

The λ-PI method [cf. Eq. (2.34)] was introduced by Bertsekas and Ioffe [BeI96], and was also presented in the book [BeT96], Section 2.3.1. It is the basis of the LSPE(λ) policy evaluation method, described by Nedić and Bertsekas [NeB03], and by Bertsekas, Borkar, and Nedić [BBN04]. It was studied further in approximate DP contexts by Thierry and Scherrer [ThS10a], Bertsekas [Ber11b], and Scherrer [Sch11]. An extension of λ-PI, called Λ-PI, uses a different parameter λ_i for each state i, and is discussed in Section 5 of the paper by Yu and Bertsekas [YuB12]. Based on the discussion of Section 1.2.5 and Exercise 1.2, Λ-PI may be viewed as a diagonally scaled version of the proximal algorithm, i.e., one that uses a different penalty parameter for each proximal term.

When the state and control spaces are finite, and cost approximation over a subspace $\{\Phi r \mid r \in \Re^s\}$ is used (cf. Section 1.2.4), a prominent approximate PI approach is to replace the exact policy evaluation equation

$$J_{\mu^k} = T_{\mu^k} J_{\mu^k}$$

with an approximate version of the form

$$\Phi r_k = W T_{\mu^k}(\Phi r_k), \tag{2.79}$$

where W is some $n \times n$ matrix whose range space is the subspace spanned by the columns of Φ, where n is the number of states. For example the projected and aggregation equations, described in Section 1.2.4, have this form. The next policy μ^{k+1} is obtained using the policy improvement equation

$$T_{\mu^{k+1}}(\Phi r_k) = T(\Phi r_k). \tag{2.80}$$

A critical issue for the validity of such a method is whether the approximate Bellman equations

$$\Phi r = W T(\Phi r)$$

and

$$\Phi r = W T_\mu(\Phi r), \quad \mu \in \mathcal{M},$$

have a unique solution. This is true if the composite mappings $W \circ T$ and $W \circ T_\mu$ are contractions over $\Re^n$. In particular, in the case of an aggregation equation, where $W = \Phi D$, the rows of Φ and D are probability distributions, and T_μ, $\mu \in \mathcal{M}$, are monotone sup-norm contractions, the mappings $W \circ T$ and $W \circ T_\mu$ are also monotone sup-norm contractions. However, in other cases, including when policy evaluation is done using the projected equation, $W \circ T$ need not be monotone or be a contraction of any kind, and the approximate PI algorithm (2.79)-(2.80) may lead to systematic oscillations, involving cycles of policies (see related discussions in [BeT96], [Ber11c], and [Ber12a]). This phenomenon has been known

since the early days of approximate DP ([Ber96] and the book [BeT96]), but its practical implications have not been fully assessed. Generally, the line of analysis of Section 2.5.3, which does not require monotonicity or sup-norm contraction properties of the composite mappings $W \circ T$ and $W \circ T_\mu$, can be applied to the approximate PI algorithm (2.79)-(2.80), but only in the case where these mappings are contractions over $\Re^n$ with respect to a common norm $\|\cdot\|$; see Exercise 2.6 for further discussion.

Section 2.6: Asynchronous VI (Section 2.6.1) for finite-state discounted MDP and games, shortest path problems, and abstract DP models, was proposed in the author's paper on distributed DP [Ber82]. The asynchronous convergence theorem (Prop. 2.6.1) was first given in the author's paper [Ber83], where it was applied to a variety of algorithms, including VI for discounted and undiscounted DP, and gradient methods for unconstrained optimization (see also Bertsekas and Tsitsiklis [BeT89], where a textbook account is presented). The key convergence mechanism, which underlies the proof of Prop. 2.6.1, is that while the algorithm iterates asynchronously on the components J_ℓ of J, an iteration with any one component does not impede the progress made by iterations with the other components, thanks to the box condition. At the same time, progress towards the solution is continuing thanks to the synchronous convergence condition.

Earlier references on distributed asynchronous iterative algorithms include the work of Chazan and Miranker [ChM69] on Gauss-Seidel methods for solving linear systems of equations (who attributed the original algorithmic idea to Rosenfeld [Ros67]), and also Baudet [Bau78] on sup-norm contractive iterations. We refer to [BeT89] for detailed references.

Asynchronous algorithms have also been studied and applied to simulation-based DP, particularly in the context of Q-learning, first proposed by Watkins [Wat89], which may be viewed as a stochastic version of VI, and is a central algorithmic concept in approximate DP and reinforcement learning. Two principal approaches for the convergence analysis of asynchronous stochastic algorithms have been suggested.

The first approach, initiated in the paper by Tsitsiklis [Tsi94], considers the totally asynchronous computation of fixed points of abstract sup-norm contractive mappings and monotone mappings, which are defined in terms of an expected value. The algorithm of [Tsi94] contains as special cases Q-learning algorithms for finite-spaces discounted MDP and SSP problems. The analysis of [Tsi94] shares some ideas with the theory of Section 2.6.1, and also relies on the theory of stochastic approximation methods. For a subsequent analysis of the convergence of Q-learning for SSP, which addresses the issue of boundedness of the iterates, we refer to Yu and Bertsekas [YuB13b].

The second approach, treats asynchronous algorithms of the stochastic approximation type under some restrictions on the size of the communi-

cation delays or on the time between consecutive updates of a typical component. This approach was initiated in the paper by Tsitsiklis, Bertsekas, and Athans [TBA86], and was also developed in the book by Bertsekas and Tsitsiklis [BeT89] for stochastic gradient optimization methods. A related analysis that uses the ODE approach for more general fixed point problems was given in the paper by Borkar [Bor98], and was refined in the papers by Abounadi, Bertsekas, and Borkar [ABB02], and Borkar and Meyn [BoM00], which also considered applications to Q-learning. We refer to the monograph by Borkar [Bor08] for a more comprehensive discussion.

The convergence of asynchronous PI for finite-state discounted MDP under the condition

$$J^0 \geq T_{\mu^0} J^0$$

was shown by Williams and Baird [WiB93], who also gave examples showing that without this condition, cycling of the algorithm may occur. The asynchronous PI algorithm with a uniform fixed point (Section 2.6.3) was introduced in the papers by Bertsekas and Yu [BeY10], [BeY12], [YuB13a], in order to address this difficulty. Our analysis follows the analysis of these papers.

In addition to resolving the asynchronous convergence issue, the asynchronous PI algorithm of Section 2.6.3, obviates the need for minimization over all controls at every iteration (this is the generic computational efficiency advantage that optimistic PI typically holds over VI). Moreover, the algorithm admits a number of variations thanks to the fact that Prop. 2.6.4 asserts the contraction property of the mapping G_μ *for all* μ. This can be used to prove convergence in variants of the algorithm where the policy μ^t is updated more or less arbitrarily, with the aim to promote some objective. We refer to the paper [BeY12], which also derives related asynchronous simulation-based Q-learning algorithms with and without cost function approximation, where μ^t is replaced by a randomized policy to enhance exploration.

The randomized asynchronous optimistic PI algorithm of Section 2.6.2, introduced in the first edition of this book, also resolves the asynchronous convergence issue. The fact that this algorithm does not require the monotonicity assumption may be useful in nonDP algorithmic contexts (see [Ber16b] and Exercise 2.6).

In addition to discounted stochastic optimal control, the results of this chapter find application in the context of the stochastic shortest path problem of Example 1.2.6, when all policies are proper. Then, under some additional assumptions, it can be shown that T and T_μ are weighted sup-norm contractions with respect to a special norm. It follows that the analysis and algorithms of this chapter apply in this case. For a detailed discussion, we refer to the monograph [BeT96] and the survey [Ber12b]. For extensions to the case of countable state space, see the textbook [Ber12a], Section 3.6, and Hinderer and Waldmann [HiW05].

EXERCISES

2.1 (Periodic Policies)

Consider the multistep mappings $\overline{T}_\nu = T_{\mu_0} \cdots T_{\mu_{m-1}}$, $\nu \in \mathcal{M}_m$, defined in Exercise 1.1 of Chapter 1, where $\mathcal{M}_m$ is the set of m-tuples $\nu = (\mu_0, \ldots, \mu_{m-1})$, with $\mu_k \in \mathcal{M}$, $k = 1, \ldots, m-1$, and m is a positive integer. Assume that the mappings T_μ satisfy the monotonicity and contraction Assumptions 2.1.1 and 2.1.2, so that the same is true for the mappings $\overline{T}_\nu$ (with the contraction modulus of $\overline{T}_\nu$ being α^m, cf. Exercise 1.1).

(a) Show that the unique fixed point of $\overline{T}_\nu$ is J_π, where π is the nonstationary but periodic policy $\pi = \{\mu_0, \ldots, \mu_{m-1}, \mu_0, \ldots, \mu_{m-1}, \ldots\}$.

(b) Show that the multistep mappings $T_{\mu_0} \cdots T_{\mu_{m-1}}$, $T_{\mu_1} \cdots T_{\mu_{m-1}} T_{\mu_0}$, $\ldots$, $T_{\mu_{m-1}} T_{\mu_0} \cdots T_{\mu_{m-2}}$, have unique corresponding fixed points $J_0, J_1, \ldots$, J_{m-1}, which satisfy

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

Hint: Apply T_{μ_0} to the fixed point relation

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

to show that $T_{\mu_0} J_1$ is the fixed point of $T_{\mu_0} \cdots T_{\mu_{m-1}}$, i.e., is equal to J_0. Similarly, apply T_{μ_1} to the fixed point relation

$$J_2 = T_{\mu_2} \cdots T_{\mu_{m-1}} T_{\mu_0} T_{\mu_1} J_2,$$

to show that $T_{\mu_1} J_2$ is the fixed point of $T_{\mu_1} \cdots T_{\mu_{m-1}} T_{\mu_0}$, etc.

Solution: (a) Let us define

$$J_0 = \lim_{k \to \infty} \overline{T}_\nu^k J', \quad J_1 = \lim_{k \to \infty} \overline{T}_\nu^k (T_{\mu_0} J'), \ \ldots, \ J_{m-1} = \lim_{k \to \infty} \overline{T}_\nu^k (T_{\mu_0} \cdots T_{\mu_{m-2}} J'),$$

where J' is some function in $\mathcal{B}(X)$. Since $\overline{T}_\nu$ is a contraction mapping, $J_0, \ldots, J_{m-1}$ are all equal to the unique fixed point of $\overline{T}_\nu$. Since $J_0, \ldots, J_{m-1}$ are all equal, they are also equal to J_π (by the definition of J_π). Thus J_π is the unique fixed point of $\overline{T}_\nu$.

(b) Follow the hint.

2.2 (Asynchronous Convergence Theorem for Time-Varying Maps)

In reference to the framework of Section 2.6.1, let $\{T_t\}$ be a sequence of mappings from $\mathcal{R}(X)$ to $\mathcal{R}(X)$ that have a common unique fixed point J^*, let Assumption 2.6.1 hold, and assume that there is a sequence of nonempty subsets $\big\{S(k)\big\} \subset \mathcal{R}(X)$ with $S(k+1) \subset S(k)$ for all k, and with the following properties:

(1) *Synchronous Convergence Condition:* Every sequence $\{J^k\}$ with $J^k \in S(k)$ for each k, converges pointwise to J^*. Moreover, we have

$$T_t J \in S(k+1), \qquad \forall\ J \in S(k),\ k,\ t = 0, 1, \ldots.$$

(2) *Box Condition:* For all k, $S(k)$ is a Cartesian product of the form

$$S(k) = S_1(k) \times \cdots \times S_m(k),$$

where $S_\ell(k)$ is a set of real-valued functions on X_ℓ, $\ell = 1, \ldots, m$.

Then for every $J^0 \in S(0)$, the sequence $\{J^t\}$ generated by the asynchronous algorithm

$$J_\ell^{t+1}(x) = \begin{cases} T_t(J_1^{\tau_{\ell 1}(t)}, \ldots, J_m^{\tau_{\ell m}(t)})(x) & \text{if } t \in \mathcal{R}_\ell,\ x \in X_\ell, \\ J_\ell^t(x) & \text{if } t \notin \mathcal{R}_\ell,\ x \in X_\ell, \end{cases}$$

[cf. Eq. (2.61)] converges pointwise to J^*.

Solution: A straightforward adaptation of the proof of Prop. 2.6.1.

2.3 (Nonmonotonic Contractive Models – Fixed Points of Concave Sup-Norm Contractions [Ber16b])

The purpose of this exercise is to make a connection between our abstract DP model and the problem of finding the fixed point of a (not necessarily monotone) mapping that is a sup-norm contraction and has concave components. Let $T : \Re^n \mapsto \Re^n$ be a real-valued function whose n scalar components are concave. Then the components of T can be represented as

$$(TJ)(x) = \inf_{u \in U(x)} \big\{F(x,u) - J'u\big\}, \qquad x = 1, \ldots, n, \tag{2.81}$$

where $u \in \Re^n$, $J'u$ denotes the inner product of J and u, $F(x,\cdot)$ is the conjugate convex function of the convex function $-(TJ)(x)$, and $U(x) = \big\{u \in \Re^n \mid F(x,u) < \infty\big\}$ is the effective domain of $F(x,\cdot)$ (for the definition of these terms, we refer to books on convex analysis, such as [Roc70] and [Ber09]). Assuming that the infimum in Eq. (2.81) is attained for all x, show how the VI algorithm of Section 2.6.1 and the PI algorithm of Section 2.6.3 can be used to find the fixed point of T in the case where T is a sup-norm contraction, but not necessarily

monotone. *Note*: For algorithms that relate to the context of this exercise and are inspired by approximate PI, see [Ber16b], [Ber18c].

Solution: The analysis of Sections 2.6.1 and 2.6.3 does not require monotonicity of the mapping T_μ given by

$$(T_\mu J)(x) = F\big(x, \mu(x)\big) - J'\mu(x).$$

2.4 (Discounted Problems with Unbounded Cost per Stage)

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

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

$$G_x = \sup_{u \in U(x)} \big|g(x, u)\big|, \qquad x \in X,$$

belongs to $\mathcal{B}(X)$.

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

$$V_x = \sup_{u \in U(x)} \sum_{y \in X} p_{xy}(u)\, v(y), \qquad x \in X,$$

belongs to $\mathcal{B}(X)$.

(3) We have

$$\frac{\sum_{y \in X} p_{xy}(u) v(y)}{v(x)} \leq 1, \qquad \forall\ x \in X.$$

Consider the monotone mappings T_μ and T, given by

$$(T_\mu J)(x) = g\big(x, \mu(x)\big) + \alpha \sum_{y \in X} p_{xy}\big(\mu(x)\big) J(y), \qquad x \in X,$$

$$(TJ)(x) = \inf_{u \in U(x)} \left[g(x, u) + \alpha \sum_{y \in X} p_{xy}(u) J(y) \right], \qquad x \in X.$$

Show that T_μ and T map $\mathcal{B}(X)$ into $\mathcal{B}(X)$, and are contraction mappings with modulus α.

Solution: We have

$$\frac{\big|(T_\mu J)(x)\big|}{v(x)} \leq \frac{G_x}{v(x)} + \alpha \sum_{y \in X} \frac{p_{xy}\big(\mu(x)\big)\, v(y)}{v(x)} \frac{\big|J(y)\big|}{v(y)}, \qquad \forall\, x \in X,\, \mu \in \mathcal{M},$$

from which, using assumptions (1) and (2),

$$\frac{\big|(T_\mu J)(x)\big|}{v(x)} \leq \|G\| + \|V\|\, \|J\|, \qquad \forall\, x \in X,\, \mu \in \mathcal{M}.$$

A similar argument shows that

$$\frac{\big|(TJ)(x)\big|}{v(x)} \leq \|G\| + \|V\|\, \|J\|, \qquad \forall\, x \in X.$$

It follows that $T_\mu J \in \mathcal{B}(X)$ and $TJ \in \mathcal{B}(X)$ if $J \in \mathcal{B}(X)$.

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

$$\begin{aligned}
\|T_\mu J - T_\mu J'\| &= \sup_{x \in X} \frac{\Big|\alpha \sum_{y \in X} p_{xy}\big(\mu(x)\big)\big(J(x) - J'(x)\big)\Big|}{v(x)} \\
&\leq \sup_{x \in X} \frac{\Big|\alpha \sum_{y \in X} p_{xy}\big(\mu(x)\big) v(y) \big(|J(y) - J'(y)|/v(y)\big)\Big|}{v(x)} \\
&\leq \sup_{x \in X} \alpha\, \frac{\Big|\sum_{y \in X} p_{xy}\big(\mu(x)\big) v(y)\Big|}{v(x)} \|J - J'\| \\
&\leq \alpha \|J - J'\|,
\end{aligned}$$

where the last inequality follows from assumption (3). Hence T_μ is a contraction of modulus α.

To show that T is a contraction, we note that

$$\frac{(T_\mu J)(x)}{v(x)} \leq \frac{(T_\mu J')(x)}{v(x)} + \alpha \|J - J'\|, \qquad x \in X,\, \mu \in \mathcal{M},$$

so by taking infimum over $\mu \in \mathcal{M}$, we obtain

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

Similarly,

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

and by combining the last two relations the contraction property of T follows.

2.5 (Solution by Mathematical Programming)

Let the monotonicity and contraction Assumptions 2.1.1 and 2.1.2 hold. Show that if $J \leq TJ$ and $J \in \mathcal{B}(X)$, then $J \leq J^*$. Use this fact to show that if $X = \{1, \ldots, n\}$ and $U(i)$ is finite for each $i = 1, \ldots, n$, then $J^*(1), \ldots, J^*(n)$ solves the following problem (in $z_1, \ldots, z_n$):

$$\begin{aligned} &\text{maximize} \quad \sum_{i=1}^{n} z_i \\ &\text{subject to} \quad z_i \leq H(i, u, z), \qquad i = 1, \ldots, n, \quad u \in U(i), \end{aligned}$$

where $z = (z_1, \ldots, z_n)$. *Note*: This is a linear or nonlinear program (depending on whether H is linear in J or not) with n variables and as many as $n \times m$ constraints, where m is the maximum number of elements in the sets $U(i)$.

Solution: If $J \leq TJ$, by monotonicity we have $J \leq \lim_{k\to\infty} T^k J = J^*$. Any feasible solution z of the given optimization problem satisfies $z_i \leq H(i, u, z)$ for all $i = 1, \ldots, n$ and $u \in U(i)$, so that $z \leq Tz$. It follows that $z \leq J^*$, which implies that J^* solves the optimization problem.

2.6 (Conditions for Convergence of PI with Cost Function Approximation [Ber11c])

Let the monotonicity and contraction Assumptions 2.1.1 and 2.1.2 hold, and assume that there are n states, and that $U(x)$ is finite for every x. Consider a PI method that aims to approximate a fixed point of T on a subspace $S = \{\Phi r \mid r \in \Re^s\}$, where Φ is an $n \times s$ matrix, and evaluates a policy $\mu \in \mathcal{M}$ with a solution $\tilde{J}_\mu$ of the following fixed point equation in the vector $J \in \Re^n$:

$$J = WT_\mu J \tag{2.82}$$

where $W : \Re^n \mapsto \Re^n$ is some mapping (possibly nonlinear, but independent of μ), whose range is contained in S. Examples where W is linear include policy evalution using the projected and aggregation equations; see Section 1.2.4. The algorithm is given by

$$\Phi r_k = WT_{\mu^k}(\Phi r_k), \qquad T_{\mu^{k+1}}(\Phi r_k) = T(\Phi r_k); \tag{2.83}$$

[cf. Eqs. (2.79)-(2.80)]. We assume the following:

(1) For each $J \in \Re^n$, there exists $\mu \in \mathcal{M}$ such that $T_\mu J = TJ$.

(2) For each $\mu \in \mathcal{M}$, Eq. (2.82) has a unique solution that belongs to S and is denoted $\tilde{J}_\mu$. Moreover, for all J such that $WT_\mu J \leq J$, we have

$$\tilde{J}_\mu = \lim_{k\to\infty} (WT_\mu)^k J.$$

(3) For each $\mu \in \mathcal{M}$, the mappings W and WT_μ are monotone in the sense that

$$WJ \leq WJ', \quad WT_\mu J \leq WT_\mu J', \qquad \forall\ J, J' \in \Re^n \text{ with } J \leq J'. \tag{2.84}$$

Note that conditions (1) and (2) guarantee that the iterations (2.83) are well-defined. Assume that the method is initiated with some policy in $\mathcal{M}$, and it is operated so that it terminates when a policy $\overline{\mu}$ is obtained such that $T_{\overline{\mu}}\tilde{J}_{\overline{\mu}} = T\tilde{J}_{\overline{\mu}}$. Show that the method terminates in a finite number of iterations, and the vector $\tilde{J}_{\overline{\mu}}$ obtained upon termination is a fixed point of WT. *Note*: Condition (2) is satisfied if WT_μ is a contraction, while condition (b) is satisfied if W is a matrix with nonnegative components and T_μ is monotone for all μ. For counterexamples to convergence when the conditions (2) and/or (3) are not satisfied, see [BeT96], Section 6.4.2, and [Ber12a], Section 2.4.3.

Solution: Similar to the standard proof of convergence of (exact) PI, we use the policy improvement equation $T_{\overline{\mu}}\tilde{J}_\mu = T\tilde{J}_\mu$, the monotonicity of W, and the policy evaluation equation to write

$$WT_{\overline{\mu}}\tilde{J}_\mu = WT\tilde{J}_\mu \leq WT_\mu\tilde{J}_\mu = \tilde{J}_\mu.$$

By iterating with the monotone mapping $WT_{\overline{\mu}}$ and by using condition (2), we obtain

$$\tilde{J}_{\overline{\mu}} = \lim_{k\to\infty}(WT_{\overline{\mu}})^k\tilde{J}_\mu \leq \tilde{J}_\mu.$$

There are finitely many policies, so we must have $\tilde{J}_{\overline{\mu}} = \tilde{J}_\mu$ after a finite number of iterations, which using the policy improvement equation $T_{\overline{\mu}}\tilde{J}_\mu = T\tilde{J}_\mu$, implies that $T_{\overline{\mu}}\tilde{J}_{\overline{\mu}} = T\tilde{J}_{\overline{\mu}}$. Thus the algorithm terminates with $\overline{\mu}$, and since $\tilde{J}_{\overline{\mu}} = WT_{\overline{\mu}}\tilde{J}_{\overline{\mu}}$, it follows that $\tilde{J}_{\overline{\mu}}$ is a fixed point of WT.

3

Semicontractive Models

Contents

We will now consider abstract DP models that are intermediate between the contractive models of Chapter 2, where all stationary policies involve a contraction mapping, and noncontractive models to be discussed in Chapter 4, where there are no contraction-like assumptions (although there are some compensating conditions, including monotonicity).

A representative instance of such an intermediate model is the deterministic shortest path problem of Example 1.2.7, where we can distinguish between two types of stationary policies: those that terminate at the destination from every starting node, and those that do not. A more general instance is the stochastic shortest path (SSP for short) problem of Example 1.2.6. In this problem, the analysis revolves around two types of stationary policies μ: those with a mapping T_μ that is a contraction with respect to some norm, and those with a mapping T_μ that is not a contraction with respect to any norm (it can be shown that the former are the ones that terminate with probability 1 starting from any state).

In the models of this chapter, like in SSP problems, we divide policies into two groups, one of which has favorable characteristics. We loosely refer to such models as *semicontractive* to indicate that these favorable characteristics include contraction-like properties of the mapping T_μ. To develop a more broadly applicable theory, we replace the notion of contractiveness of T_μ with a notion of *S-regularity of μ*, where S is an appropriate set of functions of the state (roughly, this is a form of "local stability" of T_μ, which ensures that the cost function J_μ is the unique fixed point of T_μ within S, and that $T_\mu^k J$ converges to J_μ regardless of the choice of J from within S). We allow that some policies are S-regular while others are not.

Note that the term "semicontractive" is not used in a precise mathematical sense here. Rather it refers qualitatively to a collection of models where some policies have a regularity/contraction-like property but others do not. Moreover, regularity is a relative property: the division of policies into "regular" and "irregular" depends on the choice of the set S. On the other hand, typically in practical applications an appropriate choice of S is fairly evident.

Our analysis will involve two types of assumptions:

(a) *Favorable assumptions*, under which we obtain results that are nearly as strong as those available for the contractive models of Chapter 2. In particular, we show that J^* is a fixed point of T, that the Bellman equation $J = TJ$ has a unique solution, at least within a suitable class of functions, and that variants of the VI and PI algorithms are valid. Some of the VI and PI approaches are suitable for distributed asynchronous computation, similar to their Section 2.6 counterparts for contractive models.

(a) *Less favorable assumptions*, under which serious difficulties may occur: J^* may not be a fixed point of T, and even when it is, it may not be found using the VI and PI algorithms. These anomalies may ap-

pear in simple problems, such as deterministic and stochastic shortest path problems with some zero length cycles. To address the difficulties, we will consider *a restricted problem, where the only admissible policies are the ones that are S-regular*. Under reasonable conditions we show that this problem is better-behaved. In particular, J_S^*, the optimal cost function over the S-regular policies only, is the unique solution of Bellman's equation among functions $J \in S$ with $J \geq J_S^*$, while VI converges to J_S^* starting from any $J \in S$ with $J \geq J_S^*$. We will also derive a variety of PI approaches for finding J_S^* and an S-regular policy that is optimal within the class of S-regular policies.

We will illustrate our analysis in Section 3.5, both under favorable and unfavorable assumptions, by means of four classes of practical problems. Some of these problems relate to finding a path to a destination in a graph under stochastic or set membership uncertainty, while others relate to the control of a continuous-state system to a terminal state. In particular, we will consider SSP problems, affine monotonic problems, including problems with multiplicative or risk-sensitive exponential cost function, minimax-type shortest path problems, and continuous-state deterministic problems with nonnegative cost, such as linear-quadratic problems.

The chapter is organized as follows. In Section 3.1, we illustrate the pathologies regarding solutions of Bellman's equation, and the VI and PI algorithms. To this end, we use four simple examples, ranging from finite-state shortest path problems, to continuous-state linear-quadratic problems. These examples provide orientation and motivation for S-regular policies later. In Section 3.2, we formally introduce our abstract DP model, and the notion of an S-regular policy. We then develop some of the basic associated results relating to Bellman's equation, and the convergence of VI and PI, based primarily on the ideas underlying the PI algorithm. In Section 3.3 we refine the results of Section 3.2 under favorable conditions, obtaining results and algorithms that are almost as powerful as the ones for contractive models. In Section 3.4 we develop a complementary analytical approach, which is based on the use of perturbations and applies under less favorable assumptions. In Section 3.5, we discuss in detail the application and refinement of the results of Sections 3.2-3.4 in some important shortest path-type practical contexts. In Section 3.6, we focus on variants of VI and PI-type algorithms for semicontractive DP models, including some that are suitable for asynchronous distributed computation.

3.1 PATHOLOGIES OF NONCONTRACTIVE DP MODELS

In this section we provide a general overview of the analytical and computational difficulties in noncontractive DP models, using for the most part shortest path-type problems. For illustration we will first use two of the simplest and most widely encountered finite-state DP problems: deter-

ministic and SSP problems, whereby we are aiming to reach a destination state at minimum cost.† We will also discuss an example of continuous-state shortest path problem that involves a linear system and a quadratic cost function.

We will adopt the general abstract DP model of Section 1.2. We give a brief description that is adequate for the purposes of this section, and defer a more formal definition to Section 3.2. In particular, we introduce a set of states X, and for each $x \in X$, the nonempty control constraint set $U(x)$. For each policy μ, the mapping T_μ is given by

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

where H is a suitable function of (x, u, J). The mapping T is given by

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

The cost function of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ is

$$J_\pi(x) = \limsup_{N \to \infty} J_{\pi,N}(x) = \limsup_{N \to \infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}} \bar{J})(x), \quad x \in X,$$

where $\bar{J}$ is some function.‡ We want to minimize J_π over π, i.e., to find

$$J^*(x) = \inf_\pi J_\pi(x), \qquad x \in X,$$

and a policy that attains the infimum.

For orientation purposes, we recall from Chapter 1 (Examples 1.2.1 and 1.2.2) that for a stochastic optimal control problem involving a finite-state Markov chain with state space $X = \{1, \ldots, n\}$, transition probabilities $p_{xy}(u)$, and expected one-stage cost function g, the mapping H is given by

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

and $\bar{J}(x) \equiv 0$. The SSP problem arises when there is an additional termination state that is cost-free, and corresponding transition probabilities $p_{xt}(u)$, $x \in X$.

---

† These problems are naturally undiscounted, and cannot be readily addressed by introducing a discount factor close to 1, because then the optimal policies may exhibit undesirable behavior. In particular, in the presence of discounting, they may involve moving initially along a small-length cycle in order to postpone the use of an optimal but unavoidably costly path until later, when the discount factor will reduce substantially the cost of that path.

‡ In the contractive models of Chapter 2, the choice of $\bar{J}$ is immaterial, as we discussed in Section 2.1. Here, however, the choice of $\bar{J}$ is important, and affects important characteristics of the model, as we will see later.

A more general undiscounted stochastic optimal control problem involves a stationary discrete-time dynamic system where the state is an element of a space X, and the control is an element of a space U. The control u_k is constrained to take values in a given set $U(x_k) \subset U$, which depends on the current state x_k [$u_k \in U(x_k)$, for all $x_k \in X$]. For a policy $\pi = \{\mu_0, \mu_1, \ldots\}$, the state evolves according to a system equation

$$x_{k+1} = f\big(x_k, \mu_k(x_k), w_k\big), \qquad k = 0, 1, \ldots, \tag{3.1}$$

where w_k is a random disturbance that takes values from a space W. We assume that w_k, $k = 0, 1, \ldots$, are characterized by probability distributions $P(\cdot \mid x_k, u_k)$ that are identical for all k, where $P(w_k \mid x_k, u_k)$ is the probability of occurrence of w_k, when the current state and control are x_k and u_k, respectively. Here, we allow infinite state and control spaces, as well as problems with discrete (finite or countable) state space (in which case the underlying system is a Markov chain). However, for technical reasons that relate to measure-theoretic issues, we assume that W is a countable set. †

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

$$J_\pi(x_0) = \limsup_{k\to\infty} E\left\{\sum_{t=0}^{k} g\big(x_t, \mu_t(x_t), w_t\big)\right\}, \tag{3.2}$$

subject to the system equation constraint (3.1), where g is the one-stage cost function. The corresponding mapping of the abstract DP problem is

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

and $\bar{J}(x) \equiv 0$. Again here, $(T_{\mu_0} \cdots T_{\mu_k} \bar{J})(x)$ is the expected cost of the first $k+1$ periods using π starting from x, and with terminal cost 0.

A discounted version of the problem is defined by the mapping

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

where $\alpha \in (0, 1)$ is the discount factor. It corresponds to minimization of

$$J_\pi(x_0) = \limsup_{k\to\infty} E\left\{\sum_{t=0}^{k} \alpha^t g\big(x_t, \mu_t(x_t), w_t\big)\right\}.$$

If the cost per stage g is bounded, then a problem that fits the contractive framework of Chapter 2 is obtained, and can be analyzed using the methods of that chapter. However, there are interesting infinite-state discounted optimal control problems where g is not bounded.

† Measure-theoretic issues are not addressed at all in this third edition of the book. The first edition addressed some of these issues within an abstract DP context in its Chapter 5 and Appendix C (this material is posted at the book's web site); see also the monograph by Bertsekas and Shreve [BeS78], and the paper by Yu and Bertsekas [YuB15]. An orientation summary is given in Appendix A of the author's textbook [Ber12a].

A Summary of Pathologies

The four examples to be discussed in Sections 3.1.1-3.1.4 are special cases of deterministic and stochastic optimal control problems of the type just described. In each of these examples, we will introduce a subclass of "well-behaved" policies and a restricted optimization problem, which is to minimize the cost over the "well-behaved" subclass (in Section 3.2 the property of being "well-behaved" will be formalized through the notion of S-regularity). The optimal cost function over just the "well-behaved" policies is denoted $\hat{J}$ (we will also use the notation J_S^* later). Here is a summary of the examples and the pathologies that they reveal:

(a) *A finite-state, finite-control deterministic shortest path problem (Section 3.1.1).* Here the mapping T can have infinitely many fixed points, including J^* and $\hat{J}$. There exist policies that attain the optimal costs J^* and $\hat{J}$. Depending on the starting point, the VI algorithm may converge to J^* or to $\hat{J}$ or to a third fixed point of T (for cases where $J^* \neq \hat{J}$, VI converges to $\hat{J}$ starting from any $J \geq \hat{J}$). The PI algorithm can oscillate between two policies that attain J^* and $\hat{J}$, respectively.

(b) *A finite-state, finite-control stochastic shortest path problem (Section 3.1.2).* The salient feature of this example is that J^* is not a fixed point of the mapping T. By contrast $\hat{J}$ is a fixed point of T. The VI algorithm converges to $\hat{J}$ starting from any $J \geq \hat{J}$, while it does not converge otherwise.

(c) *A finite-state, continuous-control stochastic shortest path problem (Section 3.1.3).* We give three variants of this example. In the first variant (a classical problem known as the "blackmailer's dilemma"), all the policies are "well-behaved," so $J^* = \hat{J}$, and VI converges to J^* starting from any real-valued initial condition, while PI also succeeds in finding J^* as the limit of the generated sequence $\{J_{\mu^k}\}$. However, PI cannot find an optimal policy, because there is no optimal stationary policy. In a second variant of this example, PI generates a sequence of "well-behaved" policies $\{\mu^k\}$ such that $J_{\mu^k} \downarrow \hat{J}$, but $\{\mu^k\}$ converges to a policy that is either infeasible or is strictly suboptimal. In the third variant of this example, the problem data can strongly affect the multiplicity of the fixed points of T, and the behavior of the VI and PI algorithms.

(d) *A continuous-state, continuous-control deterministic linear-quadratic problem (Section 3.1.4).* Here the mapping T has exactly two fixed points, J^* and $\hat{J}$, within the class of positive semidefinite quadratic functions. The VI algorithm converges to $\hat{J}$ starting from all positive initial conditions, and to J^* starting from all other initial conditions. Moreover, starting with a "well-behaved" policy (one that is stable),

the PI algorithm converges to $\hat{J}$ and to an optimal policy within the class of "well-behaved" (stable) policies.

It can be seen that the examples exhibit wide-ranging pathological behavior. In Section 3.2, we will aim to construct a theoretical framework that explains this behavior. Moreover, in Section 3.3, we will derive conditions guaranteeing that much of this type of behavior does not occur. These conditions are natural and broadly applicable. They are used to exclude from optimality the policies that are not "well-behaved," and to obtain results that are nearly as powerful as their counterparts for the contractive models of Chapter 2.

3.1.1 Deterministic Shortest Path Problems

Let us consider the classical deterministic shortest path problem, discussed in Example 1.2.7. Here, we have a graph of n nodes $x = 1, \ldots, n$, plus a destination t, and an arc length a_{xy} for each directed arc (x, y). The objective is to find for each x a directed path that starts at x, ends at t, and has minimum length (the length of a path is defined as the sum of the lengths of its arcs). A standard assumption, which we will adopt here, is that every node x is connected to the destination, i.e., there exists a path from every x to t.

To formulate this shortest path problem as a DP problem, we embed it within a "larger" problem, whereby we view all paths as admissible, including those that do not terminate at t. We also view t as a cost-free and absorbing node. Of course, we need to deal with the presence of policies that do not terminate, and the most common way to do this is to assume that all cycles have strictly positive length, in which case policies that do not terminate cannot be optimal. However, it is not uncommon to encounter shortest path problems with zero length cycles, and even negative length cycles. Thus we will not impose any assumption on the sign of the cycle lengths, particularly since we aim to use the shortest path problem to illustrate behavior that arises in a broader undiscounted/noncontractive DP setting.

As noted in Section 1.2, we can formulate the problem in terms of an abstract DP model where the states are the nodes $x = 1, \ldots, n$, and the controls available at x can be identified with the outgoing neighbors of x [the nodes u such that (x, u) is an arc]. The mapping H that defines the corresponding abstract DP problem is

$$H(x, u, J) = \begin{cases} a_{xu} + J(u) & \text{if } u \neq t, \\ a_{xt} & \text{if } u = t, \end{cases} \qquad x = 1, \ldots, n.$$

A stationary policy μ defines the subgraph whose arcs are $\big(x, \mu(x)\big)$, $x = 1, \ldots, n$. We say that μ is *proper* if this graph is acyclic, i.e., it consists of a tree of paths leading from each node to the destination. If μ is not

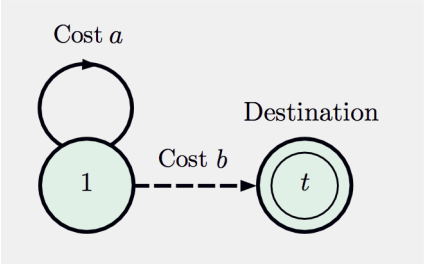


Figure 3.1.1. A deterministic shortest path problem with a single node 1 and a termination node t. At 1 there are two choices; a self-transition, which costs a, and a transition to t, which costs b.

proper, it is called *improper*. Thus there exists a proper policy if and only if each node is connected to t with a path. Furthermore, an improper policy has cost greater than $-\infty$ starting from every initial state if and only if all the cycles of the corresponding subgraph have nonnegative cycle cost.

Let us now get a sense of what may happen by considering the simple one-node example shown in Fig. 3.1.1. Here there is a single state 1 in addition to the termination state t. At state 1 there are two choices: a self-transition, which costs a, and a transition to t, which costs b. The mapping H, abbreviating $J(1)$ with just the scalar J, is

$$H(1,u,J) = \begin{cases} a+J & \text{if } u\text{: self transition,} \\ b & \text{if } u\text{: transition to } t, \end{cases} \qquad J \in \Re.$$

There are two policies here: the policy μ that transitions from 1 to t, which is proper, and the policy μ' that self-transitions at state 1, which is improper. We have

$$T_\mu J = b, \quad T_{\mu'} J = a + J, \qquad J \in \Re,$$

and

$$TJ = \min\{b,\, a+J\}, \qquad J \in \Re.$$

Note that for the proper policy μ, the mapping $T_\mu : \Re \mapsto \Re$ is a contraction. For the improper policy μ', the mapping $T_{\mu'} : \Re \mapsto \Re$ is not a contraction, and it has a fixed point within $\Re$ only if $a = 0$, in which case every $J \in \Re$ is a fixed point.

We now consider the optimal cost J^*, the fixed points of T within $\Re$, and the behavior of the VI and PI methods for different combinations of values of a and b.

(a) If $a > 0$, the optimal cost, $J^* = b$, is the unique fixed point of T, and the proper policy is optimal.

(b) If $a = 0$, the set of fixed points of T (within $\Re$) is the interval $(-\infty, b]$. Here the improper policy is optimal if $b \geq 0$, and the proper policy is optimal if $b \leq 0$ (both policies are optimal if $b = 0$).

(c) If $a = 0$ and $b > 0$, the proper policy is strictly suboptimal, yet its cost at state 1 (which is b) is a fixed point of T. The optimal cost, $J^* = 0$, lies in the interior of the set of fixed points of T, which is $(-\infty, b]$. Thus the VI method that generates $\{T^k J\}$ starting with $J \neq J^*$ cannot find J^*. In particular if J is a fixed point of T, VI stops at J, while if J is not a fixed point of T (i.e., $J > b$), VI terminates in two iterations at $b \neq J^*$. Moreover, the standard PI method is unreliable in the sense that starting with the suboptimal proper policy μ, it may stop with that policy because $T_\mu J_\mu = b = \min\{b, J_\mu\} = TJ_\mu$ (the improper/optimal policy μ' also satisfies $T_{\mu'} J_\mu = TJ_\mu$, so a rule for breaking the tie in favor of μ is needed but such a rule may not be obvious in general).

(d) If $a = 0$ and $b < 0$, the improper policy is strictly suboptimal, and we have $J^* = b$. Here it can be seen that the VI sequence $\{T^k J\}$ converges to J^* for all $J \geq b$, but stops at J for all $J < b$, since the set of fixed points of T is $(-\infty, b]$. Moreover, starting with either the proper policy or the improper policy, the standard form of PI may oscillate, since $T_\mu J_{\mu'} = TJ_{\mu'}$ and $T_{\mu'} J_\mu = TJ_\mu$, as can be easily verified [the optimal policy μ also satisfies $T_\mu J_\mu = TJ_\mu$ but it is not clear how to break the tie; compare also with case (c) above].

(e) If $a < 0$, the improper policy is optimal and we have $J^* = -\infty$. There are no fixed points of T within $\Re$, but J^* is the unique fixed point of T within the set $[-\infty, \infty]$. The VI method will converge to J^* starting from any $J \in [-\infty, \infty]$. The PI method will also converge to the optimal policy starting from either policy.

3.1.2 Stochastic Shortest Path Problems

We consider the SSP problem, which was described in Example 1.2.6 and will be revisited in Section 3.5.1. Here a policy is associated with a stationary Markov chain whose states are $1, \ldots, n$, plus the cost-free termination state t. The cost of a policy starting at a state x is the sum of the expected cost of its transitions up to reaching t. A policy is said to be *proper*, if in its Markov chain, every state is connected with t with a path of positive probability transitions, and otherwise it is called *improper*. Equivalently, a policy is proper if its Markov chain has t as its unique ergodic state, with all other states being transient.

In deterministic shortest path problems, it turns out that J_μ is always a fixed point of T_μ, and J^* is always a fixed point of T. This is a generic feature of deterministic problems, which was illustrated in Section 1.1 (see Exercise 3.1 for a rigorous proof). However, in SSP problems where the

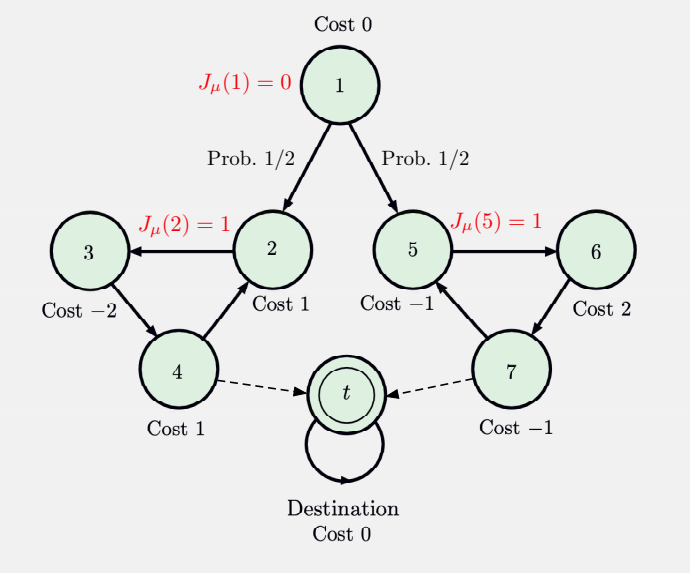


Figure 3.1.2. An example of an improper policy μ, where J_μ is not a fixed point of T_μ. All transitions under μ are shown with solid lines. These transitions are deterministic, except at state 1 where the next state is 2 or 5 with equal probability 1/2. There are additional high cost transitions from nodes 1, 4, and 7 to the destination (shown with broken lines), which create a suboptimal proper policy. We have $J^* = J_\mu$ and J^* is not a fixed point of T.

cost per stage can take both positive and negative values this need not be so, as we will now show with an example due to [BeY16].

Let us consider the problem of Fig. 3.1.2. It involves an improper policy μ, whose transitions are shown with solid lines in the figure, and form the two zero length cycles shown. All the transitions under μ are deterministic, except at state 1 where the successor state is 2 or 5 with equal probability 1/2. The problem has been deliberately constructed so that corresponding costs at the nodes of the two cycles are negatives of each other. As a result, the expected cost at each time period starting from state 1 is 0, implying that the total cost over any number or even infinite number of periods is 0.

Indeed, to verify that $J_\mu(1) = 0$, let c_k denote the cost incurred at time k, starting at state 1, and let $s_N(1) = \sum_{k=0}^{N-1} c_k$ denote the N-step accumulation of c_k starting from state 1. We have

$$s_N(1) = 0 \qquad \text{if } N = 1 \text{ or } N = 4 + 3t,\ t = 0, 1, \ldots,$$

$$s_N(1) = 1 \text{ or } s_N(1) = -1 \text{ with probability } 1/2 \text{ each} \qquad \text{if } N = 2 + 3t \text{ or } N = 3 + 3t,\ t = 0, 1, \ldots.$$

Thus $E\{s_N(1)\} = 0$ for all N, and

$$J_\mu(1) = \limsup_{N\to\infty} E\{s_N(1)\} = 0.$$

On the other hand, using the definition of J_μ in terms of lim sup, we have

$$J_\mu(2) = J_\mu(5) = 1,$$

(the sequence of N-stage costs undergoes a cycle $\{1, -1, 0, 1, -1, 0, \ldots\}$ when starting from state 2, and undergoes a cycle $\{-1, 1, 0, -1, 1, 0, \ldots\}$ when starting from state 5). Thus the Bellman equation at state 1,

$$J_\mu(1) = \tfrac{1}{2}\big(J_\mu(2) + J_\mu(5)\big),$$

is not satisfied, and J_μ is not a fixed point of T_μ.

The mathematical reason why Bellman's equation $J_\mu = T_\mu J_\mu$ may not hold for stochastic problems is that lim sup may not commute with the expected value operation that is inherent in T_μ, and the proof argument given for deterministic problems in Section 1.1 breaks down. We can also modify this example so that there are multiple policies. To this end, we can add for $i = 1, 4, 7$, another control that leads from i to t with a cost $c > 1$ (cf. the broken line arcs in Fig. 3.1.2). Then we create a proper policy that is strictly suboptimal, while not affecting J^*, which again is not a fixed point of T.

Let us finally note an anomaly around randomized policies in noncontractive models. The improper policy shown in Fig. 3.1.2 may be viewed as a randomized policy for a deterministic shortest path problem: this is the problem for which at state 1 we must (deterministically) choose one of the two successor states 2 and 5. For this deterministic problem, J^* takes the same values as before for all $i \neq 1$, but it takes the value $J^*(1) = 1$ rather than $J^*(1) = 0$. Thus, remarkably, once we allow randomized policies into the problem, the optimal cost function ceases to be a solution of Bellman's equation and simultaneously the optimal cost at state 1 is improved!

In subsequent sections we will see that favorable results hold in SSP problems where the restricted optimal cost function over just the proper policies is equal to the overall optimal J^*. This can be guaranteed by assumptions that essentially imply that improper polices cannot be optimal (see Sections 3.3 and 3.5.1). We will then see that not only is J^* a fixed point of T, but it is also the unique fixed point (within the class of real-valued functions), and that the VI and PI algorithms yield J^* and an optimal proper policy in the limit.

3.1.3 The Blackmailer's Dilemma

This is a classical example involving a profit maximizing blackmailer. We formulate it as an SSP problem involving cost minimization, with a single state $x = 1$, in addition to the termination state t.

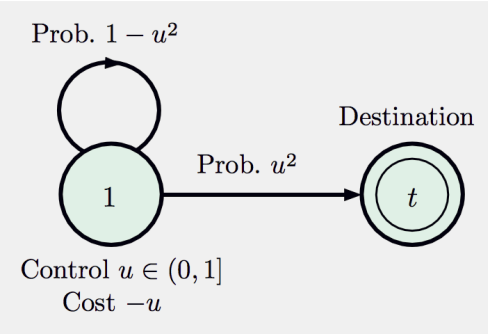


Figure 3.1.3. Transition diagram for the first variant of the blackmailer problem. At state 1, the blackmailer may demand any amount $u \in (0, 1]$. The victim will comply with probability $1 - u^2$ and will not comply with probability u^2, in which case the process will terminate.

In a first variant of the problem, at state 1, we can choose a control $u \in (0, 1]$, while incurring a cost $-u$; we then move to state t with probability u^2, and stay in state 1 with probability $1-u^2$; see Fig. 3.1.3. We may regard u as a demand made by the blackmailer, and state 1 as the situation where the victim complies. State t is arrived at when the victim (permanently) refuses to yield to the blackmailer's demand. The problem then can be viewed as one where the blackmailer tries to maximize his expected total gain by balancing his desire for increased demands (large u) with keeping his victim compliant (small u).

For notational simplicity, let us abbreviate $J(1)$ and $\mu(1)$ with just the scalars J and μ, respectively. Then in terms of abstract DP we have

$$X = \{1\}, \qquad U = (0, 1], \qquad \bar{J} = 0, \qquad H(1, u, J) = -u + (1 - u^2)J,$$

and for every stationary policy μ, we have

$$T_\mu J = -\mu + (1 - \mu^2)J. \tag{3.3}$$

Clearly T_μ, viewed as a mapping from $\Re$ to $\Re$, is a contraction with modulus $1 - \mu^2$, and its unique fixed point within $\Re$, J_μ, is the solution of

$$J_\mu = T_\mu J_\mu = -\mu + (1 - \mu^2)J_\mu,$$

which yields

$$J_\mu = -\frac{1}{\mu}.$$

Here all policies are proper in the sense that they lead asymptotically to t with probability 1, and the infimum of J_μ over μ is $-\infty$, implying also

that $J^* = -\infty$. However, there is no optimal stationary policy within the class of proper policies. †

Another interesting fact about this problem is that T_μ is a contraction for all μ. However the theory of contractive models does not apply because there is no uniform modulus of contraction ($\alpha < 1$) that applies simultaneously to all $\mu \in (0, 1]$ [cf. Eq. (3.3)]. As a result, the contraction Assumption 2.1.2 of Section 2.1 does not hold.

Let us now consider Bellman's equation. The mapping T is given by

$$TJ = \inf_{0<u\leq 1} \big\{ -u + (1-u^2)J \big\},$$

and Bellman's equation is written as

$$J = J - \sup_{0<u\leq 1} \{u + u^2 J\}.$$

It can be verified that this equation has no real-valued solution. However, $J^* = -\infty$ is a solution within the set of extended real numbers $[-\infty, \infty]$. Moreover the VI method will converge to J^* starting from any $J \in [-\infty, \infty)$. The PI method, starting from any policy μ^0, will produce the ever improving sequence of policies $\{\mu^k\}$ with $\mu^{k+1} = \mu^k/2$ and $J_{\mu^k} \downarrow J^*$, while μ^k will converge to 0, which is not a feasible policy.

A Second Problem Variant

Consider next a variant of the problem where at state 1, we terminate at no cost with probability u, and stay in state 1 at a cost $-u$ with probability $1 - u$. The control constraint is still $u \in (0, 1]$.

Here we have

$$H(1, u, J) = (1-u)(-u) + (1-u)J.$$

It can be seen that for every policy μ, T_μ is again a contraction and we have $J_\mu = \mu - 1$. Thus $J^* = -1$, but again there is no optimal policy, stationary or not. Moreover, T has multiple fixed points: its set of fixed points within $\Re$ is $\{J \mid J \leq -1\}$. Here the VI method will converge to J^* starting from any $J \in [-1, \infty)$. The PI method will produce an ever improving sequence of policies $\{\mu^k\}$ with $J_{\mu^k} \downarrow J^*$, starting from any policy μ^0, while μ^k will converge to 0, which is not a feasible policy.

† An unusual fact about this problem is that there exists a *nonstationary* policy π^* that is optimal in the sense that $J_{\pi^*} = J^* = -\infty$ (for a proof see [Ber12a], Section 3.2). The underlying intuition is that when the amount demanded u is decreased toward 0, the probability of noncompliance, u^2, decreases much faster. This fact, however, will not be significant in the context of our analysis.

A Third Problem Variant

Finally, let us again assume that

$$H(1,u,J) = (1-u)(-u) + (1-u)J, \qquad \forall\ u \in (0,1],$$

but also allow, in addition to $u \in (0,1]$, the choice $u = 0$ that self-transitions to state 1 at a cost c (this is the choice where the blackmailer can forego blackmail for a single period in exchange for a fixed payment $-c$). Here there is the extra (improper) policy μ' that chooses $\mu'(1) = 0$. We have

$$T_{\mu'}J = c + J,$$

and the mapping T is given by

$$TJ = \min\left\{c + J,\ \inf_{0<u\leq 1}\left\{-u + u^2 + (1-u)J\right\}\right\}. \tag{3.4}$$

Let us consider the optimal policies and the fixed points of T in the two cases where $c \geq 0$ and $c < 0$.

When $c \geq 0$, we have $J^* = -1$, while $J_{\mu'} = \infty$ (if $c > 0$) or $J_{\mu'} = 0$ (if $c = 0$). It can be seen that there is no optimal policy, and that all $J \in (-\infty, -1]$ are fixed points of T, including J^*. Here the VI method will converge to J^* starting from any $J \in [-1, \infty)$. The PI method will produce an ever improving sequence of policies $\{\mu^k\}$, with $J_{\mu^k} \downarrow J^*$. However, μ^k will converge to 0, which is a feasible but strictly suboptimal policy.

When $c < 0$, we have $J_{\mu'} = -\infty$, and the improper policy μ' is optimal. Here the optimal cost over just the proper policies is $\hat{J} = -1$, while $J^* = -\infty$. Moreover $\hat{J}$ is not a fixed point of T, and in fact T has no real-valued fixed points, although J^* is a fixed point. It can be verified that the VI algorithm will converge to J^* starting from any scalar J. Furthermore, starting with a proper policy, the PI method will produce the optimal (improper) policy within a finite number of iterations.

3.1.4 Linear-Quadratic Problems

One of the most important optimal control problems involves a linear system and a cost per stage that is positive semidefinite quadratic in the state and the control. The objective here is roughly to bring the system at or close to the origin, which can be viewed as a cost-free and absorbing state. Thus the problem has a shortest path character, even though the state space is continuous.

Under reasonable assumptions (involving the notions of system controllability and observability; see e.g., [Ber17a], Section 3.1), the problem admits a favorable analysis and an elegant solution: the optimal cost function is positive semidefinite quadratic and the optimal policy is a linear

function of the state. Moreover, Bellman's equation can be equivalently written as an algebraic Riccati equation, which admits a unique solution within the class of nonnegative cost functions.

On the other hand, the favorable results just noted depend on the assumptions and the structure of the linear-quadratic problem. There is no corresponding analysis for more general deterministic continuous-state optimal control problems. Moreover, even for linear-quadratic problems, when the aforementioned controllability and observability assumptions do not hold, the favorable results break down and pathological behavior can occur. This suggests analytical difficulties in more general continuous-state contexts, which we will discuss later in Section 3.5.5.

To illustrate what can happen, consider the scalar system

$$x_{k+1} = \gamma x_k + u_k, \qquad x_k \in \Re,\ u_k \in \Re,$$

with $X = U(x) = \Re$, and a cost per stage equal to u^2. Here we have $J^*(x) = 0$ for all $x \in \Re$, while the policy that applies control $u = 0$ at every state x is optimal. This is reminiscent of the deterministic shortest path problem of Section 3.1.1, for the case where $a = 0$ and there is a zero length cycle. Bellman's equation has the form

$$J(x) = \min_{u \in \Re} \big\{u^2 + J(\gamma x + u)\big\}, \qquad x \in \Re,$$

and it is seen that J^* is a solution. We will now show that there is another solution, which has an interesting interpretation.

Let us assume that $\gamma > 1$ so the system is unstable (the instability of the system is important for the purpose of this example). It is well-known that for linear-quadratic problems the class of quadratic cost functions,

$$S = \big\{J \mid J(x) = px^2,\ p \geq 0\big\},$$

plays a special role. Linear policies of the form

$$\mu(x) = rx,$$

where r is a scalar, also play a special role, particularly the subclass $\mathcal{L}$ of linear policies that are *stable*, in the sense that the closed-loop system

$$x_{k+1} = (\gamma + r)x_k$$

is stable, i.e., $|\gamma + r| < 1$. For such a policy, the generated system trajectory $\{x_k\}$, starting from an initial state x_0, is $\big\{(\gamma+r)^k x_0\big\}$, and the corresponding cost function is quadratic as shown by the following calculation,

$$J_\mu(x_0) = \sum_{k=0}^{\infty} \big(\mu(x_k)\big)^2 = \sum_{k=0}^{\infty} r^2 x_k^2 = \sum_{k=0}^{\infty} r^2(\gamma + r)^{2k} x_0^2 = \frac{r^2}{1 - (\gamma + r)^2} x_0^2. \tag{3.5}$$

Note that there is no policy in $\mathcal{L}$ that is optimal, since the optimal policy $\mu^*(x) \equiv 0$ is unstable and does not belong to $\mathcal{L}$.

Let us consider fixed points of the mapping T,

$$(TJ)(x) = \inf_{u \in \Re} \big\{u^2 + J(\gamma x + u)\big\},$$

within the class of nonnegative quadratic functions S. For $J(x) = px^2$ with $p \geq 0$, we have

$$(TJ)(x) = \inf_{u \in \Re} \big\{u^2 + p(\gamma x + u)^2\big\},$$

and by setting to 0 the derivative with respect to u, we see that the infimum is attained at

$$u^* = -\frac{p\gamma}{1+p}x.$$

By substitution into the formula for TJ, we obtain

$$(TJ)(x) = \frac{p\gamma^2}{1+p}x^2. \tag{3.6}$$

Thus the function $J(x) = px^2$ is a fixed point of T if and only if p solves the equation

$$p = \frac{p\gamma^2}{1+p}.$$

This equation has two solutions:

$$p = 0 \qquad \text{and} \qquad p = \gamma^2 - 1,$$

as shown in Fig. 3.1.4. Thus there are exactly two fixed points of T within S: the functions

$$J^*(x) \equiv 0 \qquad \text{and} \qquad \hat{J}(x) = (\gamma^2 - 1)x^2.$$

The fixed point $\hat{J}$ has some significance. It turns out to be *the optimal cost function within the subclass $\mathcal{L}$ of linear policies that are stable*. This can be verified by minimizing the expression (3.5) over the parameter r. In particular, by setting to 0 the derivative with respect to r of

$$\frac{r^2}{1-(\gamma+r)^2},$$

we obtain after a straightforward calculation that it is minimized for $r = (1-\gamma^2)/\gamma$, which corresponds to the policy

$$\hat{\mu}(x) = \frac{(1-\gamma^2)}{\gamma}x,$$

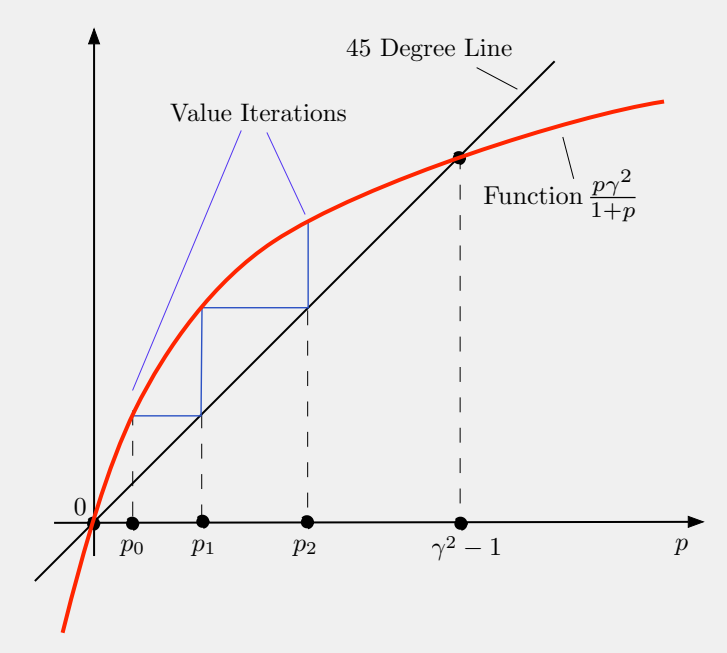


Figure 3.1.4. Illustrating the fixed points of T, and the convergence of the VI algorithm for the one-dimensional linear-quadratic problem.

while from Eq. (3.5), we can verify that

$$J_{\hat{\mu}}(x) = (\gamma^2 - 1)x^2.$$

Thus, we have

$$J_{\hat{\mu}}(x) = \inf_{\mu \in \mathcal{L}} J_\mu(x) = \hat{J}(x), \qquad x \in \Re.$$

Let us turn now to the VI algorithm starting from a function in S. Using Eq. (3.6), we see that it generates a sequence of functions $J_k \in S$ of the form

$$J_k(x) = p_k x^2,$$

where the sequence $\{p_k\}$ is generated by

$$p_{k+1} = \frac{p_k \gamma^2}{1 + p_k}, \qquad k = 0, 1, \ldots.$$

From Fig. 3.1.4 it can be seen that starting with $p_0 > 0$, the sequence $\{p_k\}$ converges to

$$\hat{p} = \gamma^2 - 1,$$

which corresponds to $\hat{J}$. In summary, starting from any nonzero function in S, the VI algorithm converges to the optimal cost function $\hat{J}$ over the linear stable policies $\mathcal{L}$, while starting from the zero function, it converges to the optimal cost function J^*.

Finally, let us consider the PI algorithm starting from a linear stable policy. We first note that given any $\mu \in \mathcal{L}$, i.e.,

$$\mu(x) = rx \qquad \text{with} \qquad |\gamma + r| < 1,$$

we can compute J_μ as the limit of the VI sequence $\{T_\mu^k J\}$, where J is any function in S, i.e.,

$$J(x) = px^2 \qquad \text{with} \qquad p \geq 0.$$

This can be verified by writing

$$(T_\mu J)(x) = \big(r^2 + p(\gamma + r)^2\big)x^2,$$

and noting that the iteration that maps p to $r^2 + p(\gamma + r)^2$ converges to

$$p_\mu = \frac{r^2}{1 - (\gamma + r)^2},$$

in view of $|\gamma + r| < 1$. Thus,

$$T_\mu^k J \to J_\mu, \qquad \forall\ \mu \in \mathcal{L},\ J \in S.$$

Moreover, we have $J_\mu = T_\mu J_\mu$.

We now use a standard proof argument to show that PI generates a sequence of linear stable policies starting from a linear stable policy. Indeed, we have for all k,

$$J_{\mu^0} = T_{\mu^0} J_{\mu^0} \geq T J_{\mu^0} = T_{\mu^1} J_{\mu^0} \geq T_{\mu^1}^k J_{\mu^0} \geq T^k \hat{J} = \hat{J},$$

where the second inequality follows by the monotonicity of T_{μ^1} and the third inequality follows from the fact $J_{\mu^0} \geq \hat{J}$. By taking the limit as $k \to \infty$, we obtain

$$J_{\mu^0} \geq T J_{\mu^0} \geq J_{\mu^1} \geq \hat{J}.$$

It can be verified that μ_1 is a nonzero linear policy, so the preceding relation implies that μ^1 is linear stable. Continuing similarly, it follows that the policies μ^k generated by PI are linear stable and satisfy for all k,

$$J_{\mu^k} \geq T J_{\mu^k} \geq J_{\mu^{k+1}} \geq \hat{J}.$$

By taking the limit as $k \to \infty$, we see that the sequence of quadratic functions $\{J_{\mu^k}\}$ converges monotonically to a quadratic function J_∞, which

is a fixed point of T and satisfies $J_\infty \geq \hat{J}$. Since we have shown that $\hat{J}$ is the only fixed point of T in the range $[\hat{J}, \infty)$, it follows that $J_\infty = \hat{J}$. In summary, the PI algorithm starting from a linear stable policy converges to $\hat{J}$, the optimal cost function over linear stable policies.

In Section 3.5.4, we will consider a more general multidimensional version of the linear-quadratic problem, using in part the analysis of Section 3.4. We will then explain the phenomena described in this section within a more general setting. We will also see there that the unusual behavior in the present example is due to the fact that there is no penalty for a nonzero state. For example, if the cost per stage is $\delta x^2 + u^2$, where $\delta > 0$, rather than u^2, then the corresponding Bellman equation has a unique solution with the class of positive semidefinite quadratic functions. We will analyze this case within a more general setting of deterministic optimal control problems in Section 3.5.5.

3.1.5 An Intuitive View of Semicontractive Analysis

In the preceding sections we have demonstrated various aspects of the character of semicontractive analysis in the context of several examples. The salient feature is a class of "well-behaved" policies (e.g., proper policies in shortest path problems, stable policies in linear-quadratic problems), and the restricted optimal cost function $\hat{J}$ over just these policies. The main results we typically derived were that $\hat{J}$ is a fixed point of T, and that the VI and PI algorithms are attracted to $\hat{J}$, at least from within some suitable class of initial conditions. In the favorable case where $\hat{J} = J^*$, these results hold also for J^*, but in general J^* need not be a fixed point of T.

The central issue of semicontractive analysis is *the choice of a class of "well-behaved" policies $\widehat{\mathcal{M}} \subset \mathcal{M}$ such that the corresponding restricted optimal cost function $\hat{J}$ is a fixed point of T*. Such a choice is often fairly evident, but there are also several systematic approaches to identify a suitable class $\widehat{\mathcal{M}}$ and to show its fixed point property; see the end of Section 3.2.2 for a discussion of various alternatives. As an example, let us introduce a class of policies $\widehat{\mathcal{M}} \subset \mathcal{M}$ for which we assume the following:

(a) *$\widehat{\mathcal{M}}$ is well-behaved with respect to VI*: For all $\mu \in \widehat{\mathcal{M}}$ and real-valued functions J, we have

$$J_\mu = T_\mu J_\mu, \qquad J_\mu = \lim_{k \to \infty} T_\mu^k J. \tag{3.7}$$

Moreover J_μ is real-valued.

(b) *$\widehat{\mathcal{M}}$ is well-behaved with respect to PI*: For each $\mu \in \widehat{\mathcal{M}}$, any policy μ' such that

$$T_{\mu'} J_\mu = T J_\mu$$

belongs to $\widehat{\mathcal{M}}$, and there exists at least one such μ'.

We can show that $\hat{J}$ is a fixed point of T and obtain our main results with the following line of argument. The first step in this argument is to show that the cost functions of a PI-generated sequence $\{\mu^k\} \subset \widehat{\mathcal{M}}$ (starting from a $\mu^0 \in \widehat{\mathcal{M}}$) are monotonically nonincreasing. Indeed, we have using Eq. (3.7),

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

Using the monotonicity property of $T_{\mu^{k+1}}$, it follows that

$$J_{\mu^k} \geq T J_{\mu^k} \geq \lim_{k\to\infty} T^k_{\mu^{k+1}} J_{\mu^k} = J_{\mu^{k+1}} \geq \hat{J}, \tag{3.8}$$

where the equality holds by Eq. (3.7), and the rightmost inequality holds since $\mu^{k+1} \in \widehat{\mathcal{M}}$ [by assumption (b) above]. Thus we obtain $J_{\mu^k} \downarrow J_\infty \geq \hat{J}$, for some function J_∞.

Now by taking the limit as $k \to \infty$ in the relation $J_{\mu^k} \geq T J_{\mu^k} \geq J_{\mu^{k+1}}$ [cf. Eq. (3.8)], it follows (under a mild continuity assumption) that J_∞ is a fixed point of T with $J_\infty \geq \hat{J}$.† We claim that $J_\infty = \hat{J}$. Indeed we have

$$\hat{J} \leq J_\infty = T^k J_\infty \leq T^k_\mu J_\infty \leq T^k_\mu J_{\mu^0}, \qquad \forall\ \mu \in \widehat{\mathcal{M}},\ k = 0, 1, \ldots.$$

By taking the limit as $k \to \infty$, and using the fact $\mu \in \widehat{\mathcal{M}}$ [cf. Eq. (3.7)], we obtain $\hat{J} \leq J_\infty \leq J_\mu$ for all $\mu \in \widehat{\mathcal{M}}$. By taking the infimum over $\mu \in \widehat{\mathcal{M}}$, it follows that $J_\infty = \hat{J}$.

Finally, let J be real-valued and satisfy $J \geq \hat{J}$. We claim that $T^k J \to \hat{J}$. Indeed, since we have shown that $\hat{J}$ is a fixed point of T, we have

$$T^k_\mu J \geq T^k J \geq T^k \hat{J} = \hat{J}, \qquad \forall\ \mu \in \widehat{\mathcal{M}},\ k \geq 0,$$

† We elaborate on this argument; see also the proof of Prop. 3.2.4 in the next section. From Eq. (3.8), we have $J_{\mu^k} \geq T J_{\mu^k} \geq T J_\infty$, so by letting $k \to \infty$, we obtain $J_\infty \geq T J_\infty$. For the reverse inequality, we assume that H has the property that $H(x, u, J) = \lim_{m\to\infty} H(x, u, J_m)$ for all $x \in X$, $u \in U(x)$, and sequence $\{J_m\}$ of real-valued functions with $J_m \downarrow J$. Thus we have

$$H(x, u, J_\infty) = \lim_{k\to\infty} H(x, u, J_{\mu^k}) \geq \lim_{k\to\infty} (T J_{\mu^k})(x), \qquad x \in X,\ u \in U(x).$$

By taking the limit in Eq. (3.8), we obtain

$$\lim_{k\to\infty} (T J_{\mu^k})(x) \geq \lim_{k\to\infty} J_{\mu^{k+1}}(x) = J_\infty(x), \qquad x \in X,$$

and from the preceding two relations we have $H(x, u, J_\infty) \geq J_\infty(x)$. By taking the infimum over $u \in U(x)$, it follows that $T J_\infty \geq J_\infty$. Combined with the relation $J_\infty \geq T J_\infty$ shown earlier, this implies that J_∞ is a fixed point of T.

so by taking the limit as $k \to \infty$ and using Eq. (3.7), we obtain

$$J_\mu \geq \lim_{k\to\infty} T^k J \geq \hat{J}, \qquad \forall\ \mu \in \widehat{\mathcal{M}}.$$

By taking the infimum over $\mu \in \widehat{\mathcal{M}}$, it follows that $T^k J \to \hat{J}$, i.e., that VI converges to $\hat{J}$ stating from all initial conditions $J \geq \hat{J}$.

The analysis of the following two sections will be based to a large extent on refinements of the preceding argument. Note that in this argument we have not assumed that $\hat{J} = J^*$, which leaves open the possibility that J^* is not a fixed point of T. Indeed this can happen, as we have seen in the SSP example of Section 3.1.2. Moreover, we have not assumed that $\hat{J}$ is real-valued. In fact $\hat{J}$ may not be real-valued even though all J_μ, $\mu \in \widehat{\mathcal{M}}$, are; see the first variant of the blackmailer problem of Section 3.1.3.

An alternative analytical approach, which does not rely on $\widehat{\mathcal{M}}$ being well-behaved with respect to PI, is given in Section 3.4. The idea there is to introduce a small δ-perturbation to the mapping H and a corresponding "δ-perturbed" problem. The perturbation is chosen so that the cost function of some policies, the "well-behaved" ones, is minimally affected [say by $O(\delta)$], while the cost function of the policies that are not "well-behaved" is driven to ∞ for some initial states, thereby excluding these policies from optimality. Thus as $\delta \downarrow 0$, the optimal cost function $\hat{J}_\delta$ of the δ-perturbed problem approaches $\hat{J}$ (not J^*). Assuming that $\hat{J}_\delta$ is a solution of the δ-perturbed Bellman equation, and we can then use a limiting argument to show that $\hat{J}$ is a fixed point of T, as well as other results relating to the VI and PI algorithms. The perturbation approach will become more prominent in our semicontractive analysis of Chapter 4 (Sections 4.5 and 4.6), where we will consider "well-behaved" policies that are nonstationary, and thus do not lend themselves to a PI-based analysis.

3.2 SEMICONTRACTIVE MODELS AND REGULAR POLICIES

In the preceding section we illustrated a general pattern of pathologies in noncontractive models, involving the solutions of Bellman's equation, and the convergence of the VI and PI algorithms. To summarize:

(a) Bellman's equation may have multiple solutions (equivalently, T may have multiple fixed points). Often but not always, J^* is a fixed point of T. Moreover, a restricted problem, involving policies that are "well-behaved" (proper in shortest path problems, or linear stable in the linear-quadratic case), may be meaningful and play an important role.

(b) The optimal cost function over all policies, J^*, may differ from $\hat{J}$, the optimal cost function over the "well-behaved" policies. Furthermore, it may be that $\hat{J}$ (not J^*) is "well-behaved" from the algorithmic point of view. In particular, $\hat{J}$ is often a fixed point of T, in which

case it is the likely limit of the VI and the PI algorithms, starting from an appropriate set of initial conditions.

In this section we will provide an analytical framework that explains this type of phenomena, and develops the kind of assumptions needed in order to avoid them. We will introduce a concept of regularity that formalizes mathematically the notion of "well-behaved" policy, and we will consider a restricted optimization problem that involves regular policies only. We will show that the optimal cost function of the restricted problem is a fixed point of T under several types of fairly natural assumptions. Moreover, we will show that it can be computed by versions of VI and PI, starting from suitable initial conditions.

Problem Formulation

Let us first introduce formally the model that we will use in this chapter. Compared to the contractive model of Chapter 2, it maintains the monotonicity assumption, but not the contraction assumption.

We introduce the set X of states and the set U of controls, and for each $x \in X$, the nonempty control constraint set $U(x) \subset U$. Since in the absence of the contraction assumption, the cost function J_μ of some policies μ may take infinite values for some states, we will use the set of extended real numbers $\Re^* = \Re \cup \{\infty, -\infty\} = [-\infty, \infty]$. The mathematical operations with ∞ and $-\infty$ are standard and are summarized in Appendix A. We consider the set of all extended real-valued functions $J : X \mapsto \Re^*$, which we denote by $\mathcal{E}(X)$. We also denote by $\mathcal{R}(X)$ the set of real-valued functions $J : X \mapsto \Re$.

As earlier, when we write lim, lim sup, or lim inf of a sequence of functions we mean it to be pointwise. We also write $J_k \to J$ to mean that $J_k(x) \to J(x)$ for each $x \in X$; see Appendix A.

We denote by $\mathcal{M}$ the set of all functions $\mu : X \mapsto U$ with $\mu(x) \in U(x)$, for all $x \in X$, and by Π the set of policies $\pi = \{\mu_0, \mu_1, \ldots\}$, where $\mu_k \in \mathcal{M}$ for all k. We refer to a stationary policy $\{\mu, \mu, \ldots\}$ simply as μ. We introduce a mapping $H : X \times U \times \mathcal{E}(X) \mapsto \Re^*$ that satisfies the following.

Assumption 3.2.1: (Monotonicity) If $J, J' \in \mathcal{E}(X)$ and $J \leq J'$, then

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

The preceding monotonicity assumption will be in effect throughout this chapter. Consequently, *we will not mention it explicitly in various propositions*. We define the mapping $T : \mathcal{E}(X) \mapsto \mathcal{E}(X)$ by

$$(TJ)(x) = \inf_{u \in U(x)} H(x, u, J), \qquad \forall\ x \in X,\ J \in \mathcal{E}(X),$$

and for each $\mu \in \mathcal{M}$ the mapping $T_\mu : \mathcal{E}(X) \mapsto \mathcal{E}(X)$ by

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

The monotonicity assumption implies the following properties for all $J, J' \in \mathcal{E}(X)$ and $k = 0, 1, \ldots,$

$$J \leq J' \qquad \Rightarrow \qquad T^k J \leq T^k J', \qquad T_\mu^k J \leq T_\mu^k J', \quad \forall\ \mu \in \mathcal{M},$$

$$J \leq TJ \qquad \Rightarrow \qquad T^k J \leq T^{k+1} J, \qquad T_\mu^k J \leq T_\mu^{k+1} J, \quad \forall\ \mu \in \mathcal{M},$$

which we will use extensively in various proof arguments.

We now define cost functions associated with T_μ and T. In Chapter 2 our starting point was to define J_μ and J^* as the unique fixed points of T_μ and T, respectively, based on the contraction assumption used there. However, under our assumptions in this chapter this is not possible, so we use a different definition, which nonetheless is consistent with the one of Chapter 2 (see the discussion of Section 2.1, following Prop. 2.1.2). We introduce a function $\bar{J} \in \mathcal{E}(X)$, and we define the infinite horizon cost of a policy in terms of the limit of its finite horizon costs with $\bar{J}$ being the cost function at the end of the horizon. Note that in the case of the optimal control problems of the preceding section we have taken $\bar{J}$ to be the zero function, $\bar{J}(x) \equiv 0$ [cf. Eq. (3.2)].

Definition 3.2.1: Given a function $\bar{J} \in \mathcal{E}(X)$, for a policy $\pi \in \Pi$ with $\pi = \{\mu_0, \mu_1, \ldots\}$, we define the cost function of π by

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

In the case of a stationary policy μ, the cost function of μ is denoted by J_μ and is given by

$$J_\mu(x) = \limsup_{k \to \infty} (T_\mu^k \bar{J})(x), \qquad \forall\ x \in X.$$

The optimal cost function J^* is given by

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

An optimal policy $\pi^* \in \Pi$ is one for which $J_{\pi^*} = J^*$.

Note two important differences from Chapter 2:

(1) J_μ is defined in terms of a pointwise lim sup rather than lim, since we don't know whether the limit exists.

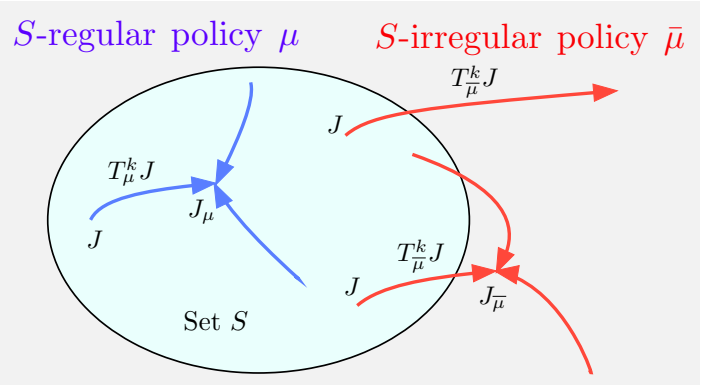


Figure 3.2.1. Illustration of S-regular and S-irregular policies. Policy μ is S-regular because $J_\mu \in S$ and $T_\mu^k J \to J_\mu$ for all $J \in S$. Policy $\overline{\mu}$ is S-irregular.

(2) J_π and J_μ in general depend on $\bar{J}$, so $\bar{J}$ becomes an important part of the problem definition.

Similar to Chapter 2, under the assumptions to be introduced in this chapter, stationary policies will typically turn out to be "sufficient" in the sense that the optimal cost obtained with nonstationary policies that depend on the initial state is matched by the one obtained by stationary ones.

3.2.1 S-Regular Policies

Our objective in this chapter is to construct an analytical framework with a strong connection to fixed point theory, based on the idea of separating policies into those that have "favorable" characteristics and those that do not. Clearly, a favorable property for a policy μ is that J_μ is a fixed point of T_μ. However, J_μ may depend on $\bar{J}$, even though T_μ does not depend on $\bar{J}$. It would thus appear that a related favorable property for μ is that J_μ stays the same if $\bar{J}$ is changed arbitrarily within some set S. We express these two properties with the following definition (see Fig. 3.2.1).

Definition 3.2.2: Given a set of functions $S \subset \mathcal{E}(X)$, we say that a stationary policy μ is *S-regular* if:

(a) $J_\mu \in S$ and $J_\mu = T_\mu J_\mu$.

(b) $T_\mu^k J \to J_\mu$ for all $J \in S$.

A policy that is not S-regular is called *S-irregular*.

Thus a policy μ is S-regular if the VI algorithm corresponding to μ, $J_{k+1} = T_\mu J_k$, represents a dynamic system that has J_μ as its unique

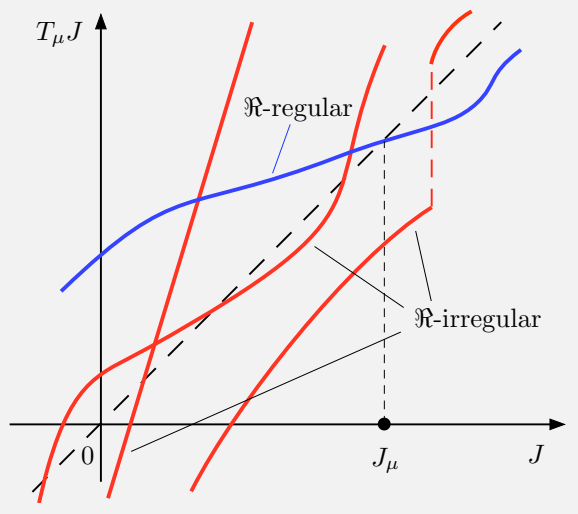


Figure 3.2.2. Illustration of S-regular and S-irregular policies for the case where there is only one state and $S = \Re$. There are three mappings T_μ corresponding to S-irregular policies: one crosses the 45-degree line at multiple points, another crosses at a single point but at an angle greater than 45 degrees, and the third is discontinuous and does not cross at all. The mapping T_μ of the $\Re$-regular policy has J_μ as its unique fixed point and satisfies $T_\mu^k J \to J_\mu$ for all $J \in \Re$.

equilibrium within S, and is asymptotically stable in the sense that the iteration converges to J_μ, starting from any $J \in S$.

For orientation purposes, we note the distinction between the set S and the problem data: *S is not part of the problem's definition*. Its choice, however, can enable analysis and clarify properties of J_μ and J^*. For example, we will later prove local fixed point statements such as

"J^* is the unique fixed point of T within S"

or local region of attraction assertions such as

"the VI sequence $\{T^k J\}$ converges to J^* starting from any $J \in S$."

Results of this type and their proofs depend on the choice of S: they may hold for some choices but not for others.

Generally, with our selection of S we will aim to differentiate between S-regular and S-irregular policies in a manner that produces useful results for the given problem and does not necessitate restrictive assumptions. Examples of sets S that we will use are $\mathcal{R}(X)$, $\mathcal{B}(X)$, $\mathcal{E}(X)$, and subsets of $\mathcal{R}(X)$, $\mathcal{B}(X)$, and $\mathcal{E}(X)$ involving functions J satisfying $J \geq J^*$ or $J \geq \bar{J}$. However, there is a diverse range of other possibilities, so it makes sense to postpone making the choice of S more specific. Figure 3.2.2 illustrates the mappings T_μ of some S-regular and S-irregular policies for the case where there is a single state and $S = \Re$. Figure 3.2.3 illustrates the mapping

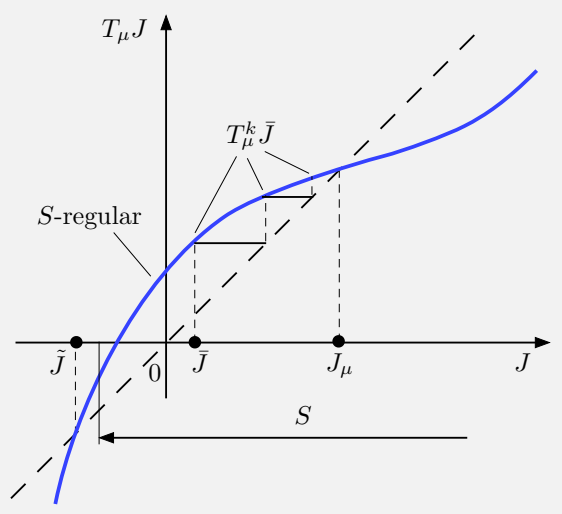


Figure 3.2.3. Illustration of a mapping T_μ where there is only one state and S is a subset of the real line. Here T_μ has two fixed points, J_μ and $\tilde{J}$. If S is as shown, μ is S-regular. If S is enlarged to include $\tilde{J}$, μ becomes S-irregular.

T_μ of an S-regular policy μ, where T_μ has multiple fixed points, and upon changing S, the policy may become S-irregular.

3.2.2 Restricted Optimization over S-Regular Policies

We will now introduce a restricted optimization framework where S-regular policies are central. Given a nonempty set $S \subset \mathcal{E}(X)$, let $\mathcal{M}_S$ denote the set of policies that are S-regular, and consider optimization over just the set $\mathcal{M}_S$. The corresponding optimal cost function is denoted J_S^*:

$$J_S^*(x) = \inf_{\mu \in \mathcal{M}_S} J_\mu(x), \qquad \forall\, x \in X. \tag{3.9}$$

We say that μ^* is $\mathcal{M}_S$*-optimal* if †

$$\mu^* \in \mathcal{M}_S \qquad \text{and} \qquad J_{\mu^*} = J_S^*.$$

An important question is whether J_S^* is a fixed point of T and can be obtained by the VI algorithm. Naturally, this depends on the choice of S, but it turns out that reasonable choices can be readily found in several important contexts, so the consequences of J_S^* being a fixed point of T are

† Note that while S is assumed nonempty, it is possible that $\mathcal{M}_S$ is empty. In this case our results will not be useful, but J_S^* is still defined by Eq. (3.9) as $J_S^*(x) \equiv \infty$. This is convenient in various proof arguments.

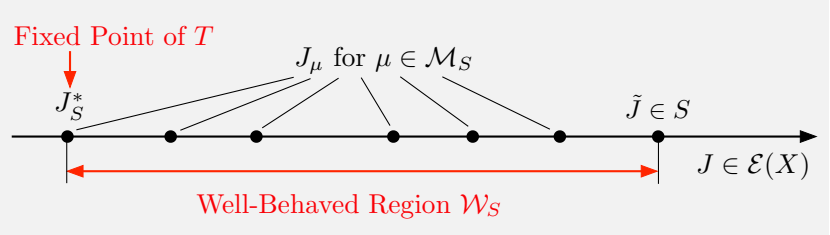


Figure 3.2.4. Interpretation of Prop. 3.2.1, where for illustration purposes, $\mathcal{E}(X)$ is represented by the extended real line. A set $S \subset \mathcal{E}(X)$ such that J_S^* is a fixed point of T, demarcates the well-behaved region $\mathcal{W}_S$ [cf. Eq. (3.10)], within which T has a unique fixed point, and starting from which the VI algorithm converges to J_S^*.

interesting. The next proposition shows that if J_S^* is a fixed point of T, then the VI algorithm is convergent starting from within the set

$$\mathcal{W}_S = \{J \in \mathcal{E}(X) \mid J_S^* \leq J \leq \tilde{J} \text{ for some } \tilde{J} \in S\}, \tag{3.10}$$

which we refer to as the *well-behaved region* (see Fig. 3.2.4). Note that by the definition of S-regularity, the cost functions J_μ, $\mu \in \mathcal{M}_S$, belong to S and hence also to $\mathcal{W}_S$. The proposition also provides a necessary and sufficient condition for an S-regular policy μ^* to be $\mathcal{M}_S$-optimal.

Proposition 3.2.1: (Well-Behaved Region Theorem) Given a set $S \subset \mathcal{E}(X)$, assume that J_S^* is a fixed point of T. Then:

(a) (*Uniqueness of Fixed Point*) If J' is a fixed point of T and there exists $\tilde{J} \in S$ such that $J' \leq \tilde{J}$, then $J' \leq J_S^*$. In particular, if $\mathcal{W}_S$ is nonempty, J_S^* is the unique fixed point of T within $\mathcal{W}_S$.

(b) (*VI Convergence*) We have $T^k J \to J_S^*$ for every $J \in \mathcal{W}_S$.

(c) (*Optimality Condition*) If μ is S-regular, $J_S^* \in S$, and $T_\mu J_S^* = TJ_S^*$, then μ is $\mathcal{M}_S$-optimal. Conversely, if μ is $\mathcal{M}_S$-optimal, then $T_\mu J_S^* = TJ_S^*$.

Proof: (a) For every $\mu \in \mathcal{M}_S$, we have using the monotonicity of T_μ,

$$J' = TJ' \leq T_\mu J' \leq \cdots \leq T_\mu^k J' \leq T_\mu^k \tilde{J}, \qquad k = 1, 2, \ldots.$$

Taking limit as $k \to \infty$, and using the S-regularity of μ, we obtain $J' \leq J_\mu$ for all $\mu \in \mathcal{M}_S$. Taking the infimum over $\mu \in \mathcal{M}_S$, we have $J' \leq J_S^*$.

Assume that $\mathcal{W}_S$ is nonempty. Then J_S^* is a fixed point of T that belongs to $\mathcal{W}_S$. To show its uniqueness, let J' be another fixed point that

belongs to $\mathcal{W}_S$, so that $J_S^* \leq J'$ and there exists $\tilde{J} \in S$ such that $J' \leq \tilde{J}$. By what we have shown so far, $J' \leq J_S^*$, implying that $J' = J_S^*$.

(b) Let $J \in \mathcal{W}_S$, so that $J_S^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$. We have for all $k \geq 1$ and $\mu \in \mathcal{M}_S$,

$$J_S^* = T^k J_S^* \leq T^k J \leq T^k \tilde{J} \leq T_\mu^k \tilde{J},$$

where the equality follows from the fixed point property of J_S^*, while the inequalities follow from the monotonicity and the definition of T. The right-hand side tends to J_μ as $k \to \infty$, since μ is S-regular and $\tilde{J} \in S$. Hence the infimum over $\mu \in \mathcal{M}_S$ of the limit of the right-hand side tends to the left-hand side J_S^*. It follows that $T^k J \to J_S^*$.

(c) From the assumptions $T_\mu J_S^* = T J_S^*$ and $T J_S^* = J_S^*$, we have $T_\mu J_S^* = J_S^*$, and since $J_S^* \in S$ and μ is S-regular, we have $J_S^* = J_\mu$. Thus μ is $\mathcal{M}_S$-optimal. Conversely, if μ is M_S-optimal, we have $J_\mu = J_S^*$, so that the fixed point property of J_S^* and the S-regularity of μ imply that

$$T J_S^* = J_S^* = J_\mu = T_\mu J_\mu = T_\mu J_S^*.$$

Q.E.D.

Some useful extensions and modified versions of the preceding proposition are given in Exercises 3.2-3.5. Let us illustrate the proposition in the context of the deterministic shortest path example of Section 3.1.1.

Example 3.2.1

Consider the deterministic shortest path example of Section 3.1.1 for the case where there is a zero length cycle ($a = 0$), and let S be the real line $\Re$. There are two policies: μ which moves from state 1 to the destination at cost b, and μ' which stays at state 1 at cost 0. We use $X = \{1\}$ (i.e., we do not include t in X, since all function values of interest are 0 at t). Then by abbreviating function values $J(1)$ with J, we have

$$J_\mu = b, \qquad J_{\mu'} = 0, \qquad J^* = \min\{b, 0\};$$

cf. Fig. 3.2.5. The corresponding mappings T_μ, $T_{\mu'}$, and T are

$$T_\mu J = b, \quad T_{\mu'} J = J, \quad J = TJ = \min\{b, J\}, \qquad J \in \mathcal{E}(X),$$

and the initial function $\bar{J}$ is taken to be 0. It can be seen from the definition of S-regularity that μ is S–regular, while the policy μ' is not. The cost functions J_μ, $J_{\mu'}$, and J^* are fixed points of the corresponding mappings, but the sets of fixed points of $T_{\mu'}$ and T within S are $\Re$ and $(-\infty, b]$, respectively. Moreover, $J_S^* = J_\mu = b$, so J_S^* is a fixed point of T and Prop. 3.2.1 applies.

The figure also shows the well-behaved regions for the two cases $b > 0$ and $b < 0$. It can be seen that the results of Prop. 3.2.1 are consistent with the discussion of Section 3.1.1. In particular, the VI algorithm fails when

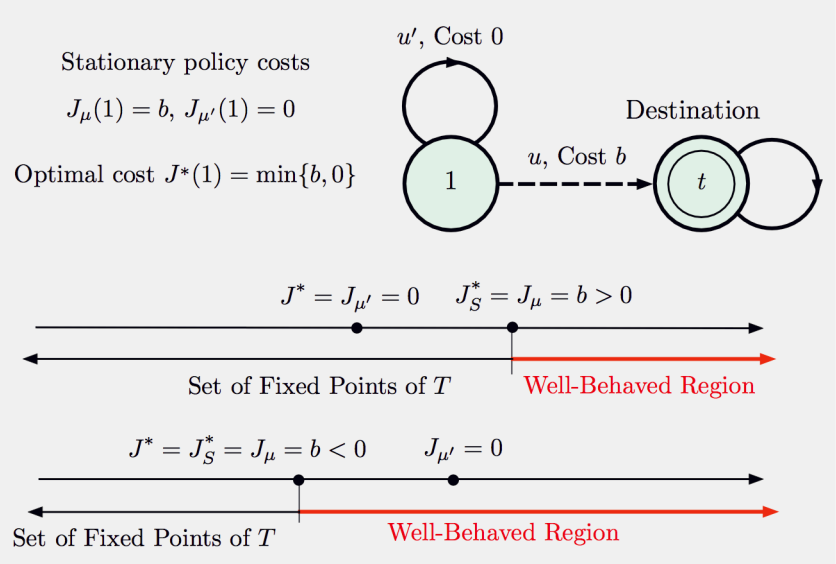


Figure 3.2.5. The well-behaved region of Eq. (3.10) for the deterministic shortest path example of Section 3.1.1 when where there is a zero length cycle ($a = 0$). For $S = \Re$, the policy μ is S-regular, while the policy μ' is not. The figure illustrates the two cases where $b > 0$ and $b < 0$.

started outside the well-behaved region, while when started from within the region, it is attracted to J_S^* rather than to J^*.

Let us now discuss some of the fine points of Prop. 3.2.1. The salient assumption of the proposition is that J_S^* is a fixed point of T. Depending on the choice of S, this may or may not be true, and much of the subsequent analysis in this chapter is geared towards the development of approaches to choose S so that J_S^* is a fixed point of T and has some other interesting properties. As an illustration of the range of possibilities, consider the three variants of the blackmailer problem of Section 3.1.3 for the choice $S = \Re$:

(a) In the first variant, we have $J^* = J_S^* = -\infty$, and J_S^* is a fixed point of T that lies outside S. Here parts (a) and (b) of Prop. 3.2.1 apply. However, part (c) does not apply (even though we have $T_\mu J_S^* = T J_S^*$ for all policies μ) because $J_S^* \notin S$, and in fact there is no $\mathcal{M}_S$-optimal policy. In the subsequent analysis, we will see that the condition $J_S^* \in S$ plays an important role in being able to assert existence of an $\mathcal{M}_S$-optimal policy (see the subsequent Props. 3.2.5 and 3.2.6).

(b) In the second variant, we have $J^* = J_S^* = -1$, and J_S^* is a fixed point of T that lies within S. Here parts (a) and (b) of Prop. 3.2.1 apply, but part (c) still does not apply because there is no S-regular μ such

that $T_\mu J_S^* = T J_S^*$, and in fact there is no M_S-optimal policy.

(c) In the third variant with $c < 0$, we have $J^* = -\infty$, $J_S^* = -1$, and J_S^* is not a fixed point of T. Thus Prop. 3.2.1 does not apply, and in fact we have $T^k J \to J^*$ for every $J \in \mathcal{W}_S$ (and not $T^k J \to J_S^*$).

Another fine point is that Prop. 3.2.1(b) asserts convergence of the VI algorithm to J_S^* only for initial conditions J satisfying $J_S^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$. For an illustrative example of an S-regular μ, where $\{T_\mu^k J\}$ does not converge to J_μ starting from some $J \geq J_\mu$ that lies outside S, consider a case where there is a single state and a single policy μ that is S-regular, so $J_S^* = J_\mu$. Suppose that $T_\mu : \Re \mapsto \Re$ has two fixed points: J_μ and another fixed point $J' > J_\mu$. Let

$$\tilde{J} = (J_\mu + J')/2, \qquad S = (-\infty, \tilde{J}],$$

and assume that T_μ is a contraction mapping within S (an example of this type can be easily constructed graphically). Then starting from any $J \in S$, we have $T^k J \to J_\mu$, so that μ is S-regular. However, since J' is a fixed point of T, the sequence $\{T^k J'\}$ stays at J' and does not converge to J_μ. The difficulty here is that $\mathcal{W}_S = [J_\mu, \tilde{J}]$ and $J' \notin \mathcal{W}_S$.

Still another fine point is that if there exists an $\mathcal{M}_S$-optimal policy μ, we have $J_S^* = T_\mu J_S^*$ (since $J_S^* = J_\mu$ and μ is S-regular), but this does not guarantee that J_S^* is a fixed point of T, which is essential for Prop. 3.2.1. This can be seen from an example given in Fig. 3.2.6, where there exists an $\mathcal{M}_S$-optimal policy, but both J_S^* and J^* are not fixed points of T (in this example the $\mathcal{M}_S$-optimal policy is also overall optimal so $J_S^* = J^*$). In particular, starting from J_S^*, the VI algorithm converges to some $J' \neq J_S^*$ that is a fixed point of T.

Convergence Rate when a Contractive Policy is $\mathcal{M}_S$-Optimal

In many contexts where Prop. 3.2.1 applies, there exists an $\mathcal{M}_S$-optimal policy μ such that T_μ is a contraction with respect to a weighted sup-norm. This is true for example in the shortest path problem to be discussed in Section 3.5.1. In such cases, the rate of convergence of VI to J_S^* is linear, as shown in the following proposition.

Proposition 3.2.2: (Convergence Rate of VI) Let S be equal to $\mathcal{B}(X)$, the space of all functions over X that are bounded with respect to a weighted sup-norm $\|\cdot\|_v$ corresponding to a positive function $v : X \mapsto \Re$. Assume that J_S^* is a fixed point of T, and that there exists an $\mathcal{M}_S$-optimal policy μ such that T_μ is a contraction with respect to $\|\cdot\|_v$, with modulus of contraction β. Then $J_S^* \in \mathcal{B}(X)$, $\mathcal{W}_S \subset \mathcal{B}(X)$, and

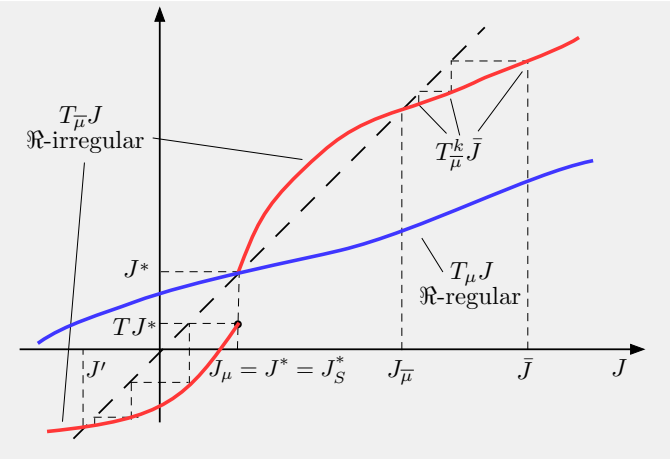


Figure 3.2.6. Illustration of why the assumption that J^*_S is a fixed point of T is essential for Prop. 3.2.1. In this example there is only one state and $S = \Re$. There are two stationary policies: μ for which T_μ is a contraction, so μ is $\Re$-regular, and $\overline{\mu}$ for which $T_{\overline{\mu}}$ has multiple fixed points, so $\overline{\mu}$ is $\Re$-irregular. Moreover, $T_{\overline{\mu}}$ is discontinuous from above at J_μ as shown. Here, it can be verified that $T_{\mu_0} \cdots T_{\mu_k} \bar{J} \geq J_\mu$ for all $\mu_0, \ldots, \mu_k$ and k, so that $J_\pi \geq J_\mu$ for all π and the S-regular policy μ is optimal, so $J^*_S = J^*$. However, as can be seen from the figure, we have $J^*_S = J^* \neq TJ^* = TJ^*_S$. Moreover, starting at J^*_S, the VI sequence $T^k J^*_S$ converges to J', the fixed point of T shown in the figure, and all parts of Prop. 3.2.1 fail.

$$\big\|TJ - J^*_S\|_v \leq \beta \|J - J^*_S\|_v, \qquad \forall\ J \in \mathcal{W}_S. \tag{3.11}$$

Moreover, we have

$$\|J - J^*_S\|_v \leq \frac{1}{1-\beta} \sup_{x \in X} \frac{J(x) - (TJ)(x)}{v(x)}, \qquad \forall\ J \in \mathcal{W}_S. \tag{3.12}$$

Proof: Since μ is S-regular and $S = \mathcal{B}(X)$, we have $J^*_S = J_\mu \in \mathcal{B}(X)$ as well as $\mathcal{W}_S \subset \mathcal{B}(X)$. By using the $\mathcal{M}_S$-optimality of μ and Prop. 3.2.1(c),

$$J^*_S = T_\mu J^*_S = TJ^*_S,$$

so for all $x \in X$ and $J \in \mathcal{W}_S$,

$$\frac{(TJ)(x) - J^*_S(x)}{v(x)} \leq \frac{(T_\mu J)(x) - (T_\mu J^*_S)(x)}{v(x)} \leq \beta \max_{x \in X} \frac{J(x) - J^*_S(x)}{v(x)},$$

where the second inequality holds by the contraction property of T_μ. By taking the supremum of the left-hand side over $x \in X$, and by using the fact $TJ \geq TJ^*_S = J^*_S$ for all $J \in \mathcal{W}_S$, we obtain Eq. (3.11).

By using again the relation $T_\mu J_S^* = T J_S^*$, we have for all $x \in X$ and all $J \in \mathcal{W}_S$,

$$\begin{aligned}\frac{J(x) - J_S^*(x)}{v(x)} &= \frac{J(x) - (TJ)(x)}{v(x)} + \frac{(TJ)(x) - J_S^*(x)}{v(x)} \\ &\leq \frac{J(x) - (TJ)(x)}{v(x)} + \frac{(T_\mu J)(x) - (T_\mu J_S^*)(x)}{v(x)} \\ &\leq \frac{J(x) - (TJ)(x)}{v(x)} + \beta \|J - J_S^*\|_v.\end{aligned}$$

By taking the supremum of both sides over x, we obtain Eq. (3.12). **Q.E.D.**

Approaches to Show that J_S^* is a Fixed Point of T

The critical assumption of Prop. 3.2.1 is that J_S^* is a fixed point of T. For a specific application, this must be proved with a separate analysis after a suitable set S is chosen. To this end, we will provide several approaches that guide the choice of S and facilitate the analysis.

One approach applies to problems where J^* is generically a fixed point of T, in which case for every set S such that $J_S^* = J^*$, Prop. 3.2.1 applies and shows that J^* can be obtained by the VI algorithm starting from any $J \in \mathcal{W}_S$. Exercise 3.1 provides some conditions that guarantee that J^* is a fixed point of T. These conditions can be verified in wide classes of problems such as deterministic models. Sections 3.5.4 and 3.5.5 illustrate this approach. Other important models where J^* is guaranteed to be a fixed point of T are the monotone increasing and monotone decreasing models of Section 4.3. We will discuss the application of Prop. 3.2.1 and other related results to these models in Chapter 4.

In the present chapter the approach for showing that J_S^* is a fixed point of T will be mostly based on the PI algorithm; cf. the discussion of Section 3.1.5. An alternative and complementary approach is the perturbation-based analysis to be given in Section 3.4. This approach will be applied to a variety of problems in Section 3.5, and will also be prominent in Sections 4.5 and 4.6 of the next chapter.

3.2.3 Policy Iteration Analysis of Bellman's Equation

We will develop a PI-based approach for showing that J_S^* is a fixed point of T. The approach is applicable under assumptions that guarantee that there is a sequence $\{\mu^k\}$ of S-regular policies that can be generated by PI. The significance of S-regularity of all μ^k lies in that *the corresponding cost function sequence $\{J_{\mu^k}\}$ belongs to the well-behaved region of Eq. (3.10), and is monotonically nonincreasing* (see the subsequent Prop. 3.2.3). Under an additional mild technical condition, the limit of this sequence is a fixed point of T and is in fact equal to J_S^* (see the subsequent Prop. 3.2.4).

Let us consider the standard form of the PI algorithm, which starts with a policy μ^0 and generates a sequence $\{\mu^k\}$ of stationary policies according to

$$T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k}, \qquad k = 0, 1, \ldots. \tag{3.13}$$

This iteration embodies both the policy evaluation step, which computes J_{μ^k} in some way, and the policy improvement step, which computes $\mu^{k+1}(x)$ as a minimum over $u \in U(x)$ of $H(x, u, J_{\mu^k})$ for each $x \in X$. Of course, to be able to carry out the policy improvement step, there should be enough assumptions to guarantee that the minimum is attained for every x. One such assumption is that $U(x)$ is a finite set for each $x \in X$. A more general assumption, which applies to the case where the constraint sets $U(x)$ are infinite, will be given in Section 3.3.

The evaluation of the cost function J_μ of a policy μ may be done by solving the equation $J_\mu = T_\mu J_\mu$, which holds when μ is an S-regular policy. An important fact is that if the PI algorithm generates a sequence $\{\mu^k\}$ consisting exclusively of S-regular policies, then not only the policy evaluation is facilitated through the equation $J_\mu = T_\mu J_\mu$, but also *the sequence of cost functions* $\{J_{\mu^k}\}$ *is monotonically nonincreasing*, as we will show next.

Proposition 3.2.3: (Policy Improvement Under S-Regularity) Given a set $S \subset \mathcal{E}(X)$, assume that $\{\mu^k\}$ is a sequence generated by the PI algorithm (3.13) that consists of S-regular policies. Then

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

Proof: Using the S-regularity of μ^k and Eq. (3.13), we have

$$J_{\mu^k} = T_{\mu^k} J_{\mu^k} \geq T J_{\mu^k} = T_{\mu^{k+1}} J_{\mu^k}. \tag{3.14}$$

By using the monotonicity of $T_{\mu^{k+1}}$, we obtain

$$J_{\mu^k} \geq T J_{\mu^k} \geq \lim_{m \to \infty} T^m_{\mu^{k+1}} J_{\mu^k} = J_{\mu^{k+1}}, \tag{3.15}$$

where the equation on the right holds since μ^{k+1} is S-regular and $J_{\mu^k} \in S$ (in view of the S-regularity of μ^k). **Q.E.D.**

The preceding proposition shows that if a sequence of S-regular policies $\{\mu^k\}$ is generated by PI, the corresponding cost function sequence $\{J_{\mu^k}\}$ is monotonically nonincreasing and hence converges to a limit J_∞. Under mild conditions, we will show that J_∞ is a fixed point of T and is equal to J^*_S. This is important as it brings to bear Prop. 3.2.1, and the

associated results on VI convergence and optimality conditions. Let us first formalize the property that the PI algorithm can generate a sequence of S-regular policies.

Definition 3.2.3: (Weak PI Property) We say that a set $S \subset \mathcal{E}(X)$ has the *weak PI property* if there exists a sequence of S-regular policies that can be generated by the PI algorithm [i.e., a sequence $\{\mu^k\}$ that satisfies Eq. (3.13) and consists of S-regular policies].

Note a fine point here. For a given starting policy μ^0, there may be many different sequences $\{\mu^k\}$ that can be generated by PI [i.e., satisfy Eq. (3.13)]. While the weak PI property guarantees that some of these consist of S-regular policies exclusively, there may be some that do not. The policy improvement property shown in Prop. 3.2.3 holds for the former sequences, but not necessarily for the latter. The following proposition provides the basis for showing that J_S^* is a fixed point of T based on the weak PI property.

Proposition 3.2.4: (Weak PI Property Theorem) Given a set $S \subset \mathcal{E}(X)$, assume that:

(1) S has the weak PI property.

(2) For each sequence $\{J_m\} \subset S$ with $J_m \downarrow J$ for some $J \in \mathcal{E}(X)$, we have

$$H(x,u,J) = \lim_{m\to\infty} H(x,u,J_m), \qquad \forall\, x \in X,\ u \in U(x). \quad (3.16)$$

Then:

(a) J_S^* is a fixed point of T and the conclusions of Prop. 3.2.1 hold.

(b) (*PI Convergence*) Every sequence of S-regular policies $\{\mu^k\}$ that can be generated by PI satisfies $J_{\mu^k} \downarrow J_S^*$. If in addition the set of S-regular policies is finite, there exists $\bar{k} \geq 0$ such that $\mu^{\bar{k}}$ is $\mathcal{M}_S$-optimal.

Proof: (a) Let $\{\mu^k\}$ be a sequence of S-regular policies generated by the PI algorithm (there exists such a sequence by the weak PI property). Then by Prop. 3.2.3, the sequence $\{J_{\mu^k}\}$ is monotonically nonincreasing and must converge to some $J_\infty \geq J_S^*$.

We first show that J_∞ is a fixed point of T. Indeed, from Eq. (3.14),

we have

$$J_{\mu^k} \geq TJ_{\mu^k} \geq TJ_\infty,$$

so by letting $k \to \infty$, we obtain $J_\infty \geq TJ_\infty$. From Eq. (3.15) we also have $TJ_{\mu^k} \geq J_{\mu^{k+1}}$ Taking the limit in this relation as $k \to \infty$, we obtain

$$\lim_{k\to\infty} (TJ_{\mu^k})(x) \geq \lim_{k\to\infty} J_{\mu^{k+1}}(x) = J_\infty(x), \qquad x \in X.$$

By using Eq. (3.16) we also have

$$H(x, u, J_\infty) = \lim_{k\to\infty} H(x, u, J_{\mu^k}) \geq \lim_{k\to\infty} (TJ_{\mu^k})(x), \qquad x \in X,\ u \in U(x).$$

By combining the preceding two relations, we obtain

$$H(x, u, J_\infty) \geq J_\infty(x), \qquad x \in X,\ u \in U(x),$$

and by taking the infimum of the left-hand side over $u \in U(x)$, it follows that $TJ_\infty \geq J_\infty$. Thus J_∞ is a fixed point of T.

Finally, we show that $J_\infty = J_S^*$. Indeed, since $J_S^* \leq J_{\mu^k}$, we have

$$J_S^* \leq J_\infty = T^k J_\infty \leq T_\mu^k J_\infty \leq T_\mu^k J_{\mu^0}, \qquad \forall\ \mu \in \mathcal{M}_S,\ k = 0, 1, \ldots.$$

By taking the limit as $k \to \infty$, and using the fact $\mu \in \mathcal{M}_S$ and $J_{\mu^0} \in S$, it follows that $J_S^* \leq J_\infty \leq J_\mu$, for all $\mu \in \mathcal{M}_S$. By taking the infimum over $\mu \in \mathcal{M}_S$, it follows that $J_\infty = J_S^*$, so J_S^* is a fixed point of T.

(b) The limit of $\{J_{\mu^k}\}$ was shown to be equal to J_S^* in the preceding proof. Moreover, the finiteness of $\mathcal{M}_S$ and the policy improvement property of Prop. 3.2.3 imply that some $\mu^{\bar{k}}$ is $\mathcal{M}_S$-optimal. **Q.E.D.**

Note that under the weak PI property, the preceding proposition shows convergence of the PI-generated cost functions J_{μ^k} to J_S^* but not necessarily to J^*. An example of this type of behavior was seen in the linear-quadratic problem of Section 3.1.4 (where S is the set of nonnegative quadratic functions). Let us describe another example, which shows in addition that under the weak PI property, it is possible for the PI algorithm to generate a nonmonotonic sequence of policy cost functions that includes both optimal and strictly suboptimal policies.

Example 3.2.2: (Weak PI Property and the Deterministic Shortest Path Example)

Consider the deterministic shortest path example of Section 3.1.1 for the case where there is a zero length cycle ($a = 0$), and let S be the real line $\Re$, as in Example 3.2.1. There are two policies: μ which moves from state 1 to the destination at cost b, and μ' which stays at state 1 at cost 0. Starting with the S-regular policy μ, the PI algorithm generates the policy that corresponds

to the minimum in $TJ_\mu = \min\{b, J_\mu\} = \min\{b, b\}$. Thus both the S-regular policy μ and the S-irregular μ' can be generated at the first iteration. This means that the weak PI property holds (although the strong PI property, which will be introduced shortly, does not hold). Indeed, consistent with Prop. 3.2.4, we have that $J_S^* = J_\mu = b$ is a fixed point of T, in fact the only fixed point of T in the well-behaved region $\{J \mid J \geq b\}$.

An interesting fact here is that when $b < 0$, and PI is started with the optimal S-regular policy μ, then it may generate the S-irregular policy μ', and from that policy, it will generate μ again. Thus the weak PI property does not preclude the PI algorithm from generating a policy sequence that includes S-irregular policies, with corresponding policy cost functions that are oscillating.

Let us also revisit the blackmailer example of Section 3.1.3. In the first variant of that example, when $S = \Re$, all policies are S-regular, the weak PI property holds, and Prop. 3.2.4 applies. In this case, PI will generate a sequence of S-regular policies that converges to $J_S^* = -\infty$, which is a fixed point of T, consistent with Prop. 3.2.4 (even though $J_S^* \notin S$ and there is no $\mathcal{M}_S$-optimal policy).

Analysis Under the Strong PI Property

Proposition 3.2.4(a) does not guarantee that *every* sequence $\{\mu^k\}$ generated by the PI algorithm satisfies $J_{\mu^k} \downarrow J_S^*$. This is true only for the sequences that consist of S-regular policies. We know that when the weak PI property holds, there exists at least one such sequence, but PI can also generate sequences that contain S-irregular policies, even when started with an S-regular policy, as we have seen in Example 3.2.2. We thus introduce a stronger type of PI property, which will guarantee stronger conclusions.

Definition 3.2.4: (Strong PI Property) We say that a set $S \subset \mathcal{E}(X)$ has the *strong PI property* if:

(a) There exists at least one S-regular policy.

(b) For every S-regular policy μ, any policy μ' such that $T_{\mu'}J_\mu = TJ_\mu$ is S-regular, and there exists at least one such μ'.

The strong PI property implies that every sequence that can be generated by PI starting from an S-regular policy consists exclusively of S-regular policies. Moreover, there exists at least one such sequence. Hence the strong PI property implies the weak PI property. Thus if the strong PI property holds together with the mild continuity condition (2) of Prop. 3.2.4, it follows that J_S^* is a fixed point of T and Prop. 3.2.1 applies. We will see that the strong PI property implies additional results, relating to the uniqueness of the fixed point of T.

The following proposition provides conditions guaranteeing that S has the strong PI property. The salient feature of these conditions is that they preclude optimality of an S-irregular policy [see condition (4) of the proposition].

Proposition 3.2.5: (Verifying the Strong PI Property) Given a set $S \subset \mathcal{E}(X)$, assume that:

(1) $J(x) < \infty$ for all $J \in S$ and $x \in X$.

(2) There exists at least one S-regular policy.

(3) For every $J \in S$ there exists a policy μ such that $T_\mu J = TJ$.

(4) For every $J \in S$ and S-irregular policy μ, there exists a state $x \in X$ such that

$$\limsup_{k\to\infty} (T_\mu^k J)(x) = \infty. \tag{3.17}$$

Then:

(a) A policy μ satisfying $T_\mu J \leq J$ for some function $J \in S$ is S-regular.

(b) S has the strong PI property.

Proof: (a) By the monotonicity of T_μ, we have $\limsup_{k\to\infty} T_\mu^k J \leq J$, and since by condition (1), $J(x) < \infty$ for all x, it follows from Eq. (3.17) that μ is S-regular.

(b) In view of condition (3), it will suffice to show that for every S-regular policy μ, any policy μ' such that $T_{\mu'} J_\mu = TJ_\mu$ is also S-regular. Indeed

$$T_{\mu'} J_\mu = TJ_\mu \leq T_\mu J_\mu = J_\mu,$$

so μ' is S-regular by part (a). **Q.E.D.**

A representative example where the preceding proposition applies is a deterministic shortest path problem where all cycles have positive length (see the subsequent Example 3.2.3, and other examples later that involve SSP problems; see Sections 3.3 and 3.5). For an example where the assumptions of the proposition fail, consider the linear-quadratic problem of Section 3.1.4. Here S is the set of nonnegative quadratic functions, but the optimal policy μ^* that applies control $u = 0$ at all states is S-irregular, since we do not have $T_{\mu^*}^k J \to J_{\mu^*} = 0$ for J equal to a positive quadratic function, while condition (4) of the proposition does not hold. Thus we cannot conclude that the strong PI property holds in the absence of additional analysis.

We next derive some of the implications of the strong PI property regarding fixed properties of J_S^*. In particular, we show that if $J_S^* \in S$, then J_S^* is the unique fixed point of T within S. This result will be the starting point for the analysis of Section 3.3.

Proposition 3.2.6: (Strong PI Property Theorem) Let S satisfy the conditions of Prop. 3.2.5.

(a) (*Uniqueness of Fixed Point*) If T has a fixed point within S, then this fixed point is equal to J_S^*.

(b) (*Fixed Point Property and Optimality Condition*) If $J_S^* \in S$, then J_S^* is the unique fixed point of T within S and the conclusions of Prop. 3.2.1 hold. Moreover, every policy μ that satisfies $T_\mu J_S^* = TJ_S^*$ is $\mathcal{M}_S$-optimal and there exists at least one such policy.

(c) (*PI Convergence*) If for each sequence $\{J_m\} \subset S$ with $J_m \downarrow J$ for some $J \in \mathcal{E}(X)$, we have

$$H(x,u,J) = \lim_{m\to\infty} H(x,u,J_m), \qquad \forall\ x \in X,\ u \in U(x),$$

then J_S^* is a fixed point of T, and every sequence $\{\mu^k\}$ generated by the PI algorithm starting from an S-regular policy μ^0 satisfies $J_{\mu^k} \downarrow J_S^*$. Moreover, if the set of S-regular policies is finite, there exists $\bar{k} \geq 0$ such that $\mu^{\bar{k}}$ is $\mathcal{M}_S$-optimal.

Proof: (a) Let $J' \in S$ be a fixed point of T. Then for every $\mu \in \mathcal{M}_S$ and $k \geq 1$, we have $J' = T^k J' \leq T_\mu^k J'$. By taking the limit as $k \to \infty$, we have $J' \leq J_\mu$, and by taking the infimum over $\mu \in \mathcal{M}_S$, we obtain $J' \leq J_S^*$. For the reverse inequality, let μ' be such that $J' = TJ' = T_{\mu'}J'$ [cf. condition (3) of Prop. 3.2.5]. Then by Prop. 3.2.5(a), it follows that μ' is S-regular, and since $J' \in S$, by the definition of S-regularity, we have $J' = J_{\mu'} \geq J_S^*$, showing that $J' = J_S^*$.

(b) For every $\mu \in \mathcal{M}_S$ we have $J_\mu \geq J_S^*$, so that

$$J_\mu = T_\mu J_\mu \geq T_\mu J_S^* \geq TJ_S^*.$$

Taking the infimum over all $\mu \in \mathcal{M}_S$, we obtain $J_S^* \geq TJ_S^*$. Let μ be a policy such that $TJ_S^* = T_\mu J_S^*$, [there exists one by condition (3) of Prop. 3.2.5, since we assume that $J_S^* \in S$]. The preceding relations yield $J_S^* \geq T_\mu J_S^*$, so by Prop. 3.2.5(a), μ is S-regular. Therefore, we have

$$J_S^* \geq TJ_S^* = T_\mu J_S^* \geq \lim_{k\to\infty} T_\mu^k J_S^* = J_\mu \geq J_S^*,$$

where the second equality holds since μ was proved to be S-regular, and $J_S^* \in S$ by assumption. Hence equality holds throughout in the above

relation, which proves that J_S^* is a fixed point of T (implying the conclusions of Prop. 3.2.1) and that μ is $\mathcal{M}_S$-optimal.

(c) Since the strong PI property [which holds by Prop. 3.2.5(b)] implies the weak PI property, the result follows from Prop. 3.2.4(b). **Q.E.D.**

The preceding proposition does not address the question whether J^* is a fixed point of T, and does not guarantee that VI converges to J_S^* or J^* starting from every $J \in S$. We will consider both of these issues in the next section. Note, however, a consequence of part (a): if J^* is known to be a fixed point of T and $J^* \in S$, then $J^* = J_S^*$.

Let us now illustrate with examples some of the fine points of the analysis. For an example where the preceding proposition does not apply, consider the first two variants of the blackmailer problem of Section 3.1.3. Let us take $S = \Re$, so that all policies are S-regular and the strong PI property holds. In the first variant of the problem, we have $J^* = J_S^* = -\infty$, and consistent with Prop. 3.2.4, J_S^* is a fixed point of T. However, $J_S^* \notin S$, and T has no fixed points within S. On the other hand if we change S to be $[-\infty, \infty)$, there are no S-regular policies at all, since for $J = -\infty \in S$, we have $T_\mu^k J = -\infty < J_\mu$ for all μ. As noted earlier, both Props. 3.2.1 and 3.2.4 do apply. In the second variant of the problem, we have $J^* = J_S^* = -1$, while the set of fixed points of T within S is $(-\infty, -1]$, so Prop. 3.2.6(a) fails. The reason is that the condition (3) of Prop. 3.2.5 is violated.

The next example, when compared with Example 3.2.2, illustrates the difference in PI-related results obtained under the weak and the strong PI properties. Moreover it highlights a generic difficulty in applying PI, even if the strong PI property holds, namely that an initial S-regular policy must be available.

Example 3.2.3: (Strong PI Property and the Deterministic Shortest Path Example)

Consider the deterministic shortest path example of Section 3.1.1 for the case where the cycle has positive length ($a > 0$), and let S be the real line $\Re$, as in Example 3.2.1. The two policies are: μ which moves from state 1 to the destination at cost b and is S-regular, and μ' which stays at state 1 at cost a, which is S-irregular. However, μ' has infinite cost and satisfies Eq (3.17). As a result, Prop. 3.2.5 applies and the strong PI property holds. Consistent with Prop. 3.2.6, J_S^* is the unique fixed point of T within S.

Turning now to the PI algorithm, we see that starting from the S-regular μ, which is optimal, it stops at μ, consistent with Prop. 3.2.6(c). However, starting from the S-irregular policy μ' the policy evaluation portion of the PI algorithm must be able to deal with the infinite cost values associated with μ'. This is a generic difficulty in applying PI to problems where there are irregular policies: we either need to know an initial S-regular policy, or

appropriately modify the PI algorithm. See the discussions in Sections 3.5.1 and 3.6.2.

3.2.4 Optimistic Policy Iteration and λ-Policy Iteration

We have already shown the validity of the VI and PI algorithms for computing J_S^* (subject to various assumptions, and restrictions involving the starting points). In this section and the next one we will consider some additional algorithmic approaches that can be justified based on the preceding analysis.

An Optimistic Form of PI

Let us consider an optimistic variant of PI, where policies are evaluated inexactly, with a finite number of VIs. In particular, this algorithm starts with some $J_0 \in \mathcal{E}(X)$ such that $J_0 \geq TJ_0$, and generates a sequence $\{J_k, \mu^k\}$ according to

$$T_{\mu^k} J_k = TJ_k, \qquad J_{k+1} = T_{\mu^k}^{m_k} J_k, \qquad k = 0, 1, \ldots, \tag{3.18}$$

where m_k is a positive integer for each k.

The following proposition shows that optimistic PI converges under mild assumptions to a fixed point of T, independently of any S-regularity framework. However, when such a framework is introduced, and the sequence generated by optimistic PI generates a sequence of S-regular policies, then the algorithm converges to J_S^*, which is in turn a fixed point of T, similar to the PI convergence result under the weak PI property; cf. Prop. 3.2.4(b).

Proposition 3.2.7: (Convergence of Optimistic PI) Let $J_0 \in \mathcal{E}(X)$ be a function such that $J_0 \geq TJ_0$, and assume that:

(1) For all $\mu \in \mathcal{M}$, we have $J_\mu = T_\mu J_\mu$, and for all $J \in \mathcal{E}(X)$ with $J \leq J_0$, there exists $\bar{\mu} \in \mathcal{M}$ such that $T_{\bar{\mu}} J = TJ$.

(2) For each sequence $\{J_m\} \subset \mathcal{E}(X)$ with $J_m \downarrow J$ for some $J \in \mathcal{E}(X)$, we have

$$H(x, u, J) = \lim_{m \to \infty} H(x, u, J_m), \quad \forall\, x \in X,\ u \in U(x).$$

Then the optimistic PI algorithm (3.18) is well defined and the following hold:

(a) The sequence $\{J_k\}$ generated by the algorithm satisfies $J_k \downarrow J_\infty$, where J_∞ is a fixed point of T.

(b) If for a set $S \subset \mathcal{E}(X)$, the sequence $\{\mu^k\}$ generated by the algorithm consists of S-regular policies, and we have $J_k \in S$ for all k, then $J_k \downarrow J_S^*$ and J_S^* is a fixed point of T.

Proof: (a) Condition (1) guarantees that the sequence $\{J_k, \mu^k\}$ is well defined in the following argument. We have

$$\begin{aligned} J_0 \geq TJ_0 = T_{\mu^0} J_0 \geq T_{\mu^0}^{m_0} J_0 = J_1 \\ \geq T_{\mu^0}^{m_0+1} J_0 = T_{\mu^0} J_1 \geq TJ_1 = T_{\mu^1} J_1 \geq \cdots \geq J_2, \end{aligned} \tag{3.19}$$

and continuing similarly, we obtain

$$J_k \geq TJ_k \geq J_{k+1}, \qquad k = 0, 1, \ldots. \tag{3.20}$$

Thus $J_k \downarrow J_\infty$ for some J_∞.

The proof that J_∞ is a fixed point of T is similar to the case of the PI algorithm (3.13) in Prop. 3.2.4. In particular, from Eq. (3.20), we have $J_k \geq TJ_\infty$, and by taking the limit as $k \to \infty$,

$$J_\infty \geq TJ_\infty.$$

For the reverse inequality, we use Eq. (3.20) to write

$$H(x, u, J_k) \geq (TJ_k)(x) \geq J_\infty(x), \qquad \forall\ x \in X,\ u \in U(x).$$

By taking the limit as $k \to \infty$ and using condition (2), we have that

$$H(x, u, J_\infty) \geq J_\infty(x), \qquad \forall\ x \in X,\ u \in U(x).$$

By taking the infimum over $u \in U(x)$, we obtain

$$TJ_\infty \geq J_\infty,$$

thus showing that $TJ_\infty = J_\infty$.

(b) In the case where all the policies μ^k are S-regular and $\{J_k\} \subset S$, from Eq. (3.19), we have $J_{k+1} \geq J_{\mu^k}$ for all k, so it follows that

$$J_\infty = \lim_{k\to\infty} J_k \geq \liminf_{k\to\infty} J_{\mu^k} \geq J_S^*.$$

We will also show that the reverse inequality holds, so that $J_\infty = J_S^*$. Indeed, for every S-regular policy μ and all $k \geq 0$, we have

$$J_\infty = T^k J_\infty \leq T_\mu^k J_\infty \leq T_\mu^k J_0,$$

from which by taking limit as $k \to \infty$ and using the assumption $J_0 \in S$, we obtain

$$J_\infty \leq \lim_{k\to\infty} T_\mu^k J_0 = J_\mu, \qquad \forall\ \mu \in \mathcal{M}_S.$$

Taking infimum over $\mu \in \mathcal{M}_S$, we have $J_\infty \leq J_S^*$. Thus, $J_\infty = J_S^*$, and by using the properties of J_∞ proved in part (a), the result follows. **Q.E.D.**

Note that, in general, the fixed point J_∞ in Prop. 3.2.7(a) need not be equal to J_S^* or J^*. As an illustration, consider the shortest path Example 3.2.1 with $S = \Re$, and $a = 0$, $b > 0$. Then if $0 < J_0 < b$, it can be seen that $J_k = J_0$ for all k, so $J^* = 0 < J_\infty$ and $J_\infty < J_S^* = b$.

λ-Policy Iteration

We next consider λ-policy iteration (λ-PI for short), which was described in Section 2.5. It involves a scalar $\lambda \in (0, 1)$ and it is defined by

$$T_{\mu^k} J_k = TJ_k, \qquad J_{k+1} = T_{\mu^k}^{(\lambda)} J_k, \tag{3.21}$$

where for any policy μ and scalar $\lambda \in (0, 1)$, $T_\mu^{(\lambda)}$ is the multistep mapping discussed in Section 1.2.5:

$$(T_\mu^{(\lambda)} J)(x) = (1-\lambda) \sum_{t=0}^{\infty} \lambda^t (T_\mu^{t+1} J)(x), \qquad x \in X. \tag{3.22}$$

Here we assume that the limit of the series above is well-defined as a function in $\mathcal{E}(X)$ for all $x \in X$, $\mu \in \mathcal{M}$, and $J \in \mathcal{E}(X)$.

We will also assume that T_μ and $T_\mu^{(\lambda)}$ commute, i.e.,

$$T_\mu(T_\mu^{(\lambda)} J) = T_\mu^{(\lambda)}(T_\mu J), \qquad \forall\ \mu \in \mathcal{M},\ J \in \mathcal{E}(X). \tag{3.23}$$

This assumption is commonly satisfied in DP problems where T_μ is linear, such as the stochastic optimal control problem of Example 1.2.1.

To compare the λ-PI method (3.21) with the exact PI algorithm (3.13), note that by the analysis of Section 1.2.5 (see also Exercise 1.2), the mapping $T_{\mu^k}^{(\lambda)}$ is an extrapolated version of the proximal mapping for solving the fixed point equation $J = T_{\mu^k} J$. *Thus in λ-PI, the policy evaluation phase is done approximately with a single iteration of the (extrapolated) proximal algorithm.*

As noted in Section 2.5, the λ-PI and the optimistic PI methods are related. The reason is that both mappings $T_{\mu^k}^{(\lambda)}$ and $T_{\mu^k}^{m_k}$ involve multiple applications of the VI mapping T_{μ^k}: a fixed number m_k in the latter case, and a geometrically weighted infinite number in the former case [cf. Eq. (3.22)]. *Thus λ-PI and optimistic PI use VI in alternative ways to evaluate J_{μ^k} approximately.*

Since λ-PI and optimistic PI are related, it is not surprising that they have the same type of convergence properties. We have the following proposition, which is similar to Prop. 3.2.7.

Proposition 3.2.8: (Convergence of λ-PI) Let $J_0 \in \mathcal{E}(X)$ be a function such that $J_0 \geq TJ_0$, assume that the limit in the series (3.22) is well defined and Eq. (3.23) holds. Assume further that:

(1) For all $\mu \in \mathcal{M}$, we have $J_\mu = T_\mu J_\mu$, and for all $J \in \mathcal{E}(X)$ with $J \leq J_0$, there exists $\bar{\mu} \in \mathcal{M}$ such that $T_{\bar{\mu}} J = TJ$.

(2) For each sequence $\{J_m\} \subset \mathcal{E}(X)$ with $J_m \downarrow J$ for some $J \in \mathcal{E}(X)$, we have

$$H(x,u,J) = \lim_{m\to\infty} H(x,u,J_m), \quad \forall\ x \in X,\ u \in U(x).$$

Then the λ-PI algorithm (3.21) is well defined and the following hold:

(a) A sequence $\{J_k\}$ generated by the algorithm satisfies $J_k \downarrow J_\infty$, where J_∞ is a fixed point of T.

(b) If for a set $S \subset \mathcal{E}(X)$, the sequence $\{\mu^k\}$ generated by the algorithm consists of S-regular policies, and we have $J_k \in S$ for all k, then $J_k \downarrow J_S^*$ and J_S^* is a fixed point of T.

Proof: (a) We first note that for all $\mu \in \mathcal{M}$ and $J \in \mathcal{E}(X)$ such that $J \geq T_\mu J$, we have

$$T_\mu J \geq T_\mu^{(\lambda)} J.$$

This follows from the power series expansion (3.22) and the fact that $J \geq T_\mu J$ implies that

$$T_\mu J \geq T_\mu^2 J \geq \cdots \geq T_\mu^{m+1} J, \qquad \forall\ m \geq 1.$$

Using also the monotonicity of T_μ and $T_\mu^{(\lambda)}$, and Eq. (3.23), we have that

$$J \geq T_\mu J \qquad \Rightarrow \qquad T_\mu J \geq T_\mu^{(\lambda)} J \geq T_\mu^{(\lambda)}(T_\mu J) = T_\mu(T_\mu^{(\lambda)} J).$$

The preceding relation and our assumptions imply that

$$\begin{aligned} J_0 \geq TJ_0 = T_{\mu_0} J_0 \geq T_{\mu_0}^{(\lambda)} J_0 = J_1 \\ \geq T_{\mu^0}(T_{\mu_0}^{(\lambda)} J_0) = T_{\mu^0} J_1 \geq TJ_1 = T_{\mu^1} J_1 \geq \cdots \geq J_2. \end{aligned}$$

Continuing similarly, we obtain $J_k \geq TJ_k \geq J_{k+1}$ for all k. Thus $J_k \downarrow J_\infty$ for some J_∞. From this point, the proof that J_∞ is a fixed point of T is similar to the one of Prop. 3.2.7(a).

(b) Similar to the proof of Prop. 3.2.7(b). **Q.E.D.**

3.2.5 A Mathematical Programming Approach

Let us finally consider an alternative to the VI and PI approaches. It is based on the fact that J_S^* is an upper bound to all functions $J \in S$ that satisfy $J \leq TJ$, as we will show shortly. We will exploit this fact to obtain a method to compute J_S^* that is based on solution of a related mathematical programming problem. We have the following proposition.

Proposition 3.2.9: Given a set $S \subset \mathcal{E}(X)$, for all functions $J \in S$ satisfying $J \leq TJ$, we have $J \leq J_S^*$.

Proof: If $J \in S$ and $J \leq TJ$, by repeatedly applying T to both sides and using the monotonicity of T, we obtain $J \leq T^k J \leq T_\mu^k J$ for all k and S-regular policies μ. Taking the limit as $k \to \infty$, we obtain $J \leq J_\mu$, so by taking the infimum over $\mu \in \mathcal{M}_S$, we obtain $J \leq J_S^*$. **Q.E.D.**

Thus if J_S^* is a fixed point of T, it is the "largest" fixed point of T, and we can use the preceding proposition to compute J_S^* by maximizing an appropriate monotonically increasing function of J subject to the constraints $J \in S$ and $J \leq TJ$. † This approach, when applied to finite-spaces Markovian decision problems, is usually referred to as the *linear programming solution method*, since then the resulting optimization problem is a linear program (see e.g., see Exercise 2.5 for the case of contractive problems or [Ber12a], Ch. 2).

Suppose now that $X = \{1, \ldots, n\}$, $S = \Re^n$, and J_S^* is a fixed point of T. Then Prop. 3.2.9 shows that $J_S^* = \big(J_S^*(1), \ldots, J_S^*(n)\big)$ is the unique solution of the following optimization problem in the vector $J = \big(J(1), \ldots, J(n)\big)$:

$$\begin{aligned} &\text{maximize} \quad \sum_{i=1}^{n} \beta_i J(i) \\ &\text{subject to} \quad J(i) \leq H(i, u, J), \quad i = 1, \ldots, n, \quad u \in U(i), \end{aligned}$$

where $\beta_1, \ldots, \beta_n$ are any positive scalars. If H is linear in J and each $U(i)$ is a finite set, this is a linear program, which can be solved by using standard linear programming methods.

† For the mathematical programming approach to apply, it is sufficient that $J_S^* \leq TJ_S^*$. However, we generally have $J_S^* \geq TJ_S^*$ (this follows by writing

$$J_\mu = T_\mu J_\mu \geq TJ_\mu \geq TJ_S^*, \qquad \forall\, \mu \in \mathcal{M}_S,$$

and taking the infimum over all $\mu \in \mathcal{M}_S$), so the condition $J_S^* \leq TJ_S^*$ is equivalent to J_S^* being a fixed point of T.

3.3 IRREGULAR POLICIES/INFINITE COST CASE

The results of the preceding section guarantee (under various conditions) that J_S^* is a fixed point of T, and can be found by the VI and PI algorithms, but they do not assert that J^* is a fixed point of T or that $J^* = J_S^*$. In this section we address these issues by carrying the strong PI property analysis further with some additional assumptions. A critical part of the analysis is based on the strong PI property theorem of Prop. 3.2.6. We first collect all of our assumptions. We will verify these assumptions in the context of several applications in Section 3.5.

Assumption 3.3.1: We have a subset $S \subset \mathcal{R}(X)$ satisfying the following:

(a) S contains $\bar{J}$, and has the property that if J_1, J_2 are two functions in S, then S contains all functions J with $J_1 \leq J \leq J_2$.

(b) The function $J_S^* = \inf_{\mu \in \mathcal{M}_S} J_\mu$ belongs to S.

(c) For each S-irregular policy μ and each $J \in S$, there is at least one state $x \in X$ such that

$$\limsup_{k \to \infty} (T_\mu^k J)(x) = \infty.$$

(d) The control set U is a metric space, and the set

$$\big\{u \in U(x) \mid H(x,u,J) \leq \lambda\big\}$$

is compact for every $J \in S$, $x \in X$, and $\lambda \in \Re$.

(e) For each sequence $\{J_m\} \subset S$ with $J_m \uparrow J$ for some $J \in S$,

$$\lim_{m \to \infty} H(x,u,J_m) = H\left(x,u,J\right), \qquad \forall\ x \in X,\ u \in U(x).$$

(f) For each function $J \in S$, there exists a function $J' \in S$ such that $J' \leq J$ and $J' \leq TJ'$.

An important restriction of the preceding assumption is that S *consists of real-valued functions*. This underlies the mechanism of differentiating between S-regular and S-irregular policies that is embodied in Assumption 3.3.1(c).

The conditions (b) and (c) of the preceding assumption have been introduced in Props. 3.2.5 and 3.2.6 in the context of the strong PI property-related analysis. New conditions, not encountered earlier, are (a), (e), and

(f). They will be used to assert that $J^* = J_S^*$, that J^* is the unique fixed point of T within S, and that the VI and PI algorithms have improved convergence properties compared with the ones of Section 3.2.

Note that in the case where S is the set of real-valued functions $\mathcal{R}(X)$ and $\bar{J} \in \mathcal{R}(X)$, condition (a) is automatically satisfied, while condition (e) is typically verified easily. The verification of condition (f) may be nontrivial in some cases. We postpone the discussion of this issue for later (see the subsequent Prop. 3.3.2).

The main result of this section is the following proposition, which provides results that are almost as strong as the ones for contractive models.

Proposition 3.3.1: Let Assumption 3.3.1 hold. Then:

(a) The optimal cost function J^* is the unique fixed point of T within the set S.

(b) We have $T^k J \to J^*$ for all $J \in S$.

(c) A policy μ is optimal if and only if $T_\mu J^* = TJ^*$. Moreover, there exists an optimal policy that is S-regular.

(d) For any $J \in S$, if $J \leq TJ$ we have $J \leq J^*$, and if $J \geq TJ$ we have $J \geq J^*$.

(e) If in addition for each sequence $\{J_m\} \subset S$ with $J_m \downarrow J$ for some $J \in S$, we have

$$H(x,u,J) = \lim_{m\to\infty} H(x,u,J_m), \qquad \forall\ x \in X,\ u \in U(x),$$

then every sequence $\{\mu^k\}$ generated by the PI algorithm starting from an S-regular policy μ^0 satisfies $J_{\mu^k} \downarrow J^*$. Moreover, if the set of S-regular policies is finite, there exists $\bar{k} \geq 0$ such that $\mu^{\bar{k}}$ is optimal.

We will prove Prop. 3.3.1 through a sequence of lemmas, which delineate the assumptions that are needed for each part of the proof. Our first lemma guarantees that starting from an S-regular policy, the PI algorithm is well defined.

Lemma 3.3.1: Let Assumption 3.3.1(d) hold. For every $J \in S$, there exists a policy μ such that $T_\mu J = TJ$.

Proof: For any $x \in X$ with $(TJ)(x) < \infty$, let $\{\lambda_m(x)\}$ be a decreasing

scalar sequence with

$$\lambda_m(x) \downarrow \inf_{u \in U(x)} H(x,u,J).$$

The set

$$U_m(x) = \big\{u \in U(x) \mid H(x,u,J) \leq \lambda_m(x)\big\},$$

is nonempty, and by assumption it is compact. The set of points attaining the infimum of $H(x,u,J)$ over $U(x)$ is $\cap_{m=0}^{\infty} U_m(x)$, and is therefore nonempty. Let u_x be a point in this intersection. Then we have

$$H(x,u_x,J) \leq \lambda_m(x), \qquad \forall\ m \geq 0. \tag{3.24}$$

Consider now a policy μ, which is formed by the point u_x for x with $(TJ)(x) < \infty$, and by any point $u_x \in U(x)$ for x with $(TJ)(x) = \infty$. Taking the limit in Eq. (3.24) as $m \to \infty$ shows that μ satisfies $(T_\mu J)(x) = (TJ)(x)$ for x with $(TJ)(x) < \infty$. For x with $(TJ)(x) = \infty$, we also have trivially $(T_\mu J)(x) = (TJ)(x)$, so $T_\mu J = TJ$. **Q.E.D.**

The next two lemmas follow from the analysis of the preceding section.

Lemma 3.3.2: Let Assumption 3.3.1(c) hold. A policy μ that satisfies $T_\mu J \leq J$ for some $J \in S$ is S-regular.

Proof: This is Prop. 3.2.5(a). **Q.E.D.**

Lemma 3.3.3: Let Assumption 3.3.1(b),(c),(d) hold. Then:

(a) The function J_S^* of Assumption 3.3.1(b) is the unique fixed point of T within S.

(b) Every policy μ satisfying $T_\mu J_S^* = TJ_S^*$ is optimal within the set of S-regular policies, i.e., μ is S-regular and $J_\mu = J_S^*$. Moreover, there exists at least one such policy.

Proof: This is Prop. 3.2.6(b) [Assumption 3.3.1(d) guarantees that for every $J \in S$, there exists a policy μ such that $T_\mu J = TJ$ (cf. Lemma 3.3.1), which is part of the assumptions of Prop. 3.2.6]. **Q.E.D.**

Let us also prove the following technical lemma, which makes use of the additional part (e) of Assumption 3.3.1.

Lemma 3.3.4: Let Assumption 3.3.1(b),(c),(d),(e) hold. Then if $J \in S$, $\{T^k J\} \subset S$, and $T^k J \uparrow J_\infty$ for some $J_\infty \in S$, we have $J_\infty = J_S^*$.

Proof: We fix $x \in X$, and consider the sets

$$U_k(x) = \Big\{ u \in U(x) \mid H(x, u, T^k J) \leq J_\infty(x) \Big\}, \qquad k = 0, 1, \ldots, \quad (3.25)$$

which are compact by assumption. Let $u_k \in U(x)$ be such that

$$H(x, u_k, T^k J) = \inf_{u \in U(x)} H(x, u, T^k J) = (T^{k+1} J)(x) \leq J_\infty(x)$$

(such a point exists by Lemma 3.3.1). Then $u_k \in U_k(x)$.

For every k, consider the sequence $\{u_i\}_{i=k}^\infty$. Since $T^k J \uparrow J_\infty$, it follows using the monotonicity of H, that for all $i \geq k$,

$$H(x, u_i, T^k J) \leq H(x, u_i, T^i J) \leq J_\infty(x).$$

Therefore from the definition (3.25), we have $\{u_i\}_{i=k}^\infty \subset U_k(x)$. Since $U_k(x)$ is compact, all the limit points of $\{u_i\}_{i=k}^\infty$ belong to $U_k(x)$ and at least one limit point exists. Hence the same is true for the limit points of the whole sequence $\{u_i\}$. Thus if $\tilde{u}$ is a limit point of $\{u_i\}$, we have

$$\tilde{u} \in \cap_{k=0}^\infty U_k(x).$$

By Eq. (3.25), this implies that

$$H\big(x, \tilde{u}, T^k J\big) \leq J_\infty(x), \qquad k = 0, 1, \ldots.$$

Taking the limit as $k \to \infty$ and using Assumption 3.3.1(e), we obtain

$$(T J_\infty)(x) \leq H(x, \tilde{u}, J_\infty) \leq J_\infty(x).$$

Thus, since x was chosen arbitrarily within X, we have $T J_\infty \leq J_\infty$. To show the reverse inequality, we write $T^k J \leq J_\infty$, apply T to this inequality, and take the limit as $k \to \infty$, so that $J_\infty = \lim_{k \to \infty} T^{k+1} J \leq T J_\infty$. It follows that $J_\infty = T J_\infty$. Since $J_\infty \in S$ by assumption, by applying Lemma 3.3.3(a) we have $J_\infty = J_S^*$. **Q.E.D.**

We are now ready to prove Prop. 3.3.1 by making use of the additional parts (a) and (f) of Assumption 3.3.1.

Proof of Prop. 3.3.1: (a), (b) We will first prove that $T^k J \to J_S^*$ for all $J \in S$, and we will use this to prove that $J_S^* = J^*$ and that there exists

an optimal S-regular policy. Thus parts (a) and (b), together with the existence of an optimal S-regular policy, will be shown simultaneously.

We fix $J \in S$, and choose $J' \in S$ such that $J' \leq J$ and $J' \leq TJ'$ [cf. Assumption 3.3.1(f)]. By the monotonicity of T, we have $T^k J' \uparrow J_\infty$ for some $J_\infty \in \mathcal{E}(X)$. Let μ be an S-regular policy such that $J_\mu = J_S^*$ [cf. Lemma 3.3.3(b)]. Then we have, using again the monotonicity of T,

$$J_\infty = \lim_{k\to\infty} T^k J' \leq \limsup_{k\to\infty} T^k J \leq \lim_{k\to\infty} T_\mu^k J = J_\mu = J_S^*. \tag{3.26}$$

Since J' and J_S^* belong to S, and $J' \leq T^k J' \leq J_\infty \leq J_S^*$, Assumption 3.3.1(a) implies that $\{T^k J'\} \subset S$, and $J_\infty \in S$. From Lemma 3.3.4, it then follows that $J_\infty = J_S^*$. Thus equality holds throughout in Eq. (3.26), proving that $\lim_{k\to\infty} T^k J = J_S^*$.

There remains to show that $J_S^* = J^*$ and that there exists an optimal S-regular policy. To this end, we note that by the monotonicity Assumption 3.2.1, for any policy $\pi = \{\mu_0, \mu_1, \ldots\}$, we have

$$T_{\mu_0} \cdots T_{\mu_{k-1}} \bar{J} \geq T^k \bar{J}.$$

Taking the limit of both sides as $k \to \infty$, we obtain

$$J_\pi \geq \lim_{k\to\infty} T^k \bar{J} = J_S^*,$$

where the equality follows since $T^k J \to J_S^*$ for all $J \in S$ (as shown earlier), and $\bar{J} \in S$ [cf. Assumption 3.3.1(a)]. Thus for all $\pi \in \Pi$, $J_\pi \geq J_S^* = J_\mu$, implying that the policy μ that is optimal within the class of S-regular policies is optimal over all policies, and that $J_S^* = J^*$.

(c) If μ is optimal, then $J_\mu = J^* \in S$, so by Assumption 3.3.1(c), μ is S-regular and therefore $T_\mu J_\mu = J_\mu$. Hence,

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

Conversely, if

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

μ is S-regular (cf. Lemma 3.3.2), so $J^* = \lim_{k\to\infty} T_\mu^k J^* = J_\mu$. Therefore, μ is optimal.

(d) If $J \in S$ and $J \leq TJ$, by repeatedly applying T to both sides and using the monotonicity of T, we obtain $J \leq T^k J$ for all k. Taking the limit as $k \to \infty$ and using the fact $T^k J \to J^*$ [cf. part (b)], we obtain $J \leq J^*$. The proof that $J \geq TJ$ implies $J \geq J^*$ is similar.

(e) As in the proof of Prop. 3.2.4(b), the sequence $\{J_{\mu^k}\}$ converges monotonically to a fixed point of T, call it J_∞. Since J_∞ lies between $J_{\mu^0} \in S$ and $J_S^* \in S$, it must belong to S, by Assumption 3.3.1(a). Since the only

fixed point of T within S is J^* [cf. part (a)], it follows that $J_\infty = J^*$. **Q.E.D.**

Note that Prop. 3.3.1(d) provides the basis for a solution method based on mathematical programming; cf. the discussion following Prop. 3.2.9. Here is an example where Prop. 3.3.1 does not apply, because the compactness condition of Assumption 3.3.1(d) fails.

Example 3.3.1

Consider the third variant of the blackmailer problem (Section 3.1.3) for the case where $c > 0$ and $S = \Re$. Then the (nonoptimal) S-irregular policy $\bar{\mu}$ whereby at each period, the blackmailer may demand no payment ($u = 0$) and pay cost $c > 0$, has infinite cost ($J_{\bar{\mu}} = \infty$). However, T has multiple fixed points within the real line, namely the set $(-\infty, -1]$. By choosing $S = \Re$, we see that the uniqueness of fixed point part (a) of Prop. 3.3.1 fails because the compactness part (d) of Assumption 3.3.1 is violated (all other parts of the assumption are satisfied). In this example, the results of Prop. 3.2.1 apply with $S = \Re$, because J_S^* is a fixed point of T.

In various applications, the verification of part (f) of Assumption 3.3.1 may not be simple. The following proposition is useful in several contexts, including some that we will encounter in Section 3.5.

Proposition 3.3.2: Let S be equal to $R_b(X)$, the subset of $\mathcal{R}(X)$ that consists of functions J that are bounded above and below, in the sense that for some $b \in \Re$, we have $\big|J(x)\big| \leq b$ for all $x \in X$. Let parts (b), (c), and (d) of Assumption 3.3.1 hold, and assume further that for all scalars $r > 0$, we have

$$TJ_S^* - re \leq T(J_S^* - re), \tag{3.27}$$

where e is the unit function, $e(x) \equiv 1$. Then part (f) of Assumption 3.3.1 also holds.

Proof: Let $J \in R_b(x)$, and let $r > 0$ be a scalar such that $J_S^* - re \leq J$ [such a scalar exists since $J_S^* \in R_b(x)$ by Assumption 3.3.1(b)]. Define $J' = J_S^* - re$, and note that by Lemma 3.3.3, J_S^* is a fixed point of T. By using Eq. (3.27), we have

$$J' = J_S^* - re = TJ_S^* - re \leq T(J_S^* - re) = TJ',$$

while $J' \in R_b(x)$, thus proving part (f) of Assumption 3.3.1. **Q.E.D.**

The relation (3.27) is satisfied among others in stochastic optimal control problems (cf. Example 1.2.1), where

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

with $\alpha \in (0,1]$. Note that application of the preceding proposition is facilitated when X is a finite set, in which case $R_b(X) = \mathcal{R}(X)$. This fact will be used in the context of some of the applications of Sections 3.5.1-3.5.4.

3.4 IRREGULAR POLICIES/FINITE COST CASE - A PERTURBATION APPROACH

In this section, we address problems where some S-irregular policies may have finite cost for all states [thus violating Assumption 3.3.1(c)], so Prop. 3.3.1 cannot be used. Our approach instead will be to assert that J_S^* is a fixed point of T, so that Prop. 3.2.1 applies and can be used to guarantee convergence of VI to J_S^* starting from $J_0 \geq J_S^*$.

Our line of analysis is quite different from the one of Sections 3.2.3 and 3.3, which was based on PI ideas. Instead, we *add a perturbation to the mapping* H, designed to provide adequate differentiation between S-regular and S-irregular policies. Using a limiting argument, as the size of the perturbation diminishes to 0, we are able to prove that J_S^* is a fixed point of T. Moreover, we provide a perturbation-based PI algorithm that may be more reliable than the standard PI algorithm, which can fail for problems where irregular policies may have finite cost for all states; cf. Example 3.2.2. We will also use the perturbation approach in Sections 4.5 and 4.6, where we will extend the notion of S-regularity to nonstationary policies that do not lend themselves to a PI-based analysis.

An example where the approach of this section will be shown to apply is an SSP problem where Assumption 3.3.1 is violated while $J^*(x) > -\infty$ for all x (see also Section 3.5.1). Here is a classical problem of this type.

Example 3.4.1 (Search Problem)

Consider a situation where the objective is to move within a finite set of states searching for a state to stop while minimizing the expected cost. We formulate this as a DP problem with finite state space X, and two controls at each $x \in X$: *stop*, which yields an immediate cost $s(x)$, and *continue*, in which case we move to a state $f(x,w)$ at cost $g(x,w)$, where w is a random variable with given distribution that may depend on x. The mapping H is

$$H(x,u,J) = \begin{cases} s(x) & \text{if } u = \text{stop}, \\ E\big\{g(x,w) + J\big(f(x,w)\big)\big\} & \text{if } u = \text{continue}, \end{cases}$$

and the function $\bar{J}$ is identically 0.

Letting $S = \mathcal{R}(X)$, we note that the policy $\overline{\mu}$ that stops nowhere is S-irregular, since $T_{\overline{\mu}}$ cannot have a unique fixed point within S (adding any unit function multiple to J adds to $T_{\overline{\mu}}J$ the same multiple). This policy may violate Assumption 3.3.1(c) of the preceding section, because its cost may be

finite for all states. A special case where this occurs is when $g(x,w) \equiv 0$ for all x. Then the cost function of $\overline{\mu}$ is identically 0.

Note that case (b) of the deterministic shortest path problem of Section 3.1.1, which involves a zero length cycle, is a special case of the search problem just described. Therefore, the anomalous behavior we saw there (nonconvergence of VI to J^* and oscillation of PI; cf. Examples 3.2.1 and 3.2.2) may also arise in the context of the present example. We will see that by adding a small positive constant to the length of the cycle we can rectify the difficulties of VI and PI, at least partially; this is the idea behind the perturbation approach that we will use in this section.

We will address the finite cost issue for irregular policies by introducing a perturbation that makes their cost infinite for some states. We can then use Prop. 3.3.1 of the preceding section. The idea is that with a perturbation, the cost functions of S-irregular policies may increase disproportionately relative to the cost functions of the S-regular policies, thereby making the problem more amenable to analysis.

We introduce a nonnegative "forcing function" $p : X \mapsto [0,\infty)$, and for each $\delta > 0$ and policy μ, we consider the mappings

$$(T_{\mu,\delta}J)(x) = H\big(x,\mu(x),J\big) + \delta p(x), \quad x \in X, \qquad T_\delta J = \inf_{\mu\in\mathcal{M}} T_{\mu,\delta}J.$$

We refer to the problem associated with the mappings $T_{\mu,\delta}$ as the δ-*perturbed problem*. The cost functions of policies $\pi = \{\mu_0,\mu_1,\ldots\} \in \Pi$ and $\mu \in \mathcal{M}$ for this problem are

$$J_{\pi,\delta} = \limsup_{k\to\infty} T_{\mu_0,\delta}\cdots T_{\mu_k,\delta}\bar{J}, \qquad J_{\mu,\delta} = \limsup_{k\to\infty} T^k_{\mu,\delta}\bar{J},$$

and the optimal cost function is $\hat{J}_\delta = \inf_{\pi\in\Pi} J_{\pi,\delta}$.

The following proposition shows that if the δ-perturbed problem is "well-behaved" with respect to a subset of S-regular policies, then its cost function $\hat{J}_\delta$ can be used to approximate the optimal cost function over this subset of policies only. Moreover J^*_S is a fixed point of T. Note that the unperturbed problem need not be as well-behaved, and indeed J^* need not be a fixed point of T.

Proposition 3.4.1: Given a set $S \subset \mathcal{E}(X)$, let $\widehat{\mathcal{M}}$ be a subset of S-regular policies, and let $\hat{J}$ be the optimal cost function over the policies in $\widehat{\mathcal{M}}$ only, i.e.,

$$\hat{J} = \inf_{\mu\in\widehat{\mathcal{M}}} J_\mu.$$

Assume that for every $\delta > 0$:

(1) The optimal cost function $\hat{J}_\delta$ of the δ-perturbed problem satisfies the corresponding Bellman equation $\hat{J}_\delta = T_\delta\hat{J}_\delta$.

(2) We have $\inf_{\mu\in\widehat{\mathcal{M}}} J_{\mu,\delta} = \hat{J}_\delta$, i.e., for every $x \in X$ and $\epsilon > 0$, there exists a policy $\mu_{x,\epsilon} \in \widehat{\mathcal{M}}$ such that $J_{\mu_{x,\epsilon},\delta}(x) \le \hat{J}_\delta(x) + \epsilon$.

(3) For every $\mu \in \widehat{\mathcal{M}}$, we have

$$J_{\mu,\delta} \le J_\mu + w_{\mu,\delta},$$

where $w_{\mu,\delta}$ is a function such that $\lim_{\delta\downarrow 0} w_{\mu,\delta} = 0$.

(4) For every sequence $\{J_m\} \subset S$ with $J_m \downarrow J$, we have

$$\lim_{m\to\infty} H(x,u,J_m) = H(x,u,J), \qquad \forall\ x \in X,\ u \in U(x).$$

Then J_S^* is a fixed point of T and the conclusions of Prop. 3.2.1 hold. Moreover, we have

$$J_S^* = \hat{J} = \lim_{\delta\downarrow 0} \hat{J}_\delta.$$

Proof: For every $x \in X$, using conditions (2) and (3), we have for all $\delta > 0$, $\epsilon > 0$, and $\mu \in \widehat{\mathcal{M}}$,

$$\hat{J}(x) - \epsilon \le J_{\mu_{x,\epsilon}}(x) - \epsilon \le J_{\mu_{x,\epsilon},\delta}(x) - \epsilon \le \hat{J}_\delta(x) \le J_{\mu,\delta}(x) \le J_\mu(x) + w_{\mu,\delta}(x).$$

By taking the limit as $\epsilon \downarrow 0$, we obtain for all $\delta > 0$ and $\mu \in \widehat{\mathcal{M}}$,

$$\hat{J} \le \hat{J}_\delta \le J_{\mu,\delta} \le J_\mu + w_{\mu,\delta}.$$

By taking the limit as $\delta \downarrow 0$ and then the infimum over all $\mu \in \widehat{\mathcal{M}}$, it follows [using also condition (3)] that

$$\hat{J} \le \lim_{\delta\downarrow 0} \hat{J}_\delta \le \inf_{\mu\in\widehat{\mathcal{M}}} \lim_{\delta\downarrow 0} J_{\mu,\delta} \le \inf_{\mu\in\widehat{\mathcal{M}}} J_\mu = \hat{J},$$

so that $\hat{J} = \lim_{\delta\downarrow 0} \hat{J}_\delta$.

Next we prove that $\hat{J}$ is a fixed point of T and use this fact to show that $\hat{J} = J_S^*$, thereby concluding the proof. Indeed, from condition (1) and the fact $\hat{J}_\delta \ge \hat{J}$ shown earlier, we have for all $\delta > 0$,

$$\hat{J}_\delta = T_\delta \hat{J}_\delta \ge T\hat{J}_\delta \ge T\hat{J},$$

and by taking the limit as $\delta \downarrow 0$ and using part (a), we obtain $\hat{J} \ge T\hat{J}$. For the reverse inequality, let $\{\delta_m\}$ be a sequence with $\delta_m \downarrow 0$. Using condition (1) we have for all m,

$$H(x,u,\hat{J}_{\delta_m}) + \delta_m p(x) \ge (T_{\delta_m}\hat{J}_{\delta_m})(x) = \hat{J}_{\delta_m}(x), \quad \forall\ x \in X,\ u \in U(x).$$

Taking the limit as $m \to \infty$, and using condition (4) and the fact $\hat{J}_{\delta_m} \downarrow \hat{J}$ shown earlier, we have

$$H(x, u, \hat{J}) \geq \hat{J}(x), \qquad \forall\ x \in X,\ u \in U(x),$$

so that $T\hat{J} \geq \hat{J}$. Thus $\hat{J}$ is a fixed point of T.

Finally, to show that $\hat{J} = J_S^*$, we first note that $J_S^* \leq \hat{J}$ since every policy in $\widehat{\mathcal{M}}$ is S-regular. For the reverse inequality, let μ be S-regular. We have $\hat{J} = T\hat{J} \leq T_\mu \hat{J} \leq T_\mu^k \hat{J}$ for all $k \geq 1$, so that for all $\mu' \in \widehat{\mathcal{M}}$,

$$\hat{J} \leq \lim_{k\to\infty} T_\mu^k \hat{J} \leq \lim_{k\to\infty} T_\mu^k J_{\mu'} = J_\mu,$$

where the equality follows since μ and μ' are S-regular (so $J_{\mu'} \in S$). Taking the infimum over all S-regular μ, we obtain $\hat{J} \leq J_S^*$, so that $J_S^* = \hat{J}$. **Q.E.D.**

Aside from S-regularity of the set $\widehat{\mathcal{M}}$, a key assumption of the preceding proposition is that $\inf_{\mu\in\widehat{\mathcal{M}}} J_{\mu,\delta} = \hat{J}_\delta$, i.e., that with a perturbation added, the subset of policies $\widehat{\mathcal{M}}$ is sufficient (the optimal cost of the δ-perturbed problem can be achieved using the policies in $\widehat{\mathcal{M}}$). This is the key insight to apply when selecting $\widehat{\mathcal{M}}$.

Note that the preceding proposition applies even if

$$\lim_{\delta\downarrow 0} \hat{J}_\delta(x) > J^*(x)$$

for some $x \in X$. This is illustrated by the deterministic shortest path example of Section 3.1.1, for the zero-cycle case where $a = 0$ and $b > 0$. Then for $S = \Re$, we have $J_S^* = b > 0 = J^*$, while the proposition applies because its assumptions are satisfied with $p(x) \equiv 1$. Consistently with the conclusions of the proposition, we have $\hat{J}_\delta = b + \delta$, so $J_S^* = \hat{J} = \lim_{\delta\downarrow 0} \hat{J}_\delta$ and J_S^* is a fixed point of T.

Proposition 3.4.1 also applies to Example 3.4.1. In particular, it can be used to assert that J_S^* is a fixed point of T, and hence also that the conclusions of Prop. 3.2.1 hold. These conclusions imply that J_S^* is the unique fixed point of T within the set $\{J \mid J \geq J_S^*\}$ and that the VI algorithm converges to J_S^* starting from within this set.

We finally note that while Props. 3.3.1 and 3.4.1 relate to qualitatively different problems, they can often be used synergistically. In particular, Prop. 3.3.1 may be applied to the δ-perturbed problem in order to verify the assumptions of Prop. 3.4.1.

A Policy Iteration Algorithm with Perturbations

We now consider a subset $\widehat{\mathcal{M}}$ of S-regular policies, and introduce a version of the PI algorithm that uses perturbations and generates a sequence $\{\mu^k\} \subset \widehat{\mathcal{M}}$ such that $J_{\mu^k} \to J_S^*$. We assume the following.

Assumption 3.4.1: The subset of S-regular policies $\widehat{\mathcal{M}}$ is such that:

(a) The conditions of Prop. 3.4.1 are satisfied.

(b) Every policy $\mu \in \widehat{\mathcal{M}}$ is S-regular for all the δ-perturbed problems, $\delta > 0$.

(c) Given a policy $\mu \in \widehat{\mathcal{M}}$ and a scalar $\delta > 0$, every policy μ' such that
$$T_{\mu'} J_{\mu,\delta} = T J_{\mu,\delta}$$
belongs to $\widehat{\mathcal{M}}$, and at least one such policy exists.

The perturbed version of the PI algorithm is defined as follows. Let $\{\delta_k\}$ be a positive sequence with $\delta_k \downarrow 0$, and let μ^0 be a policy in $\widehat{\mathcal{M}}$. At iteration k, we have a policy $\mu^k \in \widehat{\mathcal{M}}$, and we generate $\mu^{k+1} \in \widehat{\mathcal{M}}$ according to
$$T_{\mu^{k+1}} J_{\mu^k,\delta_k} = T J_{\mu^k,\delta_k}. \tag{3.28}$$
Note that by Assumption 3.4.1(c) the algorithm is well-defined, and is guaranteed to generate a sequence of policies $\{\mu^k\} \subset \widehat{\mathcal{M}}$. We have the following proposition.

Proposition 3.4.2: Let Assumption 3.4.1 hold. Then J_S^* is a fixed point of T and for a sequence of S-regular policies $\{\mu^k\}$ generated by the perturbed PI algorithm (3.28), we have $J_{\mu^k,\delta_k} \downarrow J_S^*$ and $J_{\mu^k} \to J_S^*$.

Proof: We have that J_S^* is a fixed point of T by Prop. 3.4.1. The algorithm definition (3.28) implies that for all $m \geq 1$ we have
$$T^m_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k} \leq T_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k} = T J_{\mu^k,\delta_k} + \delta_k\, p \leq J_{\mu^k,\delta_k}.$$
From this relation it follows that
$$J_{\mu^{k+1},\delta_{k+1}} \leq J_{\mu^{k+1},\delta_k} = \lim_{m\to\infty} T^m_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k} \leq J_{\mu^k,\delta_k},$$
where the equality holds because μ^{k+1} and μ^k are S-regular for all the δ-perturbed problems. It follows that $\{J_{\mu^k,\delta_k}\}$ is monotonically nonincreasing, so that $J_{\mu^k,\delta_k} \downarrow J_\infty$ for some J_∞. Moreover, we must have $J_\infty \geq J_S^*$ since $J_{\mu^k,\delta_k} \geq J_{\mu^k} \geq J_S^*$. Thus
$$J_S^* \leq J_\infty = \lim_{k\to\infty} T J_{\mu^k,\delta_k}. \tag{3.29}$$

We also have

$$\begin{aligned}\inf_{u\in U(x)} H(x,u,J_\infty) &\le \lim_{k\to\infty} \inf_{u\in U(x)} H\big(x,u,J_{\mu^k,\delta_k}\big)\\ &\le \inf_{u\in U(x)} \lim_{k\to\infty} H\big(x,u,J_{\mu^k,\delta_k}\big)\\ &= \inf_{u\in U(x)} H\big(x,u,\lim_{k\to\infty} J_{\mu^k,\delta_k}\big)\\ &= \inf_{u\in U(x)} H(x,u,J_\infty),\end{aligned}$$

where the first inequality follows from the fact $J_\infty \le J_{\mu^k,\delta_k}$, which implies that $H(x,u,J_\infty) \le H\big(x,u,J_{\mu^k,\delta_k}\big)$, and the first equality follows from the continuity property that is assumed in Prop. 3.4.1. Thus equality holds throughout above, so that

$$\lim_{k\to\infty} TJ_{\mu^k,\delta_k} = TJ_\infty. \tag{3.30}$$

Combining Eqs. (3.29) and (3.30), we obtain $J_S^* \le J_\infty = TJ_\infty$. By replacing $\hat{J}$ with J_∞ in the last part of the proof of Prop. 3.4.1, we obtain $J_S^* = J_\infty$. Thus $J_{\mu^k,\delta_k} \downarrow J_S^*$, which in view of the fact $J_{\mu^k,\delta_k} \ge J_{\mu^k} \ge J_S^*$, implies that $J_{\mu^k} \to J_S^*$. **Q.E.D.**

When the control space U is finite, Prop. 3.4.2 also implies that the generated policies μ^k will be optimal for all k sufficiently large. The reason is that the set of policies is finite and there exists a sufficiently small $\epsilon > 0$, such that for all nonoptimal μ there is some state x such that $J_\mu(x) \ge \hat{J}(x)+\epsilon$. This convergence behavior should be contrasted with the behavior of PI without perturbations, which may lead to oscillations, as noted earlier.

However, when the control space U is infinite, the generated sequence $\{\mu^k\}$ may exhibit some serious pathologies in the limit. If $\{\mu^k\}_{\mathcal{K}}$ is a subsequence of policies that converges to some $\bar{\mu}$, in the sense that

$$\lim_{k\to\infty,\, k\in\mathcal{K}} \mu^k(x) = \bar{\mu}(x), \qquad \forall\ x = 1,\ldots,n,$$

it does not follow that $\bar{\mu}$ is S-regular. In fact it is possible that the generated sequence of S-regular policies $\{\mu^k\}$ satisfies $\lim_{k\to\infty} J_{\mu^k} \to J_S^* = J^*$, yet $\{\mu^k\}$ may converge to an S-irregular policy whose cost function is strictly larger than J_S^*, as illustrated by the following example.

Example 3.4.2

Consider the third variant of the blackmailer problem (Section 3.1.3) for the case where $c = 0$ (the blackmailer may forgo demanding a payment at cost $c = 0$); see Fig. 3.4.1. Here the mapping T is given by

$$TJ = \min\left\{J,\ \inf_{0<u\le 1}\big\{-u+u^2+(1-u)J\big\}\right\},$$

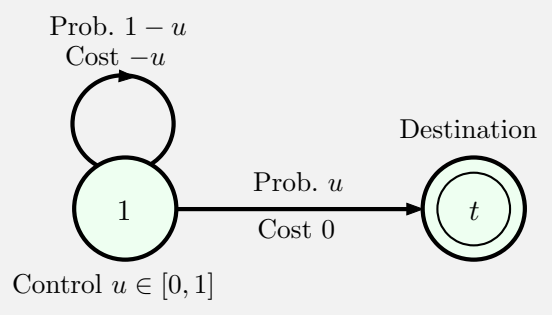


Figure 3.4.1. Transition diagram for a blackmailer problem (the third variant of Section 3.1.3 in the case where $c = 0$). At state 1, the blackmailer may demand any amount $u \in [0, 1]$. The victim will comply with probability $1 - u$ and will not comply with probability u, in which case the process will terminate.

[cf. Eq. (3.4)], and can be written as

$$TJ = \min_{0 \le u \le 1} \Big\{ -u + u^2 + (1 - u)J \Big\}.$$

Letting $S = \Re$, it can be seen that the set of fixed points of T within S is $(-\infty, -1]$. Here the policy whereby the blackmailer demands no payment ($u = 0$) and pays no cost at each period, is S-irregular and strictly suboptimal, yet has finite (zero) cost, so part (c) of Assumption 3.3.1 is violated (all other parts of the assumption are satisfied).

It can be seen that

$$J^* = J_S^* = -1,$$

J_S^* is a fixed point of T, Prop. 3.2.1 applies, and VI converges to J^* starting from any $J \ge J^*$. Moreover, starting from any policy (including the S-irregular one that applies $u = 0$), the PI algorithm (3.28) generates a sequence of S-regular policies $\{\mu^k\}$ with $J_{\mu^k} \to J_S^*$. However, $\{\mu^k\}$ converges to the S-irregular and strictly suboptimal policy that applies $u = 0$.

Here a phenomenon of "oscillation in the limit" is observed: starting with the S-irregular policy that applies $u = 0$, we generate a sequence of S-regular policies that converges to the S-irregular policy we started from! The perturbation-based PI algorithm of this section cannot rectify this type of behavior; it can only guarantee that a sequence of S-regular policies with $J_{\mu^k} \to J_S^*$ is generated.

3.5 APPLICATIONS IN SHORTEST PATH AND OTHER CONTEXTS

In this section we will apply the results of the preceding sections to various problems with a semicontractive character, including shortest path and deterministic optimal control problems of various types.

As we are about to apply the theory developed so far in this chapter, it may be helpful to summarize our results. Given a suitable set of functions S, we have been dealing with two problems. These are the original problem whose optimal cost function is J^*, and the restricted problem whose optimal cost function is J_S^*, the optimal cost over the S-regular policies. In summary, the aims of our analysis have been the following:

(a) *To establish the fixed point properties of T.* We have showed under various conditions (cf. Prop. 3.2.1) that J_S^* is the unique fixed point of T within the well-behaved region $\mathcal{W}_S$, and moreover the VI algorithm converges from above to J_S^*. Related analyses involve the use of infinite cost assumptions for S-irregular policies (Section 3.3), possibly in conjunction with the use of perturbations (Section 3.4). A favorable case is when $J_S^* = J^*$. However, we may also have $J_S^* \neq J^*$. Generally, proving that J^* is a fixed point of T is a separate issue, which may either be addressed in conjunction with the analysis of properties of J_S^* as in Section 3.3 (cf. Prop. 3.3.1), or independently of J_S^* (for example J^* is generically a fixed point of T in deterministic problems, among other classes of problems; see Exercise 3.1).

(b) *To delineate the initial conditions under which the VI and PI algorithms are guaranteed to converge to J_S^* or to J^*.* This was done in conjunction with the analysis of the fixed point properties of T. For example, a major line of analysis for establishing that J_S^* is a fixed point of T is based on the PI algorithm (cf. Sections 3.2.3 and 3.3). We have also obtained several other results relating to the convergence of variants of PI (the optimistic version, cf. Prop. 3.2.7, the λ-PI version, cf. Prop. 3.2.8, and the perturbation-based version, cf. Prop. 3.4.2), and to the mathematical programming-based solution, cf. Section 3.2.5.

(c) *To establish the existence of optimal policies for the original or for the restricted problem, and the associated optimality conditions.* This was accomplished in conjunction with the analysis of the fixed points of T, and under special compactness-like conditions (cf. Props. 3.2.1, 3.2.6, and 3.3.1).

As we apply our analysis to various specific contexts in this section, we will make frequent reference to the pathological behavior that we witnessed in the examples of Section 3.1. In particular, we will explain this behavior through our theoretical results, and we will discuss how to preclude this behavior through appropriate assumptions.

3.5.1 Stochastic Shortest Path Problems

Let us consider the SSP problem that we discussed in Section 1.3.2. It involves a directed graph with nodes $x = 1, \ldots, n$, plus a destination node

t that is cost-free and absorbing. At each node x, we must select a control $u \in U(x)$, which defines a probability distribution $p_{xy}(u)$ over all possible successor nodes $y = 1, \ldots, n, t$, while a cost $g(x, u)$ is incurred. We wish to minimize the expected cost of the traversed path, with cost accumulated up to reaching the destination.

Note that if for every feasible control the corresponding probability distribution assigns probability 1 to a single successor node, we obtain the deterministic shortest path problem of Section 3.1.1. This problem admits a relatively simple analysis, yet exhibits pathological behavior that we have described. The pathologies exhibited by SSP problems are more severe, and were illustrated in Sections 3.1.2 and 3.1.3.

We formulate the SSP problem as an abstract DP problem where:

(a) The state space is $X = \{1, \ldots, n\}$ and the control constraint set is $U(x)$ for all $x \in X$. (For technical reasons, it is convenient to exclude from X the destination t; we know that the optimal cost starting from t is 0, and including t within X would just complicate the notation and the analysis, with no tangible benefit.)

(b) The mapping H is given by

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

(c) The function $\bar{J}$ is identically 0, $\bar{J}(x) = 0$ for all x.

We continue to denote by $\mathcal{E}(X)$ the set of all extended real-valued functions $J : X \mapsto \Re^*$, and by $\mathcal{R}(X)$ the set of real-valued functions $J : X \mapsto \Re$. Note that since $X = \{1, \ldots, n\}$, $\mathcal{R}(X)$ is essentially the n-dimensional space R^n.

Here the mapping T_μ corresponding to a policy μ maps $\mathcal{R}(X)$ to $\mathcal{R}(X)$, and is given by

$$(T_\mu J)(x) = g\big(x, \mu(x)\big) + \sum_{y=1}^{n} p_{xy}\big(\mu(x)\big)J(y), \qquad x = 1, \ldots, n.$$

The corresponding cost for a given initial state $x_0 \in \{1, \ldots, n\}$ is

$$J_\mu(x_0) = \limsup_{k\to\infty} (T_\mu^k \bar{J})(x_0) = \limsup_{k\to\infty} \sum_{m=0}^{k-1} E\Big\{ g\big(x_m, \mu(x_m)\big) \Big\},$$

where $\{x_m\}$ is the (random) state trajectory generated under policy μ, starting from initial state x_0. The expected value $E\big\{g(x_m, \mu(x_m))\big\}$ above is defined in the natural way: it is the weighted sum of the numerical values $g\big(x, \mu(x)\big)$, $x = 1, \ldots, n$, weighted by the probabilities $p(x_m = x \mid x_0, \mu)$

that $x_m = x$ given that the initial state is x_0 and policy μ is used. Thus $J_\mu(x_0)$ is the upper limit as $k \to \infty$ of the cost for the first k steps or up to reaching the destination, whichever comes first.

A stationary policy μ is said to be *proper* if for every initial state there is positive probability that the destination will be reached under that policy after at most n stages. A stationary policy that is not proper is said to be *improper*. The relation between proper policies and S-regularity is given in the following proposition.

Proposition 3.5.1: (Proper Policies and Regularity) A policy is proper if and only if it is $\mathcal{R}(X)$-regular.

Proof: Clearly μ is $\mathcal{R}(X)$-regular if and only if the $n \times n$ matrix P_μ, whose components are $p_{xy}\big(\mu(x)\big)$, $x, y = 1, \ldots, n$, is a contraction (since T_μ is a linear mapping with matrix P_μ). If μ is proper then P_μ is a contraction mapping with respect to some weighted sup-norm; this is a classical result, given for example in [BeT89], Section 4.2. Conversely, it can be seen that if μ is improper, P_μ is not a contraction mapping since the Markov chain corresponding to μ has multiple ergodic classes and hence the equilibrium equation $\xi' = \xi' P_\mu$ has multiple solutions. **Q.E.D.**

Looking back to the shortest path examples of Sections 3.1.1-3.1.3, we can make some observations. In deterministic shortest path problems, $\mu(x)$ can be identified with the single successor node of node x. Thus μ is proper if and only if the corresponding graph of arcs $\big(x, \mu(x)\big)$ is acyclic. Moreover, there exists a proper policy if and only if each node is connected to the destination with a sequence of arcs. Every improper policy involves at least one cycle. Depending on the sign of the length of their cycle(s), improper policies can be strictly suboptimal (if all cycles have positive length), or may be optimal (possibly together with some proper policies, if all cycles have nonnegative length). Moreover, if there are cycles with negative length, no proper policy can be optimal and for the states x that lie on some negative length cycle we have $J^*(x) = -\infty$.

A further characterization of the optimal solution is possible in deterministic shortest path problems. Since the sets $U(x)$ are finite, there exists an optimal policy, which can be separated into a "proper" part consisting of arcs that form an acyclic subgraph, and an "improper" part consisting of cycles that have negative or zero length. These facts can be proved with simple arguments, which will not be given here (deterministic shortest path theory and algorithms are developed in detail in the author's text [Ber98]).

In SSP problems, the situation is more complicated. In particular, the cost function of an improper policy μ may not be a fixed point of T_μ while J^* may not be a fixed point of T (cf. the example of Section

3.1.2). Moreover, there may not exist an optimal stationary policy even if all policies are proper (cf. the three variants of the blackmailer example of Section 3.1.3).

In this section we will use various assumptions, which we will in turn translate into the conditions and corresponding results of Sections 3.2-3.4. Throughout this section we will assume the following.

Assumption 3.5.1: There exists at least one proper policy.

Depending on the circumstances, we will also consider the use of one or both of the following assumptions.

Assumption 3.5.2: The control space U is a metric space. Moreover, for each state x, the set $U(x)$ is a compact subset of U, the functions $p_{xy}(\cdot)$, $y = 1, \ldots, n$, are continuous over $U(x)$, and the function $g(x, \cdot)$ is lower semicontinuous over $U(x)$.

Assumption 3.5.3: For every improper policy μ and function $J \in \mathcal{R}(X)$, there exists at least one state $x \in X$ such that $J_\mu(x) = \infty$.

An important consequence of Assumption 3.5.2 is that it implies the compactness condition (d) of Assumption 3.3.1. We will also see from the proof of the following proposition that Assumption 3.5.3 implies the infinite cost condition (c) of Assumption 3.3.1.

Analysis Under the Strong SSP Conditions

The preceding three assumptions, referred to as the *strong SSP conditions*,† were introduced in the paper [BeT91], and they were used to show strong results for the SSP problem. In particular, the following proposition was shown.

Proposition 3.5.2: Let the strong SSP conditions hold. Then:

(a) The optimal cost function J^* is the unique solution of Bellman's equation $J = TJ$ within $\mathcal{R}(X)$.

† The strong SSP conditions and the weak SSP conditions, which will be introduced shortly, relate to the strong and weak PI properties of Section 3.2.

(b) The VI sequence $\{T^k J\}$ converges to J^* starting from any $J \in \mathcal{R}(X)$.

(c) A policy μ is optimal if and only if $T_\mu J^* = TJ^*$. Moreover, there exists an optimal policy that is proper.

(d) The PI algorithm, starting from any proper policy, is valid in the sense described by the conclusions of Prop. 3.3.1(e).

We will prove the proposition by using the strong SSP conditions to verify Assumption 3.3.1 for $S = \mathcal{R}(X)$, and then by applying Prop. 3.3.1. To this end, we first state without proof the following result relating to proper policies from [BeT91].

Proposition 3.5.3: Under the strong SSP conditions, the optimal cost function $\hat{J}$ over proper policies only,

$$\hat{J}(x) = \inf_{\mu:\,\text{proper}} J_\mu(x), \qquad x \in X,$$

is real-valued.

The preceding proposition holds trivially if the control space U is finite (since then the set of all policies is finite), or if J^* is somehow known to be real-valued [for example if $g(x,u) \geq 0$ for all (x,u)]. The three variants of the blackmailer problem of Section 3.1.3 provide examples illustrating what can happen if U is infinite. In particular, in the first variant of the blackmailer problem all policies are proper (and hence Assumptions 3.5.1 and 3.5.3 are satisfied), but $\hat{J}$ is not real-valued. The proof of Prop. 3.5.3 in the case of an infinite control space U was given as part of Prop. 2 of the paper [BeT91]. Despite the intuitive nature of Prop. 3.5.3, the proof embodies a fairly complicated argument (see Lemma 3 of [BeT91]).

Another related result is that if all policies are proper, then for all $\mu \in \mathcal{M}$, T_μ is a contraction mapping with respect to a common weighted sup-norm, so the contractive model analysis and algorithms of Chapter 2 apply (see [BeT96], Prop. 2.2). However, this fact will not be useful to us in this section.

Proof of Prop. 3.5.2: In the context of Section 3.3, let us choose $S = \mathcal{R}(X)$, so the proper policies are identified with the S-regular policies by Prop. 3.5.1. We will verify Assumption 3.3.1.

Indeed parts (a) and (e) are trivially satisfied, part (b) is satisfied by Prop. 3.5.3, part (d) can be easily verified by using Assumption 3.5.2. To verify part (f), we use Prop. 3.3.2, which applies because $S = \mathcal{R}(X) =$

$R_b(X)$ (since X is finite) and Eq. (3.27) clearly holds. Finally, to verify part (c) we must show that given an improper policy μ, for every $J \in \mathcal{R}(X)$ there exists an $x \in X$ such that $\limsup_{k\to\infty}(T_\mu^k J)(x) = \infty$. This follows since by Assumption 3.5.3, $J_\mu(x) = \limsup_{k\to\infty}(T_\mu^k \bar{J})(x) = \infty$, for some $x \in X$, and $(T_\mu^k J)(x)$ and $(T_\mu^k \bar{J})(x)$ differ by $E\{J(x_k)\}$, an amount that is finite since J is real-valued and has a finite number of components $J(x)$. Thus Assumption 3.3.1 holds and the result follows from Prop. 3.3.1. **Q.E.D.**

Analysis Under the Weak SSP Conditions

Under the strong SSP conditions, we showed in Prop. 3.5.2 that J^* is the unique fixed point of T within $\mathcal{R}(X)$. Moreover, we showed that a policy μ^* is optimal if and only if $T_{\mu^*}J^* = TJ^*$, and an optimal proper policy exists (so in particular J^*, being the cost function of a proper policy, is real-valued). In addition, J^* can be computed by the VI algorithm starting with any $J \in \Re^n$.

We will now replace Assumption 3.5.3 (improper policies have cost ∞ for some initial states) with the following weaker assumption:

Assumption 3.5.4: The optimal cost function J^* is real-valued.

We will refer to the Assumptions 3.5.1, 3.5.2, and 3.5.4 as the *weak SSP conditions*. The examples of Sections 3.1.1 and 3.1.2 show that under these assumptions, it is possible that

$$J^* \neq \hat{J} = \inf_{\mu:\,\text{proper}} J_\mu,$$

while J^* need not be a fixed point of T (Section 3.1.2). The key fact is that under Assumption 3.5.4, we can use the perturbation approach of Section 3.4, whereby adding $\delta > 0$ to the mapping T_μ makes all improper policies have infinite cost for some initial states, so the results of Prop. 3.5.2 can be used for the δ-perturbed problem. In particular, Prop. 3.5.1 implies that $J_S^* = \hat{J}$, so from Prop. 3.4.1 it follows that $\hat{J}$ is a fixed point of T and the conclusions of Prop. 3.2.1 hold. We thus obtain the following proposition, which provides additional results, not implied by Prop. 3.2.1; see Fig. 3.5.1.

Proposition 3.5.4: Let the weak SSP conditions hold. Then:

(a) The optimal cost function over proper policies, $\hat{J}$, is the largest solution of Bellman's equation $J = TJ$ within $\mathcal{R}(X)$, i.e., $\hat{J}$ is a solution that belongs to $\mathcal{R}(X)$, and if $J' \in \mathcal{R}(X)$ is another solution, we have $J' \leq \hat{J}$.

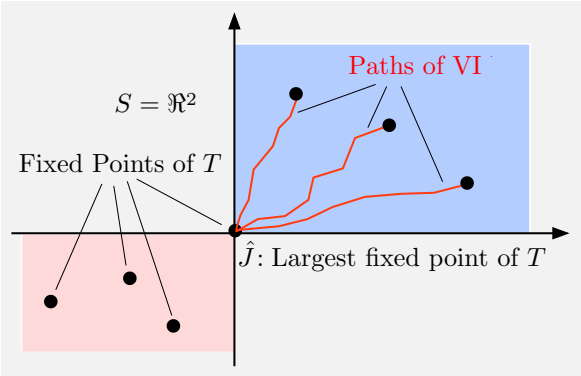


Figure 3.5.1. Schematic illustration of Prop. 3.5.4 for a problem with two states, so $\mathcal{R}(X) = \Re^2 = S$. We have that $\hat{J}$ is the largest solution of Bellman's equation, while VI converges to $\hat{J}$ starting from $J \geq \hat{J}$. As shown in Section 3.1.2, J^* need not be a solution of Bellman's equation.

(b) The VI sequence $\{T^k J\}$ converges linearly to $\hat{J}$ starting from any $J \in \mathcal{R}(X)$ with $J \geq \hat{J}$.

(c) Let μ be a proper policy. Then μ is optimal within the class of proper policies (i.e., $J_\mu = \hat{J}$) if and only if $T_\mu \hat{J} = T\hat{J}$.

(d) For every $J \in \mathcal{R}(X)$ such that $J \leq TJ$, we have $J \leq \hat{J}$.

Proof: (a), (b) Let $S = \mathcal{R}(X)$, so the proper policies are identified with the S-regular policies by Prop. 3.5.1. We use the perturbation framework of Section 3.4 with forcing function $p(x) \equiv 1$. From Prop. 3.5.2 it follows that Prop. 3.4.1 applies so that $\hat{J}$ is a fixed point of T, and the conclusions of Prop. 3.2.1 hold, so $T^k J \to \hat{J}$ starting from any $J \in \mathcal{R}(X)$ with $J \geq \hat{J}$. The convergence rate of VI is linear in view of Prop. 3.2.2 and the existence of an optimal proper policy to be shown in part (c). Finally, let $J' \in \mathcal{R}(X)$ be another solution of Bellman's equation, and let $J \in \mathcal{R}(X)$ be such that $J \geq \hat{J}$ and $J \geq J'$. Then $T^k J \to \hat{J}$, while $T^k J \geq T^k J' = J'$. It follows that $\hat{J} \geq J'$.

(c) If the proper policy μ satisfies $J_\mu = \hat{J}$, we have $\hat{J} = J_\mu = T_\mu J_\mu = T_\mu \hat{J}$, so, using also the relation $\hat{J} = T\hat{J}$ [cf. part (a)], we obtain $T_\mu \hat{J} = T\hat{J}$. Conversely, if μ satisfies $T_\mu \hat{J} = T\hat{J}$, then using part (a), we have $T_\mu \hat{J} = \hat{J}$ and hence $\lim_{k\to\infty} T_\mu^k \hat{J} = \hat{J}$. Since μ is proper, we have $J_\mu = \lim_{k\to\infty} T_\mu^k \hat{J}$, so $J_\mu = \hat{J}$.

(d) Let $J \leq TJ$ and $\delta > 0$. We have $J \leq TJ + \delta e = T_\delta J$, and hence $J \leq T_\delta^k J$ for all k. Since the strong SSP conditions hold for the δ-perturbed

problem, it follows that $T_\delta^k J \to \hat{J}_\delta$, so $J \leq \hat{J}_\delta$. By taking $\delta \downarrow 0$ and using Prop. 3.4.1, it follows that $J \leq \hat{J}$. **Q.E.D.**

The first variant of the blackmailer Example 3.4.2 shows that under the weak SSP conditions there may not exist an optimal policy or an optimal policy within the class of proper policies if the control space is infinite. This is consistent with Prop. 3.5.4(c). Another interesting fact is provided by the third variant of this example in the case where $c < 0$. Then $J^*(1) = -\infty$ (violating Assumption 3.5.4), but $\hat{J}$ is real-valued and does not solve Bellman's equation, contrary to the conclusion of Prop. 3.5.4(a).

Part (d) of Prop. 3.5.4 shows that $\hat{J}$ is the unique solution of the problem of maximizing $\sum_{i=1}^n \beta_i J(i)$ over all $J = \big(J(1), \ldots, J(n)\big)$ such that $J \leq TJ$, where $\beta_1, \ldots, \beta_n$ are any positive scalars (cf. Prop. 3.2.9). This problem can be written as

$$\text{maximize} \quad \sum_{i=1}^n J(i)$$

$$\text{subject to} \quad J(x) \leq g(x,u) + \sum_{y=1}^n p_{ij}(u) J(j), \quad i = 1, \ldots, n, \quad u \in U(i),$$

and is a linear program if each $U(i)$ is a finite set.

Generally, under the weak SSP conditions the strong PI property may not hold, so a sequence generated by PI starting from a proper policy need not have the cost improvement property. An example is the deterministic shortest path problem of Section 3.1.1, when there is a zero length cycle ($a = 0$) and the only optimal policy is proper ($b = 0$). Then the PI algorithm may oscillate between the optimal proper policy and the strictly suboptimal improper policy. We will next consider the modified version of the PI algorithm that is based on the use of perturbations (Section 3.4).

Policy Iteration with Perturbations

To deal with the oscillatory behavior of PI, which was illustrated in the deterministic shortest path Example 3.2.2, we may use the perturbed version of the PI algorithm of Section 3.4, with forcing function $p(x) \equiv 1$. Thus, we have

$$(T_{\mu,\delta} J)(x) = H\big(x, \mu(x), J\big) + \delta, \quad x \in X, \qquad T_\delta J = \inf_{\mu \in \mathcal{M}} T_{\mu,\delta} J.$$

The algorithm generates the sequence $\{\mu^k\}$ as follows.

Let $\{\delta_k\}$ be a positive sequence with $\delta_k \downarrow 0$, and let μ^0 be any proper policy. At iteration k, we have a proper policy μ^k, and we generate μ^{k+1} according to

$$T_{\mu^{k+1}} J_{\mu^k, \delta_k} = T J_{\mu^k, \delta_k}, \tag{3.31}$$

where J_{μ^k,δ_k} is computed as the unique fixed point of the mapping T_{μ^k,δ_k} given by

$$T_{\mu^k,\delta_k} J = T_{\mu^k} J + \delta_k e.$$

The policy μ^{k+1} of Eq. (3.31) exists by the compactness Assumption 3.5.2. We claim that μ^{k+1} is proper. To see this, note that

$$T_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k} = T J_{\mu^k,\delta_k} + \delta_k\, e \leq T_{\mu^k} J_{\mu^k,\delta_k} + \delta_k\, e = J_{\mu^k,\delta_k},$$

so that by the monotonicity of T_μ^{k+1},

$$T^m_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k} \leq T_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k} = T J_{\mu^k,\delta_k} + \delta_k\, e \leq J_{\mu^k,\delta_k}, \qquad \forall\ m \geq 1.$$

Since J_{μ^k,δ_k} forms an upper bound to $T^m_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k}$, it follows that μ^{k+1} is proper [if it were improper, we would have $(T^m_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k})(x) \to \infty$ for some x, because of the perturbation δ_k]. Thus the sequence $\{\mu^k\}$ generated by the perturbed PI algorithm (3.31) is well-defined and consists of proper policies. We have the following proposition.

Proposition 3.5.5: Let the weak SSP conditions hold. Then the sequence $\{J_{\mu^k}\}$ generated by the perturbed PI algorithm (3.31) satisfies $J_{\mu^k} \to \hat{J}$.

Proof: We apply the perturbation framework of Section 3.4 with $S = \mathcal{R}(X)$, $\widehat{\mathcal{M}}$ equal to the set of proper policies, and the forcing function $p(x) \equiv 1$. Clearly Assumption 3.4.1 holds, so Prop. 3.4.2 applies. **Q.E.D.**

When the control space U is finite, the generated policies μ^k will be optimal for all k sufficiently large, as noted following Prop. 3.4.2. However, when the control space U is infinite, the generated sequence $\{\mu^k\}$ may exhibit some serious pathologies in the limit, as we have seen in Example 3.4.2.

3.5.2 Affine Monotonic Problems

In this section, we consider a class of semicontractive models, called *affine monotonic*, where the abstract mapping T_μ associated with a stationary policy μ is affine and maps nonnegative functions to nonnegative functions. These models include as special cases stochastic undiscounted nonnegative cost problems, and multiplicative cost problems, such as risk-averse problems with exponentiated additive cost and a termination state (see Example 1.2.8). Here we will focus on the special case where the state space is finite and a certain compactness condition holds.

We consider a finite state space $X = \{1, \ldots, n\}$ and a (possibly infinite) control constraint set $U(x)$ for each state x. For each $\mu \in \mathcal{M}$ the mapping T_μ is given by

$$T_\mu J = b_\mu + A_\mu J,$$

where b_μ is a vector of $\Re^n$ with components $b\big(x, \mu(x)\big)$, $x = 1, \ldots, n$, and A_μ is an $n \times n$ matrix with scalar components $A_{xy}\big(\mu(x)\big)$, $x, y = 1, \ldots, n$. We assume that $b(x, u)$ and $A_{xy}(u)$ are nonnegative,

$$b(x, u) \geq 0, \qquad A_{xy}(u) \geq 0, \qquad \forall\ x, y = 1, \ldots, n,\ u \in U(x).$$

Thus T_μ maps $\mathcal{E}^+(X)$ into $\mathcal{E}^+(X)$, where $\mathcal{E}^+(X)$ denotes the set of nonnegative extended real-valued functions $J : X \mapsto [0, \infty]$. Moreover T_μ also maps $\mathcal{R}^+(X)$ to $\mathcal{R}^+(X)$, where $\mathcal{R}^+(X)$ denotes the set of nonnegative real-valued functions $J : X \mapsto [0, \infty)$.

The mapping $T : \mathcal{E}^+(X) \mapsto \mathcal{E}^+(X)$ is given by

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

or equivalently,

$$(TJ)(x) = \inf_{u \in U(x)} \left[b(x, u) + \sum_{y=1}^{n} A_{xy}(u) J(y) \right], \qquad x \in X.$$

Multiplicative and Exponential Cost SSP Problems

Affine monotonic models appear in several contexts. In particular, finite-state sequential stochastic control problems (including SSP problems) with nonnegative cost per stage (see, e.g., [Ber12a], Chapter 3, and Section 4.1) are special cases where $\bar{J}$ is the identically zero function [$\bar{J}(x) \equiv 0$]. We will describe another type of SSP problem, where the cost function of a policy accumulates over time multiplicatively, rather than additively.

As in the SSP problems of the preceding section, we assume that there are n states $x = 1, \ldots, n$, and a cost-free and absorbing state t. There are probabilistic state transitions among the states $x = 1, \ldots, n$, up to the first time a transition to state t occurs, in which case the state transitions terminate. We denote by $p_{xt}(u)$ and $p_{xy}(u)$ the probabilities of transition under u from x to t and to y, respectively, so that

$$p_{xt}(u) + \sum_{y=1}^{n} p_{xy}(u) = 1, \qquad x = 1, \ldots, n,\ u \in U(x).$$

We introduce nonnegative scalars $h(x, u, t)$ and $h(x, u, y)$,

$$h(x, u, t) \geq 0, \qquad h(x, u, y) \geq 0, \qquad \forall\ x, y = 1, \ldots, n,\ u \in U(x),$$

and we consider the affine monotonic problem where the scalars $A_{xy}(u)$ and $b(x,u)$ are defined by

$$A_{xy}(u) = p_{xy}(u)h(x,u,y), \qquad x,y = 1,\ldots,n,\ u \in U(x),$$

and

$$b(x,u) = p_{xt}(u)h(x,u,t), \quad x = 1,\ldots,n,\ u \in U(x),$$

and the vector $\bar{J}$ is the unit vector,

$$\bar{J}(x) = 1, \qquad x = 1,\ldots,n.$$

The cost function of this problem has a multiplicative character as we show next.

Indeed, with the preceding definitions of $A_{xy}(u)$, $b(x,u)$, and $\bar{J}$, we will prove that the expression for the cost function of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$,

$$J_\pi(x_0) = \limsup_{N\to\infty} (T_{\mu_0}\cdots T_{\mu_{N-1}}\bar{J})(x_0), \qquad x_0 = 1,\ldots,n,$$

can be written in the multiplicative form

$$J_\pi(x_0) = \limsup_{N\to\infty} E\left\{\prod_{k=0}^{N-1} h\big(x_k,\mu_k(x_k),x_{k+1}\big)\right\}, \qquad x_0 = 1,\ldots,n, \tag{3.32}$$

where:

(a) $\{x_0, x_1, \ldots\}$ is the random state trajectory generated starting from x_0, using π.

(b) The expected value is with respect to the probability distribution of that trajectory.

(c) We use the notation

$$h\big(x_k,\mu_k(x_k),x_{k+1}\big) = 1, \qquad \text{if } x_k = x_{k+1} = t,$$

(so that the multiplicative cost accumulation stops once the state reaches t).

Thus, we claim that $J_\pi(x_0)$ *can be viewed as the expected value of cost accumulated multiplicatively, starting from* x_0 *up to reaching the termination state* t *(or indefinitely accumulated multiplicatively, if* t *is never reached).*

To verify the formula (3.32) for J_π, we use the definition $T_\mu J = b_\mu + A_\mu J$, to show by induction that for every $\pi = \{\mu_0, \mu_1, \ldots\}$, we have

$$T_{\mu_0}\cdots T_{\mu_{N-1}}\bar{J} = A_{\mu_0}\cdots A_{\mu_{N-1}}\bar{J} + b_{\mu_0} + \sum_{k=1}^{N-1} A_{\mu_0}\cdots A_{\mu_{k-1}}b_{\mu_k}. \tag{3.33}$$

We then interpret the n components of each vector on the right as conditional expected values of the expression

$$\prod_{k=0}^{N-1} h\big(x_k, \mu_k(x_k), x_{k+1}\big) \tag{3.34}$$

multiplied with the appropriate conditional probability. In particular:

(a) The ith component of the vector $A_{\mu_0} \cdots A_{\mu_{N-1}} \bar{J}$ in Eq. (3.33) is the conditional expected value of the expression (3.34), given that $x_0 = i$ and $x_N \neq t$, multiplied with the conditional probability that $x_N \neq t$, given that $x_0 = i$.

(b) The ith component of the vector b_{μ_0} in Eq. (3.33) is the conditional expected value of the expression (3.34), given that $x_0 = i$ and $x_1 = t$, multiplied with the conditional probability that $x_1 = t$, given that $x_0 = i$.

(c) The ith component of the vector $A_{\mu_0} \cdots A_{\mu_{k-1}} b_{\mu_k}$ in Eq. (3.33) is the conditional expected value of the expression (3.34), given that $x_0 = i$, $x_1, \ldots, x_{k-1} \neq t$, and $x_k = t$, multiplied with the conditional probability that $x_1, \ldots, x_{k-1} \neq t$, and $x_k = t$, given that $x_0 = i$.

By adding these conditional probability expressions, we obtain the ith component of the unconditional expected value

$$E\left\{ \prod_{k=0}^{N-1} h\big(x_k, \mu_k(x_k), x_{k+1}\big) \right\},$$

thus verifying the formula (3.32).

A special case of multiplicative cost problem is the *risk-sensitive SSP problem with exponential cost function*, where for all $x = 1, \ldots, n$, and $u \in U(x)$,

$$h(x, u, y) = \exp\big(g(x, u, y)\big), \quad y = 1, \ldots, n, t,$$

and the function g can take both positive and negative values. The mapping T_μ has the form

$$\begin{aligned}(T_\mu J)(x) = {} & p_{xt}\big(\mu(x)\big)\exp\big(g(x, \mu(x), t)\big) \\ & + \sum_{y=1}^{n} p_{xy}\big(\mu(x)\big)\exp\big(g(x, \mu(x), y)\big) J(y), \qquad x = 1, \ldots, n,\end{aligned} \tag{3.35}$$

where $p_{xy}(u)$ is the probability of transition from x to y under u, and $g(x, u, y)$ is the cost of the transition. The Bellman equation is

$$J(x) = \inf_{u \in U(x)} \left[p_{xt}(u)\exp\big(g(x, u, t)\big) + \sum_{y=1}^{n} p_{xy}(u)\exp\big(g(x, u, y)\big) J(y) \right].$$

Based on Eq. (3.32), we have that $J_\pi(x_0)$ is the limit superior of the expected value of the exponential of the N-step additive finite horizon cost up to termination, i.e., $\sum_{k=0}^{\bar{k}} g\big(x_k, \mu_k(x_k), x_{k+1}\big)$, where $\bar{k}$ is equal to the first index prior to $N-1$ such that $x_{\bar{k}+1} = t$, or is equal to $N-1$ if there is no such index. The use of the exponential introduces risk aversion, by assigning a strictly convex increasing penalty for large rather than small cost of a trajectory up to termination (and hence a preference for small variance of the additive cost up to termination).

The deterministic version of the exponential cost problem where for each $u \in U(x)$, one of the transition probabilities $p_{xt}(u), p_{x1}(u), \ldots, p_{xn}(u)$ is equal to 1 and all others are equal to 0, is mathematically equivalent to the classical deterministic shortest path problem (since minimizing the exponential of a deterministic expression is equivalent to minimizing that expression). For this problem a standard assumption is that there are no cycles that have negative total length to ensure that the shortest path length is finite. However, it is interesting that this assumption is not required for the analysis of the present section: when there are paths that travel perpetually around a negative length cycle we simply have $J^*(x) = 0$ for all states x on the cycle, which is permissible within our context.

Assumptions on Policies - Contractive Policies

Let us now derive an expression for the cost function of a policy. By repeatedly applying the mapping T_μ to the equation $T_\mu J = b_\mu + A_\mu J$, we have

$$T_\mu^N J = A_\mu^N J + \sum_{k=0}^{N-1} A_\mu^k b_\mu, \qquad \forall\ J \in \mathcal{E}^+(X),\ N = 1, 2, \ldots,$$

and hence

$$J_\mu = \limsup_{N\to\infty} T_\mu^N \bar{J} = \limsup_{N\to\infty} A_\mu^N \bar{J} + \sum_{k=0}^{\infty} A_\mu^k b_\mu \tag{3.36}$$

(the series converges since A_μ and b_μ have nonnegative components).

We say that μ *is contractive* if A_μ has eigenvalues that are strictly within the unit circle. In this case T_μ is a contraction mapping with respect to some weighted sup-norm (see Prop. B.3 in Appendix B). If μ is contractive, then $A_\mu^N \bar{J} \to 0$ and from Eq. (3.36), it follows that

$$J_\mu = \sum_{k=0}^{\infty} A_\mu^k b_\mu = (I - A_\mu)^{-1} b_\mu,$$

and J_μ is real-valued as well as nonnegative, i.e., $J_\mu \in \mathcal{R}^+(X)$. Moreover, a contractive μ is also $\mathcal{R}^+(X)$-regular, since J_μ does not depend on the initial function $\bar{J}$. The reverse is also true as shown by the following proposition.

Proposition 3.5.6: A policy μ is contractive if and only if it is $\mathcal{R}^+(X)$-regular. Moreover, if μ is noncontractive and all the components of b_μ are strictly positive, there exists a state x such that the corresponding component of the vector $\sum_{k=0}^{\infty} A_\mu^k b_\mu$ is ∞.

Proof: As noted earlier, if μ is contractive it is $\mathcal{R}^+(X)$-regular. It will thus suffice to show that for a noncontractive μ and strictly positive components of b_μ, some component of $\sum_{k=0}^{\infty} A_\mu^k b_\mu$ is ∞. Indeed, according to the Perron-Frobenius Theorem, the nonnegative matrix A_μ has a real eigenvalue λ, which is equal to its spectral radius, and an associated nonnegative eigenvector $\xi \neq 0$ [see Prop. B.3(a) in Appendix B]. Choose $\gamma > 0$ to be such that $b_\mu \geq \gamma\xi$, so that

$$\sum_{k=0}^{\infty} A_\mu^k b_\mu \geq \gamma \sum_{k=0}^{\infty} A_\mu^k \xi = \gamma \left(\sum_{k=0}^{\infty} \lambda^k \right) \xi.$$

Since some component of ξ is positive while $\lambda \geq 1$ (since μ is noncontractive), the corresponding component of the infinite sum on the right is infinite, and the same is true for the corresponding component of the vector $\sum_{k=0}^{\infty} A_\mu^k b_\mu$ on the left. **Q.E.D.**

Let us introduce some assumptions that are similar to the ones of the preceding section.

Assumption 3.5.5: There exists at least one contractive policy.

Assumption 3.5.6: (Compactness and Continuity) The control space U is a metric space, and $A_{xy}(\cdot)$ and $b(x,\cdot)$ are continuous functions of u over $U(x)$, for all x and y. Moreover, for each state x, the sets

$$\left\{ u \in U(x) \;\middle|\; b(x,u) + \sum_{y=1}^{n} A_{xy}(u) J(y) \leq \lambda \right\}$$

are compact subsets of U for all scalars $\lambda \in \Re$ and $J \in \mathcal{R}^+(X)$.

Case of Infinite Cost Noncontractive Policies

We now turn to questions relating to Bellman's equation, the convergence of the VI and PI algorithms, as well as conditions for optimality of a stationary

policy. We first consider the following assumption, which parallels the infinite cost Assumption 3.5.3 for SSP problems.

Assumption 3.5.7: (Infinite Cost Condition) For every noncontractive policy μ, there is at least one state such that the corresponding component of the vector $\sum_{k=0}^{\infty} A_\mu^k b_\mu$ is equal to ∞.

We will now show that for $S = \mathcal{R}^+(X)$, Assumptions 3.5.5, 3.5.6, and 3.5.7 imply all the parts of Assumption 3.3.1 of Section 3.3, so Prop. 3.3.1 can be applied to the affine monotonic model. Indeed parts (a), (e) of Assumption 3.3.1 clearly hold. Part (b) also holds, since by Assumption 3.5.5 there exists a contractive and hence S-regular policy, so we have $J_S^* \in \mathcal{R}^+(X)$. Moreover Assumption 3.5.6 implies part (d), while Assumption 3.5.7 implies part (c). Finally part (f) holds since for every $J \in \mathcal{R}^+(X)$, the zero function, $J'(x) \equiv 0$, lies in $\mathcal{R}^+(X)$, and satisfies $J' \leq J$ and $J' \leq TJ'$. Thus Prop. 3.3.1 yields the following result.

Proposition 3.5.7: (Bellman's Equation, Policy Iteration, Value Iteration, and Optimality Conditions) Let Assumptions 3.5.5, 3.5.6, and 3.5.7 hold.

(a) The optimal cost vector J^* is the unique fixed point of T within $\mathcal{R}^+(X)$.

(b) We have $T^k J \to J^*$ for all $J \in \mathcal{R}^+(X)$.

(c) A policy μ is optimal if and only if $T_\mu J^* = TJ^*$. Moreover there exists an optimal policy that is contractive.

(d) For any $J \in \mathcal{R}^+(X)$, if $J \leq TJ$ we have $J \leq J^*$, and if $J \geq TJ$ we have $J \geq J^*$.

(e) Every sequence $\{\mu^k\}$ generated by the PI algorithm starting from a contractive policy μ^0 satisfies $J_{\mu^k} \downarrow J^*$. Moreover, if the set of contractive policies is finite, there exists $\bar{k} \geq 0$ such that $\mu^{\bar{k}}$ is optimal.

Example 3.5.1 (Exponential Cost Shortest Path Problem)

Consider the deterministic shortest path example of Section 3.1.1, but with the exponential cost function of the present subsection; cf. Eq. (3.35). There are two policies denoted μ and μ'; see Fig. 3.5.2. The corresponding mappings and costs are shown in the figure, and Bellman's equation is given by

$$J(1) = (TJ)(1) = \min\big\{\exp(b),\, \exp(a)J(1)\big\}.$$

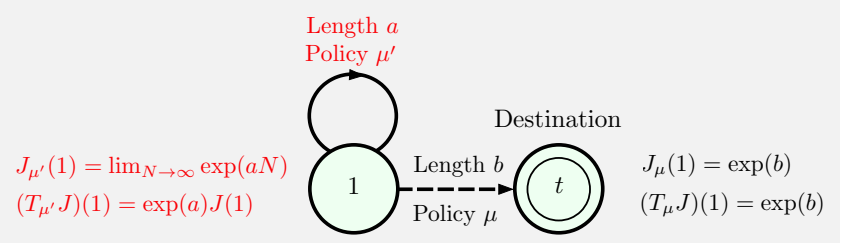


Figure 3.5.2. Shortest path problem with exponential cost function.

We consider three cases:

(a) $a > 0$: Here the proper policy μ is optimal, and the improper policy μ' is $\mathcal{R}^+(X)$-irregular (noncontractive) and has infinite cost, $J_{\mu'}(1) = \infty$. The assumptions of Prop. 3.5.7 hold, and consistently with the conclusions of the proposition, $J^*(1) = \exp(b)$ is the unique solution of Bellman's equation.

(b) $a = 0$: Here the improper policy μ' is $\mathcal{R}^+(X)$-irregular (noncontractive) and has finite cost, $J_{\mu'}(1) = 1$, so the assumptions of Prop. 3.5.7 are violated. The set of solutions of Bellman's equation within $S = \mathcal{R}^+(X)$ is the interval $\big[0, \exp(b)\big]$.

(c) $a < 0$: Here both policies are contractive, including the improper policy μ'. The assumptions of Prop. 3.5.7 hold, and consistently with the conclusions of the proposition, $J^*(1) = 0$ is the unique solution of Bellman's equation.

The reader may also verify that in the cases where $a \neq 0$, the assumptions and the results of Prop. 3.5.7 hold.

Case of Finite Cost Noncontractive Policies

We will now eliminate Assumption 3.5.7, thus allowing noncontractive policies with real-valued cost functions, similar to the corresponding case of the preceding section, under the weak SSP conditions. Let us denote by $\hat{J}$ the optimal cost function that can be achieved with contractive policies only,

$$\hat{J}(x) = \inf_{\mu:\,\text{contractive}} J_\mu(x), \qquad x = 1, \ldots, n. \tag{3.37}$$

We use the perturbation approach of Section 3.4 and Prop. 3.4.1 to show that $\hat{J}$ is a solution of Bellman's equation. In particular, we add a constant $\delta > 0$ to all components of b_μ. By using arguments that are entirely analogous to the ones for the SSP case of Section 3.5.1, we obtain the following proposition, which is illustrated in Fig. 3.5.3. A detailed analysis and proof is given in the exercises.

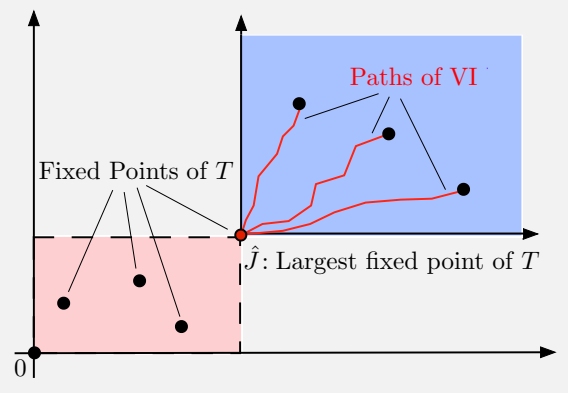


Figure 3.5.3. Schematic illustration of Prop. 3.5.8 for a problem with two states. The optimal cost function over contractive policies, $\hat{J}$, is the largest solution of Bellman's equation, while VI converges to $\hat{J}$ starting from $J \geq \hat{J}$.

Proposition 3.5.8: (Bellman's Equation, Value Iteration, and Optimality Conditions) Let Assumptions 3.5.5 and 3.5.6 hold. Then:

(a) The optimal cost function over contractive policies, $\hat{J}$, is the largest solution of Bellman's equation $J = TJ$ within $\mathcal{R}^+(X)$, i.e., $\hat{J}$ is a solution that belongs to $\mathcal{R}^+(X)$, and if $J' \in \mathcal{R}^+(X)$ is another solution, we have $J' \leq \hat{J}$.

(b) We have $T^k J \to \hat{J}$ for every $J \in \mathcal{R}^+(X)$ with $J \geq \hat{J}$.

(c) Let μ be a contractive policy. Then μ is optimal within the class of contractive policies (i.e., $J_\mu = \hat{J}$) if and only if $T_\mu \hat{J} = T\hat{J}$.

(d) For every $J \in \mathcal{R}^+(X)$ such that $J \leq TJ$, we have $J \leq \hat{J}$.

The other results of Section 3.5.1 for SSP problems also have straightforward analogs. Moreover, there is an adaptation of the example of Section 3.1.2, which provides an affine monotonic model for which J^* is not a fixed point of T (see the author's paper [Ber16a], to which we refer for further discussion).

Example 3.5.2 (Deterministic Shortest Path Problem with Exponential Cost - Continued)

Consider the problem of Fig. 3.5.2, for the case $a = 0$. This is the case where the noncontractive policy μ' has finite cost, so Assumption 3.5.7 is violated and Prop. 3.5.7 does not apply. However, it can be seen that the assumptions of Prop. 3.5.8 hold. Consistent with part (a) of the proposition, the optimal

cost over contractive policies, $\hat{J}(1) = \exp(b)$, is the largest of the fixed points of T. The other parts of Prop. 3.5.8 may also be easily verified.

We note that in the absence of the infinite cost Assumption 3.5.7, it is possible that the only optimal policy is noncontractive, even if the compactness Assumption 3.5.6 holds and $\hat{J} = J^*$. This is shown in the following example.

Example 3.5.3 (A Counterexample on the Existence of an Optimal Contractive Policy)

Consider the exponential cost version of the blackmailer problem of Example 3.4.2 (cf. Fig. 3.4.1). Here there is a single state 1, at which we must choose $u \in [0, 1]$. Then, we terminate at no cost [$g(1, u, t) = 0$ in Eq. (3.35)] with probability u, and we stay at state 1 at cost $-u$ [i.e., $g(1, u, 1) = -u$ in Eq. (3.35)] with probability $1 - u$. We have

$$b(i, u) = u \exp(0) = u, \qquad A_{11}(u) = (1 - u) \exp(-u),$$

so that

$$H(1, u, J) = u + (1 - u) \exp(-u) J.$$

Here there is a unique noncontractive policy μ': it chooses $u = 0$ at state 1, and has cost $J_{\mu'}(1) = 1$. Every policy μ with $\mu(1) \in (0, 1]$ is contractive, and J_μ can be obtained by solving the equation $J_\mu = T_\mu J_\mu$, i.e.,

$$J_\mu(1) = \mu(1) + \big(1 - \mu(1)\big) \exp\big(-\mu(1)\big) J_\mu(1).$$

We thus obtain

$$J_\mu(1) = \frac{\mu(1)}{1 - \big(1 - \mu(1)\big) \exp\big(-\mu(1)\big)}.$$

By minimizing over $\mu(1) \in (0, 1]$ this expression, it can be seen that $\hat{J}(1) = J^*(1) = \frac{1}{2}$, but there exists no optimal policy, and no optimal policy within the class of contractive policies [$J_\mu(1)$ decreases monotonically to $\frac{1}{2}$ as $\mu(1) \to 0$].

3.5.3 Robust Shortest Path Planning

We will now discuss how the analysis of Sections 3.3 and 3.4 applies to minimax shortest path-type problems, following the author's paper [Ber19c], to which we refer for further discussion. To formally describe the problem, we consider a graph with a finite set of nodes $X \cup \{t\}$ and a finite set of directed arcs $\mathcal{A} \subset \big\{(x, y) \mid x, y \in X \cup \{t\}\big\}$, where t is a special node called the *destination*. At each node $x \in X$ we may choose a control u from a

nonempty set $U(x)$, which is a subset of a finite set U. Then a successor node y is selected by an antagonistic opponent from a nonempty set $Y(x,u) \subset X \cup \{t\}$ and a cost $g(x,u,y)$ is incurred. The destination node t is absorbing and cost-free, in the sense that the only outgoing arc from t is (t,t), and we have $Y(t,u) = \{t\}$ and $g(t,u,t) = 0$ for all $u \in U(t)$.

As earlier, we denote the set of all policies by Π, and the finite set of all stationary policies by $\mathcal{M}$. Also, we denote the set of functions $J : X \mapsto [-\infty,\infty]$ by $\mathcal{E}(X)$, and the set of functions $J : X \mapsto (-\infty,\infty)$ by $\mathcal{R}(X)$. We introduce the mapping $H : X \times U \times \mathcal{E}(X) \mapsto [-\infty,\infty]$ given by

$$H(x,u,J) = \max_{y \in Y(x,u)} \big[g(x,u,y) + \tilde{J}(y) \big], \qquad x \in X, \tag{3.38}$$

where for any $J \in \mathcal{E}(X)$ we denote by $\tilde{J}$ the function given by

$$\tilde{J}(y) = \begin{cases} J(y) & \text{if } y \in X, \\ 0 & \text{if } y = t. \end{cases} \tag{3.39}$$

We consider the mapping $T : \mathcal{E}(X) \mapsto \mathcal{E}(X)$ defined by

$$(TJ)(x) = \min_{u \in U(x)} H(x,u,J), \qquad x \in X, \tag{3.40}$$

and for each policy μ, the mapping $T_\mu : \mathcal{E}(X) \mapsto \mathcal{E}(X)$, defined by

$$(T_\mu J)(x) = H\big(x, \mu(x), J\big), \qquad x \in X. \tag{3.41}$$

We let $\bar{J}$ be the zero function,

$$\bar{J}(x) = 0, \qquad \forall\ x \in X.$$

The cost function of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ is

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

and $J^*(x) = \inf_{\pi \in \Pi} J_\pi(x)$, cf. Definition 3.2.1.

For a policy $\mu \in \mathcal{M}$, we define a *possible path under μ starting at node $x_0 \in X$* to be an arc sequence of the form

$$p = \big\{ (x_0,x_1), (x_1,x_2), \ldots \big\},$$

such that $x_{k+1} \in Y\big(x_k, \mu(x_k)\big)$ for all $k \geq 0$. The set of all possible paths under μ starting at x_0 is denoted by $P(x_0,\mu)$. The length of a path $p \in P(x_0,\mu)$ is defined by

$$L_\mu(p) = \limsup_{m \to \infty} \sum_{k=0}^{m} g\big(x_k, \mu(x_k), x_{k+1}\big).$$

Using Eqs. (3.38)-(3.41), we see that for any $\mu \in \mathcal{M}$ and $x \in X$, $(T_\mu^k \bar{J})(x)$ is the result of the k-stage DP algorithm that computes the length of the *longest path* under μ that starts at x and consists of k arcs.

For completeness, we also define the length of a portion

$$\big\{(x_i, x_{i+1}), (x_{i+1}, x_{i+2}), \ldots, (x_m, x_{m+1})\big\}$$

of a path $p \in P(x_0, \mu)$, consisting of a finite number of consecutive arcs, by

$$\sum_{k=i}^{m} g\big(x_k, \mu(x_k), x_{k+1}\big).$$

When confusion cannot arise we will also refer to such a finite-arc portion as a path. Of special interest are *cycles*, i.e., paths of the form $\big\{(x_i, x_{i+1}), (x_{i+1}, x_{i+2}), \ldots, (x_{i+m}, x_i)\big\}$. Paths that do not contain any cycle other than the self-cycle (t, t) are called *simple*.

For a given policy $\mu \in \mathcal{M}$ and $x_0 \neq t$, a path $p \in P(x_0, \mu)$ is said to be *terminating* if it has the form

$$p = \big\{(x_0, x_1), (x_1, x_2), \ldots, (x_m, t), (t, t), \ldots\big\}, \tag{3.42}$$

where m is a positive integer, and $x_0, \ldots, x_m$ are distinct nondestination nodes. Since $g(t, u, t) = 0$ for all $u \in U(t)$, the length of a terminating path p of the form (3.42), corresponding to μ, is given by

$$L_\mu(p) = g\big(x_m, \mu(x_m), t\big) + \sum_{k=0}^{m-1} g\big(x_k, \mu(x_k), x_{k+1}\big),$$

and is equal to the finite length of its initial portion that consists of the first $m + 1$ arcs.

An important characterization of a policy $\mu \in \mathcal{M}$ is provided by the subset of arcs

$$\mathcal{A}_\mu = \cup_{x \in X} \big\{(x, y) \mid y \in Y\big(x, \mu(x)\big)\big\}.$$

Thus $\mathcal{A}_\mu \cup (t, t)$ can be viewed as the set of all possible paths under μ, $\cup_{x \in X} P(x, \mu)$, in the sense that it contains this set of paths and no other paths. We refer to $\mathcal{A}_\mu$ as the *characteristic graph of* μ. We say that $\mathcal{A}_\mu$ is *destination-connected* if for each $x \in X$ there exists a terminating path in $P(x, \mu)$.

We say that μ is *proper* if the characteristic graph $\mathcal{A}_\mu$ is acyclic (i.e., contains no cycles). Thus μ is proper if and only if all the paths in $\cup_{x \in X} P(x, \mu)$ are simple and hence terminating (equivalently μ is proper if and only if $\mathcal{A}_\mu$ is destination-connected and has no cycles). The term "proper" is consistent with the one used in Section 3.5.1 for SSP problems, where it indicates a policy under which the destination is reached

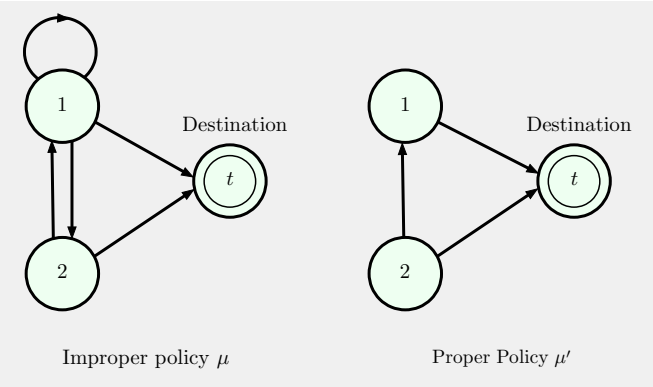


Figure 3.5.4. A robust shortest path problem with $X = \{1, 2\}$, two controls at node 1, and one control at node 2. The two policies, μ and μ', correspond to the two controls at node 1. The figure shows the characteristic graphs $\mathcal{A}_\mu$ and $\mathcal{A}_{\mu'}$.

with probability 1. If μ is not proper, it is called *improper*, in which case the characteristic graph $\mathcal{A}_\mu$ must contain a cycle; see the examples of Fig. 3.5.4. Intuitively, a policy is improper, if and only if under that policy there are initial states such that the antagonistic opponent can force movement along a cycle without ever reaching the destination.

The following proposition clarifies the properties of J_μ when μ is improper.

Proposition 3.5.9: Let μ be an improper policy.

(a) If all cycles in the characteristic graph $\mathcal{A}_\mu$ have nonpositive length, $J_\mu(x) < \infty$ for all $x \in X$.

(b) If all cycles in the characteristic graph $\mathcal{A}_\mu$ have nonnegative length, $J_\mu(x) > -\infty$ for all $x \in X$.

(c) If all cycles in the characteristic graph $\mathcal{A}_\mu$ have zero length, J_μ is real-valued.

(d) If there is a positive length cycle in the characteristic graph $\mathcal{A}_\mu$, we have $J_\mu(x) = \infty$ for at least one node $x \in X$. More generally, for each $J \in \mathcal{R}(X)$, we have $\limsup_{k\to\infty}(T_\mu^k J)(x) = \infty$ for at least one $x \in X$.

Proof: Any path with a finite number of arcs, can be decomposed into a simple path, and a finite number of cycles (see e.g., the path decomposition theorem of [Ber98], Prop. 1.1, and Exercise 1.4). Since there is only a finite number of simple paths under μ, their length is bounded above and below. Thus in part (a) the length of all paths with a finite number of

arcs is bounded above, and in part (b) it is bounded below, implying that $J_\mu(x) < \infty$ for all $x \in X$ or $J_\mu(x) > -\infty$ for all $x \in X$, respectively. Part (c) follows by combining parts (a) and (b).

To show part (d), consider a path p, which consists of an infinite repetition of the positive length cycle that is assumed to exist. Let $C_\mu^k(p)$ be the length of the path that consists of the first k cycles in p. Then $C_\mu^k(p) \to \infty$ and $C_\mu^k(p) \leq J_\mu(x)$ for all k, where x is the first node in the cycle, thus implying that $J_\mu(x) = \infty$. Moreover for every $J \in \mathcal{R}(X)$ and all k, $(T_\mu^k J)(x)$ is the maximum over the lengths of the k-arc paths that start at x, plus a terminal cost that is equal to either $J(y)$ (if the terminal node of the k-arc path is $y \in X$), or 0 (if the terminal node of the k-arc path is the destination). Thus we have,

$$(T_\mu^k \bar{J})(x) + \min\left\{0, \min_{x \in X} J(x)\right\} \leq (T_\mu^k J)(x).$$

Since $\limsup_{k\to\infty}(T_\mu^k \bar{J})(x) = J_\mu(x) = \infty$ as shown earlier, it follows that $\limsup_{k\to\infty}(T_\mu^k J)(x) = \infty$ for all $J \in \mathcal{R}(X)$. **Q.E.D.**

Note that if there is a negative length cycle in the characteristic graph $\mathcal{A}_\mu$, it is not necessarily true that for some $x \in X$ we have $J_\mu(x) = -\infty$. Even for x on the negative length cycle, the value of $J_\mu(x)$ is determined by the *longest* path in $P(x, \mu)$, which may be simple in which case $J_\mu(x)$ is a real number, or contain an infinite repetition of a positive length cycle in which case $J_\mu(x) = \infty$.

Properness and Regularity

We will now make a formal connection between the notions of properness and $\mathcal{R}(X)$-regularity. We recall that μ is $\mathcal{R}(X)$-regular if $J_\mu \in \mathcal{R}(X)$, $J_\mu = T_\mu J_\mu$, and $T_\mu^k J \to J_\mu$ for all $J \in \mathcal{R}(X)$ (cf. Definition 3.2.2). Clearly if μ is proper, we have $J_\mu \in \mathcal{R}(X)$ and the equation $J_\mu = T_\mu J_\mu$ holds (this is Bellman's equation for the longest path problem involving the acyclic graph $\mathcal{A}_\mu$). We will also show that $T_\mu^k J \to J_\mu$ for all $J \in \mathcal{R}(X)$, so that a proper policy is $\mathcal{R}(X)$-regular. However, the following proposition shows that there may be some $\mathcal{R}(X)$-regular policies that are improper, depending on the sign of the lengths of their associated cycles.

Proposition 3.5.10: The following are equivalent for a policy μ:

(i) μ is $\mathcal{R}(X)$-regular.

(ii) The characteristic graph $\mathcal{A}_\mu$ is destination-connected and all its cycles have negative length.

(iii) μ is either proper or else it is improper, all the cycles of the characteristic graph $\mathcal{A}_\mu$ have negative length, and $J_\mu \in \mathcal{R}(X)$.

Proof: To show that (i) implies (ii), let μ be $\mathcal{R}(X)$-regular and to arrive at a contradiction, assume that $\mathcal{A}_\mu$ contains a nonnegative length cycle. Let x be a node on the cycle, consider the path p that starts at x and consists of an infinite repetition of this cycle, and let $L^k_\mu(p)$ be the length of the first k arcs of that path. Let also J be a constant function, $J(x) \equiv r$, where r is a scalar. Then we have

$$L^k_\mu(p) + r \leq (T^k_\mu J)(x),$$

since from the definition of T_μ, we have that $(T^k_\mu J)(x)$ is the maximum over the lengths of all k-arc paths under μ starting at x, plus r, if the last node in the path is not the destination. Since μ is $\mathcal{R}(X)$-regular, we have $J_\mu \in \mathcal{R}(X)$ and $\limsup_{k\to\infty}(T^k_\mu J)(x) = J_\mu(x) < \infty$, so that for all scalars r,

$$\limsup_{k\to\infty} \left(L^k_\mu(p) + r\right) \leq J_\mu(x) < \infty.$$

Taking supremum over $r \in \Re$, it follows that $\limsup_{k\to\infty} L^k_\mu(p) = -\infty$, which contradicts the nonnegativity of the cycle of p. Thus all cycles of $\mathcal{A}_\mu$ have negative length. To show that $\mathcal{A}_\mu$ is destination-connected, assume the contrary. Then there exists some node $x \in X$ such that all paths in $P(x,\mu)$ contain an infinite number of cycles. Since the length of all cycles is negative, as just shown, it follows that $J_\mu(x) = -\infty$, which contradicts the $\mathcal{R}(X)$-regularity of μ.

To show that (ii) implies (iii), we assume that μ is improper and show that $J_\mu \in \mathcal{R}(X)$. By (ii) $\mathcal{A}_\mu$ is destination-connected, so the set $P(x,\mu)$ contains a simple path for all $x \in X$. Moreover, since by (ii) the cycles of $\mathcal{A}_\mu$ have negative length, each path in $P(x,\mu)$ that is not simple has smaller length than some simple path in $P(x,\mu)$. This implies that $J_\mu(x)$ is equal to the largest path length among simple paths in $P(x,\mu)$, so $J_\mu(x)$ is a real number for all $x \in X$.

To show that (iii) implies (i), we note that if μ is proper, it is $\mathcal{R}(X)$-regular, so we focus on the case where μ is improper. Then by (iii), $J_\mu \in \mathcal{R}(X)$, so to show $\mathcal{R}(X)$-regularity of μ, we must show that $(T^k_\mu J)(x) \to J_\mu(x)$ for all $x \in X$ and $J \in \mathcal{R}(X)$, and that $J_\mu = T_\mu J_\mu$. Indeed, from the definition of T_μ, we have

$$(T^k_\mu J)(x) = \sup_{p\in P(x,\mu)} \left[L^k_\mu(p) + J(x^k_p)\right], \tag{3.43}$$

where $L^k_\mu(p)$ is the length of the first k arcs of path p, x^k_p is the node reached after k arcs along the path p, and $J(t)$ is defined to be equal to 0. Thus as $k \to \infty$, for every path p that contains an infinite number of cycles (each necessarily having negative length), the sequence $L^k_p(\mu) + J(x^k_p)$ approaches $-\infty$. It follows that for sufficiently large k, the supremum in Eq. (3.43) is attained by one of the simple paths in $P(x,\mu)$, so $x^k_p = t$ and $J(x^k_p) = 0$. Thus the limit of $(T^k_\mu J)(x)$ does not depend on J, and is equal to the limit

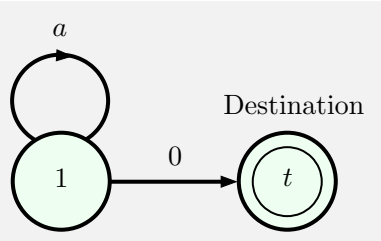


Figure 3.5.5. The characteristic graph $\mathcal{A}_\mu$ corresponding to an improper policy, for the case of a single node 1 and a destination node t. The arcs lengths are shown in the figure.

of $(T_\mu^k \bar{J})(x)$, i.e., $J_\mu(x)$. To show that $J_\mu = T_\mu J_\mu$, we note that by the preceding argument, $J_\mu(x)$ is the length of the longest path among paths that start at x and terminate at t. Moreover, we have

$$(T_\mu J_\mu)(x) = \max_{y \in Y(x,\mu(x))} \big[g(x,\mu(x),y) + J_\mu(y)\big],$$

where we denote $J_\mu(t) = 0$. Thus $(T_\mu J_\mu)(x)$ is also the length of the longest path among paths that start at x and terminate at t, and hence it is equal to $J_\mu(x)$. **Q.E.D.**

We illustrate the preceding proposition, in relation to the infinite cost condition of Assumption 3.3.1, with a two-node example involving an improper policy with a cycle that may have positive, zero, or negative length.

Example 3.5.4:

Let $X = \{1\}$, and consider the policy μ where at state 1, the antagonistic opponent may force either staying at 1 or terminating, i.e., $Y\big(1,\mu(1)\big) = \{1, t\}$; cf. Fig. 3.5.5. Then μ is improper since its characteristic graph $\mathcal{A}_\mu$ contains the self-cycle $(1,1)$. Let

$$g\big(1,\mu(1),1\big) = a, \qquad g\big(1,\mu(1),t\big) = 0.$$

Then,

$$(T_\mu J_\mu)(1) = \max\big[0,\, a + J_\mu(1)\big],$$

and

$$J_\mu(1) = \begin{cases} \infty & \text{if } a > 0, \\ 0 & \text{if } a \leq 0. \end{cases}$$

Consistently with Prop. 3.5.10, the following hold:

(a) For $a > 0$, the cycle $(1,1)$ has positive length, and μ is $\mathcal{R}(X)$-irregular. Here we have $J_\mu(1) = \infty$, and the infinite cost condition of Assumption 3.3.1 is satisfied.

(b) For $a = 0$, the cycle $(1, 1)$ has zero length, and μ is $\mathcal{R}(X)$-irregular. Here we have $J_\mu(1) = 0$, and the infinite cost condition of Assumption 3.3.1 is violated because for a function $J \in \mathcal{R}(X)$ with $J(1) > 0$,

$$\limsup_{k\to\infty}(T_\mu^k J)(x) = J(1) > 0 = J_\mu(1).$$

(c) For $\alpha < 0$, the cycle $(1, 1)$ has negative length, and μ is $\mathcal{R}(X)$-regular. Here we have $J_\mu \in \mathcal{R}(X)$, $J_\mu(1) = \max\big[0,\, a + J_\mu(1)\big] = (T_\mu J_\mu)(1)$, and for all $J \in \mathcal{R}(X)$,

$$\lim_{k\to\infty}(T_\mu^k J)(1) = 0 = J_\mu(1).$$

We will now apply the regularity results of Sections 3.2-3.4 with $S = \mathcal{R}(X)$. To this end, we introduce assumptions that will allow the use of Prop. 3.3.1.

Assumption 3.5.8:

(a) There exists at least one $\mathcal{R}(X)$-regular policy.

(b) For every $\mathcal{R}(X)$-irregular policy μ, some cycle in the characteristic graph $\mathcal{A}_\mu$ has positive length.

Assumption 3.5.8 is implied by the weaker conditions given in the following proposition. These conditions may be more easily verifiable in some contexts.

Proposition 3.5.11: Assumption 3.5.8 holds if anyone of the following two conditions is satisfied.

(1) There exists at least one proper policy, and for every improper policy μ, all cycles in the characteristic graph $\mathcal{A}_\mu$ have positive length.

(2) Every policy μ is either proper or else it is improper and its characteristic graph $\mathcal{A}_\mu$ is destination-connected with all cycles having negative length, and $J_\mu \in \mathcal{R}(X)$.

Proof: Under condition (1), by Prop. 3.5.10, a policy is $\mathcal{R}(X)$-regular if and only if it is proper. Moreover, since each $\mathcal{R}(X)$-irregular and hence improper policy μ has cycles with positive length, it follows that for all $J \in \mathcal{R}(X)$, we have

$$\limsup_{k\to\infty}(T_\mu^k J)(x) = \infty$$

for some $x \in X$. The proof under condition (2) is similar, using Prop. 3.5.10. **Q.E.D.**

We now show our main result for the problem of this section.

Proposition 3.5.12: Let Assumption 3.5.8 hold. Then:

(a) The optimal cost function J^* is the unique fixed point of T within $\mathcal{R}(X)$.

(b) We have $T^k J \to J^*$ for all $J \in \mathcal{R}(X)$.

(c) A policy μ^* is optimal if and only if $T_{\mu^*} J^* = TJ^*$. Moreover, there exists an optimal proper policy.

(d) For any $J \in \mathcal{R}(X)$, if $J \leq TJ$ we have $J \leq J^*$, and if $J \geq TJ$ we have $J \geq J^*$.

Proof: We verify the parts (a)-(f) of Assumption 3.3.1 with $S = \mathcal{R}(X)$, and we then use Prop. 3.3.1. To this end we argue as follows:

(1) Part (a) is satisfied since $S = \mathcal{R}(X)$.

(2) Part (b) is satisfied since by Assumption 3.5.8(a), there exists at least one $\mathcal{R}(X)$-regular policy. Moreover, for each $\mathcal{R}(X)$-regular policy μ, we have $J_\mu \in \mathcal{R}(X)$. Since the number of all policies is finite, it follows that $J_S^* \in \mathcal{R}(X)$.

(3) To show that part (c) is satisfied, note that by Prop. 3.5.10 every $\mathcal{R}(X)$-irregular policy μ must be improper, so by Assumption 3.5.8(b), the characteristic graph $\mathcal{A}_\mu$ contains a cycle of positive length. By Prop. 3.5.9(d), this implies that for each $J \in \mathcal{R}(X)$ and for at least one $x \in X$, we have $\limsup_{k\to\infty}(T_\mu^k J)(x) = \infty$.

(4) Part (d) is satisfied since $U(x)$ is a finite set.

(5) Part (e) is satisfied since X is finite and T_μ is a continuous function that maps the finite-dimensional space $\mathcal{R}(X)$ into itself.

(6) Part (f) follows from Prop. 3.3.2, which applies because $S = \mathcal{R}(X) = R_b(X)$ (since X is finite) and Eq. (3.27) clearly holds.

Thus all parts of Assumption 3.3.1 are satisfied, and Prop. 3.3.1 applies with $S = \mathcal{R}(X)$. The conclusions of this proposition are precisely the results we want to prove [since improper policies have infinite cost for some initial states, as argued earlier, optimal S-regular policies must be proper; cf. the conclusion of part (c)]. **Q.E.D.**

The following example illustrates what may happen in the absence of Assumption 3.5.8(b), when there may exist improper policies that involve

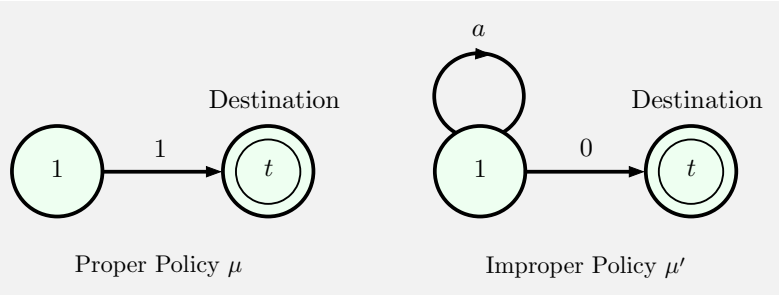


Figure 3.5.6. A counterexample involving a single node 1 in addition to the destination t. There are two policies, μ and μ', with corresponding characteristic graphs $\mathcal{A}_\mu$ and $\mathcal{A}_{\mu'}$, and arc lengths shown in the figure. The improper policy μ' is optimal when $a \leq 0$. It is $\mathcal{R}(X)$-irregular if $a = 0$, and it is $\mathcal{R}(X)$-regular if $a < 0$.

a nonpositive length cycle.

Example 3.5.5:

Let $X = \{1\}$, and consider the proper policy μ with $Y\big(1, \mu(1)\big) = \{t\}$ and the improper policy μ' with $Y\big(1, \mu'(1)\big) = \{1, t\}$ (cf. Fig. 3.5.6). Let

$$g\big(1, \mu(1), t\big) = 1, \qquad g\big(1, \mu'(1), 1\big) = a \leq 0, \qquad g\big(1, \mu'(1), t\big) = 0.$$

The improper policy is the same as the one of Example 3.5.4. It can be seen that under both policies, the longest path from 1 to t consists of the arc $(1, t)$. Thus,

$$J_\mu(1) = 1, \qquad J_{\mu'}(1) = 0,$$

so the improper policy μ' is optimal, and strictly dominates the proper policy μ. To explain what is happening here, we consider two different cases:

(1) $a = 0$: In this case, the optimal policy μ' is both improper and $\mathcal{R}(X)$-irregular, but with finite cost $J_{\mu'}(1) < \infty$. Thus the conditions of Props. 3.3.1 and 3.5.12 do not hold because Assumptions 3.3.1(c) and 3.5.9(b) are violated.

(2) $a < 0$: In this case, μ' is improper but $\mathcal{R}(X)$-regular, so there are no $\mathcal{R}(X)$-irregular policies. Then all the conditions of Assumption 3.5.8 are satisfied, and Prop. 3.5.12 applies. Consistent with this proposition, there exists an optimal $\mathcal{R}(X)$-regular policy (i.e., optimal over both proper and improper policies), which however is improper.

For further analysis and algorithms for the robust shortest path planning problem, we refer to the paper [Ber19c]. In particular, this paper applies the perturbation approach of Section 3.4 to the case where it may be easier to guarantee nonnegativity rather than positivity of the lengths

of cycles corresponding to improper policies, which is required by Assumption 3.5.8(b). The paper shows that the VI algorithm terminates in a finite number of iterations starting from the initial function J with $J(x) = \infty$ for all $x \in X$. Moreover the paper provides a Dijkstra-like algorithm for problems with nonnegative arc lengths.

3.5.4 Linear-Quadratic Optimal Control

In this subsection, we consider a classical problem from control theory, which involves the deterministic linear system

$$x_{k+1} = Ax_k + Bu_k, \qquad k = 0, 1, \ldots,$$

where $x_k \in \Re^n$, $u_k \in \Re^m$ for all k, and A and B are given matrices. The cost function of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ has the form

$$J_\pi(x_0) = \lim_{N\to\infty} \sum_{k=0}^{N-1} \big(x_k' Q x_k + \mu_k(x_k)' R \mu_k(x_k)\big),$$

where x' denotes the transpose of a column vector x, Q is a positive semidefinite symmetric $n \times n$ matrix, and R is a positive definite symmetric $m \times m$ matrix. This is a special case of the deterministic optimal control problem of Section 1.1, and was discussed briefly in the context of the one-dimensional example of Section 3.1.4.

The theory of this problem is well-known and is discussed in various forms in many sources, including the textbooks [AnM79] and [Ber17a] (Section 3.1). The solution revolves around stationary policies μ that are *linear*, in the sense that

$$\mu(x) = Lx,$$

where L is some $m \times n$ matrix, and *stable*, in the sense that the matrix $A + BL$ has eigenvalues that are strictly within the unit circle. Thus for a linear stable policy, the closed loop system

$$x_{k+1} = (A + BL)x_k$$

is stable. We assume that *there exists at least one linear stable policy*. Among others, this guarantees that the optimal cost function J^* is real-valued (it is bounded above by the real-valued cost function of every linear stable policy).

The solution also revolves around the *algebraic matrix Riccati equation*, which is given by

$$P = A'\big(P - PB(B'PB + R)^{-1}B'P\big)A + Q,$$

where the unknown is P, a symmetric $n \times n$ matrix. It is well-known that if Q is positive definite, then the Riccati equation has a unique solution P^*

within the class of positive semidefinite symmetric matrices, and that the optimal cost function has the form

$$J^*(x) = x'P^*x.$$

Moreover, there is a unique optimal policy, and this policy is linear stable of the form

$$\mu^*(x) = Lx, \qquad L = -(B'P^*B + R)^{-1}B'P^*A.$$

The existence of an optimal linear stable policy can be extended to the case where Q is instead positive semidefinite, but satisfies a certain "detectability" condition; see the textbooks cited earlier.

However, in the general case where Q is positive semidefinite without further assumptions (e.g., $Q = 0$), the example of Section 3.1.4 shows that the optimal policy need not be stable, and in fact the optimal cost function over just the linear stable policies may be different than J^*.† We will discuss this case by using the perturbation-based approach of Section 3.4, and provide results that are consistent with the behavior observed in the example of Section 3.1.4.

To convert the problem to our abstract format, we let

$$X = \Re^n, \qquad U(x) = \Re^m, \qquad \bar{J}(x) = 0, \qquad \forall\ x \in X,$$

$$H(x, u, J) = x'Qx + u'Ru + J(Ax + Bu).$$

Let S be the set of positive semidefinite quadratic functions, i.e.,

$$S = \big\{J \mid J(x) = x'Px,\ P : \text{positive semidefinite symmetric}\big\}.$$

Let $\widehat{\mathcal{M}}$ be the set of linear stable policies, and note that every linear stable policy is S-regular. This is due to the fact that for every quadratic function $J(x) = x'Px$ and linear stable policy $\mu(x) = Lx$, the k-stage costs $(T_\mu^k J)(x)$ and $(T_\mu^k \bar{J})(x)$ differ by the term

$$x'(A + BL)^{k'}P(A + BL)^k x,$$

which vanishes in the limit as $k \to \infty$, since μ is stable.

Consider the perturbation framework of Section 3.4, with forcing function

$$p(x) = \|x\|^2.$$

† This is also true in the discounted version of the example of Section 3.1.4, where there is a discount factor $\alpha \in (0, 1)$. The Riccati equation then takes the form $P = A'\big(\alpha P - \alpha^2 PB(\alpha B'PB + R)^{-1}B'P\big)A + Q$, and for the given system and cost per stage, it has two solutions, $P^* = 0$ and $\hat{P} = \frac{\alpha\gamma^2 - 1}{\alpha}$. The VI algorithm converges to $\hat{P}$ starting from any $P > 0$.

Then for $\delta > 0$, the mapping $T_{\mu,\delta}$ has the form

$$(T_{\mu,\delta}J)(x) = x'(Q + \delta I)x + \mu(x)'R\mu(x) + J\big(Ax + B\mu(x)\big),$$

where I is the identity, and corresponds to the linear-quadratic problem where Q is replaced by the positive definite matrix $Q + \delta I$. This problem admits a quadratic positive definite optimal cost $\hat{J}_\delta(x) = x'P_\delta^* x$, and an optimal linear stable policy. Moreover, all the conditions of Prop. 3.4.1 can be verified. It follows that J_S^* is equal to the optimal cost over just the linear stable policies $\hat{J}$, and is obtained as $\lim_{\delta\to 0}\hat{J}_\delta$, which also implies that $\hat{J}(x) = x'\hat{P}x$ where $\hat{P} = \lim_{\delta\to 0} P_\delta^*$.

The perturbation line of analysis of the linear-quadratic problem will be generalized in Section 4.5. This generalization will address a deterministic discrete-time infinite horizon optimal control problem involving the system

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

a nonnegative cost per stage $g(x, u)$, and a cost-free termination state. We will introduce there a notion of stability, and we will show that the optimal cost function over the stable policies is the largest solution of Bellman's equation. Moreover, we will show that the VI algorithm and several versions of the PI algorithm are valid for suitable initial conditions.

3.5.5 Continuous-State Deterministic Optimal Control

In this section, we consider an optimal control problem, where the objective is to steer a deterministic system towards a cost-free and absorbing set of states. The system equation is

$$x_{k+1} = f(x_k, u_k), \qquad k = 0, 1, \ldots, \tag{3.44}$$

where x_k and u_k are the state and control at stage k, belonging to sets X and U, respectively, and f is a function mapping $X \times U$ to X. The control u_k must be chosen from a constraint set $U(x_k)$. No restrictions are placed on the nature of X and U: for example, they may be finite sets as in deterministic shortest path problems, or they may be continuous spaces as in classical problems of control to the origin or some other terminal set, including the linear-quadratic problem of Section 3.5.4. The cost per stage is denoted by $g(x, u)$, and is assumed to be a real number. †

Because the system is deterministic, given an initial state x_0, a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ when applied to the system (3.44), generates a unique sequence of state-control pairs $\big(x_k, \mu_k(x_k)\big)$, $k = 0, 1, \ldots$. The corresponding

† In Section 4.5, we will consider a similar problem where the cost per stage will be assumed to be nonnegative, but some other assumptions from the present section (e.g., the subsequent Assumption 3.5.9) will be relaxed.

cost function is

$$J_\pi(x_0) = \limsup_{N\to\infty} \sum_{k=0}^{N-1} g\big(x_k, \mu_k(x_k)\big), \qquad x_0 \in X.$$

We assume that there is a nonempty *stopping set* $X_0 \subset X$, which consists of cost-free and absorbing states in the sense that

$$g(x,u) = 0, \qquad x = f(x,u), \qquad \forall\ x \in X_0,\ u \in U(x). \tag{3.45}$$

Based on our assumptions to be introduced shortly, the objective will be roughly to reach or asymptotically approach the set X_0 at minimum cost.

To formulate a corresponding abstract DP problem, we introduce the mapping $T_\mu : \mathcal{R}(X) \mapsto \mathcal{R}(X)$ by

$$(T_\mu J)(x) = g\big(x, \mu(x)\big) + J\big(f\big(x, \mu(x)\big)\big), \qquad x \in X,$$

and the mapping $T : \mathcal{E}(X) \mapsto \mathcal{E}(X)$ given by

$$(TJ)(x) = \inf_{u\in U(x)} \big\{g(x,u) + J\big(f(x,u)\big)\big\}, \qquad x \in X.$$

Here as earlier, we denote by $\mathcal{R}(X)$ the set of real-valued functions over X, and by $\mathcal{E}(X)$ the set of extended real-valued functions over X. The initial function $\bar{J}$ is the zero function $[\bar{J}(x) \equiv 0]$. An important fact is that because the problem is deterministic, J^* *is a fixed point of* T (cf. Exercise 3.1).

The analysis of the linear-quadratic problem of the preceding section has revealed two distinct types of behavior for the case where $g \geq 0$:

(a) J^* is the unique fixed point of T within the set S (the set of nonnegative definite quadratic functions).

(b) J^* and the optimal cost function $\hat{J}$ over a restricted subset of S-regular policies (the linear stable policies) are both fixed points of T within the set S, but $J^* \neq \hat{J}$, and the VI algorithm converges to $\hat{J}$ when started with a function $J \geq \hat{J}$.

In what follows we will introduce assumptions that preclude case (b); we will postpone the discussion of of problems where we can have $J^* \neq \hat{J}$ for Section 4.5, where we will use a perturbation-based line of analysis. Similar to the linear-quadratic problem, the restricted set of policies that we will consider have some "stability" property: they are either terminating (reach X_0 in a finite number of steps), or else they asymptotically approach X_0 in a manner to be made precise later.

As a first step in the analysis, let us introduce the effective domain of J^*, i.e., the set

$$X^* = \big\{x \in X \mid J^*(x) < \infty\big\}.$$

Ordinarily, in practical applications, the states in X^* are those from which one can reach the stopping set X_0, at least asymptotically. We say that a policy μ is *terminating* if starting from any $x_0 \in X^*$, the state sequence $\{x_k\}$ generated using μ reaches X_0 in finite time, i.e., satisfies $x_{\bar{k}} \in X_0$ for some index $\bar{k}$. The set of terminating policies is denoted by $\widehat{\mathcal{M}}$.

Our key assumption in this section is that for all $x \in X^*$, the optimal cost $J^*(x)$ can be approximated arbitrarily closely by using terminating policies. In Section 4.5 we will relax this assumption.

Assumption 3.5.9: (Near-Optimal Termination) For every pair (x, ϵ) with $x \in X^*$ and $\epsilon > 0$, there exists a terminating policy μ [possibly dependent on (x, ϵ)] that satisfies $J_\mu(x) \leq J^*(x) + \epsilon$.

This assumption implies in particular that the optimal cost function over terminating policies,

$$\hat{J}(x) = \inf_{\mu \in \widehat{\mathcal{M}}} J_\mu(x), \qquad x \in X,$$

is equal to J^*. Note that Assumption 3.5.9 is equivalent to a seemingly weaker assumption where nonstationary policies can be used for termination (see Exercise 3.7).

Specific and easily verifiable conditions that imply Assumption 3.5.9 are given in the exercises. A prominent case is when X and U are finite, so the problem becomes a deterministic shortest path problem. If all cycles of the state transition graph have positive length, then for every π and x with $J_\pi(x) < \infty$ the generated path starting from x and using π must reach the destination, and this implies that there exists an optimal policy that terminates from all $x \in X^*$. Thus, in this case Assumption 3.5.9 is naturally satisfied.

Another interesting case arises when $g(x, u) = 0$ for all (x, u) except if $x \notin X_0$ and the next state $f(x, u)$ is a termination state, in which case the cost of the stage is strictly negative, i.e., $g(x, u) < 0$ only when $f(x, u) \in X_0$. Thus no cost is incurred except for a negative cost upon termination. Intuitively, this is the problem of trying to find the best state from which to terminate, out of all states that are reachable from the initial state x_0. Then, assuming that X_0 can be reached from all states, Assumption 3.5.9 is satisfied.

When X is the n-dimensional Euclidean space $\Re^n$, it may easily happen that the optimal policies are not terminating from some $x \in X^*$, but instead the optimal state trajectories may approach X_0 asymptotically. This is true for example in the linear-quadratic problem of the preceding section, where $X = \Re^n$, $X_0 = \{0\}$, $U = \Re^m$, the system is linear of the form $x_{k+1} = Ax_k + Bu_k$, where A and B are given matrices, and the optimal cost

function is positive definite quadratic. There the optimal policy is linear stable of the form $\mu^*(x) = Lx$, where L is some matrix obtained through the steady-state solution of the Riccati equation. Since the optimal closed-loop system has the form $x_{k+1} = (A+BL)x_k$, the state will typically never reach the termination set $X_0 = \{0\}$ in finite time, although it will approach it asymptotically. However, the Assumption 3.5.9 is satisfied under some natural and easily verifiable conditions (see Exercise 3.8).

Let us consider the set of functions

$$S = \big\{J \in \mathcal{E}(X) \mid \ J(x) = 0, \forall\ x \in X_0,\ J(x) \in \Re, \forall\ x \in X^*\big\}.$$

Since X_0 consists of cost-free and absorbing states [cf. Eq. (3.45)], and $J^*(x) > -\infty$ for all $x \in X$ (by Assumption 3.5.9), the set S contains the cost functions J_μ of all terminating policies μ, as well as J^*. Moreover it can be seen that every terminating policy is S-regular, i.e., $\widehat{\mathcal{M}} \subset \mathcal{M}_S$, implying that $J_S^* = J^*$. The reason is that the terminal cost is zero after termination for any terminal cost function $J \in S$, i.e.,

$$(T_\mu^k J)(x) = (T_\mu^k \bar{J})(x) = J_\mu(x),$$

for $\mu \in \widehat{\mathcal{M}}$, $x \in X^*$, and k sufficiently large.

The following proposition is a consequence of the well-behaved region theorem (Prop. 3.2.1), the deterministic character of the problem (which guarantees that J^* is a fixed point of T; Exercise 3.1), and Assumption 3.5.9 (which guarantees that $J_S^* = J^*$).

Proposition 3.5.13: Let Assumption 3.5.9 hold. Then:

(a) J^* is the unique solution of the Bellman equation $J = TJ$ within the set of all $J \in S$ such that $J \geq J^*$.

(b) We have $T^k J \to J^*$ for every $J \in S$ such that $J \geq J^*$.

(c) If μ^* is terminating and $T_{\mu^*} J^* = TJ^*$, then μ^* is optimal. Conversely, if μ^* is terminating and is optimal, then $T_{\mu^*} J^* = TJ^*$.

Generally, the convergence $T^k J \to J^*$ for every $J \in S$ [Prop. 3.5.13(b)] cannot be shown except in special cases, such as finite-state problems (see Prop. 1.1(b), Ch. 4, of the book by Bertsekas and Tsitsiklis [BeT89]). To see what may happen in the absence of Assumption 3.5.9, consider the deterministic shortest path example of Section 3.1.1 with $a = 0$, $b > 0$, and $S = \Re$. Here Assumption 3.5.9 is violated and we have $0 = J^* < \hat{J} = b$, while the set of fixed points of T is the interval $(-\infty, b]$. However, for the same example, but with $b \leq 0$ instead of $b > 0$, Assumption 3.5.9 is satisfied and Prop. 3.5.13 applies. Consider also the linear-quadratic example of

Section 3.1.4. Here Assumption 3.5.9 is violated. This results in multiple fixed points of T within S: the functions $J^*(x) \equiv 0$ and $\hat{J}(x) = (\gamma^2 - 1)x^2$. In Section 4.5, we will reconsider this example, as well as the problem of this section for the case $g(x, u) \geq 0$ for all (x, u), but under assumptions that are much weaker than Assumption 3.5.9. There, we will make a connection between regularity, perturbations like the ones of Section 3.4, and traditional notions of stability.

Another interesting fact is that when the model of this section is extended in the natural way to a stochastic model with infinite state space, then under the analog of Assumption 3.5.9, J^* need not be the unique solution of Bellman's equation within the set of all $J \in S$ such that $J \geq J^*$. Indeed, we will show this in Section 4.6.1 with a stochastic example that involves a single control per state and nonnegative but unbounded cost per stage (if the cost per stage is nonnegative and bounded, and the optimal cost over the proper policies only is equal to J^*, then J^* will be proved to be the unique solution of Bellman's equation within the set of all bounded J such that $J \geq 0$). This is a striking difference between deterministic and stochastic optimal control problems with infinite state space. Another striking difference is that J^* is always a solution of Bellman's equation in deterministic problems (cf. Exercise 3.1), but this is not so in stochastic problems, even when the state space is finite (cf. Section 3.1.2).

3.6 ALGORITHMS

We have already discussed some VI and PI algorithms for finding J^* and an optimal policy as part of our analysis under the weak and strong PI properties in Section 3.2. Moreover, we have shown that the VI algorithm converges to the optimal cost function J^* for any starting function $J \in S$ in the case of Assumption 3.3.1 (cf. Prop. 3.3.1), or to the restricted optimal cost function J_S^* under the assumptions of Prop. 3.4.1(b).

In this section, we will introduce additional algorithms. In Section 3.6.1, we will discuss asynchronous versions of VI and will prove satisfactory convergence properties under reasonable assumptions. In Section 3.6.2, we will focus on a modified version of PI that is unaffected by the presence of S-irregular policies. This algorithm is similar to the optimistic PI algorithm with uniform fixed point (cf. Section 2.6.3), and can also be implemented in a distributed asynchronous computing environment.

3.6.1 Asynchronous Value Iteration

Let us consider the model of Section 2.6.1 for asynchronous distributed computation of the fixed point of a mapping T, and the asynchronous distributed VI method described there. The model involves a partition of X into disjoint nonempty subsets $X_1, \ldots, X_m$, and a corresponding partition of J as $J = (J_1, \ldots, J_m)$, where J_ℓ is the restriction of J on the set X_ℓ.

We consider a network of m processors, each updating asynchronously corresponding components of J. In particular, we assume that J_ℓ is updated only by processor ℓ, and only for times t in a selected subset $\mathcal{R}_\ell$ of iterations. Moreover, as in Section 2.6.1, processor ℓ uses components J_j supplied by other processors $j \neq \ell$ with communication "delays" $t - \tau_{\ell j}(t) \geq 0$:

$$J_\ell^{t+1}(x) = \begin{cases} T\left(J_1^{\tau_{\ell 1}(t)}, \ldots, J_m^{\tau_{\ell m}(t)}\right)(x) & \text{if } t \in \mathcal{R}_\ell,\ x \in X_\ell, \\ J_\ell^t(x) & \text{if } t \notin \mathcal{R}_\ell,\ x \in X_\ell. \end{cases} \tag{3.46}$$

We can prove convergence within the frameworks of Sections 3.3 and 3.4 by using the asynchronous convergence theorem (cf. Prop. 2.6.1), and the fact that T is monotone and has J^* as its unique fixed point within the appropriate set. We assume that the continuous updating and information renewal Assumption 2.6.1 holds. For simplicity we restrict attention to the framework of Section 3.3, under Assumption 3.3.1 with $S = \mathcal{B}(X)$. Assume further that we have two functions $\underline{V}, \overline{V} \in S$ such that

$$\underline{V} \leq T\underline{V} \leq T\overline{V} \leq \overline{V}, \tag{3.47}$$

so that, by Prop. 3.3.1, $T^k\underline{V} \leq J^* \leq T^k\overline{V}$ for all k, and

$$T^k\underline{V} \uparrow J^*, \qquad T^k\overline{V} \downarrow J^*.$$

Then we can show asynchronous convergence of the VI algorithm (3.46), starting from any function J^0 with $\underline{V} \leq J^0 \leq \overline{V}$.

Indeed, let us apply Prop. 2.6.1 with the sets $S(k)$ given by

$$S(k) = \left\{J \in S \mid T^k\underline{V} \leq J \leq T^k\overline{V}\right\}, \qquad k = 0, 1, \ldots.$$

The sets $S(k)$ satisfy $S(k+1) \subset S(k)$ in view of Eq. (3.47) and the monotonicity of T. Using Prop. 3.3.1, we also see that $S(k)$ satisfy the synchronous convergence and box conditions of Prop. 2.6.1. Thus, together with Assumption 2.6.1, all the conditions of Prop. 2.6.1 are satisfied, and the convergence of the algorithm follows starting from any $J^0 \in S(0)$.

3.6.2 Asynchronous Policy Iteration

In this section, we focus on PI methods, under Assumption 3.3.1 and some additional assumptions to be introduced shortly. We first discuss briefly a natural form of PI algorithm, which generates S-regular policies exclusively. Let μ^0 be an initial S-regular policy [there exists one by Assumption 3.3.1(b)]. At the typical iteration k, we have an S-regular policy μ^k, and we compute a policy μ^{k+1} such that $T_{\mu^{k+1}}J_{\mu^k} = TJ_{\mu^k}$ (this is possible by Lemma 3.3.1). Then μ^{k+1} is S-regular, by Lemma 3.3.2, and we have

$$J_{\mu^k} = T_{\mu^k}J_{\mu^k} \geq TJ_{\mu^k} = T_{\mu^{k+1}}J_{\mu^k} \geq \lim_{m\to\infty} T^k_{\mu^{k+1}}J_{\mu^k} = J_{\mu^{k+1}}.$$

We can thus construct a sequence of S-regular policies $\{\mu^k\}$ and a corresponding nonincreasing sequence $\{J_{\mu^k}\}$. Under some additional mild conditions it is then possible to show that $J_{\mu^k} \downarrow J^*$, cf. Prop. 3.3.1(e).

Unfortunately, when there are S-irregular policies, the preceding PI algorithm is somewhat limited, because an initial S-regular policy may not be known. Moreover, when asynchronous versions of the algorithm are implemented, it is difficult to guarantee that all the generated policies are S-regular.

In what follows in this section, we will discuss a PI algorithm that works in the presence of S-irregular policies, and can operate in a distributed asynchronous environment, like the PI algorithm for contractive models of Section 2.6.3. The main assumption is that J^* is the unique fixed point of T within $\mathcal{R}(X)$, the set of real-valued functions over X. This assumption holds under Assumption 3.3.1 with $S = \mathcal{R}(X)$, but it also holds under weaker conditions. Our assumptions also include finiteness of U, which among others facilitates the policy evaluation and policy improvement operations, and ensures that the algorithm generates iterates that lie in $\mathcal{R}(X)$. The algorithm and its analysis also go through if $\mathcal{R}(X)$ is replaced by $\mathcal{R}^+(X)$ (the set of all nonnegative real-valued functions) in the following assumptions, arguments, and propositions.

Assumption 3.6.1: In addition to the monotonicity Assumption 3.2.1, the following hold.

(a) $H(x, u, J)$ is real-valued for all $J \in \mathcal{R}(X)$, $x \in X$, and $u \in U(x)$.

(b) U is a finite set.

(c) For each sequence $\{J_m\} \subset \mathcal{R}(X)$ with either $J_m \uparrow J$ or $J_m \downarrow J$ for some $J \in \mathcal{R}(X)$, we have

$$\lim_{m\to\infty} H(x, u, J_m) = H(x, u, J), \qquad \forall\ x \in X,\ u \in U(x).$$

(d) For all scalars $r > 0$ and functions $J \in \mathcal{R}(X)$, we have

$$H(x, u, J + r\,e) \le H(x, u, J) + r\,e, \qquad \forall\ x \in X,\ u \in U(x), \tag{3.48}$$

where e is the unit function.

(e) J^* is the unique fixed point of T within $\mathcal{R}(X)$.

Part (d) of the preceding assumption is a nonexpansiveness condition for $H(x, u, \cdot)$, and can be easily verified in many DP models, including deterministic, minimax, and stochastic optimal control problems. It is not readily satisfied, however, in the affine monotonic model of Section 3.5.2.

Similar to Section 2.6.3, we introduce a new mapping that is parametrized by μ and can be shown to have a common fixed point for all μ. It operates on a pair (V, Q) where:

- V is a real-valued function with a component denoted $V(x)$ for each $x \in X$.
- Q is a real-valued function with a component denoted $Q(x, u)$ for each pair (x, u) with $x \in X$, $u \in U(x)$.

The mapping produces a pair

$$\big(MF_\mu(V,Q), F_\mu(V,Q)\big),$$

where

- $F_\mu(V, Q)$ is a function with a component $F_\mu(V, Q)(x, u)$ for each (x, u), defined by
$$F_\mu(V,Q)(x,u) = H\big(x,u,\min\{V,Q_\mu\}\big), \tag{3.49}$$
where for any Q and μ, we denote by Q_μ the function of x defined by
$$Q_\mu(x) = Q\big(x,\mu(x)\big), \qquad x \in X,$$
and for any two functions V_1 and V_2, we denote by $\min\{V_1, V_2\}$ the function of x given by
$$\min\{V_1,V_2\}(x) = \min\big\{V_1(x), V_2(x)\big\}, \qquad x \in X.$$

- $MF_\mu(V, Q)$ is a function with a component $\big(MF_\mu(V,Q)\big)(x)$ for each x, where M is the operator of pointwise minimization over u:
$$(MQ)(x) = \min_{u\in U(x)} Q(x,u),$$
so that
$$\big(MF_\mu(V,Q)\big)(x) = \min_{u\in U(x)} F_\mu(V,Q)(x,u).$$
Note that under Assumption 3.6.1, M maps real-valued functions to real-valued functions, since by part (b) of that assumption, U is assumed finite.

We consider an algorithm that is similar to the asynchronous PI algorithm given in Section 2.6.3 for contractive models. It applies asynchronously the mapping $MF_\mu(V, Q)$ for local policy improvement and update of V and μ, and the mapping $F_\mu(V, Q)$ for local policy evaluation and update of Q. The algorithm involves a partition of the state space into sets $X_1, \ldots, X_m$, and assignment of each subset X_ℓ to a processor $\ell \in \{1, \ldots, m\}$. For each ℓ, there are two infinite disjoint subsets of times

$\mathcal{R}_\ell, \overline{\mathcal{R}}_\ell \subset \{0, 1, \ldots\}$, corresponding to policy improvement and policy evaluation iterations, respectively. At time t, each processor ℓ operates on $V^t(x)$, $Q^t(x,u)$, and $\mu^t(x)$, only for x in its "local" state space X_ℓ. In particular, at each time t, each processor ℓ does one of the following:

(a) *Local policy improvement*: If $t \in \mathcal{R}_\ell$, processor ℓ sets for all $x \in X_\ell$,

$$V^{t+1}(x) = \min_{u \in U(x)} H\big(x, u, \min\{V^t, Q^t_{\mu^t}\}\big) = \big(MF_{\mu^t}(V^t, Q^t)\big)(x), \tag{3.50}$$

sets $\mu^{t+1}(x)$ to a u that attains the minimum, and leaves Q unchanged, i.e., $Q^{t+1}(x,u) = Q^t(x,u)$ for all $x \in X_\ell$ and $u \in U(x)$.

(b) *Local policy evaluation*: If $t \in \overline{\mathcal{R}}_\ell$, processor ℓ sets for all $x \in X_\ell$ and $u \in U(x)$,

$$Q^{t+1}(x,u) = H\big(x, u, \min\{V^t, Q^t_{\mu^t}\}\big) = F_{\mu^t}(V^t, Q^t)(x,u), \tag{3.51}$$

and leaves V and μ unchanged, i.e., $V^{t+1}(x) = V^t(x)$ and $\mu^{t+1}(x) = \mu^t(x)$ for all $x \in X_\ell$.

(c) *No local change*: If $t \notin \mathcal{R}_\ell \cup \overline{\mathcal{R}}_\ell$, processor ℓ leaves Q, V, and μ unchanged, i.e., $Q^{t+1}(x,u) = Q^t(x,u)$ for all $x \in X_\ell$ and $u \in U(x)$, $V^{t+1}(x) = V^t(x)$, and $\mu^{t+1}(x) = \mu^t(x)$ for all $x \in X_\ell$.

Under Assumption 3.6.1, the algorithm generates real-valued functions if started with real-valued V^0 and Q^0. We will prove that it converges to (J^*, Q^*), where J^* is the unique fixed point of T within $\mathcal{R}(X)$ [cf. Assumption 3.6.1(e)], and Q^* is defined by

$$Q^*(x,u) = H(x,u,J^*), \qquad x \in X,\ u \in U(x). \tag{3.52}$$

To this end, we introduce the mapping F defined by

$$(FQ)(x,u) = H\big(x,u,MQ\big), \qquad x \in X,\ u \in U(x), \tag{3.53}$$

and we show the following proposition.

Proposition 3.6.1: Let Assumption 3.6.1 hold. Then Q^* is the unique fixed point of F within the class of real-valued functions.

Proof: By minimizing over $u \in U(x)$ in Eq. (3.52) and noting that J^* is a fixed point of T, we have $MQ^* = TJ^* = J^*$. Thus, by applying Eq. (3.53) and then Eq. (3.52), we obtain

$$(FQ^*)(x,u) = H(x,u,J^*) = Q^*(x,u), \qquad \forall\ x \in X,\ u \in U(x).$$

Thus Q^* is a fixed point of F, and it is real-valued since J^* is real-valued and H is real-valued.

To show uniqueness, let Q' be any real-valued fixed point of F. Then $Q'(x,u) = H(x,u,MQ')$ for all $x \in X$, $u \in U(x)$, and by minimization over $u \in U(x)$, we have $MQ' = T(MQ')$. Hence MQ' is equal to the unique fixed point J^* of T, so that the equation $Q' = FQ'$ yields $Q'(x,u) = H(x,u,MQ') = H(x,u,J^*)$, for all (x,u). From the definition (3.52) of Q^*, it then follows that $Q' = Q^*$. **Q.E.D.**

We introduce the μ-dependent mapping

$$L_\mu(V,Q) = \big(MQ, F_\mu(V,Q)\big), \tag{3.54}$$

where $F_\mu(V,Q)$ is given by Eq. (3.49). For this mapping and other related mappings to be defined shortly, we implicitly assume that it operates on real-valued functions, so by Assumption 3.6.1(a),(b), it produces real-valued functions. Note that the policy evaluation part of the algorithm [cf. Eq. (3.51)] amounts to applying the second component of L_μ, while the policy improvement part of the algorithm [cf. Eq. (3.50)] amounts to applying the second component of L_μ, *and* then applying the first component of L_μ. The following proposition shows that (J^*,Q^*) is the common fixed point of the mappings L_μ, for all μ.

Proposition 3.6.2: Let Assumption 3.6.1 hold. Then for all $\mu \in \mathcal{M}$, the mapping L_μ of Eq. (3.54) is monotone, and (J^*,Q^*) is its unique fixed point within the class of real-valued functions.

Proof: Monotonicity of L_μ follows from the monotonicity of the operators M and F_μ. To show that L_μ has (J^*,Q^*) as its unique fixed point, we first note that $J^* = MQ^*$ and $Q^* = FQ^*$; cf. Prop. 3.6.1. Then, using also the definition of F_μ, we have

$$J^* = MQ^*, \qquad Q^* = FQ^* = F_\mu(J^*,Q^*),$$

which shows that (J^*,Q^*) is a fixed point of L_μ.

To show uniqueness, let (V',Q') be a real-valued fixed point of L_μ, i.e., $V' = MQ'$ and $Q' = F_\mu(V',Q')$. Then

$$Q' = F_\mu(V',Q') = FQ',$$

where the last equality follows from $V' = MQ'$. Thus Q' is a fixed point of F, and since Q^* is the unique fixed point of F (cf. Prop. 3.6.1), we have $Q' = Q^*$. It follows that $V' = MQ^* = J^*$, so (J^*,Q^*) is the unique fixed point of L_μ within the class of real-valued functions. **Q.E.D.**

The uniform fixed point property of L_μ just shown is, however, insufficient for the convergence proof of the asynchronous algorithm, in the absence of a contraction property. For this reason, we introduce two mappings $\underline{L}$ and $\overline{L}$ that are associated with the mappings L_μ and satisfy

$$\underline{L}(V,Q) \leq L_\mu(V,Q) \leq \overline{L}(V,Q), \qquad \forall\ \mu \in \mathcal{M}. \tag{3.55}$$

These are the mappings defined by

$$\underline{L}(V,Q) = \Big(MQ, \min_{\mu\in\mathcal{M}} F_\mu(V,Q)\Big), \quad \overline{L}(V,Q) = \Big(MQ, \max_{\mu\in\mathcal{M}} F_\mu(V,Q)\Big), \tag{3.56}$$

where the min and max over μ are attained in view of the finiteness of $\mathcal{M}$ [cf. Assumption 3.6.1(b)]. We will show that $\underline{L}$ and $\overline{L}$ also have (J^*,Q^*) as their unique fixed point. Note that there exists $\bar{\mu}$ that attains the maximum in Eq. (3.56), uniformly for all V and (x,u), namely a policy $\bar{\mu}$ for which

$$Q\big(x,\bar{\mu}(x)\big) = \max_{u\in U(x)} Q(x,u), \qquad \forall\ x \in X,$$

[cf. Eq. (3.49)]. Similarly, there exists $\underline{\mu}$ that attains the minimum in Eq. (3.56), uniformly for all V and (x,u). Thus for any given (V,Q), we have

$$\underline{L}(V,Q) = L_{\underline{\mu}}(V,Q), \qquad \overline{L}(V,Q) = L_{\bar{\mu}}(V,Q), \tag{3.57}$$

where $\underline{\mu}$ and $\bar{\mu}$ are some policies. The following proposition shows that (J^*,Q^*), the common fixed point of the mappings L_μ, for all μ, is also the unique fixed point of $\underline{L}$ and $\overline{L}$.

Proposition 3.6.3: Let Assumption 3.6.1 hold. Then the mappings $\underline{L}$ and $\overline{L}$ of Eq. (3.56) are monotone, and have (J^*,Q^*) as their unique fixed point within the class of real-valued functions.

Proof: Monotonicity of $\underline{L}$ and $\overline{L}$ follows from the monotonicity of the operators M and F_μ. Since (J^*,Q^*) is the common fixed point of L_μ for all μ (cf. Prop. 3.6.2), and there exists $\underline{\mu}$ such that $\underline{L}(J^*,Q^*) = L_{\underline{\mu}}(J^*,Q^*)$ [cf. Eq. (3.57)], it follows that (J^*,Q^*) is a fixed point of $\underline{L}$. To show uniqueness, suppose that (V,Q) is a fixed point, so $(V,Q) = \underline{L}(V,Q)$. Then by Eq. (3.57), we have

$$(V,Q) = \underline{L}(V,Q) = L_{\underline{\mu}}(V,Q)$$

for some $\underline{\mu} \in \mathcal{M}$. Since by Prop. 3.6.2, (J^*,Q^*) is the only fixed point of $L_{\underline{\mu}}$, it follows that $(V,Q) = (J^*,Q^*)$, so (J^*,Q^*) is the only fixed point of $\underline{L}$. Similarly, we show that (J^*,Q^*) is the unique fixed point of $\overline{L}$. **Q.E.D.**

We are now ready to construct a sequence of sets needed to apply Prop. 2.6.1 and prove convergence. For a scalar $c \geq 0$, we denote

$$J_c^- = J^* - c\,e, \qquad Q_c^- = Q^* - c\,e_Q,$$

$$J_c^+ = J^* + c\,e, \qquad Q_c^+ = Q^* + c\,e_Q,$$

with e and e_Q are the unit functions in the spaces of J and Q, respectively.

Proposition 3.6.4: Let Assumption 3.6.1 hold. Then for all $c > 0$,

$$\underline{L}^k(J_c^-, Q_c^-) \uparrow (J^*, Q^*), \qquad \overline{L}^k(J_c^+, Q_c^+) \downarrow (J^*, Q^*), \tag{3.58}$$

where $\underline{L}^k$ (or $\overline{L}^k$) denotes the k-fold composition of $\underline{L}$ (or $\overline{L}$, respectively).

Proof: For any $\mu \in \mathcal{M}$, using the assumption (3.48), we have for all (x, u),

$$\begin{aligned} F_\mu(J_c^+, Q_c^+)(x,u) &= H\big(x, u, \min\{J_c^+, Q_c^+\}\big) \\ &= H\big(x, u, \min\{J^*, Q^*\} + c\,e\big) \\ &\leq H\big(x, u, \min\{J^*, Q^*\}\big) + c \\ &= Q^*(x,u) + c \\ &= Q_c^+(x,u), \end{aligned}$$

and similarly

$$Q_c^-(x,u) \leq F_\mu(J_c^-, Q_c^-)(x,u).$$

We also have $MQ_c^+ = J_c^+$ and $MQ_c^- = J_c^-$. From these relations, the definition of L_μ, and the fact $L_\mu(J^*, Q^*) = (J^*, Q^*)$ (cf. Prop. 3.6.2), we have

$$(J_c^-, Q_c^-) \leq L_\mu(J_c^-, Q_c^-) \leq (J^*, Q^*) \leq L_\mu(J_c^+, Q_c^+) \leq (J_c^+, Q_c^+).$$

Using this relation and Eqs. (3.55) and (3.57), we obtain

$$(J_c^-, Q_c^-) \leq \underline{L}(J_c^-, Q_c^-) \leq (J^*, Q^*) \leq \overline{L}(J_c^+, Q_c^+) \leq (J_c^+, Q_c^+). \tag{3.59}$$

Denote for $k = 0, 1, \ldots,$

$$(\overline{V}_k, \overline{Q}_k) = \overline{L}^k(J_c^+, Q_c^+), \qquad (\underline{V}_k, \underline{Q}_k) = \underline{L}^k(J_c^-, Q_c^-).$$

From the monotonicity of $\overline{L}$ and $\underline{L}$ and Eq. (3.59), we have that $(\overline{V}_k, \overline{Q}_k)$ converges monotonically from above to some pair

$$(\overline{V}, \overline{Q}) \geq (J^*, Q^*),$$

while $(\underline{V}_k, \underline{Q}_k)$ converges monotonically from below to some pair

$$(\underline{V}, \underline{Q}) \leq (J^*, Q^*).$$

By taking the limit in the equation

$$(\overline{V}_{k+1}, \overline{Q}_{k+1}) = \overline{L}(\overline{V}_k, \overline{Q}_k),$$

and using the continuity from above and below property of $\overline{L}$, implied by Assumption 3.6.1(c), it follows that $(\overline{V}, \overline{Q}) = \overline{L}(\overline{V}, \overline{Q})$, so $(\overline{V}, \overline{Q})$ must be equal to (J^*, Q^*), the unique fixed point of $\overline{L}$. Thus, $\overline{L}^k(J_c^+, Q_c^+) \downarrow (J^*, Q^*)$. Similarly, $\underline{L}^k(J_c^-, Q_c^-) \uparrow (J^*, Q^*)$. **Q.E.D.**

To show asynchronous convergence of the algorithm (3.50)-(3.51), consider the sets

$$S(k) = \big\{(V, Q) \mid \underline{L}^k(J_c^-, Q_c^-) \leq (V, Q) \leq \overline{L}^k(J_c^+, Q_c^+)\big\}, \qquad k = 0, 1, \ldots,$$

whose intersection is (J^*, Q^*) [cf. Eq. (3.58)]. By Prop. 3.6.4 and Eq. (3.55), this set sequence together with the mappings L_μ satisfy the synchronous convergence and box conditions of the asynchronous convergence theorem of Prop. 2.6.1 (more precisely, its time-varying version of Exercise 2.2). This proves the convergence of the algorithm (3.50)-(3.51) for starting points $(V, Q) \in S(0)$. Since c can be chosen arbitrarily large, it follows that the algorithm is convergent from an arbitrary starting point.

Finally, let us note some variations of the asynchronous PI algorithm. One such variation is to allow "communication delays" $t - \tau_{\ell j}(t)$. Another variation, for the case where we want to calculate just J^*, is to use a reduced space implementation similar to the one discussed in Section 2.6.3. There is also a variant with interpolation, cf. Section 2.6.3.

3.7 NOTES, SOURCES, AND EXERCISES

The semicontractive model framework of this chapter was first formulated in the 2013 edition of the book, and it has since been extended through a series of papers and reports by the author: [Ber15], [Ber16a], [BeY16], [Ber17c], [Ber17d], [Ber19c]. The framework is inspired from the analysis of the SSP problem of Example 1.2.6, which involves finite state and control spaces, as well as a termination state. In the absence of a termination state, a key idea has been to generalize the notion of a proper policy from one that leads to termination with probability 1, to one that is S-regular for an appropriate set of functions S.

Section 3.1: The counterexample showing that J^* may fail to solve Bellman's equation in SSP problems is due to Bertsekas and Yu [BeY16]. The

blackmailer's dilemma is a classic problem in the DP literature. The book by Whittle [Whi82] has a substantial discussion. The set of solutions of the Riccati equation in continuous-time linear-quadratic optimal control (cf. Section 3.1.4) has been described in the paper by Willems [Wil71], which stimulated considerable further work on the subject (see the book by Lancaster and Rodman [LaR95] for an extensive account). The pathologies of infinite horizon linear-quadratic optimal control problems can be largely eliminated under some well-studied controllability and observability conditions (see, e.g., [Ber17a], Section 3.1).

Section 3.2: The PI-based analysis of Section 3.2 was developed in the author's paper [Ber15] after the 2013 edition of the book was published. The author's joint work with H. Yu [BeY16] was also influential. In particular, the SSP example of Section 3.1.2, where J^* does not satisfy Bellman's equation, and the perturbation analysis of Section 3.4 were given in the paper [BeY16]. This is also the source for the convergence rate result of Prop. 3.2.2. The λ-PI method was introduced by Bertsekas and Ioffe [BeI96] in the context of discounted and SSP problems, and subsequent work includes the papers by Nedić and Bertsekas [NeB03], and by Bertsekas, Borkar, and Nedić [BBN04] on the LSPE(λ) method. The analysis of λ-PI in Section 3.2.4 is new and is related to an analysis of a linearized form of the proximal algorithm given in the author's papers [Ber16b], [Ber18c].

Section 3.3: The central result of Section 3.3, Prop. 3.3.1, was given in the 2013 edition of the book. It is patterned after a result of Bertsekas and Tsitsiklis [BeT91] for SSP problems with finite state space and compact control constraint sets, which is reproduced in Section 3.5.1. The proof given there contains an intricate demonstration of a real-valued lower bound on the cost functions of proper policies (Lemma 3 of [BeT91], which implies Prop. 3.5.3).

Section 3.4: The perturbation approach of Section 3.4 was introduced in the 2013 edition of the book. It is presented here in somewhat stronger form, which will also be applied to nonstationary S-regular policies in the next chapter.

Section 3.5: The SSP problem analysis of Section 3.5.1 for the case of the strong SSP conditions is due to Bertsekas and Tsitsiklis [BeT91]. For the case of the weak SSP conditions it is due to Bertsekas and Yu [BeY16]. The perturbation-based PI algorithm was given in Section 3.3.3 of the 2013 edition of the book. A different PI algorithm that embodies a mechanism for breaking ties in the policy improvement step was given by Guillot and Stauffer [GuS17] for the case of finite state and control spaces.

The affine monotonic model of Section 3.5.2 was initially formulated and analyzed in the 2013 edition of the book, in a more general setting where the state space can be an infinite set. The analysis of Section 3.5.2 of the finite-state case comes from the author's paper [Ber16a], which con-

tains more details. The exponentiated cost version of the SSP problem was analyzed in the papers by Denardo and Rothblum [DeR79], and by Patek [Pat01]. The paper [DeR79] assumes that the state and control spaces are finite, that there exists at least one contractive policy (a transient policy in the terminology of [DeR79]), and that every improper policy is noncontractive and has infinite cost from some initial state. These assumptions bypass the pathologies around infinite control spaces and multiple solutions or no solution of Bellman's equation. Also the approach of [DeR79] is based on linear programming (relying on the finite control space), and is thus quite different from ours. The paper [Pat01] assumes that the state space is finite, that the control constraint set is compact, and that the expected one-stage cost is strictly positive for all state-control pairs, which is much stronger than what we have assumed. Our results of Section 3.5.2, when specialized to the exponential cost problem, are consistent with and subsume the results of Denardo and Rothblum [DeR79], and Patek [Pat01].

The discussion on robust shortest path planning in Section 3.5.3 follows the author's paper [Ber19c]. This paper contains further analysis and computational methods, including a finitely terminating Dijkstra-like algorithm for problems with nonnegative arc lengths.

The deterministic optimal control model of Section 3.5.5 is discussed in more detail in the author's paper [Ber17b] under Assumption 3.5.9 for the case where $g \geq 0$; see also Section 4.5 and the paper [Ber17c]. The analysis under the more general assumptions given here is new. Deterministic and minimax infinite-spaces optimal control problems have also been discussed by Reissig [Rei16] under assumptions different than ours.

Section 3.6: The asynchronous VI algorithm of Section 3.6.1 was first given in the author's paper on distributed DP [Ber82]. It was further formalized in the paper [Ber83], where a DP problem was viewed as a special case of a fixed point problem, involving monotonicity and possibly contraction assumptions.

The analysis of Section 3.6.2, parallels the one of Section 2.6.3, and is due to joint work of the author with H. Yu, presented in the papers [BeY12] and [YuB13a]. In particular, the algorithm of Section 3.6.2 is one of the optimistic PI algorithms in [YuB13a], which was applied to the SSP problem of Section 3.5.1 under the strong SSP conditions. We have followed the line of analysis of that paper and the related paper [BeY12], which focuses on discounted problems. These papers also analyzed asynchronous stochastic iterative versions of PI, and proved convergence results that parallel those for classical Q-learning for SSP, given in Tsitsiklis [Tsi94], and Yu and Bertsekas [YuB13b]. An earlier paper, which deals with a slightly different asynchronous abstract PI algorithm without a contraction structure, is Bertsekas and Yu [BeY10].

By allowing an infinite state space, the analysis of the present chapter applies among others to SSP problems with a countable state space. Such

problems often arise in queueing control settings where the termination state corresponds to an empty queue. The problem then is to empty the queue with minimum expected cost. Generalized forms of SSP problems, which involve an infinite (uncountable) number of states, in addition to the termination state, were analyzed by Pliska [Pli78], Hernandez-Lerma et al. [HCP99], and James and Collins [JaC06]. The latter paper allows improper policies, assumes that g is bounded and J^* is bounded below, and generalizes the results of [BeT91] to infinite (Borel) state spaces, using a similar line of proof. Infinite spaces SSP problems will also be discussed in Section 4.6.

A notable SSP problem with infinite state space arises under imperfect state information. There the problem is converted to a perfect state information problem whose states are belief states, i.e., posterior probability distributions of the original state given the observations thus far. The paper by Patek [Pat07] addresses SSP problems with imperfect state information and proves results that are similar to the ones for their perfect state information counterparts. These results can also be derived using the line of analysis of this chapter. In particular, the critical condition that the cost functions of proper policies are bounded below by some real-valued function [cf. Assumption 3.3.1(b)] is proved as Lemma 5 in [Pat07], using the fact that the cost functions of the proper policies are bounded below by the optimal cost function of a corresponding perfect state information problem.

E X E R C I S E S

3.1 (Conditions for J^* to be a Fixed Point of T)

The purpose of this exercise is to show that the optimal cost function J^* is a fixed point of T under some assumptions, which among others, are satisfied generically in deterministic optimal control problems. Let $\hat{\Pi}$ be a subset of policies such that:

(1) We have

$$(\mu, \pi) \in \hat{\Pi} \qquad \text{if and only if} \qquad \mu \in \mathcal{M},\ \pi \in \hat{\Pi},$$

where for $\mu \in \mathcal{M}$ and $\pi = \{\mu_0, \mu_1, \ldots\}$, we denote by (μ, π) the policy $\{\mu, \mu_0, \mu_1, \ldots\}$. *Note*: This condition precludes the possibility that $\hat{\Pi}$ is the set of all stationary policies (unless there is only one stationary policy).

(2) For every $\pi = \{\mu_0, \mu_1, \ldots\} \in \hat{\Pi}$, we have

$$J_\pi = T_{\mu_0} J_{\pi_1},$$

where π_1 is the policy $\pi_1 = \{\mu_1, \mu_2, \ldots\}$.

(3) We have

$$\inf_{\mu\in\mathcal{M},\,\pi\in\hat{\Pi}} T_\mu J_\pi = \inf_{\mu\in\mathcal{M}} T_\mu \hat{J},$$

where the function $\hat{J}$ is given by

$$\hat{J}(x) = \inf_{\pi\in\hat{\Pi}} J_\pi(x), \qquad x \in X.$$

Show that:

(a) $\hat{J}$ is a fixed point of T. In particular, if $\hat{\Pi} = \Pi$, then J^* is a fixed point of T.

(b) The assumptions (1)-(3) hold with $\hat{\Pi} = \Pi$ in the case of the deterministic mapping

$$H(x,u,J) = g(x,u) + J\big(f(x,u)\big), \qquad x \in X,\ u \in u(x),\ J \in \mathcal{E}(X). \quad (3.60)$$

(c) Consider the SSP example of Section 3.1.2, where J^* is not a fixed point of T. Which of the conditions (1)-(3) is violated?

Solution: (a) For every $x \in X$, we have

$$\hat{J}(x) = \inf_{\pi\in\hat{\Pi}} J_\pi(x) = \inf_{\mu\in\mathcal{M},\,\pi\in\hat{\Pi}} (T_\mu J_\pi)(x) = \inf_{\mu\in\mathcal{M}} (T_\mu \hat{J})(x) = (T\hat{J})(x),$$

where the second equality holds by conditions (1) and (2), and the third equality holds by condition (3).

(b) This is evident in the case of the deterministic mapping (3.60). *Notes*: (i) If $\hat{\Pi} = \Pi$, parts (a) and (b) show that J^*, which is equal to $\hat{J}$, is a fixed point of T. Moreover, if we choose a set S such that J_S^* can be shown to be equal to J^*, then Prop. 3.2.1 applies and shows that J^* is the unique fixed point of T with the set $\big\{J \in \mathcal{E}(X) \mid J_S^* \le J \le \tilde{J}\big\}$ for some $\tilde{J} \in S$. In addition the VI sequence $\{T^k J\}$ converges to J^* starting from every J within that set. (ii) The assumptions (1)-(3) of this exercise also hold for other choices of $\hat{\Pi}$. For example, when $\hat{\Pi}$ is the set of all *eventually stationary* policies, i.e., policies of the form $\{\mu_0, \ldots, \mu_k, \mu, \mu, \ldots\}$, where $\mu_0, \ldots, \mu_k, \mu \in \mathcal{M}$ and k is some positive integer.

(c) For the SSP problem of Section 3.1.1, condition (2) of the preceding proposition need not be satisfied (because the expected value operation need not commute with lim sup).

3.2 (Alternative Semicontractive Conditions I)

This exercise provides a different starting point for the semicontractive analysis of Section 3.2. In particular, the results of Prop. 3.2.1 are shown without assuming that J_S^* is a fixed point of T, but by making different assumptions, which include the existence of an S-regular policy that is optimal. Let S be a given subset of $\mathcal{E}(X)$. Assume that:

(1) There exists an S-regular policy μ^* that is optimal, i.e., $J_{\mu^*} = J^*$.

(2) The policy μ^* satisfies $T_{\mu^*} J^* = TJ^*$.

Show that the following hold:

(a) The optimal cost function J^* is the unique fixed point of T within the set $\{J \in S \mid J \geq J^*\}$.

(b) We have $T^k J \to J^*$ for every $J \in S$ with $J \geq J^*$.

(c) An S-regular policy μ that satisfies $T_\mu J^* = TJ^*$ is optimal. Conversely if μ is an S-regular optimal policy, it satisfies $T_\mu J^* = TJ^*$.

Note: Part (a) and the assumptions show that J_S^* is a fixed point of T (as well as that $J_S^* = J^* \in S$), so parts (b) and (c) also follow from Prop. 3.2.1.

Solution: (a) We first show that any fixed point J of T that lies in S satisfies $J \leq J^*$. Indeed, if $J = TJ$, then for the optimal S-regular policy μ^*, we have $J \leq T_{\mu^*} J$, so in view of the monotonicity of T_{μ^*} and the S-regularity of μ^*,

$$J \leq \lim_{k \to \infty} T_{\mu^*}^k J = J_{\mu^*} = J^*.$$

Thus the only function within $\{J \in S \mid J \geq J^*\}$ that can be a fixed point of T is J^*. Using the optimality and S-regularity of μ^*, and condition (2), we have

$$J^* = J_{\mu^*} = T_{\mu^*} J_{\mu^*} = T_{\mu^*} J^* = TJ^*,$$

so J^* is a fixed point of T. Finally, $J^* \in S$ since $J^* = J_{\mu^*}$ and μ^* is S-regular, so J^* is the unique fixed point of T within $\{J \in S \mid J \geq J^*\}$.

(b) For the optimal S-regular policy μ^* and any $J \in S$ with $J \geq J^*$, we have

$$T_{\mu^*}^k J \geq T^k J \geq T^k J^* = J^*, \qquad k = 0, 1, \ldots.$$

Taking the limit as $k \to \infty$, and using the fact $\lim_{k\to\infty} T_{\mu^*}^k J = J_{\mu^*} = J^*$, which holds since μ^* is S-regular and optimal, we see that $T^k J \to J^*$.

(c) If μ satisfies $T_\mu J^* = TJ^*$, then using part (a), we have $T_\mu J^* = J^*$ and hence $\lim_{k\to\infty} T_\mu^k J^* = J^*$. If μ is in addition S-regular, then $J_\mu = \lim_{k\to\infty} T_\mu^k J^* = J^*$ and μ is optimal. Conversely, if μ is optimal and S-regular, then $J_\mu = J^*$ and $J_\mu = T_\mu J_\mu$, which combined with $J^* = TJ^*$ [cf. part (a)], yields $T_\mu J^* = TJ^*$.

3.3 (Alternative Semicontractive Conditions II)

Let S be a given subset of $\mathcal{E}(X)$. Show that the assumptions of Exercise 3.2 hold if and only if $J^* \in S$, $TJ^* \leq J^*$, and there exists an S-regular policy μ such that $T_\mu J^* = TJ^*$.

Solution: Let the conditions (1) and (2) of Exercise 3.2 hold, and let μ^* be the S-regular policy that is optimal. Then condition (1) implies that $J^* = J_{\mu^*} \in S$ and $J^* = T_{\mu^*} J^* \geq TJ^*$, while condition (2) implies that there exists an S-regular policy μ such that $T_\mu J^* = TJ^*$.

Conversely, assume that $J^* \in S$, $TJ^* \leq J^*$, and there exists an S-regular policy μ such that $T_\mu J^* = TJ^*$. Then we have $T_\mu J^* = TJ^* \leq J^*$. Hence $T_\mu^k J^* \leq J^*$ for all k, and by taking the limit as $k \to \infty$, we obtain $J_\mu \leq J^*$. Hence the S-regular policy μ is optimal, and the conditions of Exercise 3.2 hold.

3.4 (Alternative Semicontractive Conditions III)

Let S be a given subset of $\mathcal{E}(X)$. Assume that:

(1) There exists an optimal S-regular policy.

(2) For every S-irregular policy $\overline{\mu}$, we have $T_{\overline{\mu}} J^* \geq J^*$.

Show that the assumptions of Exercise 3.2 hold.

Solution: It will be sufficient to show that conditions (1) and (2) imply that $J^* = TJ^*$. Assume to obtain a contradiction, that $J^* \neq TJ^*$. Then $J^* \geq TJ^*$, as can be seen from the relations

$$J^* = J_{\mu^*} = T_{\mu^*} J_{\mu^*} \geq TJ_{\mu^*} = TJ^*,$$

where μ^* is an optimal S-regular policy. Thus the relation $J^* \neq TJ^*$ implies that there exists μ' and $x \in X$ such that

$$J^*(x) \geq (T_{\mu'} J^*)(x), \qquad \forall\, x \in X,$$

with strict inequality for some x [note here that we can choose $\overline{\mu}(x) = \mu^*(x)$ for all x such that $J^*(x) = (TJ^*)(x)$, and we can choose $\overline{\mu}(x)$ to satisfy $J^*(x) > (T_{\overline{\mu}} J^*)(x)$ for all other x]. If $\overline{\mu}$ were S-regular, we would have

$$J^* \geq T_{\overline{\mu}} J^* \geq \lim_{k \to \infty} T_{\overline{\mu}}^k J^* = J_{\mu'},$$

with strict inequality for some $x \in X$, which is impossible. Hence μ' is S-irregular, which contradicts condition (2).

3.5 (Restricted Optimization over a Subset of S-Regular Policies)

This exercise provides a useful extension of Prop. 3.2.1. Given a set S, it may be more convenient to work with a subset $\widehat{\mathcal{M}} \subset \mathcal{M}_S$. Let $\hat{J}$ denote the corresponding restricted optimal value:

$$\hat{J}(x) = \inf_{\mu \in \widehat{\mathcal{M}}} J_\mu(x),$$

and assume that $\hat{J}$ is a fixed point of T. Show that the following analogs of the conclusions of Prop. 3.2.1 hold:

(a) (*Uniqueness of Fixed Point*) If J' is a fixed point of T and there exists $\tilde{J} \in S$ such that $J' \leq \tilde{J}$, then $J' \leq \hat{J}$. In particular, if the set $\widehat{\mathcal{W}}$ given by

$$\widehat{\mathcal{W}} = \big\{ J \in \mathcal{E}(X) \mid \hat{J} \leq J \leq \tilde{J} \text{ for some } \tilde{J} \in S \big\},$$

is nonempty, then $\hat{J}$ is the unique fixed point of T within $\widehat{\mathcal{W}}$.

(b) (*VI Convergence*) We have $T^k J \to \hat{J}$ for every $J \in \widehat{\mathcal{W}}$.

Solution: The proof is nearly identical to the one of Prop. 3.2.1. Let $J \in \widehat{\mathcal{W}}$, so that

$$\hat{J} \leq J \leq \tilde{J}$$

for some $\tilde{J} \in S$. We have for all $k \geq 1$ and $\mu \in \widehat{\mathcal{M}}$,

$$\hat{J} = T^k \hat{J} \leq T^k J \leq T^k \tilde{J} \leq T_\mu^k \tilde{J},$$

where the equality follows from the fixed point property of $\hat{J}$, while the inequalities follow by using the monotonicity and the definition of T. The right-hand side tends to J_μ as $k \to \infty$, since μ is S-regular and $\tilde{J} \in S$. Hence the infimum over $\mu \in \widehat{\mathcal{M}}$ of the limit of the right-hand side tends to the left-hand side $\hat{J}$. It follows that $T^k J \to \hat{J}$, proving part (b). To prove part (a), let J' be a fixed point of T that belongs to $\widehat{\mathcal{W}}$. Then J' is equal to $\lim_{k\to\infty} T^k J'$, which has been proved to be equal to $\hat{J}$.

3.6 (The Case $J_S^* \leq \bar{J}$)

Within the framework of Section 3.2, assume that $J_S^* \leq \bar{J}$. (This occurs in particular in the monotone decreasing model where $\bar{J} \geq T_\mu \bar{J}$ for all $\mu \in \mathcal{M}$; see Section 4.3.) Show that if J_S^* is a fixed point of T, then we have $J_S^* = J^*$. *Note*: This result manifests itself in the shortest path Example 3.2.1 for the case where $b < 0$.

Solution: For all k and policies $\pi = \{\mu_0, \mu_1, \ldots\}$, we have

$$J_S^* = \lim_{k\to\infty} T^k J_S^* \leq \limsup_{k\to\infty} T^k \bar{J} \leq \limsup_{k\to\infty} T_{\mu_0} \cdots T_{\mu_{k-1}} \bar{J} = J_\pi,$$

and by taking the infimum over $\pi \in \Pi$, we obtain $J_S^* \leq J^*$. Since generically we have $J_S^* \geq J^*$, it follows that $J_S^* = J^*$.

3.7 (Weakening the Near-Optimal Termination Assumption)

Consider the deterministic optimal control problem of Section 3.5.5. The purpose of this exercise is to show that the Assumption 3.5.9 is equivalent to a seemingly weaker assumption where nonstationary policies can be used for termination. Given a state $x \in X^*$, we say that a (possibly nonstationary) policy $\pi \in \Pi$ *terminates from* x if the sequence $\{x_k\}$, which is generated starting from x and using π, reaches X_0 in the sense that $x_{\bar{k}} \in X_0$ for some index $\bar{k}$. Assume that for every $x \in X^*$, there exists a policy $\pi \in \Pi$ that terminates from x. Show that:

(a) The set $\widehat{\mathcal{M}}$ of terminating stationary policies is nonempty, i.e., there exists a stationary policy that terminates from every $x \in X^*$.

(b) Assumption 3.5.9 is satisfied if for every pair (x, ϵ) with $x \in X^*$ and $\epsilon > 0$, there exists a policy $\pi \in \Pi$ that terminates from x and satisfies $J_\pi(x) \leq J^*(x) + \epsilon$.

Solution: (a) Consider the sequence of subsets of X defined for $k = 0, 1, \ldots$, by

$$X_k = \{x \in X^* \mid \text{there exists } \pi \in \Pi \text{ that terminates from } x \text{ in } k \text{ steps or less}\},$$

starting with the stopping set X_0. Note that $\cup_{k=0}^{\infty} X_k = X^*$. Define a stationary policy $\bar{\mu}$ as follows: For each $x \in X_k$ with $x \notin X_{k-1}$, let $\{\mu_0, \mu_1, \ldots\}$ be a policy that terminates from x in the minimum possible number of steps (which is k), and let $\bar{\mu} = \mu_0$. For each $x \notin X^*$, let $\bar{\mu}(x)$ be an arbitrary control in $U(x)$. It can be seen that $\bar{\mu}$ is a terminating stationary policy.

(b) Given any state $\bar{x} \in X^*$ with $\bar{x} \notin X_0$, and a nonstationary policy $\pi = \{\mu_0, \mu_1, \ldots\}$ that terminates from $\bar{x}$, we construct a stationary policy μ that terminates from every $x \in X^*$ and generates essentially the same trajectory as π starting from $\bar{x}$ (i.e., after cycles are subtracted). To construct such a μ, we consider the sequence generated by π starting from $\bar{x}$. If this sequence contains cycles, we shorten the sequence by eliminating the cycles, and we redefine π so that starting from $\bar{x}$ it generates a terminating trajectory without cycles. This redefined version of π, denoted $\pi' = \{\mu'_0, \mu'_1, \ldots\}$, terminates from $\bar{x}$ and has cost $J_{\pi'}(\bar{x}) \leq J_\pi(\bar{x})$ [since all the eliminated transitions that belonged to cycles have nonnegative cost, in view of the fact $J^*(x) > -\infty$ for all x, which is implied by Assumption 3.5.9]. We now consider the sequence of subsets of X defined by

$$X_k = \{x \in X \mid \pi' \text{ terminates from } x \text{ in } k \text{ steps or less}\}, \quad k = 0, 1, \ldots,$$

where X_0 is the stopping set. Let $\bar{k}$ be the first $k \geq 1$ such that $\bar{x} \in X_k$. Construct the stationary policy μ as follows: for $x \in \cup_{k=1}^{\bar{k}} X_k$, let

$$\mu(x) = \mu'_{\bar{k}-k}(x), \qquad \text{if } x \in X_k \text{ and } x \notin X_{k-1}, \quad k = 1, 2, \ldots,$$

and for $x \notin \cup_{k=1}^{\bar{k}} X_k$, let $\mu(x) = \bar{\mu}(x)$, where $\bar{\mu}$ is a stationary policy that terminates from every $x \in X^*$ [and was shown to exist in part (a)]. Then it is seen that μ terminates from every $x \in X^*$, and generates the same sequence as π' starting from the state $\bar{x}$, so it satisfies $J_\mu(\bar{x}) = J_{\pi'}(\bar{x}) \leq J_\pi(\bar{x})$.

3.8 (Verifying the Near-Optimal Termination Assumption)

In the context of the deterministic optimal control problem of Section 3.5.5, assume that X is a normed space with norm denoted $\|\cdot\|$. We say that π *asymptotically terminates from* x if the sequence $\{x_k\}$ generated starting from x and using π converges to X_0 in the sense that

$$\lim_{k \to \infty} \text{dist}(x_k, X_0) = 0,$$

where $\text{dist}(x, X_0)$ denotes the minimum distance from x to X_0,

$$\text{dist}(x, X_0) = \inf_{y \in X_0} \|x - y\|, \qquad x \in X.$$

The purpose of this exercise is to provide a readily verifiable condition that guarantees Assumption 3.5.9. Assume that

$$0 \leq g(x,u), \qquad x \in X,\ u \in U(x),$$

and that

$$J^*(x) > 0, \qquad \forall\ x \notin X_0.$$

Assume further the following:

(1) For every $x \in X^* = \big\{x \in X \mid J^*(x) < \infty\big\}$ and $\epsilon > 0$, there exits a policy π that asymptotically terminates from x and satisfies $J_\pi(x) \leq J^*(x) + \epsilon$.

(2) For every $\epsilon > 0$, there exists a $\delta_\epsilon > 0$ such that for each $x \in X^*$ with

$$\text{dist}(x, X_0) \leq \delta_\epsilon,$$

there is a policy π that terminates from x and satisfies $J_\pi(x) \leq \epsilon$.

Then:

(a) Show that Assumption 3.5.9 holds.

(b) Show that condition (1) holds if for each $\delta > 0$ there exists $\epsilon > 0$ such that

$$\inf_{u \in U(x)} g(x,u) \geq \epsilon, \qquad \forall\ x \in X \text{ such that } \text{dist}(x, X_0) \geq \delta.$$

Note: For further discussion, analysis, and application to the case of a linear system, see the author's paper [Ber17b].

Solution: (a) Fix $x \in X^*$ and $\epsilon > 0$. Let π be a policy that asymptotically terminates from x, and satisfies $J_\pi(x) \leq J^*(x) + \epsilon$, as per condition (1). Starting from x, this policy will generate a sequence $\{x_k\}$ such that for some index $\bar{k}$ we have $\text{dist}(x_{\bar{k}}, X_0) \leq \delta_\epsilon$, so by condition (2), there exists a policy $\bar{\pi}$ that terminates from $x_{\bar{k}}$ and is such that $J_{\bar{\pi}}(x_{\bar{k}}) \leq \epsilon$. Consider the policy π' that follows π up to index $\bar{k}$ and follows $\bar{\pi}$ afterwards. This policy terminates from x and satisfies

$$J_{\pi'}(x) = J_{\pi,\bar{k}}(x) + J_{\bar{\pi}}(x_{\bar{k}}) \leq J_\pi(x) + J_{\bar{\pi}}(x_{\bar{k}}) \leq J^*(x) + 2\epsilon,$$

where $J_{\pi,\bar{k}}(x)$ is the cost incurred by π starting from x up to reaching $x_{\bar{k}}$. From Exercise 3.7 it follows that Assumption 3.5.9 holds.

(b) For any x and policy π that does not asymptotically terminate from x, we will have $J_\pi(x) = \infty$, so that if $x \in X^*$, all policies π with $J_\pi(x) < \infty$ must be asymptotically terminating from x.

3.9 (Perturbations and S-Regular Policies)

The purpose of this exercise is to illustrate that the set of S-regular policies may be different in the perturbed and unperturbed problems of Section 3.4. Consider a single-state problem with $\bar{J} = 0$ and two policies μ and μ', where

$$T_\mu J = \min\{1, J\}, \qquad T_{\mu'} J = \beta > 0.$$

Let $S = \Re$.

(a) Verify that μ is S-irregular and $J_\mu = J^* = 0$.

(b) Verify that μ' is S-regular and $J_{\mu'} = J_S^* = \beta$.

(c) For $\delta > 0$ consider the δ-perturbed problem with $p(x) = 1$, where x is the only state. Show that both μ and μ' are S-regular for this problem. Moreover, we have $\hat{J}_\delta = \min\{1, \beta\} + \delta$.

(d) Verify that Prop. 3.4.1 applies for $\widehat{\mathcal{M}} = \{\mu'\}$ and $\beta \leq 1$, but does not apply if $\widehat{\mathcal{M}} = \{\mu, \mu'\}$ or $\beta > 1$. Which assumptions of the proposition are violated in the latter case?

Solution: Parts (a) and (b) are straightforward. It is also straightforward to verify the definition of S-regularity for both policies in the δ-perturbed problem, and that $J_{\mu,\delta} = 1 + \delta$ and $J_{\mu',\delta} = \beta + \delta$. If $\beta \leq 1$, the policy μ' is optimal for the δ-perturbed problem, and Prop. 3.4.1 applies for $\widehat{\mathcal{M}} = \{\mu'\}$ because all its assumptions are satisfied. However, when $\beta > 1$ and $\widehat{\mathcal{M}} = \{\mu'\}$ there is no ϵ-optimal policy in $\widehat{\mathcal{M}}$ for the δ-perturbed problem (contrary to the assumption of Prop. 3.4.1), and indeed we have $\beta = J_S^* > \lim_{\delta \downarrow 0} \hat{J}_\delta = 1$. Also when $\widehat{\mathcal{M}} = \{\mu, \mu'\}$, the policy μ is not S-regular, contrary to the assumption of Prop. 3.4.1.

3.10 (Perturbations in Affine Monotonic Models [Ber16a])

Consider the affine monotonic model of Section 3.5.2, and let Assumptions 3.5.5 and 3.5.6 hold. In a perturbed version of this model we add a constant $\delta > 0$ to all components of b_μ, thus obtaining what we call the *δ-perturbed affine monotonic problem*. We denote by $\hat{J}_\delta$ and $J_{\mu,\delta}$ the corresponding optimal cost function and policy cost functions, respectively.

(a) Show that for all $\delta > 0$, $\hat{J}_\delta$ is the unique solution within $\Re_+^n$ of the equation

$$J(i) = (TJ)(i) + \delta, \qquad i = 1, \ldots, n.$$

(b) Show that for all $\delta > 0$, a policy μ is optimal for the δ-perturbed problem (i.e., $J_{\mu,\delta} = \hat{J}_\delta$) if and only if $T_\mu \hat{J}_\delta = T\hat{J}_\delta$. Moreover, for the δ-perturbed problem, all optimal policies are contractive and there exists at least one contractive policy that is optimal.

(c) The optimal cost function over contractive policies $\hat{J}$ [cf. Eq. (3.37)] satisfies

$$\hat{J}(i) = \lim_{\delta \downarrow 0} \hat{J}_\delta(i), \qquad i = 1, \ldots, n.$$

(d) If the control constraint set $U(i)$ is finite for all states $i = 1, \ldots, n$, there exists a contractive policy $\hat{\mu}$ that attains the minimum over all contractive policies, i.e., $J_{\hat{\mu}} = \hat{J}$.

(e) Show Prop. 3.5.8.

Solution: (a), (b) By Prop. 3.5.6, we have that Assumption 3.3.1 holds for the δ-perturbed problem. The results follow by applying Prop. 3.5.7 [the equation of part (a) is Bellman's equation for the δ-perturbed problem].

(c) For an optimal contractive policy μ^*_δ of the δ-perturbed problem [cf. part (b)], we have

$$\hat{J} = \inf_{\mu:\,\text{contractive}} J_\mu \leq J_{\mu^*_\delta} \leq J_{\mu^*_\delta,\delta} = \hat{J}_\delta \leq J_{\mu',\delta}, \qquad \forall\ \mu' : \text{contractive}.$$

Since for every contractive policy μ', we have $\lim_{\delta\downarrow 0} J_{\mu',\delta} = J_{\mu'}$, it follows that

$$\hat{J} \leq \lim_{\delta\downarrow 0} \hat{J}_\delta \leq J_{\mu'}, \qquad \forall\ \mu' : \text{contractive}.$$

By taking the infimum over all μ' that are contractive, the result follows.

(d) Let $\{\delta_k\}$ be a positive sequence with $\delta_k \downarrow 0$, and consider a corresponding sequence $\{\mu_k\}$ of optimal contractive policies for the δ_k-perturbed problems. Since the set of contractive policies is finite [in view of the finiteness of $U(i)$], some policy $\hat{\mu}$ will be repeated infinitely often within the sequence $\{\mu_k\}$, and since $\{J^*_{\delta_k}\}$ is monotonically nonincreasing, we will have

$$\hat{J} \leq J_{\hat{\mu}} \leq J^*_{\delta_k},$$

for all k sufficiently large. Since by part (c), $J^*_{\delta_k} \downarrow \hat{J}$, it follows that $J_{\hat{\mu}} = \hat{J}$.

(e) For all contractive μ, we have $J_\mu = T_\mu J_\mu \geq T_\mu \hat{J} \geq T\hat{J}$. Taking the infimum over contractive μ, we obtain $\hat{J} \geq T\hat{J}$. Conversely, for all $\delta > 0$ and $\mu \in \mathcal{M}$, we have

$$\hat{J}_\delta = T\hat{J}_\delta + \delta e \leq T_\mu \hat{J}_\delta + \delta e.$$

Taking limit as $\delta \downarrow 0$, and using part (c), we obtain $\hat{J} \leq T_\mu \hat{J}$ for all $\mu \in \mathcal{M}$. Taking infimum over $\mu \in \mathcal{M}$, it follows that $\hat{J} \leq T\hat{J}$. Thus $\hat{J}$ is a fixed point of T.

For all $J \in \Re^n$ with $J \geq \hat{J}$ and contractive μ, we have by using the relation $\hat{J} = T\hat{J}$ just shown,

$$\hat{J} = \lim_{k\to\infty} T^k \hat{J} \leq \lim_{k\to\infty} T^k J \leq \lim_{k\to\infty} T^k_\mu J = J_\mu.$$

Taking the infimum over all contractive μ, we obtain

$$\hat{J} \leq \lim_{k\to\infty} T^k J \leq \hat{J}, \qquad \forall\ J \geq \hat{J}.$$

This proves that $T^k J \to \hat{J}$. Finally, let $J' \in \mathcal{R}(X)$ be another solution of Bellman's equation, and let $J \in \mathcal{R}(X)$ be such that $J \geq \hat{J}$ and $J \geq J'$. Then $T^k J \to \hat{J}$, while $T^k J \geq T^k J' = J'$. It follows that $\hat{J} \geq J'$.

To prove Prop. 3.5.8(c) note that if μ is a contractive policy with $J_\mu = \hat{J}$, we have $\hat{J} = J_\mu = T_\mu J_\mu = T_\mu \hat{J}$, so, using also the relation $\hat{J} = T\hat{J}$ [cf. part (a)], we obtain $T_\mu \hat{J} = T\hat{J}$. Conversely, if μ satisfies $T_\mu \hat{J} = T\hat{J}$, then from part (a), we have $T_\mu \hat{J} = \hat{J}$ and hence $\lim_{k\to\infty} T^k_\mu \hat{J} = \hat{J}$. Since μ is contractive, we obtain $J_\mu = \lim_{k\to\infty} T^k_\mu \hat{J}$, so $J_\mu = \hat{J}$.

The proof of Prop. 3.5.8(d) is nearly identical to the one of Prop. 3.5.4(d).

4

Noncontractive Models

Contents

In this chapter, we consider abstract DP models that are similar to the ones of the earlier chapters, but we do not assume any contraction-like property. We discuss both finite and infinite horizon models, and introduce just enough assumptions (including monotonicity) to obtain some minimal results, which we will strengthen as we go along.

In Section 4.2, we consider a general type of finite horizon problem. Under some reasonable assumptions, we show the standard results that one may expect in an abstract setting.

In Section 4.3, we discuss an infinite horizon problem that is motivated by the well-known *positive* and *negative* DP models (see [Ber12a], Chapter 4). These are the special cases of the infinite horizon stochastic optimal control problem of Example 1.2.1, where the cost per stage g is uniformly nonpositive or uniformly nonnegative. For these models there is interesting theory (the validity of Bellman's equation and the availability of optimality conditions in a DP context), which originated with the works of Blackwell [Bla65b] and Strauch [Str66], and is discussed in Section 4.3.1. There are also interesting computational methods, patterned after the VI and PI algorithms, which are discussed in Sections 4.3.2 and 4.3.3. However, the performance guarantees for these methods are not as powerful as in the contractive case, and their validity hinges upon certain additional assumptions.

In Section 4.4, we extend the notion of regularity of Section 3.2 so that it applies more broadly, including situations where nonstationary policies need to be considered. The mathematical reason for considering nonstationary policies is that for some of the noncontractive models of Section 4.3, stationary policies are insufficient in the sense that there may not exist ϵ-optimal policies that are stationary. In this section, we also discuss some applications, including some general types of optimal control problems with nonnegative cost per stage. Principal results here are that J^* is the unique solution of Bellman's equation within a certain class of functions, and other related results regarding the convergence of the VI algorithm.

In Section 4.5, we discuss a nonnegative cost deterministic optimal control problem, which combines elements of the noncontractive models of Section 4.3, and the semicontractive models of Chapter 3 and Section 4.4. Within this setting we explore the structure and the multiplicity of solutions of Bellman's equation. We draw inspiration from the analysis of Section 4.4, but we also use a perturbation-based line of analysis, similar to the one of Section 3.4. In particular, our starting point is a perturbed version of the mapping T_μ that defines the "stable" policies, in place of a subset S that defines the S-regular policies. Still with a proper definition of S, the "stable" policies are S-regular.

Finally, in Section 4.6, we extend the ideas of Section 4.5 to stochastic optimal control problems, by generalizing the notion of a proper policy to the case of infinite state and control spaces. This analysis is considerably more complex than the finite-spaces SSP analysis of Section 3.5.1.

4.1 NONCONTRACTIVE MODELS - PROBLEM FORMULATION

Throughout this chapter we will continue to use the model of Section 3.2, which involves the set of extended real numbers $\Re^* = \Re \cup \{\infty, -\infty\}$. To repeat some of the basic definitions, we denote by $\mathcal{E}(X)$ the set of all extended real-valued functions $J : X \mapsto \Re^*$, by $\mathcal{R}(X)$ the set of real-valued functions $J : X \mapsto \Re$, and by $\mathcal{B}(X)$ the set of real-valued functions $J : X \mapsto \Re$ that are bounded with respect to a given weighted sup-norm.

We have a set X of states and a set U of controls, and for each $x \in X$, the nonempty control constraint set $U(x) \subset U$. We denote by $\mathcal{M}$ the set of all functions $\mu : X \mapsto U$ with $\mu(x) \in U(x)$, for all $x \in X$, and by Π the set of "nonstationary policies" $\pi = \{\mu_0, \mu_1, \ldots\}$, with $\mu_k \in \mathcal{M}$ for all k. We refer to a stationary policy $\{\mu, \mu, \ldots\}$ simply as μ.

We introduce a mapping $H : X \times U \times \mathcal{E}(X) \mapsto \Re^*$, and we define the mapping $T : \mathcal{E}(X) \mapsto \mathcal{E}(X)$ by

$$(TJ)(x) = \inf_{u \in U(x)} H(x, u, J), \qquad \forall\ x \in X,\ J \in \mathcal{E}(X),$$

and for each $\mu \in \mathcal{M}$ the mapping $T_\mu : \mathcal{E}(X) \mapsto \mathcal{E}(X)$ by

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

We continue to use the following assumption throughout this chapter, without mentioning it explicitly in various propositions.

Assumption 4.1.1: (Monotonicity) If $J, J' \in \mathcal{E}(X)$ and $J \leq J'$, then

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

A fact that we will be using frequently is that for each $J \in \mathcal{E}(X)$ and scalar $\epsilon > 0$, there exists a $\mu_\epsilon \in \mathcal{M}$ such that for all $x \in X$,

$$(T_{\mu_\epsilon} J)(x) \leq \begin{cases} (TJ)(x) + \epsilon & \text{if } (TJ)(x) > -\infty, \\ -(1/\epsilon) & \text{if } (TJ)(x) = -\infty. \end{cases}$$

In particular, if J is such that

$$(TJ)(x) > -\infty, \qquad \forall\ x \in X,$$

then for each $\epsilon > 0$, there exists a $\mu_\epsilon \in \mathcal{M}$ such that

$$(T_{\mu_\epsilon} J)(x) \leq (TJ)(x) + \epsilon, \qquad \forall\ x \in X.$$

We will often use in our analysis the unit function e, defined by $e(x) \equiv 1$, so for example, we write the preceding relation in shorthand as

$$T_{\mu_\epsilon} J \leq TJ + \epsilon\, e.$$

We define cost functions for policies consistently with Chapters 2 and 3. In particular, we are given a function $\bar{J} \in \mathcal{E}(X)$, and we consider for every policy $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi$ and positive integer N the function $J_{N,\pi} \in \mathcal{E}(X)$ defined by

$$J_{N,\pi}(x) = (T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(x), \qquad \forall\ x \in X,$$

and the function $J_\pi \in \mathcal{E}(X)$ defined by

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

We refer to $J_{N,\pi}$ as the N-stage cost function of π and to J_π as the infinite horizon cost function of π (or just "cost function" if the length of the horizon is clearly implied by the context). For a stationary policy $\pi = \{\mu, \mu, \ldots\}$ we also write J_π as J_μ.

In Section 4.2, we consider the N-stage optimization problem

$$\begin{aligned} &\text{minimize} \quad J_{N,\pi}(x) \\ &\text{subject to} \quad \pi \in \Pi, \end{aligned} \tag{4.1}$$

while in Sections 4.3 and 4.4 we discuss its infinite horizon version

$$\begin{aligned} &\text{minimize} \quad J_\pi(x) \\ &\text{subject to} \quad \pi \in \Pi. \end{aligned} \tag{4.2}$$

For a fixed $x \in X$, we denote by $J_N^*(x)$ and $J^*(x)$ the optimal costs for these problems, i.e.,

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

We say that a policy $\pi^* \in \Pi$ is *N-stage optimal* if

$$J_{N,\pi^*}(x) = J_N^*(x), \qquad \forall\ x \in X,$$

and (infinite horizon) *optimal* if

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

For a given $\epsilon > 0$, we say that π_ϵ is *N-stage ϵ-optimal* if

$$J_{N,\pi_\epsilon}(x) \leq \begin{cases} J_N^*(x) + \epsilon & \text{if } J_N^*(x) > -\infty, \\ -(1/\epsilon) & \text{if } J_N^*(x) = -\infty, \end{cases}$$

and we say that π_ϵ is *ϵ-optimal* if

$$J_{\pi_\epsilon}(x) \leq \begin{cases} J^*(x) + \epsilon & \text{if } J^*(x) > -\infty, \\ -(1/\epsilon) & \text{if } J^*(x) = -\infty. \end{cases}$$

4.2 FINITE HORIZON PROBLEMS

Consider the N-stage problem (4.1), where the cost function $J_{N,\pi}$ is defined by

$$J_{N,\pi}(x) = (T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(x), \qquad \forall\ x \in X.$$

Based on the theory of finite horizon DP, we expect that (at least under some conditions) the optimal cost function J_N^* is obtained by N successive applications of the DP mapping T on the initial function $\bar{J}$, i.e.,

$$J_N^* = \inf_{\pi \in \Pi} J_{N,\pi} = T^N \bar{J}.$$

This is the analog of Bellman's equation for the finite horizon problem in a DP context.

The Case Where Uniformly N-Stage Optimal Policies Exist

A favorable case where the analysis is simplified and we can easily show that $J_N^* = T^N \bar{J}$ is when the finite horizon DP algorithm yields an optimal policy during its execution. By this we mean that the algorithm that starts with $\bar{J}$, and sequentially computes $T\bar{J}, T^2\bar{J}, \ldots, T^N\bar{J}$, also yields corresponding $\mu_{N-1}^*, \mu_{N-2}^*, \ldots, \mu_0^* \in \mathcal{M}$ such that

$$T_{\mu_k^*} T^{N-k-1} \bar{J} = T^{N-k} \bar{J}, \qquad k = 0, \ldots, N-1. \tag{4.3}$$

While $\mu_{N-1}^*, \ldots, \mu_0^* \in \mathcal{M}$ satisfying this relation need not exist (because the corresponding infimum in the definition of T is not attained), if they do exist, they both form an optimal policy and also guarantee that

$$J_N^* = T^N \bar{J}.$$

The proof is simple: we have for every $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi$

$$J_{N,\pi} = T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J} \geq T^N \bar{J} = T_{\mu_0^*} \cdots T_{\mu_{N-1}^*} \bar{J}, \tag{4.4}$$

where the inequality follows from the monotonicity assumption and the definition of T, and the last equality follows from Eq. (4.3). Thus $\{\mu_0^*, \mu_1^*, \ldots\}$ has no worse N-stage cost function than every other policy, so it is N-stage optimal and $J_N^* = T_{\mu_0^*} \cdots T_{\mu_{N-1}^*} \bar{J}$. By taking the infimum of the left-hand side over $\pi \in \Pi$ in Eq. (4.4), we obtain $J_N^* = T^N \bar{J}$.

The preceding argument can also be used to show that $\{\mu_k^*, \mu_{k+1}^*, \ldots\}$ is $(N-k)$-stage optimal for all $k = 0, \ldots, N-1$. Such a policy is called *uniformly N-stage optimal*. The fact that the finite horizon DP algorithm provides an optimal solution of *all* the k-stage problems for $k = 1, \ldots, N$, rather than just the last one, is a manifestation of the classical principle

of optimality, expounded by Bellman in the early days of DP (the tail portion of an optimal policy obtained by DP minimizes the corresponding tail portion of the finite horizon cost). Note, however, that there may exist an N-stage optimal policy that is not k-stage optimal for some $k < N$.

We state the result just derived as a proposition.

Proposition 4.2.1: Suppose that a policy $\{\mu_0^*, \mu_1^*, \ldots\}$ satisfies the condition (4.3). Then this policy is uniformly N-stage optimal, and we have $J_N^* = T^N \bar{J}$.

While the preceding result is theoretically limited, it is very useful in practice, because the existence of a policy satisfying the condition (4.3) can often be established with a simple analysis. For example, this condition is trivially satisfied if the control space is finite. The following proposition provides a generalization.

Proposition 4.2.2: Let the control space U be a metric space, and assume that for each $x \in X$, $\lambda \in \Re$, and $k = 0, 1, \ldots, N-1$, the set

$$U_k(x, \lambda) = \big\{u \in U(x) \mid H(x, u, T^k \bar{J}) \leq \lambda\big\}$$

is compact. Then there exists a uniformly N-stage optimal policy.

Proof: We will show that the infimum in the relation

$$(T^{k+1}\bar{J})(x) = \inf_{u \in U(x)} H\big(x, u, T^k \bar{J}\big)$$

is attained for all $x \in X$ and k. Indeed if $H\big(x, u, T^k \bar{J}\big) = \infty$ for all $u \in U(x)$, then every $u \in U(x)$ attains the infimum. If for a given $x \in X$,

$$\inf_{u \in U(x)} H\big(x, u, T^k \bar{J}\big) < \infty,$$

the corresponding part of the proof of Lemma 3.3.1 applies and shows that the above infimum is attained. The result now follows from Prop. 4.2.1. **Q.E.D.**

The General Case

We now consider the case where there may not exist a uniformly N-stage optimal policy. By using the definitions of J_N^* and $T^N \bar{J}$, the equation

$J_N^* = T^N \bar{J}$, which we want to prove, can be equivalently written as

$$\inf_{\mu_0,\ldots,\mu_{N-1}\in\mathcal{M}} T_{\mu_0}\cdots T_{\mu_{N-1}}\bar{J} = \inf_{\mu_0\in\mathcal{M}} T_{\mu_0}\left(\inf_{\mu_1\in\mathcal{M}} T_{\mu_1}\left(\cdots \inf_{\mu_{N-1}\in\mathcal{M}} T_{\mu_{N-1}}\bar{J}\right)\right).$$

Thus we have $J_N^* = T^N\bar{J}$ if the operations inf and T_μ can be interchanged in the preceding equation. We will introduce two alternative assumptions, which guarantee that this interchange is valid. Our first assumption is a form of continuity from above of H with respect to J.

Assumption 4.2.1: For each sequence $\{J_m\} \subset \mathcal{E}(X)$ with $J_m \downarrow J$ and $H(x,u,J_0) < \infty$ for all $x \in X$ and $u \in U(x)$, we have

$$\lim_{m\to\infty} H(x,u,J_m) = H(x,u,J), \qquad \forall\ x \in X,\ u \in U(x). \tag{4.5}$$

Note that if $\{J_m\}$ is monotonically nonincreasing, the same is true for $\{T_\mu J_m\}$. It follows that

$$\inf_m J_m = \lim_{m\to\infty} J_m, \qquad \inf_m (T_\mu J_m) = \lim_{m\to\infty}(T_\mu J_m),$$

so for all $\mu \in \mathcal{M}$, Eq. (4.5) implies that

$$\inf_m (T_\mu J_m) = \lim_{m\to\infty}(T_\mu J_m) = T_\mu\left(\lim_{m\to\infty} J_m\right) = T_\mu\left(\inf_m J_m\right).$$

This equality can be extended for any $\mu_1,\ldots,\mu_k \in \mathcal{M}$ as follows:

$$\begin{aligned}\inf_m (T_{\mu_1}\cdots T_{\mu_k} J_m) &= T_{\mu_1}\left(\inf_m (T_{\mu_1}\cdots T_{\mu_k} J_m)\right)\\ &= \cdots\\ &= T_{\mu_1}T_{\mu_1}\cdots T_{\mu_{k-1}}\left(\inf_m (T_{\mu_k} J_m)\right)\\ &= T_{\mu_1}\cdots T_{\mu_k}\left(\inf_m J_m\right).\end{aligned} \tag{4.6}$$

We use this relation to prove the following proposition.

Proposition 4.2.3: Let Assumption 4.2.1 hold, and assume further that $J_{k,\pi}(x) < \infty$, for all $x \in X$, $\pi \in \Pi$, and $k \geq 1$. Then $J_N^* = T^N\bar{J}$.

Proof: We select for each $k = 0,\ldots,N-1$, a sequence $\{\mu_k^m\} \subset \mathcal{M}$ such that

$$\lim_{m\to\infty} T_{\mu_k^m}(T^{N-k-1}\bar{J}) \downarrow T^{N-k}\bar{J}.$$

Since $J_N^* \leq T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J}$ for all $\mu_0, \ldots, \mu_{N-1} \in \mathcal{M}$, we have using also Eq. (4.6) and the assumption $J_{k,\pi}(x) < \infty$, for all k, π, and x,

$$\begin{aligned} J_N^* &\leq \inf_{m_0} \cdots \inf_{m_{N-1}} T_{\mu_0^{m_0}} \cdots T_{\mu_{N-1}^{m_{N-1}}} \bar{J} \\ &= \inf_{m_0} \cdots \inf_{m_{N-2}} T_{\mu_0^{m_0}} \cdots T_{\mu_{N-2}^{m_{N-2}}} \left(\inf_{m_{N-1}} T_{\mu_{N-1}^{m_{N-1}}} \bar{J} \right) \\ &= \inf_{m_0} \cdots \inf_{m_{N-2}} T_{\mu_0^{m_0}} \cdots T_{\mu_{N-2}^{m_{N-2}}} T\bar{J} \\ &\quad \vdots \\ &= \inf_{m_0} T_{\mu_0^{m_0}} (T^{N-1}\bar{J}) \\ &= T^N \bar{J}. \end{aligned}$$

On the other hand, it is clear from the definitions that $T^N \bar{J} \leq J_{N,\pi}$ for all N and $\pi \in \Pi$, so that $T^N \bar{J} \leq J_N^*$. Thus, $J_N^* = T^N \bar{J}$. **Q.E.D.**

We now introduce an alternative assumption, which in addition to $J_N^* = T^N \bar{J}$, guarantees the existence of an ϵ-optimal policy.

Assumption 4.2.2: We have

$$J_k^*(x) > -\infty, \qquad \forall\ x \in X,\ k = 1, \ldots, N.$$

Moreover, there exists a scalar $\alpha \in (0, \infty)$ such that for all scalars $r \in (0, \infty)$ and functions $J \in \mathcal{E}(X)$, we have

$$H(x, u, J + r\, e) \leq H(x, u, J) + \alpha\, r, \qquad \forall\ x \in X,\ u \in U(x). \quad (4.7)$$

Proposition 4.2.4: Let Assumption 4.2.2 hold. Then $J_N^* = T^N \bar{J}$, and for every $\epsilon > 0$, there exists an ϵ-optimal policy.

Proof: Note that since by assumption, $J_N^*(x) > -\infty$ for all $x \in X$, an N-stage ϵ-optimal policy $\pi_\epsilon \in \Pi$ is one for which

$$J_N^* \leq J_{N,\pi_\epsilon} \leq J_N^* + \epsilon\, e.$$

We use induction. The result clearly holds for $N = 1$. Assume that it holds for $N = k$, i.e., $J_k^* = T^k \bar{J}$ and for any given $\epsilon > 0$, there is a $\pi_\epsilon \in \Pi$

with $J_{k,\pi_\epsilon} \leq J_k^* + \epsilon\, e$. Using Eq. (4.7), we have for all $\mu \in \mathcal{M}$,

$$J_{k+1}^* \leq T_\mu J_{k,\pi_\epsilon} \leq T_\mu J_k^* + \alpha\epsilon\, e.$$

Taking the infimum over μ and then the limit as $\epsilon \to 0$, we obtain $J_{k+1}^* \leq TJ_k^*$. By using the induction hypothesis $J_k^* = T^k\bar{J}$, it follows that $J_{k+1}^* \leq T^{k+1}\bar{J}$. On the other hand, we have clearly $T^{k+1}\bar{J} \leq J_{k+1,\pi}$ for all $\pi \in \Pi$, so that $T^{k+1}\bar{J} \leq J_{k+1}^*$, and hence $T^{k+1}\bar{J} = J_{k+1}^*$.

We now turn to the existence of an ϵ-optimal policy part of the induction argument. Using the assumption $J_k^*(x) > -\infty$ for all $x \in x \in X$, for any $\overline{\epsilon} > 0$, we can choose $\overline{\pi} = \{\overline{\mu}_0, \overline{\mu}_1, \ldots\}$ such that

$$J_{k,\overline{\pi}} \leq J_k^* + \frac{\overline{\epsilon}}{2\alpha}\, e, \tag{4.8}$$

and $\overline{\mu} \in \mathcal{M}$ such that

$$T_{\overline{\mu}} J_k^* \leq TJ_k^* + \frac{\overline{\epsilon}}{2}\, e.$$

Let $\overline{\pi}_{\overline{\epsilon}} = \{\overline{\mu}, \overline{\mu}_0, \overline{\mu}_1, \ldots\}$. Then

$$J_{k+1,\overline{\pi}_{\overline{\epsilon}}} = T_{\overline{\mu}} J_{k,\overline{\pi}} \leq T_{\overline{\mu}} J_k^* + \frac{\overline{\epsilon}}{2}\, e \leq TJ_k^* + \overline{\epsilon}\, e = J_{k+1}^* + \overline{\epsilon}\, e,$$

where the first inequality is obtained by applying $T_{\overline{\mu}}$ to Eq. (4.8) and using Eq. (4.7). The induction is complete. **Q.E.D.**

We now provide some counterexamples showing that the conditions of the preceding propositions are necessary, and that for exceptional (but otherwise very simple) problems, the Bellman equation $J_N^* = T^N\bar{J}$ may not hold and/or there may not exist an ϵ-optimal policy.

Example 4.2.1 (Counterexample to Bellman's Equation I)

Let

$$X = \{0\}, \qquad U(0) = (-1, 0], \qquad \bar{J}(0) = 0,$$

$$H(0, u, J) = \begin{cases} u & \text{if } -1 < J(0), \\ J(0) + u & \text{if } J(0) \leq -1. \end{cases}$$

Then

$$(T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(0) = \mu_0(0),$$

and $J_N^*(0) = -1$, while $(T^N\bar{J})(0) = -N$ for every N. Here Assumption 4.2.1, and the condition (4.7) (cf. Assumption 4.2.2) are violated, even though the condition $J_k^*(x) > -\infty$ for all $x \in X$ (cf. Assumption 4.2.2) is satisfied.

Example 4.2.2 (Counterexample to Bellman's Equation II)

Let

$$X = \{0, 1\}, \qquad U(0) = U(1) = (-\infty, 0], \qquad \bar{J}(0) = \bar{J}(1) = 0,$$

$$H(0, u, J) = \begin{cases} u & \text{if } J(1) = -\infty, \\ 0 & \text{if } J(1) > -\infty, \end{cases} \qquad H(1, u, J) = u.$$

Then

$$(T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(0) = 0, \quad (T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(1) = \mu_0(1), \qquad \forall\ N \geq 1.$$

It can be seen that for $N \geq 2$, we have $J_N^*(0) = 0$ and $J_N^*(1) = -\infty$, but $(T^N \bar{J})(0) = (T^N \bar{J})(1) = -\infty$. Here Assumption 4.2.1, and the condition $J_k^*(x) > -\infty$ for all $x \in X$ (cf. Assumption 4.2.2) are violated, even though the condition (4.7) of Assumption 4.2.2 is satisfied.

In the preceding two examples, the anomalies are due to discontinuity of the mapping H with respect to J. In classical finite horizon DP, the mapping H is usually continuous when it takes finite values, but counterexamples arise in unusual problems where infinite values occur. The next example is a simple stochastic optimal control problem, which involves some infinite expected values of random variables and we have $J_2^* \neq T^2 \bar{J}$.

Example 4.2.3 (Counterexample to Bellman's Equation III)

Let

$$X = \{0, 1\}, \qquad U(0) = U(1) = \Re, \qquad \bar{J}(0) = \bar{J}(1) = 0,$$

let w be a real-valued random variable with $E\{w\} = \infty$, and let

$$H(x, u, J) = \begin{cases} E\big\{w + J(1)\big\} & \text{if } x = 0, \\ u + J(1) & \text{if } x = 1, \end{cases} \qquad \forall\ x \in X,\ u \in U(x).$$

Then if J_m is real-valued for all m, and $J_m(1) \downarrow J(1) = -\infty$, we have

$$\lim_{m \to \infty} H(0, u, J_m) = \lim_{m \to \infty} E\big\{w + J_m(1)\big\} = \infty,$$

while

$$H\left(0, u, \lim_{m \to \infty} J_m\right) = E\big\{w + J(1)\big\} = -\infty,$$

so Assumption 4.2.1 is violated. Indeed, the reader may verify with a straightforward calculation that $J_2^*(0) = \infty$, $J_2^*(1) = -\infty$, while $(T^2 \bar{J})(0) = -\infty$, $(T^2 \bar{J})(1) = -\infty$, so $J_2^* \neq T^2 \bar{J}$. Note that Assumption 4.2.2 is also violated because $J_2^*(1) = -\infty$.

In the next counterexample, Bellman's equation holds, but there is no ϵ-optimal policy. This is an undiscounted deterministic optimal control problem of the type discussed in Section 1.1, where $J_k^*(x) = -\infty$ for some x and k, so Assumption 4.2.2 is violated. We use the notation introduced there.

Example 4.2.4 (Counterexample to Existence of an ϵ-Optimal Policy)

Let $\alpha = 1$ and

$$N = 2, \qquad X = \{0, 1, \ldots\}, \qquad U(x) = (0, \infty), \qquad \bar{J}(x) = 0, \quad \forall\, x \in X,$$

$$f(x, u) = 0, \qquad \forall\, x \in X,\ u \in U(x),$$

$$g(x, u) = \begin{cases} -u & \text{if } x = 0, \\ x & \text{if } x \neq 0, \end{cases} \qquad \forall\, u \in U(x),$$

so that

$$H(x, u, J) = g(x, u) + J(0).$$

Then for $\pi \in \Pi$ and $x \neq 0$, we have $J_{2,\pi}(x) = x - \mu_1(0)$, so that $J_2^*(x) = -\infty$ for all $x \neq 0$. Clearly, we also have $J_2^*(0) = -\infty$. Here Assumption 4.2.1, as well as Eq. (4.7) (cf. Assumption 4.2.2) are satisfied, and indeed we have $J_2^*(x) = (T^2\bar{J})(x) = -\infty$ for all $x \in X$. However, the condition $J_k^*(x) > -\infty$ for all x and k (cf. Assumption 4.2.2) is violated, and it is seen that there does not exist a two-stage ϵ-optimal policy for any $\epsilon > 0$. The reason is that an ϵ-optimal policy $\pi = \{\mu_0, \mu_1\}$ must satisfy

$$J_{2,\pi}(x) = x - \mu_1(0) \leq -\frac{1}{\epsilon}, \qquad \forall\, x \in X,$$

[in view of $J_2^*(x) = -\infty$ for all $x \in X$], which is impossible since the left-hand side above can become positive for x sufficiently large.

4.3 INFINITE HORIZON PROBLEMS

We now turn to the infinite horizon problem (4.2), where the cost function of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ is

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

In this section one of the following two assumptions will be in effect.

Assumption I: (Monotone Increase)

(a) We have

$$-\infty < \bar{J}(x) \leq H(x, u, \bar{J}), \qquad \forall\, x \in X,\ u \in U(x).$$

(b) For each convergent sequence $\{J_m\} \subset \mathcal{E}(X)$ with $J_m \uparrow J$ and $\bar{J} \leq J_m$ for all $m \geq 0$, we have

$$\lim_{m \to \infty} H(x, u, J_m) = H(x, u, J), \qquad \forall\, x \in X,\ u \in U(x).$$

(c) There exists a scalar $\alpha \in (0,\infty)$ such that for all scalars $r \in (0,\infty)$ and functions $J \in \mathcal{E}(X)$ with $\bar{J} \leq J$, we have

$$H(x,u,J+r\,e) \leq H(x,u,J) + \alpha\, r, \qquad \forall\ x \in X,\ u \in U(x).$$

Assumption D: (Monotone Decrease)

(a) We have

$$\bar{J}(x) \geq H(x,u,\bar{J}), \qquad \forall\ x \in X,\ u \in U(x).$$

(b) For each convergent sequence $\{J_m\} \subset \mathcal{E}(X)$ with $J_m \downarrow J$ and $J_m \leq \bar{J}$ for all $m \geq 0$, we have

$$\lim_{m\to\infty} H(x,u,J_m) = H\left(x,u,J\right), \qquad \forall\ x \in X,\ u \in U(x).$$

Assumptions I and D apply to the positive and negative cost DP models, respectively (see [Ber12a], Chapter 4). These are the special cases of the infinite horizon stochastic optimal control problem of Example 1.2.1, where $\bar{J}(x) \equiv 0$ and the cost per stage g is uniformly nonnegative or uniformly nonpositive, respectively. The latter arises often when we want to maximize positive rewards.

It is important to note that Assumptions I and D allow J_π to be defined as a limit rather than as a lim sup. In particular, part (a) of the assumptions and the monotonicity of H imply that

$$\bar{J} \leq T_{\mu_0}\bar{J} \leq T_{\mu_0}T_{\mu_1}\bar{J} \leq \cdots \leq T_{\mu_0}\cdots T_{\mu_k}\bar{J} \leq \cdots$$

under Assumption I, and

$$\bar{J} \geq T_{\mu_0}\bar{J} \geq T_{\mu_0}T_{\mu_1}\bar{J} \geq \cdots \geq T_{\mu_0}\cdots T_{\mu_k}\bar{J} \geq \cdots$$

under Assumption D. Thus we have

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

with the limit being a real number or ∞ or $-\infty$, respectively.

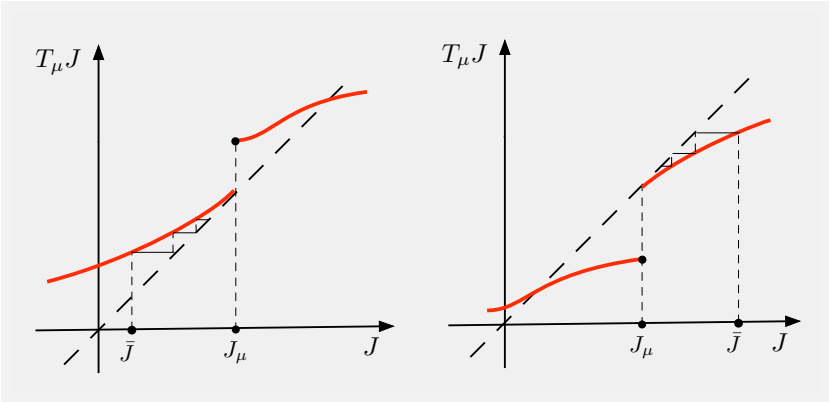


Figure 4.3.1. Illustration of the consequences of lack of continuity of T_μ from below or from above [cf. part (b) of Assumption I or D, respectively]. In the figure on the left, we have $\bar{J} \leq T_\mu \bar{J}$ but T_μ is discontinuous from below at J_μ, so Assumption I does not hold, and J_μ is not a fixed point of T_μ. In the figure on the right, we have $\bar{J} \geq T_\mu \bar{J}$ but T_μ is discontinuous from above at J_μ, so Assumption D does not hold, and J_μ is not a fixed point of T_μ.

The conditions of part (b) of Assumptions I and D are continuity assumptions designed to preclude some of the pathologies of the type encountered also in Chapter 3, and addressed with the use of S-regular policies. In particular, these conditions are essential for making a connection with fixed point theory: they ensure that J_μ is a fixed point of T_μ, as shown in the following proposition.

Proposition 4.3.1: Let Assumption I or Assumption D hold. Then for every policy $\mu \in \mathcal{M}$, we have

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

Proof: Let Assumption I hold. Then for all $k \geq 0$,

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

and by taking the limit as $k \to \infty$, and using part (b) of Assumption I, and the fact $T_\mu^k \bar{J} \uparrow J_\mu$, we have for all $x \in X$,

$$J_\mu(x) = \lim_{k\to\infty} H\big(x, \mu(x), T_\mu^k \bar{J}\big) = H\big(x, \mu(x), \lim_{k\to\infty} T_\mu^k \bar{J}\big) = H\big(x, \mu(x), J_\mu\big),$$

or equivalently $J_\mu = T_\mu J_\mu$. The proof for the case of Assumption D is similar. **Q.E.D.**

Figure 4.3.1 illustrates how J_μ may fail to be a fixed point of T_μ if part (b) of Assumption I or D is violated. Note also that continuity of T_μ does not imply continuity of T, and for example, under Assumption I, T may be discontinuous from below. We will see later that as a result, the value iteration sequence $\{T^k\bar{J}\}$ may fail to converge to J^* in the absence of additional conditions (see Section 4.3.2). Part (c) of Assumption I is a technical condition that facilitates the analysis, and assures the existence of ϵ-optimal policies.

Despite the similarities between Assumptions I and D, the corresponding results that one may obtain involve some substantial differences. An important fact, which breaks the symmetry between the two cases, is that J^* is approached by $T^k\bar{J}$ from below in the case of Assumption I and from above in the case of Assumption D. Another important fact is that since the condition $\bar{J}(x) > -\infty$ for all $x \in X$ is part of Assumption I, all the functions J encountered in the analysis under this assumption (such as $T^k\bar{J}$, J_π, and J^*) also satisfy $J(x) > -\infty$, for all $x \in X$. In particular, if $J \geq \bar{J}$, we have

$$(TJ)(x) \geq (T\bar{J})(x) > -\infty, \qquad \forall\ x \in X,$$

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

$$T_{\mu_\epsilon} J \leq TJ + \epsilon\, e.$$

This property is critical for the existence of an ϵ-optimal policy under Assumption I (see the next proposition) and is not available under Assumption D. It accounts in part for the different character of the results that can be obtained under the two assumptions.

4.3.1 Fixed Point Properties and Optimality Conditions

We first consider the question whether the optimal cost function J^* is a fixed point of T. This is indeed true, but the lines of proof are different under the Assumptions I and D. We begin with the proof under Assumption I, and as a preliminary step we show the existence of an ϵ-optimal policy, something that is of independent theoretical interest.

Proposition 4.3.2: Let Assumption I hold. Then given any $\epsilon > 0$, there exists a policy $\pi_\epsilon \in \Pi$ such that

$$J^* \leq J_{\pi_\epsilon} \leq J^* + \epsilon\, e.$$

Furthermore, if the scalar α in part (c) of Assumption I satisfies $\alpha < 1$, the policy π_ϵ can be taken to be stationary.

Proof: Let $\{\epsilon_k\}$ be a sequence such that $\epsilon_k > 0$ for all k and

$$\sum_{k=0}^{\infty} \alpha^k \epsilon_k = \epsilon. \tag{4.9}$$

For each $x \in X$, consider a sequence of policies $\{\pi_k[x]\} \subset \Pi$ of the form

$$\pi_k[x] = \{\mu_0^k[x], \mu_1^k[x], \dots\}, \tag{4.10}$$

such that for $k = 0, 1, \dots,$

$$J_{\pi_k[x]}(x) \leq J^*(x) + \epsilon_k. \tag{4.11}$$

Such a sequence exists, since we have assumed that $\bar{J}(x) > -\infty$, and therefore $J^*(x) > -\infty$, for all $x \in X$.

The preceding notation should be interpreted as follows. The policy $\pi_k[x]$ of Eq. (4.10) is associated with x. Thus $\mu_i^k[x]$ denotes for each x and k, a function in $\mathcal{M}$, while $\mu_i^k[x](z)$ denotes the value of $\mu_i^k[x]$ at an element $z \in X$. In particular, $\mu_i^k[x](x)$ denotes the value of $\mu_i^k[x]$ at $x \in X$.

Consider the functions $\overline{\mu}_k$ defined by

$$\overline{\mu}_k(x) = \mu_0^k[x](x), \qquad \forall\ x \in X, \tag{4.12}$$

and the functions $\bar{J}_k$ defined by

$$\bar{J}_k(x) = H\left(x, \overline{\mu}_k(x), \lim_{m\to\infty} T_{\mu_1^k[x]} \cdots T_{\mu_m^k[x]} \bar{J}\right), \qquad \forall\ x \in X,\ k = 0, 1, \dots. \tag{4.13}$$

By using Eqs. (4.11), (4.12), and part (b) of Assumption I, we obtain for all $x \in X$ and $k = 0, 1, \dots$

$$\begin{aligned} \bar{J}_k(x) &= \lim_{m\to\infty} \left(T_{\mu_0^k[x]} \cdots T_{\mu_m^k[x]} \bar{J}\right)(x) \\ &= J_{\pi_k[x]}(x) \\ &\leq J^*(x) + \epsilon_k. \end{aligned} \tag{4.14}$$

From Eqs. (4.13), (4.14), and part (c) of Assumption I, we have for all $x \in X$ and $k = 1, 2, \dots,$

$$\begin{aligned} (T_{\overline{\mu}_{k-1}} \bar{J}_k)(x) &= H\left(x, \overline{\mu}_{k-1}(x), \bar{J}_k\right) \\ &\leq H\left(x, \overline{\mu}_{k-1}(x), J^* + \epsilon_k\, e\right) \\ &\leq H\left(x, \overline{\mu}_{k-1}(x), J^*\right) + \alpha\epsilon_k \\ &\leq H\left(x, \overline{\mu}_{k-1}(x), \lim_{m\to\infty} T_{\mu_1^{k-1}[x]} \cdots T_{\mu_m^{k-1}[x]} \bar{J}\right) + \alpha\epsilon_k \\ &= \bar{J}_{k-1}(x) + \alpha\epsilon_k, \end{aligned}$$

and finally

$$T_{\overline{\mu}_{k-1}} \bar{J}_k \leq \bar{J}_{k-1} + \alpha \epsilon_k \, e, \qquad k = 1, 2, \ldots .$$

Using this inequality and part (c) of Assumption I, we obtain

$$\begin{aligned} T_{\overline{\mu}_{k-2}} T_{\overline{\mu}_{k-1}} \bar{J}_k &\leq T_{\overline{\mu}_{k-2}} (\bar{J}_{k-1} + \alpha \epsilon_k \, e) \\ &\leq T_{\overline{\mu}_{k-2}} \bar{J}_{k-1} + \alpha^2 \epsilon_k \, e \\ &\leq \bar{J}_{k-2} + (\alpha \epsilon_{k-1} + \alpha^2 \epsilon_k) \, e. \end{aligned}$$

Continuing in the same manner, we have for $k = 1, 2, \ldots,$

$$T_{\overline{\mu}_0} \cdots T_{\overline{\mu}_{k-1}} \bar{J}_k \leq \bar{J}_0 + (\alpha \epsilon_1 + \cdots + \alpha^k \epsilon_k) \, e \leq J^* + \left(\sum_{i=0}^{k} \alpha^i \epsilon_i \right) e.$$

Since $\bar{J} \leq \bar{J}_k$, it follows that

$$T_{\overline{\mu}_0} \cdots T_{\overline{\mu}_{k-1}} \bar{J} \leq J^* + \left(\sum_{i=0}^{k} \alpha^i \epsilon_i \right) e.$$

Denote $\pi_\epsilon = \{\overline{\mu}_0, \overline{\mu}_1, \ldots\}$. Then by taking the limit in the preceding inequality and using Eq. (4.9), we obtain

$$J_{\pi_\epsilon} \leq J^* + \epsilon \, e.$$

If $\alpha < 1$, we take $\epsilon_k = \epsilon(1-\alpha)$ for all k, and $\pi_k[x] = \{\mu_0[x], \mu_1[x], \ldots\}$ in Eq. (4.11). The stationary policy $\pi_\epsilon = \{\overline{\mu}, \overline{\mu}, \ldots\}$, where $\overline{\mu}(x) = \mu_0[x](x)$ for all $x \in X$, satisfies $J_{\pi_\epsilon} \leq J^* + \epsilon \, e$. **Q.E.D.**

Note that the assumption $\alpha < 1$ is essential in order to be able to take π_ϵ stationary in the preceding proposition. As an example, let $X = \{0\}$, $U(0) = (0, \infty)$, $\bar{J}(0) = 0$, $H(0, u, J) = u + J(0)$. Then $J^*(0) = 0$, but for any $\mu \in \mathcal{M}$, we have $J_\mu(0) = \infty$.

By using Prop. 4.3.2 we can prove the following.

Proposition 4.3.3: Let Assumption I hold. Then

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

Furthermore, if $J' \in \mathcal{E}(X)$ is such that $J' \geq \bar{J}$ and $J' \geq TJ'$, then $J' \geq J^*$.

Proof: For every $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi$ and $x \in X$, we have using part (b) of Assumption I,

$$\begin{aligned} J_\pi(x) &= \lim_{k\to\infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} \bar{J})(x) \\ &= T_{\mu_0} \left(\lim_{k\to\infty} T_{\mu_1} \cdots T_{\mu_k} \bar{J} \right)(x) \\ &\geq (T_{\mu_0} J^*)(x) \\ &\geq (TJ^*)(x). \end{aligned}$$

By taking the infimum of the left-hand side over $\pi \in \Pi$, we obtain

$$J^* \geq TJ^*.$$

To prove the reverse inequality, let ϵ_1 and ϵ_2 be any positive scalars, and let $\overline{\pi} = \{\overline{\mu}_0, \overline{\mu}_1, \ldots\}$ be such that

$$T_{\overline{\mu}_0} J^* \leq TJ^* + \epsilon_1 e, \qquad J_{\pi_1} \leq J^* + \epsilon_2 e,$$

where $\pi_1 = \{\overline{\mu}_1, \overline{\mu}_2, \ldots\}$ (such a policy exists by Prop. 4.3.2). The sequence $\{T_{\overline{\mu}_1} \cdots T_{\overline{\mu}_k} \bar{J}\}$ is monotonically nondecreasing, so by using the preceding relations and part (c) of Assumption I, we have

$$\begin{aligned} T_{\overline{\mu}_0} T_{\overline{\mu}_1} \cdots T_{\overline{\mu}_k} \bar{J} &\leq T_{\overline{\mu}_0} \left(\lim_{k\to\infty} T_{\overline{\mu}_1} \cdots T_{\overline{\mu}_k} \bar{J} \right) \\ &= T_{\overline{\mu}_0} J_{\pi_1} \\ &\leq T_{\overline{\mu}_0} J^* + \alpha\epsilon_2 e \\ &\leq TJ^* + (\epsilon_1 + \alpha\epsilon_2) e. \end{aligned}$$

Taking the limit as $k \to \infty$, we obtain

$$J^* \leq J_{\overline{\pi}} = \lim_{k\to\infty} T_{\overline{\mu}_0} T_{\overline{\mu}_1} \cdots T_{\overline{\mu}_k} \bar{J} \leq TJ^* + (\epsilon_1 + \alpha\epsilon_2) e.$$

Since ϵ_1 and ϵ_2 can be taken arbitrarily small, it follows that

$$J^* \leq TJ^*.$$

Hence $J^* = TJ^*$.

Assume that $J' \in \mathcal{E}(X)$ satisfies $J' \geq \bar{J}$ and $J' \geq TJ'$. Let $\{\epsilon_k\}$ be any sequence with $\epsilon_k > 0$ for all k, and consider a policy $\overline{\pi} = \{\overline{\mu}_0, \overline{\mu}_1, \ldots\} \in \Pi$ such that

$$T_{\overline{\mu}_k} J' \leq TJ' + \epsilon_k e, \qquad k = 0, 1, \ldots.$$

We have from part (c) of Assumption I

$$\begin{aligned}
J^* &= \inf_{\pi\in\Pi} \lim_{k\to\infty} T_{\mu_0}\cdots T_{\mu_k}\bar{J} \\
&\le \inf_{\pi\in\Pi} \liminf_{k\to\infty} T_{\mu_0}\cdots T_{\mu_k}J' \\
&\le \liminf_{k\to\infty} T_{\overline{\mu}_0}\cdots T_{\overline{\mu}_k}J' \\
&\le \liminf_{k\to\infty} T_{\overline{\mu}_0}\cdots T_{\overline{\mu}_{k-1}}(TJ'+\epsilon_k\, e) \\
&\le \liminf_{k\to\infty} T_{\overline{\mu}_0}\cdots T_{\overline{\mu}_{k-1}}(J'+\epsilon_k\, e) \\
&\le \liminf_{k\to\infty} (T_{\overline{\mu}_0}\cdots T_{\overline{\mu}_{k-1}}J'+\alpha^k\epsilon_k\, e) \\
&\vdots \\
&\le \lim_{k\to\infty}\left(TJ'+\left(\sum_{i=0}^{k}\alpha^i\epsilon_i\right)e\right) \\
&\le J'+\left(\sum_{i=0}^{k}\alpha^i\epsilon_i\right)e.
\end{aligned}$$

Since we may choose $\sum_{i=0}^{k}\alpha^i\epsilon_i$ as small as desired, it follows that $J^*\le J'$. **Q.E.D.**

The following counterexamples show that parts (b) and (c) of Assumption I are essential for the preceding proposition to hold.

Example 4.3.1 (Counterexample to Bellman's Equation I)

Let

$$X=\{0,1\},\qquad U(0)=U(1)=(-1,0],\qquad \bar{J}(0)=\bar{J}(1)=-1,$$

$$H(0,u,J)=\begin{cases} u & \text{if } J(1)\le -1,\\ 0 & \text{if } J(1)>-1,\end{cases}\qquad H(1,u,J)=u.$$

Then for $N\ge 1$,

$$(T_{\mu_0}\cdots T_{\mu_{N-1}}\bar{J})(0)=0,\qquad (T_{\mu_0}\cdots T_{\mu_{N-1}}\bar{J})(1)=\mu_0(1).$$

Thus

$$J^*(0)=0,\qquad J^*(1)=-1,\qquad (TJ^*)(0)=-1,\qquad (TJ^*)(1)=-1,$$

and hence $J^*\ne TJ^*$. Notice also that $\bar{J}$ is a fixed point of T, while $\bar{J}\le J^*$ and $\bar{J}\ne J^*$, so the second part of Prop. 4.3.3 fails when $\bar{J}=J'$. Here parts (a) and (b) of Assumption I are satisfied, but part (c) is violated, since $H(0,u,\cdot)$ is discontinuous at $J=-1$ when $u<0$.

Example 4.3.2 (Counterexample to Bellman's Equation II)

Let

$$X = \{0, 1\}, \qquad U(0) = U(1) = \{0\}, \qquad \bar{J}(0) = \bar{J}(1) = 0,$$

$$H(0, 0, J) = \begin{cases} 0 & \text{if } J(1) < \infty, \\ \infty & \text{if } J(1) = \infty, \end{cases} \qquad H(1, 0, J) = J(1) + 1.$$

Here there is only one policy, which we denote by μ. For all $N \geq 1$, we have

$$(T_\mu^N \bar{J})(0) = 0, \qquad (T_\mu^N \bar{J})(1) = N,$$

so $J^*(0) = 0$, $J^*(1) = \infty$. On the other hand, we have $(TJ^*)(0) = (TJ^*)(1) = \infty$ and $J^* \neq TJ^*$. Here parts (a) and (c) of Assumption I are satisfied, but part (b) is violated.

As a corollary to Prop. 4.3.3 we obtain the following.

Proposition 4.3.4: Let Assumption I hold. Then for every $\mu \in \mathcal{M}$, we have

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

Furthermore, if $J' \in \mathcal{E}(X)$ is such that $J' \geq \bar{J}$ and $J' \geq T_\mu J'$, then $J' \geq J_\mu$.

Proof: Consider the variant of the infinite horizon problem where the control constraint set is $U_\mu(x) = \{\mu(x)\}$ rather than $U(x)$ for all $x \in X$. Application of Prop. 4.3.3 yields the result. **Q.E.D.**

We now provide the counterpart of Prop. 4.3.3 under Assumption D. We first prove a preliminary result regarding the convergence of the value iteration method, which is of independent interest (we will see later that this result need not hold under Assumption I).

Proposition 4.3.5: Let Assumption D hold. Then $T^N \bar{J} = J_N^*$, where J_N^* is the optimal cost function for the N-stage problem. Moreover

$$J^* = \lim_{N \to \infty} J_N^*.$$

Proof: By repeating the proof of Prop. 4.2.3, we have $T^N \bar{J} = J_N^*$ [part (b) of Assumption D is essentially identical to the assumption of that proposition]. Clearly we have $J^* \leq J_N^*$ for all N, and hence $J^* \leq \lim_{N \to \infty} J_N^*$.

To prove the reverse inequality, we note that for all $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi$, we have

$$T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J} \geq J_N^*.$$

By taking the limit of both sides as $N \to \infty$, we obtain $J_\pi \geq \lim_{N\to\infty} J_N^*$, and by taking infimum over π, $J^* \geq \lim_{N\to\infty} J_N^*$. Thus $J^* = \lim_{N\to\infty} J_N^*$. **Q.E.D.**

Proposition 4.3.6: Let Assumption D hold. Then

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

Furthermore, if $J' \in \mathcal{E}(X)$ is such that $J' \leq \bar{J}$ and $J' \leq TJ'$, then $J' \leq J^*$.

Proof: For any $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi$, we have

$$J_\pi = \lim_{k\to\infty} T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} \bar{J} \geq \lim_{k\to\infty} T_{\mu_0} T^k \bar{J} \geq T_{\mu_0} J^*,$$

where the last inequality follows from the fact $T^k \bar{J} \downarrow J^*$ (cf. Prop. 4.3.5). Taking the infimum of both sides over $\pi \in \Pi$, we obtain $J^* \geq TJ^*$.

To prove the reverse inequality, we select any $\mu \in \mathcal{M}$, and we apply T_μ to both sides of the equation $J^* = \lim_{N\to\infty} T^N \bar{J}$ (cf. Prop. 4.3.5). By using part (b) of assumption D, we obtain

$$T_\mu J^* = T_\mu \left(\lim_{N\to\infty} T^N \bar{J} \right) = \lim_{N\to\infty} T_\mu T^N \bar{J} \geq \lim_{N\to\infty} T^{N+1} \bar{J} = J^*.$$

Taking the infimum of the left-hand side over $\mu \in \mathcal{M}$, we obtain $TJ^* \geq J^*$, showing that $TJ^* = J^*$.

To complete the proof, let $J' \in \mathcal{E}(X)$ be such that $J' \leq \bar{J}$ and $J' \leq TJ'$. Then we have

$$\begin{aligned} J^* &= \inf_{\pi\in\Pi} \lim_{N\to\infty} T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J} \\ &\geq \lim_{N\to\infty} \inf_{\pi\in\Pi} T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J} \\ &\geq \lim_{N\to\infty} \inf_{\pi\in\Pi} T_{\mu_0} \cdots T_{\mu_{N-1}} J' \\ &\geq \lim_{N\to\infty} T^N J' \\ &\geq J', \end{aligned}$$

where the last inequality follows from the hypothesis $J' \leq TJ'$. Thus $J^* \geq J'$. **Q.E.D.**

Counterexamples to Bellman's equation can be readily constructed if part (b) of Assumption D (continuity from above) is violated. In particular, in Examples 4.2.1 and 4.2.2, part (a) of Assumption D is satisfied but part (b) is not. In both cases we have $J^* \neq TJ^*$, as the reader can verify with a straightforward calculation.

Similar to Prop. 4.3.4, we obtain the following.

Proposition 4.3.7: Let Assumption D hold. Then for every $\mu \in \mathcal{M}$, we have

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

Furthermore, if $J' \in \mathcal{E}(X)$ is such that $J' \leq \bar{J}$ and $J' \leq T_\mu J'$, then $J' \leq J_\mu$.

Proof: Consider the variation of our problem where the control constraint set is $U_\mu(x) = \{\mu(x)\}$ rather than $U(x)$ for all $x \in X$. Application of Prop. 4.3.6 yields the result. **Q.E.D.**

An examination of the proof of Prop. 4.3.6 shows that the only point where we need part (b) of Assumption D was in establishing the relations

$$\lim_{N\to\infty} TJ_N^* = T\left(\lim_{N\to\infty} J_N^*\right)$$

and

$$J_N^* = T^N \bar{J}.$$

If these relations can be established independently, then the result of Prop. 4.3.6 follows. In this manner we obtain the following proposition.

Proposition 4.3.8: Let part (a) of Assumption D hold, assume that X is a finite set, and that $J^*(x) > -\infty$ for all $x \in X$. Assume further that there exists a scalar $\alpha \in (0, \infty)$ such that for all scalars $r \in (0, \infty)$ and functions $J \in \mathcal{E}(X)$ with $J \leq \bar{J}$, we have

$$H(x,u,J) - \alpha\, r \leq H(x,u,J - r\, e), \qquad \forall\ x \in X,\ u \in U(x). \quad (4.15)$$

Then

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

Furthermore, if $J' \in \mathcal{E}(X)$ is such that $J' \leq \bar{J}$ and $J' \leq TJ'$, then $J' \leq J^*$.

Proof: A nearly verbatim repetition of Prop. 4.2.4 shows that under our assumptions we have $J_N^* = T^N \bar{J}$ for all N. We will show that

$$\lim_{N\to\infty} H(x,u,J_N^*) \leq H\left(x,u,\lim_{N\to\infty} J_N^*\right), \qquad \forall\ x\in X,\ u\in U(x).$$

Then the result follows as in the proof of Prop. 4.3.6.

Assume the contrary, i.e., that for some $\tilde{x}\in X$, $\tilde{u}\in U(\tilde{x})$, and $\epsilon>0$, there holds

$$H(\tilde{x},\tilde{u},J_k^*) - \epsilon > H\left(\tilde{x},\tilde{u},\lim_{N\to\infty} J_N^*\right), \qquad k=1,2,\ldots.$$

From the finiteness of X and the fact

$$J^*(x) = \lim_{N\to\infty} J_N^*(x) > -\infty, \qquad \forall\ x\in X,$$

it follows that for some integer $\overline{k}>0$

$$J_k^* - (\epsilon/\alpha)e \leq \lim_{N\to\infty} J_N^*, \qquad \forall\ k\geq \overline{k}.$$

By using the condition (4.15), we obtain for all $k\geq\overline{k}$

$$H(\tilde{x},\tilde{u},J_k^*) - \epsilon \leq H\big(\tilde{x},\tilde{u},J_k^* - (\epsilon/\alpha)\,e\big) \leq H\left(\tilde{x},\tilde{u},\lim_{N\to\infty} J_N^*\right),$$

which contradicts the earlier inequality. **Q.E.D.**

Characterization of Optimal Policies

We now provide necessary and sufficient conditions for optimality of a stationary policy. These conditions are markedly different under Assumptions I and D.

Proposition 4.3.9: Let Assumption I hold. Then a stationary policy μ is optimal if and only if

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

Proof: If μ is optimal, then $J_\mu = J^*$ so that the equation $J^* = TJ^*$ (cf. Prop. 4.3.3) implies that $J_\mu = TJ_\mu$. Since $J_\mu = T_\mu J_\mu$ (cf. Prop. 4.3.4), it follows that $T_\mu J^* = TJ^*$.

Conversely, if $T_\mu J^* = TJ^*$, then since $J^* = TJ^*$, it follows that $T_\mu J^* = J^*$. By Prop. 4.3.4, it follows that $J_\mu \leq J^*$, so μ is optimal. **Q.E.D.**

Proposition 4.3.10: Let Assumption D hold. Then a stationary policy μ is optimal if and only if

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

Proof: If μ is optimal, then $J_\mu = J^*$, so that the equation $J^* = TJ^*$ (cf. Prop. 4.3.6) can be written as $J_\mu = TJ_\mu$. Since $J_\mu = T_\mu J_\mu$ (cf. Prop. 4.3.4), it follows that $T_\mu J_\mu = TJ_\mu$.

Conversely, if $T_\mu J_\mu = TJ_\mu$, then since $J_\mu = T_\mu J_\mu$, it follows that $J_\mu = TJ_\mu$. By Prop. 4.3.7, it follows that $J_\mu \leq J^*$, so μ is optimal. **Q.E.D.**

An example showing that under Assumption I, the condition $T_\mu J_\mu = TJ_\mu$ does not guarantee optimality of μ is given in Exercise 4.3. Under Assumption D, we note that by Prop. 4.3.1, we have $J_\mu = T_\mu J_\mu$ for all μ, so if μ is a stationary optimal policy, the fixed point equation

$$J^*(x) = \inf_{u \in U(x)} H(x, u, J^*), \qquad \forall\ x \in X, \tag{4.16}$$

and the optimality condition of Prop. 4.3.10, yield

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

Thus under D, a stationary optimal policy attains the infimum in the fixed point Eq. (4.16) for all x. However, there may exist nonoptimal stationary policies also attaining the infimum for all x; an example is the shortest path problem of Section 3.1.1 for the case where $a = 0$ and $b = 1$. Moreover, it is possible that this infimum is attained but no optimal policy exists, as shown by Fig. 4.3.2.

Proposition 4.3.9 shows that under Assumption I, there exists a stationary optimal policy if and only if the infimum in the optimality equation

$$J^*(x) = \inf_{u \in U(x)} H(x, u, J^*)$$

is attained for every $x \in X$. When the infimum is not attained for some $x \in X$, this optimality equation can still be used to yield an ϵ-optimal policy, which can be taken to be stationary whenever the scalar α in Assumption I(c) is strictly less than 1. This is shown in the following proposition.

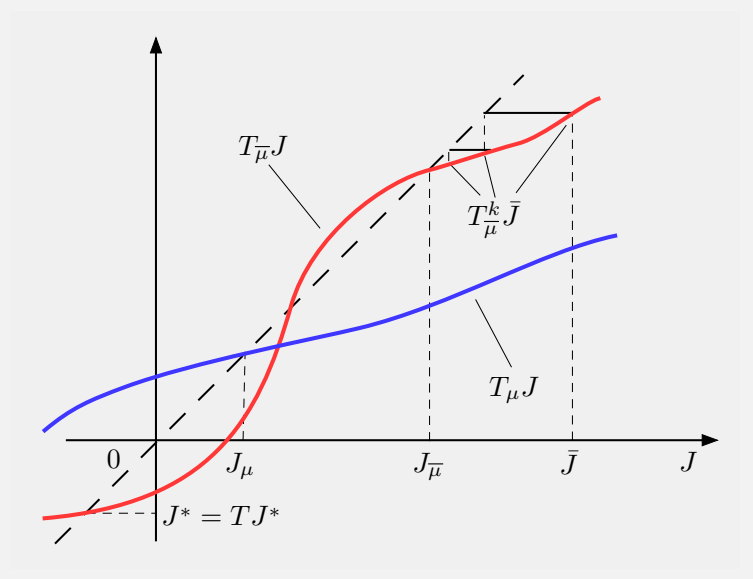


Figure 4.3.2. An example where nonstationary policies are dominant under Assumption D. Here there is only one state and $S = \Re$. There are two stationary policies μ and $\overline{\mu}$ with cost functions J_μ and $J_{\overline{\mu}}$ as shown. However, by considering a nonstationary policy of the form $\pi_k = \{\overline{\mu}, \ldots, \overline{\mu}, \mu, \mu, \ldots\}$, with a number k of policies $\overline{\mu}$, we can obtain a sequence $\{J_{\pi_k}\}$ that converges to the value J^* shown. Note that here there is no optimal policy, stationary or not.

Proposition 4.3.11: Let Assumption I hold. Then:

(a) If $\epsilon > 0$, the sequence $\{\epsilon_k\}$ satisfies $\sum_{k=0}^{\infty} \alpha^k \epsilon_k = \epsilon$, and $\epsilon_k > 0$ for all k, and the policy $\pi^* = \{\mu_0^*, \mu_1^*, \ldots\} \in \Pi$ is such that

$$T_{\mu_k^*} J^* \leq TJ^* + \epsilon_k\, e, \qquad \forall\ k = 0, 1, \ldots,$$

then

$$J^* \leq J_{\pi^*} \leq J^* + \epsilon\, e.$$

(b) If $\epsilon > 0$, the scalar α in part (c) of Assumption I is strictly less than 1, and $\mu^* \in \mathcal{M}$ is such that

$$T_{\mu^*} J^* \leq TJ^* + \epsilon(1 - \alpha)\, e,$$

then

$$J^* \leq J_{\mu^*} \leq J^* + \epsilon\, e.$$

Proof: (a) Since $TJ^* = J^*$, we have

$$T_{\mu_k^*} J^* \leq J^* + \epsilon_k \, e,$$

and applying $T_{\mu_{k-1}^*}$ to both sides, we obtain

$$T_{\mu_{k-1}^*} T_{\mu_k^*} J^* \leq T_{\mu_{k-1}^*} J^* + \alpha \epsilon_k \, e \leq J^* + (\epsilon_{k-1} + \alpha \epsilon_k) \, e.$$

Applying $T_{\mu_{k-2}^*}$ throughout and repeating the process, we obtain for every $k = 1, 2, \ldots,$

$$T_{\mu_0^*} \cdots T_{\mu_k^*} J^* \leq J^* + \left(\sum_{i=0}^{k} \alpha^i \epsilon_i \right) e, \qquad k = 1, 2, \ldots .$$

Since $\bar{J} \leq J^*$, it follows that

$$T_{\mu_0^*} \cdots T_{\mu_k^*} \bar{J} \leq J^* + \left(\sum_{i=0}^{k} \alpha^i \epsilon_i \right) e, \qquad k = 1, 2, \ldots .$$

By taking the limit as $k \to \infty$, we obtain $J_{\pi^*} \leq J^* + \epsilon \, e$.

(b) This part is proved by taking $\epsilon_k = \epsilon(1 - \alpha)$ and $\mu_k^* = \mu^*$ for all k in the preceding argument. **Q.E.D.**

Under Assumption D, the existence of an ϵ-optimal policy is harder to establish, and requires some restrictive conditions.

Proposition 4.3.12: Let Assumption D hold, and let the additional assumptions of Prop. 4.3.8 hold. Then for any $\epsilon > 0$, there exists an ϵ-optimal policy.

Proof: For each N, denote

$$\epsilon_N = \frac{\epsilon}{2(1 + \alpha + \cdots + \alpha^{N-1})},$$

and let

$$\pi_N = \{\mu_0^N, \mu_1^N, \ldots, \mu_{N-1}^N, \mu, \mu \ldots\}$$

be such that $\mu \in \mathcal{M}$, and for $k = 0, \ldots, N-1$, $\mu_k^N \in \mathcal{M}$ and

$$T_{\mu_k^N} T^{N-k-1} \bar{J} = T^{N-k} \bar{J} + \epsilon_N \, e.$$

We have $T^N_{\mu_{N-1}} \bar{J} \leq T\bar{J} + \epsilon_N\, e$, and applying $T^N_{\mu_{N-2}}$ to both sides, we obtain

$$T^N_{\mu_{N-2}} T^N_{\mu_{N-1}} \bar{J} \leq T^N_{\mu_{N-2}} T\bar{J} + \alpha\epsilon_N\, e \leq T^2\bar{J} + (1+\alpha)\epsilon_N\, e.$$

Continuing in the same manner, we have

$$T^N_{\mu_0} \cdots T^N_{\mu_{N-1}} \bar{J} \leq T^N \bar{J} + (1 + \alpha + \cdots + \alpha^{N-1})\epsilon_N\, e,$$

from which we obtain for $N = 0, 1, \ldots$,

$$J_{\pi_N} \leq T^N \bar{J} + (\epsilon/2)\, e.$$

By Prop. 4.3.5, we have $J^* = \lim_{N\to\infty} T^N \bar{J}$, so let $\bar{N}$ be such that

$$T^{\bar{N}} \bar{J} \leq J^* + (\epsilon/2)\, e$$

[such a $\bar{N}$ exists using the assumptions of finiteness of X and $J^*(x) > -\infty$ for all $x \in X$]. Then we obtain $J_{\pi_{\bar{N}}} \leq J^* + \epsilon\, e$, and $\pi_{\bar{N}}$ is the desired policy. **Q.E.D.**

4.3.2 Value Iteration

We will now discuss algorithms for abstract DP under Assumptions I and and D. We first consider the VI algorithm, which consists of successively generating $T\bar{J}, T^2\bar{J}, \ldots$. Note that because T need not be a contraction, it may have multiple fixed points J all of which satisfy $J \geq J^*$ under Assumption I (cf. Prop. 4.3.3) or $J \leq J^*$ under Assumption D (cf. Prop. 4.3.6). Thus, in the absence of additional conditions (to be discussed in Sections 4.4 and 4.5), it is essential to start VI with $\bar{J}$ or an initial J_0 such that $\bar{J} \leq J_0 \leq J^*$ under Assumption I or $\bar{J} \geq J_0 \geq J^*$ under Assumption D. In the next two propositions, we show that for such initial conditions, we have convergence of VI to J^* under Assumption D, and with an additional compactness condition, under Assumption I.

Proposition 4.3.13: Let Assumption D hold, and assume that $J_0 \in \mathcal{E}(X)$ is such that $\bar{J} \geq J_0 \geq J^*$. Then

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

Proof: The condition $\bar{J} \geq J_0 \geq J^*$ implies that $T^k\bar{J} \geq T^k J_0 \geq J^*$ for all k. By Prop. 4.3.5, $T^k\bar{J} \to J^*$, and the result follows. **Q.E.D.**

The convergence of VI under I requires an additional compactness condition, which is satisfied in particular if $U(x)$ is a finite set for all $x \in X$.

Proposition 4.3.14: Let Assumption I hold, let U be a metric space, and assume that the sets

$$U_k(x,\lambda) = \big\{u \in U(x) \big|\ H(x,u,T^k\bar{J}) \leq \lambda\big\} \tag{4.17}$$

are compact for every $x \in X$, $\lambda \in \Re$, and for all k greater than some integer $\overline{k}$. Assume that $J_0 \in \mathcal{E}(X)$ is such that $\bar{J} \leq J_0 \leq J^*$. Then

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

Furthermore, there exists a stationary optimal policy.

Proof: Similar to the proof of Prop. 4.3.13, it will suffice to show that $T^k\bar{J} \to J^*$. Since $\bar{J} \leq J^*$, we have $T^k\bar{J} \leq T^kJ^* = J^*$, so that

$$\bar{J} \leq T\bar{J} \leq \cdots \leq T^k\bar{J} \leq \cdots \leq J^*.$$

Thus we have $T^k\bar{J} \uparrow J_\infty$ for some $J_\infty \in \mathcal{E}(X)$ satisfying $T^k\bar{J} \leq J_\infty \leq J^*$ for all k. Applying T to this relation, we obtain

$$(T^{k+1}\bar{J})(x) = \min_{u\in U(x)} H(x,u,T^k\bar{J}) \leq (TJ_\infty)(x),$$

and by taking the limit as $k \to \infty$, it follows that

$$J_\infty \leq TJ_\infty.$$

Assume to arrive at a contradiction that there exists a state $\tilde{x} \in X$ such that

$$J_\infty(\tilde{x}) < (TJ_\infty)(\tilde{x}). \tag{4.18}$$

Similar to Lemma 3.3.1, there exists a point u_k attaining the minimum in

$$(T^{k+1}\bar{J})(\tilde{x}) = \inf_{u\in U(\tilde{x})} H(\tilde{x},u,T^k\bar{J});$$

i.e., u_k is such that

$$(T^{k+1}\bar{J})(\tilde{x}) = H(\tilde{x},u_k,T^k\bar{J}).$$

Clearly, by Eq. (4.18), we must have $J_\infty(\tilde{x}) < \infty$. For every k, consider the set

$$U_k\big(\tilde{x}, J_\infty(\tilde{x})\big) = \Big\{u \in U(\tilde{x}) \ \big|\ H(\tilde{x},u_k,T^k\bar{J}) \leq J_\infty(\tilde{x})\Big\},$$

and the sequence $\{u_i\}_{i=k}^{\infty}$. Since $T^k\bar{J} \uparrow J_\infty$, it follows that for all $i \geq k$,

$$H(\tilde{x}, u_i, T^k\bar{J}) \leq H(\tilde{x}, u_i, T^i\bar{J}) \leq J_\infty(\tilde{x}).$$

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 $\{u_i\}$. It follows that if $\tilde{u}$ is a limit point of $\{u_i\}$ then

$$\tilde{u} \in \cap_{k=0}^{\infty} U_k\big(\tilde{x}, J_\infty(\tilde{x})\big).$$

By Eq. (4.17), this implies that for all $k \geq \overline{k}$

$$J_\infty(\tilde{x}) \geq H(\tilde{x}, \tilde{u}, T^k\bar{J}) \geq (T^{k+1}\bar{J})(\tilde{x}).$$

Taking the limit as $k \to \infty$, and using part (b) of Assumption I, we obtain

$$J_\infty(\tilde{x}) \geq H(\tilde{x}, \tilde{u}, J_\infty) \geq (TJ_\infty)(\tilde{x}), \tag{4.19}$$

which contradicts Eq. (4.18). Hence $J_\infty = TJ_\infty$, which implies that $J_\infty \geq J^*$ in view of Prop. 4.3.3. Combined with the inequality $J_\infty \leq J^*$, which was shown earlier, we have $J_\infty = J^*$.

To show that there exists an optimal stationary policy, observe that the relation $J^* = J_\infty = TJ_\infty$ and Eq. (4.19) [whose proof is valid for all $\tilde{x} \in X$ such that $J^*(\tilde{x}) < \infty$] imply that $\tilde{u}$ attains the infimum in

$$J^*(\tilde{x}) = \inf_{u \in U(\tilde{x})} H(\tilde{x}, u, J^*)$$

for all $\tilde{x} \in X$ with $J^*(\tilde{x}) < \infty$. For $\tilde{x} \in X$ such that $J^*(\tilde{x}) = \infty$, every $u \in U(\tilde{x})$ attains the preceding minimum. Hence by Prop. 4.3.9 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 infimum in the relation

$$(T^{k+1}\bar{J})(\tilde{x}) = \inf_{u \in U(x)} H(\tilde{x}, u, T^k\bar{J}),$$

and $\mu^*(\tilde{x})$ is a limit point of $\{\mu_k(\tilde{x})\}$, for every $\tilde{x} \in X$, then the stationary policy μ^* is optimal. Furthermore, $\{\mu_k(\tilde{x})\}$ has at least one limit point for every $\tilde{x} \in X$ for which $J^*(\tilde{x}) < \infty$. Thus *the VI algorithm under the assumption of Prop. 4.3.14 yields in the limit not only the optimal cost function J^* but also an optimal stationary policy.*

On the other hand, under Assumption I but in the absence of the compactness condition (4.17), $T^k\bar{J}$ need not converge to J^*. What is happening here is that while the mappings T_μ are continuous from below as required by Assumption I(b), T may not be, and a phenomenon like the one illustrated in the left-hand side of Fig. 4.3.1 may occur, whereby

$$\lim_{k\to\infty} T^k\bar{J} \leq T\left(\lim_{k\to\infty} T^k\bar{J}\right),$$

with strict inequality for some $x \in X$. This can happen even in simple deterministic optimal control problems, as shown by the following example.

Example 4.3.3 (Counterexample to Convergence of VI)

Let

$$X = [0, \infty), \qquad U(x) = (0, \infty), \quad \bar{J}(x) = 0, \quad \forall\ x \in X,$$

and

$$H(x, u, J) = \min\big\{1,\ x + J(2x + u)\big\}, \qquad \forall\ x \in X,\ u \in U(x).$$

Then it can be verified that for all $x \in X$ and policies μ, we have $J_\mu(x) = 1$, as well as $J^*(x) = 1$, while it can be seen by induction that starting with $\bar{J}$, the VI algorithm yields

$$(T^k \bar{J})(x) = \min\big\{1,\ (1 + 2^{k-1})x\big\}, \qquad \forall\ x \in X,\ k = 1, 2 \ldots.$$

Thus we have $0 = \lim_{k\to\infty} (T^k \bar{J})(0) \neq J^*(0) = 1$.

The range of convergence of VI may be expanded under additional assumptions. In particular, in Chapter 3, under various conditions involving the existence of optimal S-regular policies, we showed that VI converges to J^* assuming that the initial condition J_0 satisfies $J_0 \geq J^*$. Thus if the assumptions of Prop. 4.3.14 hold in addition, we are guaranteed convergence of VI starting from any J satisfying $J \geq \bar{J}$. Results of this type will be obtained in Sections 4.4 and 4.5, where semicontractive models satisfying Assumption I will be discussed.

Asynchronous Value Iteration

The concepts of asynchronous VI that we developed in Section 2.6.1 apply also under the Assumptions I and D of this section. Under Assumption I, if J^* is real-valued, we may apply Prop. 2.6.1 with the sets $S(k)$ defined by

$$S(k) = \{J \mid T^k \bar{J} \leq J \leq J^*\}, \qquad k = 0, 1, \ldots.$$

Assuming that $T^k \bar{J} \to J^*$ (cf. Prop. 4.3.14), it follows that the asynchronous form of VI converges pointwise to J^* starting from any function in $S(0)$. This result can also be shown for the case where J^* is not real-valued, by using a simple extension of Prop. 2.6.1, where the set of real-valued functions $\mathcal{R}(X)$ is replaced by the set of all $J \in \mathcal{E}(X)$ with $\bar{J} \leq J \leq J^*$.

Under Assumption D similar conclusions hold for the asynchronous version of VI that starts with a function J with $J^* \leq J \leq \bar{J}$. Asynchronous pointwise convergence to J^* can be shown, based on an extension of the asynchronous convergence theorem (Prop. 2.6.1), where $\mathcal{R}(X)$ is replaced by the set of all $J \in \mathcal{E}(X)$ with $J^* \leq J \leq \bar{J}$.

4.3.3 Exact and Optimistic Policy Iteration - λ-Policy Iteration

Unfortunately, in the absence of further conditions, the PI algorithm is not guaranteed to yield the optimal cost function and/or an optimal policy under either Assumption I or D. However, there are convergence results for nonoptimistic and optimistic variants of PI under some conditions. In what follows in this section we will provide an analysis of various types of PI, mainly under Assumption D. The analysis of PI under Assumption I will be given primarily in the next two sections, as it requires different assumptions and methods of proof, and will be coupled with regularity ideas relating to the semicontractive models of Chapter 3.

Optimistic Policy Iteration Under D

A surprising fact under Assumption D is that nonoptimistic/exact PI may generate a policy that is strictly inferior over the preceding one. Moreover there may be an oscillation between nonoptimal policies even when the state and control spaces are finite. An illustrative example is the shortest path example of Section 3.1.1, where it can be verified that exact PI may oscillate between the policy that moves to the destination from node 1 and the policy that does not. For a mathematical explanation, note that under Assumption D, we may have $T_\mu J^* = TJ^*$ without μ being optimal, so starting from an optimal policy, we may obtain a nonoptimal policy by PI.

On the other hand optimistic PI under Assumption D has much better convergence properties, because it embodies the mechanism of VI, which is convergent to J^* as we saw in the preceding subsection. Indeed, let us consider an optimistic PI algorithm that generates a sequence $\{J_k, \mu^k\}$ according to †

$$T_{\mu^k} J_k = TJ_k, \qquad J_{k+1} = T_{\mu^k}^{m_k} J_k, \tag{4.20}$$

where m_k is a positive integer. We assume that the algorithm starts with a function $J_0 \in \mathcal{E}(X)$ that satisfies $\bar{J} \geq J_0 \geq J^*$ and $J_0 \geq TJ_0$. For example, we may choose $J_0 = \bar{J}$. We have the following proposition.

Proposition 4.3.15: Let Assumption D hold and let $\{J_k, \mu^k\}$ be a sequence generated by the optimistic PI algorithm (4.20), assuming that $\bar{J} \geq J_0 \geq J^*$ and $J_0 \geq TJ_0$. Then $J_k \downarrow J^*$.

Proof: We have

$$J_0 \geq T_{\mu^0} J_0 \geq T_{\mu^0}^{m_0} J_0 = J_1 \geq T_{\mu^0}^{m_0+1} J_0 = T_{\mu^0} J_1 \geq TJ_1 = T_{\mu^1} J_1 \geq \cdots \geq J_2,$$

† As with all PI algorithms in this book, we assume that the policy improvement operation is well-defined, in the sense that there exists μ^k such that $T_{\mu^k} J_k = TJ_k$ for all k.

where the first, second, and third inequalities hold because the assumption $J_0 \geq TJ_0 = T_{\mu^0} J_0$ implies that

$$T_{\mu^0}^m J_0 \geq T_{\mu^0}^{m+1} J_0, \qquad \forall\ m \geq 0.$$

Continuing similarly we obtain

$$J_k \geq TJ_k \geq J_{k+1}, \qquad \forall\ k \geq 0.$$

Moreover, we can show by induction that $J_k \geq J^*$. Indeed this is true for $k = 0$ by assumption. If $J_k \geq J^*$, we have

$$J_{k+1} = T_{\mu^k}^{m_k} J_k \geq T^{m_k} J_k \geq T^{m_k} J^* = J^*, \tag{4.21}$$

where the last equality follows from the fact $TJ^* = J^*$ (cf. Prop. 4.3.6), thus completing the induction. By combining the preceding two relations, we have

$$J_k \geq TJ_k \geq J_{k+1} \geq J^*, \qquad \forall\ k \geq 0. \tag{4.22}$$

We will now show by induction that

$$T^k J_0 \geq J_k \geq J^*, \qquad \forall\ k \geq 0. \tag{4.23}$$

Indeed this relation holds by assumption for $k = 0$, and assuming that it holds for some $k \geq 0$, we have by applying T to it and by using Eq. (4.22),

$$T^{k+1} J_0 \geq TJ_k \geq J_{k+1} \geq J^*,$$

thus completing the induction. By applying Prop. 4.3.13 to Eq. (4.23), we obtain $J_k \downarrow J^*$. **Q.E.D.**

λ-Policy Iteration Under D

We now consider the λ-PI algorithm. It involves a scalar $\lambda \in (0, 1)$ and a corresponding multistep mapping, which bears a relation to temporal differences and the proximal algorithm (cf. Section 1.2.5). It is defined by

$$T_{\mu^k} J_k = TJ_k, \qquad J_{k+1} = T_{\mu^k}^{(\lambda)} J_k, \tag{4.24}$$

where for any policy μ and scalar $\lambda \in (0, 1)$, $T_\mu^{(\lambda)}$ is the mapping defined by

$$(T_\mu^{(\lambda)} J)(x) = (1 - \lambda) \sum_{t=0}^{\infty} \lambda^t (T_\mu^{t+1} J)(x), \qquad x \in X.$$

Here we assume that T_μ maps $\mathcal{R}(X)$ to $\mathcal{R}(X)$, and that for all $\mu \in \mathcal{M}$ and $J \in \mathcal{R}(X)$, the limit of the series above is well-defined as a function in $\mathcal{R}(X)$.

We discussed the λ-PI algorithm in connection with semicontractive problems in Section 3.2.4, where we assumed that

$$T_\mu(T_\mu^{(\lambda)}J) = T_\mu^{(\lambda)}(T_\mu J), \qquad \forall\ \mu \in \mathcal{M},\ J \in \mathcal{E}(X). \tag{4.25}$$

We will show that for undiscounted finite-state MDP, the algorithm can be implemented by using matrix inversion, just like nonoptimistic PI for discounted finite-state MDP. It turns out that this can be an advantage in some settings, including approximate simulation-based implementations.

As noted earlier, λ-PI and optimistic PI are similar: they just use the mapping T_{μ^k} to apply VI in different ways. In view of this similarity, it is not surprising that it has the same type of convergence properties as the earlier optimistic PI method (4.20). Similar to Prop. 4.3.15, we have the following.

Proposition 4.3.16: Let Assumption D hold and let $\{J_k, \mu^k\}$ be a sequence generated by the λ-PI algorithm (4.24), assuming Eq. (4.25), and that $\bar{J} \geq J_0 \geq J^*$ and $J_0 \geq TJ_0$. Then $J_k \downarrow J^*$.

Proof: As in the proof of Prop. 4.3.15, by using Assumption D, the monotonicity of T_μ, and the hypothesis $J_0 \geq TJ_0$, we have

$$J_0 \geq TJ_0 = T_{\mu^0}J_0 \geq T_{\mu^0}^{(\lambda)}J_0 = J_1 \geq T_{\mu^0}J_1 \geq TJ_1 = T_{\mu^1}J_1 \geq T_{\mu^1}^{(\lambda)}J_0 = J_2,$$

where for the third inequality, we use the relation $J_0 \geq T_{\mu^0}J_0$, the definition of J_1, and the assumption (4.25). Continuing in the same manner,

$$J_k \geq TJ_k \geq J_{k+1}, \qquad \forall\ k \geq 0.$$

Similar to the proof of Prop. 4.3.15, we show by induction that $J_k \geq J^*$, using the fact that if $J_k \geq J^*$, then

$$J_{k+1} = T_{\mu^k}^{(\lambda)}J_k \geq T_{\mu^k}^{(\lambda)}J^* = (1-\lambda)\sum_{t=0}^{\infty}\lambda^t T^{t+1}J^* = J^*,$$

[cf. the induction step of Eq. (4.21)]. By combining the preceding two relations, we obtain Eq. (4.22), and the proof is completed by using the argument following that equation. **Q.E.D.**

The λ-PI algorithm has a useful property, which involves the mapping $W_k : \mathcal{R}(X) \mapsto \mathcal{R}(X)$ given by

$$W_k J = (1-\lambda)T_{\mu^k}J_k + \lambda\, T_{\mu^k}J. \tag{4.26}$$

In particular J_{k+1} *is a fixed point of* W_k. Indeed, using the definition

$$J_{k+1} = T_{\mu^k}^{(\lambda)} J_k$$

[cf. Eq. (4.24)], and the linearity assumption (4.25), we have

$$\begin{aligned} W_k J_{k+1} &= (1-\lambda) T_{\mu^k} J_k + \lambda T_{\mu^k} \left(T_{\mu^k}^{(\lambda)} J_k \right) \\ &= (1-\lambda) T_{\mu^k} J_k + \lambda T_{\mu^k}^{(\lambda)} (T_{\mu^k} J_k) \\ &= T_{\mu^k}^{(\lambda)} J_k \\ &= J_{k+1}. \end{aligned}$$

Thus J_{k+1} can be calculated as a fixed point of W_k.

Consider now the case where T_{μ^k} is nonexpansive with respect to some norm. Then from Eq. (4.26), it is seen that W_k is a contraction of modulus λ with respect to that norm, so J_{k+1} is the unique fixed point of W_k. Moreover, if the norm is a weighted sup-norm, J_{k+1} can be found using the methods of Chapter 2 for contractive models. The following example applies this idea to finite-state SSP problems. The interesting aspect of this example is that it implements the policy evaluation portion of λ-PI through solution of a system of linear equations, similar to the exact policy evaluation method of classical PI.

Example 4.3.4 (Stochastic Shortest Path Problems with Nonpositive Costs)

Consider the SSP problem of Example 1.2.6 with states $1, \ldots, n$, plus the termination state 0. For all $u \in U(x)$, the state following x is y with probability $p_{xy}(u)$ and the expected cost incurred is nonpositive. This problem arises when we wish to maximize nonnegative rewards up to termination. It includes a classical search problem where the aim, roughly speaking, is to move through the state space looking for states with favorable termination rewards.

We view the problem within our abstract framework with $\bar{J}(x) \equiv 0$ and

$$T_\mu J = g_\mu + P_\mu J, \tag{4.27}$$

with $g_\mu \in \Re^n$ being the corresponding nonpositive one-stage cost vector, and P_μ being an $n \times n$ substochastic matrix. The components of P_μ are the probabilities $p_{xy}\big(\mu(x)\big)$, $x, y = 1, \ldots, n$. Clearly Assumption D holds.

Consider the λ-PI method (4.24), with J_{k+1} computed by solving the fixed point equation $J = W_k J$, cf. Eq. (4.26). This is a nonsingular n-dimensional system of linear equations, and can be solved by matrix inversion, just like in exact PI for discounted n-state MDP. In particular, using Eqs. (4.26) and (4.27), we have

$$J_{k+1} = (I - \lambda P_{\mu^k})^{-1} \big(g_{\mu^k} + (1-\lambda) P_{\mu^k} J_k \big). \tag{4.28}$$

For a small number of states n, this matrix inversion-based policy evaluation may be simpler than the optimistic PI policy evaluation equation

$$J_{k+1} = T_{\mu^k}^{m_k} J_k$$

[cf. Eq. (4.20)], which points to an advantage of λ-PI.

Note that based on the relation between the multistep mapping $T_\mu^{(\lambda)}$ and the proximal mapping, discussed in Section 1.2.5 and Exercise 1.2, the policy evaluation Eq. (4.28) may be viewed as an extrapolated proximal iteration. Note also that as $\lambda \to 1$, the policy evaluation Eq. (4.28) resembles the policy evaluation equation

$$J_{\mu^k} = (I - \lambda P_{\mu^k})^{-1} g_{\mu^k}$$

for λ-discounted n-state MDP. An important difference, however, is that for a discounted finite-state MDP, exact PI will find an optimal policy in a finite number of iterations, while this is not guaranteed for λ-PI. Indeed λ-PI does not require that there exists an optimal policy or even that $J^*(x)$ is finite for all x.

Policy Iteration Under I

Contrary to the case of Assumption D, the important cost improvement property of PI holds under Assumption I. Thus, if μ is a policy and $\bar{\mu}$ satisfies the policy improvement equation $T_{\bar{\mu}} J_\mu = T J_\mu$, we have

$$J_\mu = T_\mu J_\mu \geq T J_\mu = T_{\bar{\mu}} J_\mu,$$

from which we obtain

$$J_\mu \geq \lim_{k \to \infty} T_{\bar{\mu}}^k J_\mu.$$

Since $J_\mu \geq \bar{J}$ and $J_{\bar{\mu}} = \lim_{k \to \infty} T_{\bar{\mu}}^k \bar{J}$, it follows that

$$J_\mu \geq T J_\mu \geq J_{\bar{\mu}}. \tag{4.29}$$

However, this cost improvement property is not by itself sufficient for the validity of PI under Assumption I (see the deterministic shortest path example of Section 3.1.1). Thus additional conditions are needed to guarantee convergence. To this end we may use the semicontractive framework of Chapter 3, and take advantage of the fact that under Assumption I, J^* is known to be a fixed point of T.

In particular, suppose that we have a set $S \subset \mathcal{E}(X)$ such that $J_S^* = J^*$. Then J_S^* is a fixed point of T and the theory of Section 3.2 comes into play. Thus, by Prop. 3.2.1 the following hold:

(a) We have $T^k J \to J^*$ for every $J \in \mathcal{E}(X)$ such that $J^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$.

(b) J^* is the only fixed point of T within the set of all $J \in \mathcal{E}(X)$ such that $J^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$.

Moreover, by Prop. 3.2.4, if S has the weak PI property and for each sequence $\{J_m\} \subset \mathcal{E}(X)$ with $J_m \downarrow J$ for some $J \in \mathcal{E}(X)$, we have

$$H(x,u,J) = \lim_{m\to\infty} H(x,u,J_m),$$

then every sequence of S-regular policies $\{\mu^k\}$ that can be generated by PI satisfies $J_{\mu^k} \downarrow J^*$. If in addition the set of S-regular policies is finite, there exists $\bar{k} \geq 0$ such that $\mu^{\bar{k}}$ is optimal.

For these properties to hold, it is of course critical that $J_S^* = J^*$. If this is not so, but J_S^* is still a fixed point of T, the VI and PI algorithms may converge to J_S^* rather than to J^* (cf. the linear quadratic problem of Section 3.5.4).

4.4 REGULARITY AND NONSTATIONARY POLICIES

In this section, we will extend the notion of regularity of Section 3.2 so that it applies more broadly. We will use this notion as our main tool for exploring the structure of the solution set of Bellman's equation. We will then discuss some applications involving mostly monotone increasing models in this section, as well as in Sections 4.5 and 4.6. We continue to focus on the infinite horizon case of the problem of Section 4.1, but we do not impose for the moment any additional assumptions, such as Assumption I or D.

We begin with the following extension of the definition of S-regularity, which we will use to prove a general result regarding the convergence properties of VI in the following Prop. 4.4.1. We will apply this result in the context of various applications in Sections 4.4.2-4.4.4, as well as in Sections 4.5 and 4.6.

Definition 4.4.1: For a nonempty set of functions $S \subset \mathcal{E}(X)$, we say that a nonempty collection $\mathcal{C}$ of policy-state pairs (π, x), with $\pi \in \Pi$ and $x \in X$, is *S-regular* if

$$J_\pi(x) = \limsup_{k\to\infty} (T_{\mu_0} \cdots T_{\mu_k} J)(x), \qquad \forall\ (\pi,x) \in \mathcal{C},\ J \in S.$$

The essence of the preceding definition of S-regularity is similar to the one of Chapter 3 for stationary policies: *for an S-regular collection of pairs (π, x), the value of $J_\pi(x)$ is not affected if the starting function is changed from $\bar{J}$ to any $J \in S$.* It is important to extend the definition of regularity to nonstationary policies because in noncontractive models, stationary policies are generally not sufficient, i.e., the optimal cost over

stationary policies may not be the same as the one over nonstationary policies (cf. Prop. 4.3.2, and the subsequent example). Generally, when referring to an S-regular collection $\mathcal{C}$, we implicitly assume that S and $\mathcal{C}$ are nonempty, although on occasion we may state explicitly this fact for emphasis.

For a given set $\mathcal{C}$ of policy-state pairs (π, x), let us consider the function $J^*_{\mathcal{C}} \in \mathcal{E}(X)$, given by

$$J^*_{\mathcal{C}}(x) = \inf_{\{\pi \mid (\pi,x)\in\mathcal{C}\}} J_\pi(x), \qquad x \in X.$$

Note that $J^*_{\mathcal{C}}(x) \geq J^*(x)$ for all $x \in X$ [for those $x \in X$ for which the set of policies $\{\pi \mid (\pi, x) \in \mathcal{C}\}$ is empty, we have by convention $J^*_{\mathcal{C}}(x) = \infty$].

For an important example, note that in the analysis of Chapter 3, the set of S-regular policies $\mathcal{M}_S$ of Section 3.2 defines the S-regular collection

$$\mathcal{C} = \big\{(\mu, x) \mid \mu \in \mathcal{M}_S,\ x \in X\big\},$$

and the corresponding restricted optimal cost function J^*_S is equal to $J^*_{\mathcal{C}}$. In Sections 3.2-3.4 we saw that when J^*_S is a fixed point of T, then favorable results are obtained. Similarly, in this section we will see that for an S-regular collection $\mathcal{C}$, when $J^*_{\mathcal{C}}$ is a fixed point of T, interesting results are obtained.

The following two propositions play a central role in our analysis on this section and the next two, and may be compared with Prop. 3.2.1, which played a pivotal role in the analysis of Chapter 3.

Proposition 4.4.1: (Well-Behaved Region Theorem) Given a nonempty set $S \subset \mathcal{E}(X)$, let $\mathcal{C}$ be a nonempty collection of policy-state pairs (π, x) that is S-regular. Then:

(a) For all $J \in \mathcal{E}(X)$ such that $J \leq \tilde{J}$ for some $\tilde{J} \in S$, we have

$$\limsup_{k\to\infty} T^k J \leq J^*_{\mathcal{C}}.$$

(b) For all $J' \in \mathcal{E}(X)$ with $J' \leq TJ'$, and all $J \in \mathcal{E}(X)$ such that $J' \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$, we have

$$J' \leq \liminf_{k\to\infty} T^k J \leq \limsup_{k\to\infty} T^k J \leq J^*_{\mathcal{C}}.$$

Proof: (a) Using the generic relation $TJ \leq T_\mu J$, $\mu \in \mathcal{M}$, and the monotonicity of T and T_μ, we have for all k

$$(T^k \tilde{J})(x) \leq (T_{\mu_0} \cdots T_{\mu_{k-1}} \tilde{J})(x), \qquad \forall\ (\pi, x) \in \mathcal{C},\ \tilde{J} \in S.$$

By letting $k \to \infty$ and by using the definition of S-regularity, it follows that for all $(\pi, x) \in \mathcal{C}$, $J \in \mathcal{E}(X)$, and $\tilde{J} \in S$ with $J \leq \tilde{J}$,

$$\limsup_{k\to\infty}(T^k J)(x) \leq \limsup_{k\to\infty}(T^k \tilde{J})(x) \leq \limsup_{k\to\infty}(T_{\mu_0}\cdots T_{\mu_{k-1}}\tilde{J})(x) = J_\pi(x),$$

and by taking infimum of the right side over $\{\pi \mid (\pi, x) \in \mathcal{C}\}$, we obtain the result.

(b) Using the hypotheses $J' \leq TJ'$, and $J' \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$, and the monotonicity of T, we have

$$J'(x) \leq (TJ')(x) \leq \cdots \leq (T^k J')(x) \leq (T^k J)(x).$$

Letting $k \to \infty$ and using part (a), we obtain the result. **Q.E.D.**

Let us discuss some interesting implications of part (b) of the proposition. Suppose we are given a set $S \subset \mathcal{E}(X)$, and a collection $\mathcal{C}$ that is S-regular. Then:

(1) $J^*_{\mathcal{C}}$ is an upper bound to every fixed point J' of T that lies below some $\tilde{J} \in S$ (i.e., $J' \leq \tilde{J}$). Moreover, for such a fixed point J', the VI algorithm, starting from any J with $J^*_{\mathcal{C}} \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$, ends up asymptotically within the region

$$\{J \in \mathcal{E}(X) \mid J' \leq J \leq J^*_{\mathcal{C}}\}.$$

Thus the convergence of VI is characterized by the *well-behaved region*

$$\mathcal{W}_{S,\mathcal{C}} = \{J \in \mathcal{E}(X) \mid J^*_{\mathcal{C}} \leq J \leq \tilde{J} \text{ for some } \tilde{J} \in S\}, \tag{4.30}$$

(cf. the corresponding definition in Section 3.2), and the *limit region*

$$\{J \in \mathcal{E}(X) \mid J' \leq J \leq J^*_{\mathcal{C}} \text{ for all fixed points } J' \text{ of } T \text{ with } J' \leq \tilde{J} \text{ for some } \tilde{J} \in S\}.$$

The VI algorithm, starting from the former, ends up asymptotically within the latter; cf. Figs. 4.4.1 and 4.4.2.

(2) If $J^*_{\mathcal{C}}$ is a fixed point of T (a common case in our subsequent analysis), then the VI-generated sequence $\{T^k J\}$ converges to $J^*_{\mathcal{C}}$ starting from any J in the well-behaved region. If $J^*_{\mathcal{C}}$ is not a fixed point of T, we only have $\limsup_{k\to\infty} T^k J \leq J^*_{\mathcal{C}}$ for all J in the well-behaved region.

(3) If the well-behaved region is unbounded above in the sense that $\mathcal{W}_{S,\mathcal{C}} = \{J \in \mathcal{E}(X) \mid J^*_{\mathcal{C}} \leq J\}$, which is true for example if $S = E(X)$, then $J' \leq J^*_{\mathcal{C}}$ for every fixed point J' of T. The reason is that for every fixed point J' of T we have $J' \leq J$ for some $J \in \mathcal{W}_{S,\mathcal{C}}$, and hence also $J' \leq \tilde{J}$ for some $\tilde{J} \in S$, so observation (1) above applies.

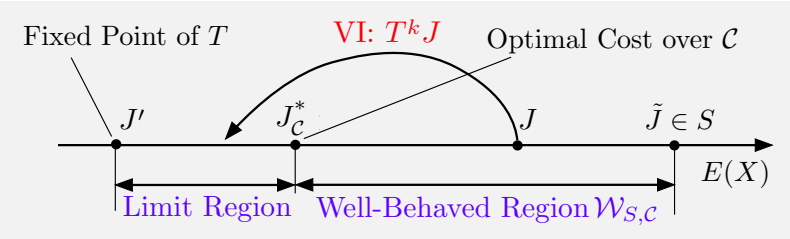


Figure 4.4.1. Schematic illustration of Prop. 4.4.1. Neither $J^*_{\mathcal{C}}$ nor J^* need to be fixed points of T, but if $\mathcal{C}$ is S-regular, and there exists $\tilde{J} \in S$ with $J^*_{\mathcal{C}} \leq \tilde{J}$, then $J^*_{\mathcal{C}}$ demarcates from above the range of fixed points of T that lie below $\tilde{J}$.

For future reference, we state these observations as a proposition, which should be compared to Prop. 3.2.1, the stationary special case where $\mathcal{C}$ is defined by the set of S-regular stationary policies, i.e., $\mathcal{C} = \big\{(\mu, x) \mid \mu \in \mathcal{M}_S,\ x \in X\big\}$. Figures 4.4.2 and 4.4.3 illustrate some of the consequences of Prop. 4.4.1 for two cases, respectively: when $S = E(X)$ while $J^*_{\mathcal{C}}$ is not a fixed point of T, and when S is a strict subset of $E(X)$ while $J^*_{\mathcal{C}}$ is a fixed point of T.

Proposition 4.4.2: (Uniqueness of Fixed Point of T and Convergence of VI) Given a set $S \subset \mathcal{E}(X)$, let $\mathcal{C}$ be a collection of policy-state pairs (π, x) that is S-regular. Then:

(a) If J' is a fixed point of T with $J' \leq \tilde{J}$ for some $\tilde{J} \in S$, then $J' \leq J^*_{\mathcal{C}}$. Moreover, $J^*_{\mathcal{C}}$ is the only possible fixed point of T within $\mathcal{W}_{S,\mathcal{C}}$.

(b) We have $\limsup_{k\to\infty} T^kJ \leq J^*_{\mathcal{C}}$ for all $J \in \mathcal{W}_{S,\mathcal{C}}$, and if $J^*_{\mathcal{C}}$ is a fixed point of T, then $T^kJ \to J^*_{\mathcal{C}}$ for all $J \in \mathcal{W}_{S,\mathcal{C}}$.

(c) If $\mathcal{W}_{S,\mathcal{C}}$ is unbounded from above in the sense that

$$\mathcal{W}_{S,\mathcal{C}} = \big\{J \in \mathcal{E}(X) \mid J^*_{\mathcal{C}} \leq J\big\},$$

then $J' \leq J^*_{\mathcal{C}}$ for every fixed point J' of T. In particular, if $J^*_{\mathcal{C}}$ is a fixed point of T, then $J^*_{\mathcal{C}}$ is the largest fixed point of T.

Proof: (a) The first statement follows from Prop. 4.4.1(b). For the second statement, let J' be a fixed point of T with $J' \in \mathcal{W}_{S,\mathcal{C}}$. Then from the definition of $\mathcal{W}_{S,\mathcal{C}}$, we have $J^*_{\mathcal{C}} \leq J'$ as well as $J' \leq \tilde{J}$ for some $\tilde{J} \in S$, so from Prop. 4.4.1(b) it follows that $J' \leq J^*_{\mathcal{C}}$. Hence $J' = J^*_{\mathcal{C}}$.

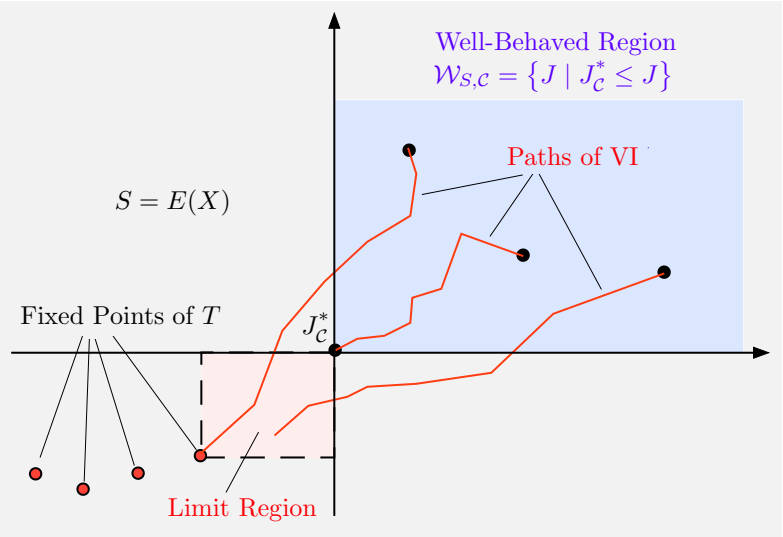


Figure 4.4.2. Schematic illustration of Prop. 4.4.2, for the case where $S = E(X)$ so that $\mathcal{W}_{S,\mathcal{C}}$ is unbounded above, i.e., $\mathcal{W}_{S,\mathcal{C}} = \{J \in E(X) \mid J^*_{\mathcal{C}} \leq J\}$. In this figure $J^*_{\mathcal{C}}$ is not a fixed point of T. The VI algorithm, starting from the well-behaved region $\mathcal{W}_{S,\mathcal{C}}$, ends up asymptotically within the limit region.

(b) The result follows from Prop. 4.4.1(a), and in the case where $J^*_{\mathcal{C}}$ is a fixed point of T, from Prop. 4.4.1(b), with $J' = J^*_{\mathcal{C}}$.

(c) See observation (3) in the discussion preceding the proposition. **Q.E.D.**

Examples and counterexamples illustrating the preceding proposition are provided by the problems of Section 3.1 for the stationary case where

$$\mathcal{C} = \{(\mu, x) \mid \mu \in \mathcal{M}_S,\ x \in X\}.$$

Similar to the analysis of Chapter 3, the preceding proposition takes special significance when J^* is a fixed point of T and $\mathcal{C}$ is rich enough so that $J^*_{\mathcal{C}} = J^*$, as for example in the case where $\mathcal{C}$ is the set $\Pi \times X$ of all (π, x), or other choices to be discussed later. It then follows that every fixed point J' of T that belongs to S satisfies $J' \leq J^*$, and that VI converges to J^* starting from any $J \in \mathcal{E}(X)$ such that $J^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$. However, there will be interesting cases where $J^*_{\mathcal{C}} \neq J^*$, as in shortest path-type problems (see Sections 3.5.1, 4.5, and 4.6).

Note that Prop. 4.4.2 does not say anything about fixed points of T that lie below $J^*_{\mathcal{C}}$, and does not give conditions under which $J^*_{\mathcal{C}}$ is a fixed point. Moreover, it does not address the question whether J^* is a fixed point of T, or whether VI converges to J^* starting from $\bar{J}$ or from below J^*. Generally, it can happen that both, only one, or none of the two

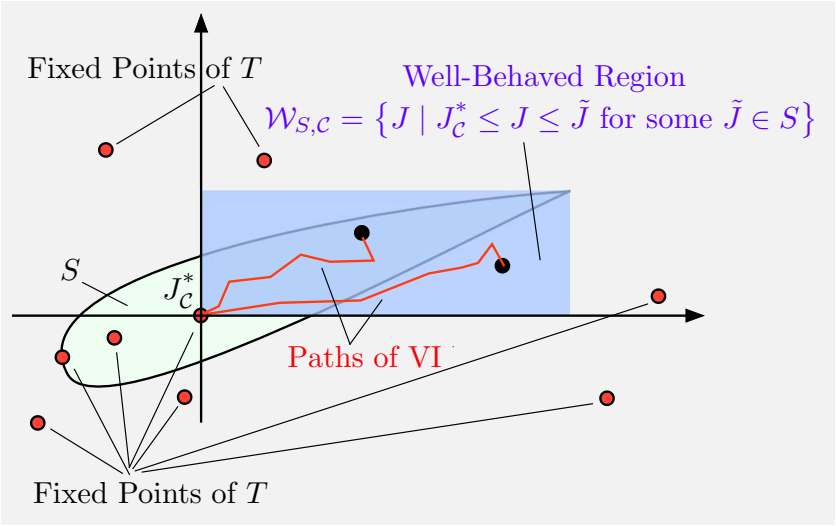


Figure 4.4.3. Schematic illustration of Prop. 4.4.2, and the set $\mathcal{W}_{S,\mathcal{C}}$ of Eq. (4.30), for a case where $J^*_{\mathcal{C}}$ is a fixed point of T and S is a strict subset of $E(X)$. Every fixed point of T that lies below some $\tilde{J} \in S$ should lie below $J^*_{\mathcal{C}}$. Also, the VI algorithm converges to $J^*_{\mathcal{C}}$ starting from within $\mathcal{W}_{S,\mathcal{C}}$. If S were unbounded from above, as in Fig. 4.4.2, $J^*_{\mathcal{C}}$ would be the largest fixed point of T.

functions $J^*_{\mathcal{C}}$ and J^* is a fixed point of T, as can be seen from the examples of Section 3.1.

The Case Where $J^*_{\mathcal{C}} \leq \bar{J}$

We have seen in Section 4.3 that the results for monotone increasing and monotone decreasing models are markedly different. In the context of S-regularity of a collection $\mathcal{C}$, it turns out that there are analogous significant differences between the cases $J^*_{\mathcal{C}} \geq \bar{J}$ and $J^*_{\mathcal{C}} \leq \bar{J}$. The following proposition establishes some favorable aspects of the condition $J^*_{\mathcal{C}} \leq \bar{J}$ in the context of VI. These can be attributed to the fact that $\bar{J}$ can always be added to S without affecting the S-regularity of $\mathcal{C}$, so $\bar{J}$ can serve as the element $\tilde{J}$ of S in Props. 4.4.1 and 4.4.2 (see the subsequent proof). The following proposition may also be compared with the result on convergence of VI under Assumption D (cf. Prop. 4.3.13).

Proposition 4.4.3: Given a set $S \subset \mathcal{E}(X)$, let $\mathcal{C}$ be a collection of policy-state pairs (π, x) that is S-regular, and assume that $J^*_{\mathcal{C}} \leq \bar{J}$. Then:

(a) For all $J' \in \mathcal{E}(X)$ with $J' \leq TJ'$, we have

$$J' \leq \liminf_{k\to\infty} T^k \bar{J} \leq \limsup_{k\to\infty} T^k \bar{J} \leq J^*_{\mathcal{C}}.$$

(b) If $J^*_{\mathcal{C}}$ is a fixed point of T, then $J^*_{\mathcal{C}} = J^*$ and we have $T^k\bar{J} \to J^*$ as well as $T^k J \to J^*$ for every $J \in \mathcal{E}(X)$ such that $J^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$.

Proof: (a) If S does not contain $\bar{J}$, we can replace S with $\bar{S} = S \cup \{\bar{J}\}$, and $\mathcal{C}$ will still be $\bar{S}$-regular. By applying Prop. 4.4.1(b) with S replaced by $\bar{S}$ and $\tilde{J} = \bar{J}$, the result follows.

(b) Assume without loss of generality that $\bar{J} \in S$ [cf. the proof of part (a)]. By using Prop. 4.4.2(b) with $\tilde{J} = \bar{J}$, we have $J^*_{\mathcal{C}} = \lim_{k\to\infty} T^k\bar{J}$. Thus for every policy $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi$,

$$J^*_{\mathcal{C}} = \lim_{k\to\infty} T^k \bar{J} \leq \limsup_{k\to\infty} T_{\mu_0} \cdots T_{\mu_{k-1}} \bar{J} = J_\pi,$$

so by taking the infimum over $\pi \in \Pi$, we obtain $J^*_{\mathcal{C}} \leq J^*$. Since generically $J^*_{\mathcal{C}} \geq J^*$, it follows that $J^*_{\mathcal{C}} = J^*$. Finally, from Prop. 4.4.2(b), $T^k J \to J^*$ for all $J \in \mathcal{W}_{S,C}$, implying the result. **Q.E.D.**

As a special case of the preceding proposition, we have that if $J^* \leq \bar{J}$ and J^* is a fixed point of T, then $J^* = \lim_{k\to\infty} T^k\bar{J}$, and for every other fixed point J' of T we have $J' \leq J^*$ (apply the proposition with $\mathcal{C} = \Pi \times X$ and $S = \{\bar{J}\}$, in which case $J^*_{\mathcal{C}} = J^* \leq \bar{J}$). This occurs, among others, in the monotone decreasing models, where $T_\mu \bar{J} \leq \bar{J}$ for all $\mu \in \mathcal{M}$. A special case is the convergence of VI under Assumption D (cf. Prop. 4.3.5).

The preceding proposition also applies to a classical type of search problem with both positive and negative costs per stage. This is the SSP problem, where at each $x \in X$ we have cost $E\{g(x,u,w)\} \geq 0$ for all u except one that leads to a termination state with probability 1 and nonpositive cost; here $\bar{J}(x) = 0$ and $J^*_{\mathcal{C}}(x) \leq 0$ for all $x \in X$, but Assumption D need not hold.

4.4.1 Regularity and Monotone Increasing Models

We will now return to the monotone increasing model, cf. Assumption I. For this model, we know from Section 4.3 that J^* is the smallest fixed point of T within the class of functions $J \geq \bar{J}$, under certain relatively mild assumptions. However, VI may not converge to J^* starting from below J^* (e.g., starting from $\bar{J}$), and also starting from above J^*. In this section

we will address the question of convergence of VI from above J^* by using regularity ideas, and in Section 4.5 we will consider the characterization of the largest fixed point of T in the context of deterministic optimal control and infinite-space shortest path problems. We summarize the results of Section 4.3 that are relevant to our development in the following proposition (cf. Props. 4.3.2, 4.3.3, 4.3.9, and 4.3.14).

Proposition 4.4.4: Let Assumption I hold. Then:

(a) $J^* = TJ^*$, and if $J' \in \mathcal{E}(X)$ is such that $J' \geq \bar{J}$ and $J' \geq TJ'$, then $J' \geq J^*$.

(b) For all $\mu \in \mathcal{M}$ we have $J_\mu = T_\mu J_\mu$, and if $J' \in \mathcal{E}(X)$ is such that $J' \geq \bar{J}$ and $J' \geq T_\mu J'$, then $J' \geq J_\mu$.

(c) $\mu^* \in \mathcal{M}$ is optimal if and only if $T_{\mu^*} J^* = TJ^*$.

(d) If U is a metric space and the sets

$$U_k(x,\lambda) = \big\{u \in U(x) \big|\ H(x,u,T^k\bar{J}) \leq \lambda\big\}$$

are compact for all $x \in X$, $\lambda \in \Re$, and k, then there exists at least one optimal stationary policy, and we have $T^k J \to J^*$ for all $J \in \mathcal{E}(X)$ with $J \leq J^*$.

(e) Given any $\epsilon > 0$, there exists a policy $\pi_\epsilon \in \Pi$ such that

$$J^* \leq J_{\pi_\epsilon} \leq J^* + \epsilon\, e.$$

Furthermore, if the scalar α in part (c) of Assumption I satisfies $\alpha < 1$, the policy π_ϵ can be taken to be stationary.

Since under Assumption I there may exist fixed points J' of T with $J^* \leq J'$, VI may not converge to J^* starting from above J^*. However, convergence of VI to J^* from above, if it occurs, is often much faster than convergence from below, so starting points $J \geq J^*$ may be desirable. One well-known such case is deterministic finite-state shortest path problems where major algorithms, such as the Bellman-Ford method or other label correcting methods have polynomial complexity, when started from J above J^*, but only pseudopolynomial complexity when started from J below J^* [see e.g., [BeT89] (Prop. 1.2 in Ch.4), [Ber98] (Exercise 2.7)].

In the next two subsections, we will consider discounted and undiscounted optimal control problems with nonnegative cost per stage, and we will establish conditions under which J^* is the unique nonnegative fixed point of T, and VI converges to J^* from above. Our analysis will proceed as follows:

(a) Define a collection $\mathcal{C}$ such that $J^*_{\mathcal{C}} = J^*$.

(b) Define a set $S \subset \mathcal{E}^+(X)$ such that $J^* \in S$ and $\mathcal{C}$ is S-regular.

(c) Use Prop. 4.4.2 (which shows that $J^*_{\mathcal{C}}$ is the largest fixed point of T within S) in conjunction with Prop. 4.4.4(a) (which shows that J^* is the smallest fixed point of T within S) to show that J^* is the unique fixed point of T within S. Use also Prop. 4.4.2(b) to show that the VI algorithm converges to J^* starting from $J \in S$ such that $J \geq J^*$.

(d) Use the compactness condition of Prop. 4.4.4(d), to enlarge the set of functions starting from which VI converges to J^*.

4.4.2 Nonnegative Cost Stochastic Optimal Control

Let us consider the undiscounted stochastic optimal control problem that involves the mapping

$$H(x,u,J) = E\big\{g(x,u,w) + J\big(f(x,u,w)\big)\big\}, \tag{4.31}$$

where g is the one-stage cost function and f is the system function. The expected value is taken with respect to the distribution of the random variable w (which takes values in a countable set W). We assume that

$$0 \leq E\big\{g(x,u,w)\big\} < \infty, \qquad \forall\ x \in X,\ u \in U(x),\ w \in W.$$

We consider the abstract DP model with H as above, and with $\bar{J}(x) \equiv 0$. Using the nonnegativity of g, we can write the cost function of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ in terms of a limit,

$$J_\pi(x_0) = \lim_{k\to\infty} E^\pi_{x_0}\left\{\sum_{m=0}^{k} g\big(x_m, \mu_m(x_m), w_m\big)\right\}, \qquad x_0 \in X, \tag{4.32}$$

where $E^\pi_{x_0}\{\cdot\}$ denotes expected value with respect to the probability distribution induced by π under initial state x_0.

We will apply the analysis of this section with

$$\mathcal{C} = \big\{(\pi, x) \mid J_\pi(x) < \infty\big\},$$

for which $J^*_{\mathcal{C}} = J^*$. We assume that $\mathcal{C}$ is nonempty, which is true if and only if J^* is not identically ∞, i.e., $J^*(x) < \infty$ for some $x \in X$. Consider the set

$$S = \Big\{J \in \mathcal{E}^+(X) \mid E^\pi_{x_0}\big\{J(x_k)\big\} \to 0,\ \forall\ (\pi, x_0) \in \mathcal{C}\Big\}. \tag{4.33}$$

One interpretation is that the functions J that are in S have the character of Lyapounov functions for the policies π for which the set $\big\{x_0 \mid J_\pi(x_0) < \infty\big\}$ is nonempty.

Note that S *is the largest set with respect to which* $\mathcal{C}$ *is regular* in the sense that $\mathcal{C}$ is S-regular and if $\mathcal{C}$ is S'-regular for some other set S', then $S' \subset S$. To see this we write for all $J \in \mathcal{E}^+(X)$, $(\pi, x_0) \in \mathcal{C}$, and k,

$$(T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x_0) = E^{\pi}_{x_0}\big\{J(x_k)\big\} + E^{\pi}_{x_0}\left\{\sum_{m=0}^{k-1} g\big(x_m, \mu_m(x_m), w_m\big)\right\},$$

where μ_m, $m = 0, 1, \ldots$, denote generically the components of π. The rightmost term above converges to $J_\pi(x_0)$ as $k \to \infty$ [cf. Eq. (4.32)], so by taking upper limit, we obtain

$$\limsup_{k\to\infty} (T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x_0) = \limsup_{k\to\infty} E^{\pi}_{x_0}\big\{J(x_k)\big\} + J_\pi(x_0). \tag{4.34}$$

In view of the definition (4.33) of S, this implies that for all $J \in S$, we have

$$\limsup_{k\to\infty} (T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x_0) = J_\pi(x_0), \qquad \forall\ (\pi, x_0) \in \mathcal{C}, \tag{4.35}$$

so $\mathcal{C}$ is S-regular. Moreover, if $\mathcal{C}$ is S'-regular and $J \in S'$, Eq. (4.35) holds, so that [in view of Eq. (4.34) and $J \in \mathcal{E}^+(X)$] $\lim_{k\to\infty} E^{\pi}_{x_0}\big\{J(x_k)\big\} = 0$ for all $(\pi, x_0) \in \mathcal{C}$, implying that $J \in S$.

From Prop. 4.4.2, the fixed point property of J^* [cf. Prop. 4.4.4(a)], and the fact $J^*_{\mathcal{C}} = J^*$, it follows that $T^k J \to J^*$ for all $J \in S$ that satisfy $J \geq J^*$. Moreover, if the sets $U_k(x, \lambda)$ of Eq. (4.17) are compact, the convergence of VI starting from below J^* will also be guaranteed. We thus have the following proposition, which in addition shows that J^* belongs to S and is the unique fixed point of T within S.

Proposition 4.4.5: (Uniqueness of Fixed Point of T and Convergence of VI) Consider the problem corresponding to the mapping (4.31) with $g \geq 0$, and assume that J^* is not identically ∞. Then:

(a) J^* belongs to S and is the unique fixed point of T within S. Moreover, we have $T^k J \to J^*$ for all $J \geq J^*$ with $J \in S$.

(b) If U is a metric space, and the sets $U_k(x, \lambda)$ of Eq. (4.17) are compact for all $x \in X$, $\lambda \in \Re$, and k, we have $T^k J \to J^*$ for all $J \in S$, and an optimal stationary policy is guaranteed to exist.

Proof: (a) We first show that $J^* \in S$. Given a policy $\pi = \{\mu_0, \mu_1, \ldots\}$, we denote by π_k the policy

$$\pi_k = \{\mu_k, \mu_{k+1}, \ldots\}.$$

We have for all $(\pi, x_0) \in \mathcal{C}$

$$J_\pi(x_0) = E_{x_0}^\pi \big\{ g\big(x_0, \mu_0(x_0), w_0\big) \big\} + E_{x_0}^\pi \big\{ J_{\pi_1}(x_1) \big\}, \tag{4.36}$$

and for all $m = 1, 2, \ldots$,

$$E_{x_0}^\pi \big\{ J_{\pi_m}(x_m) \big\} = E_{x_0}^\pi \big\{ g\big(x_m, \mu_m(x_m), w_m\big) \big\} + E_{x_0}^\pi \big\{ J_{\pi_{m+1}}(x_{m+1}) \big\}, \tag{4.37}$$

where $\{x_m\}$ is the sequence generated starting from x_0 and using π. By using repeatedly the expression (4.37) for $m = 1, \ldots, k-1$, and combining it with Eq. (4.36), we obtain for all $k = 1, 2, \ldots$,

$$J_\pi(x_0) = E_{x_0}^\pi \big\{ J_{\pi_k}(x_k) \big\} + \sum_{m=0}^{k-1} E_{x_0}^\pi \Big\{ g\big(x_m, \mu_m(x_m), w_m\big) \Big\}, \quad \forall\ (\pi, x_0) \in \mathcal{C}.$$

The rightmost term above tends to $J_\pi(x_0)$ as $k \to \infty$, so we obtain

$$E_{x_0}^\pi \big\{ J_{\pi_k}(x_k) \big\} \to 0, \qquad \forall\ (\pi, x_0) \in \mathcal{C}.$$

Since $0 \leq J^* \leq J_{\pi_k}$, it follows that

$$E_{x_0}^\pi \big\{ J^*(x_k) \big\} \to 0, \qquad \forall\ x_0 \text{ with } J^*(x_0) < \infty.$$

Thus $J^* \in S$ while J^* (which is equal to $J_{\mathcal{C}}^*$) is a fixed point of T. For every other fixed point J' of T, we have $J' \geq J^*$ [by Prop. 4.4.4(b)], so if J' belongs to S, by Prop. 4.4.2(a), $J' \leq J^*$ and thus $J' = J^*$. Hence, J^* is the unique fixed point of T within the set S. By Prop. 4.4.2(b), we also have $T^k J \to J^*$ for all $J \in S$ with $J \geq J^*$.

(b) This part follows from part (a) and Prop. 4.4.4(d). **Q.E.D.**

Note that under the assumptions of the preceding proposition, either T has a unique fixed point within $\mathcal{E}^+(X)$ (namely J^*), or else all the additional fixed points of T within $\mathcal{E}^+(X)$ lie outside S. To illustrate the limitations of this result, consider the shortest path problem of Section 3.1.1 for the case where the choice at state 1 is either to stay at 1 at cost 0, or move to the destination at cost $b > 0$. Then Bellman's equation at state 1 is $J(1) = \min\big\{b, J(1)\big\}$, and its set of nonnegative solutions is the interval $[0, b]$, while we have $J^* = 0$. The set S of Eq. (4.33) here consists of just J^* and Prop. 4.4.5 applies, but it is not very useful. Similarly, in the linear-quadratic example of Section 3.1.4, where T has the two fixed points $J^*(x) = 0$ and $\hat{J}(x) = (\gamma^2 - 1)x^2$, the set S of Eq. (4.33) consists of just J^*.

Thus the regularity framework of this section is useful primarily in the favorable case where J^* is the unique nonnegative fixed point of T. In particular, Prop. 4.4.5 cannot be used to differentiate between multiple

fixed points of T, and to explain the unusual behavior in the preceding two examples. In Sections 4.5 and 4.6, we address this issue within the more restricted contexts of deterministic and stochastic optimal control, respectively.

A consequence of Prop. 4.4.5 is the following condition for VI convergence from above, first discovered and published in the paper by Yu and Bertsekas [YuB15] (Theorem 5.1) within a broader context that also addressed universal measurability issues.

Proposition 4.4.6: Under the conditions of Prop. 4.4.5, we have $T^k J \to J^*$ for all $J \in \mathcal{E}^+(X)$ satisfying

$$J^* \leq J \leq cJ^*, \tag{4.38}$$

for some scalar $c > 1$. Moreover, J^* is the unique fixed point of T within the set

$$\big\{J \in \mathcal{E}^+(X) \mid J \leq cJ^* \text{ for some } c > 0\big\}.$$

Proof: Since $J^* \in S$ as shown in Prop. 4.4.5, any J satisfying Eq. (4.38), also belongs to the set S of Eq. (4.33), and the result follows from Prop. 4.4.5. **Q.E.D.**

Note a limitation of the preceding proposition: in order to find functions J satisfying $J^* \leq J \leq c\,J^*$ we must essentially know the sets of states x where $J^*(x) = 0$ and $J^*(x) = \infty$.

4.4.3 Discounted Stochastic Optimal Control

We will now consider a discounted version of the stochastic optimal control problem of the preceding section. For a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ we have

$$J_\pi(x_0) = \lim_{k\to\infty} E^\pi_{x_0}\left\{\sum_{m=0}^{k-1} \alpha^m g\big(x_m, \mu_m(x_m), w_m\big)\right\},$$

where $\alpha \in (0,1)$ is the discount factor, and as earlier $E^\pi_{x_0}\{\cdot\}$ denotes expected value with respect to the probability measure induced by $\pi \in \Pi$ under initial state x_0. We assume that the one-stage expected cost is nonnegative,

$$0 \leq E\big\{g(x,u,w)\big\} < \infty, \qquad \forall\ x \in X,\ u \in U(x),\ w \in W.$$

By defining the mapping H as

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

and $\bar{J}(x) \equiv 0$, we can view this problem within the abstract DP framework of this chapter where Assumption I holds.

Note that because of the discount factor, the existence of a terminal set of states is not essential for the optimal costs to be finite. Moreover, the nonnegativity of g is not essential for our analysis. Any problem where g can take both positive and negative values, but is bounded below, can be converted to an equivalent problem where g is nonnegative, by adding a suitable constant c to g. Then the cost of all policies will simply change by the constant $\sum_{k=0}^{\infty} \alpha^k c = c/(1-\alpha)$.

The line of analysis of this section makes a connection between the S-regularity notion of Definition 4.4.1 and a notion of stability, which is common in feedback control theory and will be explored further in Section 4.5. We assume that X is a normed space, so that boundedness within X is defined with respect to its norm. We introduce the set

$$X^* = \big\{x \in X \mid J^*(x) < \infty\big\},$$

which we assume to be nonempty. Given a state $x \in X^*$, we say that a policy π is *stable from* x if there exists a bounded subset of X^* [that depends on (π, x)] such that the (random) sequence $\{x_k\}$ generated starting from x and using π lies with probability 1 within that subset. We consider the set of policy-state pairs

$$\mathcal{C} = \big\{(\pi, x) \mid x \in X^*,\ \pi \text{ is stable from } x\big\},$$

and we assume that $\mathcal{C}$ is nonempty.

Let us say that a function $J \in \mathcal{E}^+(X)$ is *bounded on bounded subsets of* X^* if for every bounded subset $\tilde{X} \subset X^*$ there is a scalar b such that $J(x) \leq b$ for all $x \in \tilde{X}$. Let us also introduce the set

$$S = \big\{J \in \mathcal{E}^+(X) \mid J \text{ is bounded on bounded subsets of } X^*\big\}.$$

We assume that $\mathcal{C}$ is nonempty, $J^* \in S$, and for every $x \in X^*$ and $\epsilon > 0$, there exists a policy π that is stable from x and satisfies $J_\pi(x) \leq J^*(x) + \epsilon$ (thus implying that $J^*_{\mathcal{C}} = J^*$). We have the following proposition.

Proposition 4.4.7: Under the preceding assumptions, J^* is the unique fixed point of T within S, and we have $T^k J \to J^*$ for all $J \in S$ with $J^* \leq J$. If in addition U is a metric space, and the sets $U_k(x,\lambda)$ of Eq. (4.17) are compact for all $x \in X$, $\lambda \in \Re$, and k, we have $T^k J \to J^*$ for all $J \in S$, and an optimal stationary policy is guaranteed to exist.

Proof: We have for all $J \in \mathcal{E}(X)$, $(\pi, x_0) \in \mathcal{C}$, and k,

$$(T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x_0) = \alpha^k E^{\pi}_{x_0}\big\{J(x_k)\big\} + E^{\pi}_{x_0}\left\{\sum_{m=0}^{k-1} \alpha^m g\big(x_m, \mu_m(x_m), w_m\big)\right\}.$$

Since $(\pi, x_0) \in \mathcal{C}$, there is a bounded subset of X^* such that $\{x_k\}$ belongs to that subset with probability 1, so if $J \in S$ it follows that $\alpha^k E^{\pi}_{x_0}\big\{J(x_k)\big\} \to 0$. Thus by taking limit as $k \to \infty$ in the preceding relation, we have for all $(\pi, x_0) \in \mathcal{C}$ and $J \in S$,

$$\begin{aligned}\lim_{k\to\infty} (T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x_0) &= \lim_{k\to\infty} E^{\pi}_{x_0}\left\{\sum_{m=0}^{k-1} \alpha^m g\big(x_m, \mu_m(x_m), w_m\big)\right\} \\ &= J_\pi(x_0),\end{aligned}$$

so $\mathcal{C}$ is S-regular. Since $J^*_{\mathcal{C}}$ is equal to J^*, which is a fixed point of T, the result follows similar to the proof of Prop. 4.4.5. **Q.E.D.**

4.4.4 Convergent Models

In this section we consider a case of an abstract DP model that generalizes both the monotone increasing and the monotone decreasing models. The model is patterned after the stochastic optimal control problem of Example 1.2.1, where the cost per stage function g can take negative as well as positive values. Our main assumptions are that the cost functions of all policies are defined as limits (rather than upper limits), and that $-\infty < \bar{J}(x) \leq J^*(x)$ for all $x \in X$.

These conditions are somewhat restrictive and make the model more similar to the monotone increasing than to the monotone decreasing model, but are essential for the results of this section (for a discussion of the pathological behaviors that can occur without the condition $\bar{J} \leq J^*$, see the paper by H. Yu [Yu15]). We will show that J^* is a fixed point of T, and that there exists an ϵ-optimal policy for every $\epsilon > 0$. This will bring to bear the regularity ideas and results of Prop. 4.4.2, and will provide a convergence result for the VI algorithm.

In particular, we denote

$$\mathcal{E}_b(X) = \big\{J \in \mathcal{E}(X) \mid J(x) > -\infty,\ \forall\ x \in X\big\},$$

and we will assume the following.

Assumption 4.4.1:

(a) For all $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi$, J_π can be defined as a limit:

$$J_\pi(x) = \lim_{k\to\infty} (T_{\mu_0} \cdots T_{\mu_k} \bar{J})(x), \qquad \forall\ x \in X. \tag{4.39}$$

Furthermore, we have $\bar{J} \in \mathcal{E}_b(X)$ and

$$\bar{J} \leq J^*.$$

(b) For each sequence $\{J_m\} \subset \mathcal{E}_b(X)$ with $J_m \to J \in \mathcal{E}_b(X)$, we have

$$\lim_{m\to\infty} H(x,u,J_m) = H(x,u,J), \qquad \forall\ x \in X,\ u \in U(x).$$

(c) There exists $\alpha > 0$ such that for all $J \in \mathcal{E}_b(X)$ and $r \in \Re$,

$$H(x,u,J+re) \leq H(x,u,J) + \alpha r, \qquad \forall\ x \in X,\ u \in U(x),$$

where e is the unit function, $e(x) \equiv 1$.

For an example of a type of problem where the convergence condition (4.39) is satisfied, consider the stochastic optimal control problem of Example 1.2.1, assuming that the state space consists of two regions: X_1 where the cost per stage is nonnegative under all controls, and X_2 where the cost per stage is nonpositive. Assuming that once the system enters X_1 it can never return to X_2, the convergence condition (4.39) is satisfied for all π. The same is true for the reverse situation, where once the system enters X_2 it can never return to X_1. Optimal stopping problems and SSP problems are often of this type.

We first prove the existence of ϵ-optimal policies and then use it to establish that J^* is a fixed point of T. The proofs are patterned after the ones under Assumption I (cf. Props. 4.3.2 and 4.3.3).

Proposition 4.4.8: Let Assumption 4.4.1 hold. Given any $\epsilon > 0$, there exists a policy $\pi_\epsilon \in \Pi$ such that

$$J^* \leq J_{\pi_\epsilon} \leq J^* + \epsilon\, e.$$

Proof: Let $\{\epsilon_k\}$ be a sequence such that $\epsilon_k > 0$ for all k and

$$\sum_{k=0}^{\infty} \alpha^k \epsilon_k = \epsilon, \tag{4.40}$$

where α is the scalar of Assumption 4.4.1(c). For each $x \in X$, consider a sequence of policies $\{\pi_k[x]\} \subset \Pi$, with components of $\pi_k[x]$ (to emphasize

their dependence on x) denoted by $\mu_m^k[x]$, $m = 0, 1, \ldots$,

$$\pi_k[x] = \big\{\mu_0^k[x], \mu_1^k[x], \ldots\big\},$$

such that for $k = 0, 1, \ldots$,

$$J_{\pi_k[x]}(x) \leq J^*(x) + \epsilon_k. \tag{4.41}$$

Such a sequence exists since $J^* \in \mathcal{E}_b(X)$.

Consider the functions $\overline{\mu}_k$ defined by

$$\overline{\mu}_k(x) = \mu_0^k[x](x), \qquad \forall\ x \in X, \tag{4.42}$$

and the functions $\bar{J}_k$ defined by

$$\bar{J}_k(x) = H\left(x, \overline{\mu}_k(x), \lim_{m\to\infty} T_{\mu_1^k[x]} \cdots T_{\mu_m^k[x]} \bar{J}\right), \qquad \forall\ x \in X,\ k = 0, 1, \ldots. \tag{4.43}$$

By using Eqs. (4.41)-(4.43), and the continuity property of Assumption 4.4.1(b), we obtain for all $x \in X$ and $k = 0, 1, \ldots$,

$$\begin{aligned}
\bar{J}_k(x) &= H\left(x, \mu_0^k[x](x), \lim_{m\to\infty} T_{\mu_1^k[x]} \cdots T_{\mu_m^k[x]} \bar{J}\right) \\
&= \lim_{m\to\infty} H\left(x, \mu_0^k[x](x), T_{\mu_1^k[x]} \cdots T_{\mu_m^k[x]} \bar{J}\right) \\
&= \lim_{m\to\infty} \big(T_{\mu_0^k[x]} \cdots T_{\mu_m^k[x]} \bar{J}\big)(x) \\
&= J_{\pi_k[x]}(x) \\
&\leq J^*(x) + \epsilon_k.
\end{aligned} \tag{4.44}$$

From Eqs. (4.43), (4.44), and Assumption 4.4.1(c), we have for all $x \in X$ and $k = 1, 2, \ldots$,

$$\begin{aligned}
(T_{\overline{\mu}_{k-1}} \bar{J}_k)(x) &= H\big(x, \overline{\mu}_{k-1}(x), \bar{J}_k\big) \\
&\leq H\big(x, \overline{\mu}_{k-1}(x), J^* + \epsilon_k\, e\big) \\
&\leq H\big(x, \overline{\mu}_{k-1}(x), J^*\big) + \alpha\epsilon_k \\
&\leq H\left(x, \overline{\mu}_{k-1}(x), \lim_{m\to\infty} T_{\mu_1^{k-1}[x]} \cdots T_{\mu_m^{k-1}[x]} \bar{J}\right) + \alpha\epsilon_k \\
&= \bar{J}_{k-1}(x) + \alpha\epsilon_k,
\end{aligned}$$

and finally

$$T_{\overline{\mu}_{k-1}} \bar{J}_k \leq \bar{J}_{k-1} + \alpha\epsilon_k\, e, \qquad k = 1, 2, \ldots.$$

Using this inequality and Assumption 4.4.1(c), we obtain

$$\begin{aligned}
T_{\overline{\mu}_{k-2}} T_{\overline{\mu}_{k-1}} \bar{J}_k &\leq T_{\overline{\mu}_{k-2}} (\bar{J}_{k-1} + \alpha\epsilon_k\, e) \\
&\leq T_{\overline{\mu}_{k-2}} \bar{J}_{k-1} + \alpha^2\epsilon_k\, e \\
&\leq \bar{J}_{k-2} + (\alpha\epsilon_{k-1} + \alpha^2\epsilon_k)\, e.
\end{aligned}$$

Continuing in the same manner, we have for $k = 1, 2, \ldots,$

$$T_{\overline{\mu}_0} \cdots T_{\overline{\mu}_{k-1}} \bar{J}_k \leq \bar{J}_0 + (\alpha\epsilon_1 + \cdots + \alpha^k \epsilon_k)\, e \leq J^* + \left(\sum_{i=0}^{k} \alpha^i \epsilon_i \right) e.$$

Since by Assumption 4.4.1(c), we have $\bar{J} \leq J^* \leq \bar{J}_k$, it follows that

$$T_{\overline{\mu}_0} \cdots T_{\overline{\mu}_{k-1}} \bar{J} \leq J^* + \left(\sum_{i=0}^{k} \alpha^i \epsilon_i \right) e.$$

Denote $\pi_\epsilon = \{\overline{\mu}_0, \overline{\mu}_1, \ldots\}$. Then by taking the limit in the preceding inequality and using Eq. (4.40), we obtain

$$J_{\pi_\epsilon} \leq J^* + \epsilon\, e.$$

Q.E.D.

By using Prop. 4.4.8 we can prove the following.

Proposition 4.4.9: Let Assumption 4.4.1 hold. Then J^* is a fixed point of T.

Proof: For every $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi$ and $x \in X$, we have using the continuity property of Assumption 4.4.1(b) and the monotonicity of H,

$$\begin{aligned} J_\pi(x) &= \lim_{k\to\infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} \bar{J})(x) \\ &= T_{\mu_0} \left(\lim_{k\to\infty} T_{\mu_1} \cdots T_{\mu_k} \bar{J} \right)(x) \\ &\geq (T_{\mu_0} J^*)(x) \\ &\geq (TJ^*)(x). \end{aligned}$$

By taking the infimum of the left-hand side over $\pi \in \Pi$, we obtain

$$J^* \geq TJ^*.$$

To prove the reverse inequality, let ϵ_1 and ϵ_2 be any positive scalars, and let $\overline{\pi} = \{\overline{\mu}_0, \overline{\mu}_1, \ldots\}$ be such that

$$T_{\overline{\mu}_0} J^* \leq TJ^* + \epsilon_1\, e, \qquad J_{\pi_1} \leq J^* + \epsilon_2\, e,$$

where $\pi_1 = \{\overline{\mu}_1, \overline{\mu}_2, \ldots\}$ (such a policy exists by Prop. 4.4.8). By using the preceding relations and Assumption 4.4.1(c), we have

$$\begin{aligned} J^* &\leq J_{\overline{\pi}} \\ &= \lim_{k\to\infty} T_{\overline{\mu}_0} T_{\overline{\mu}_1} \cdots T_{\overline{\mu}_k} \bar{J} \\ &= T_{\overline{\mu}_0} \left(\lim_{k\to\infty} T_{\overline{\mu}_1} \cdots T_{\overline{\mu}_k} \bar{J} \right) \\ &= T_{\overline{\mu}_0} J_{\pi_1} \\ &\leq T_{\overline{\mu}_0} (J^* + \epsilon_2\, e) \\ &\leq T_{\overline{\mu}_0} J^* + \alpha\epsilon_2\, e \\ &\leq TJ^* + (\epsilon_1 + \alpha\epsilon_2)\, e. \end{aligned}$$

Since ϵ_1 and ϵ_2 can be taken arbitrarily small, it follows that

$$J^* \leq TJ^*.$$

Hence $J^* = TJ^*$. **Q.E.D.**

It is known that J^* may not be a fixed point of T if the convergence condition (a) of Assumption 4.4.1 is violated (see the example of Section 3.1.2). Moreover, J^* may not be a fixed point of T if either part (b) or part (c) of Assumption 4.4.1 is violated, even when the monotone increase condition $\bar{J} \leq T\bar{J}$ [and hence also the convergence condition of part (a)] is satisfied (see Examples 4.3.1 and 4.3.2). By applying Prop. 4.4.2, we have the following proposition.

Proposition 4.4.10: Let Assumption 4.4.1 hold, let $\mathcal{C}$ be a set of policy-state pairs such that $J^*_{\mathcal{C}} = J^*$, and let S be any subset of $\mathcal{E}(X)$ such that $\mathcal{C}$ is S-regular. Then:

(a) J^* is the only possible fixed point of T within the set $\{J \in S \mid J \geq J^*\}$.

(b) We have $T^k J \to J^*$ for every $J \in \mathcal{E}(X)$ such that $J^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$.

Proof: By Prop. 4.4.9, J^* is a fixed point of T. The result follows from Prop. 4.4.2. **Q.E.D.**

4.5 STABLE POLICIES AND DETERMINISTIC OPTIMAL CONTROL

In this section, we will consider the use of the regularity ideas of the preceding section in conjunction with a particularly favorable class of monotone

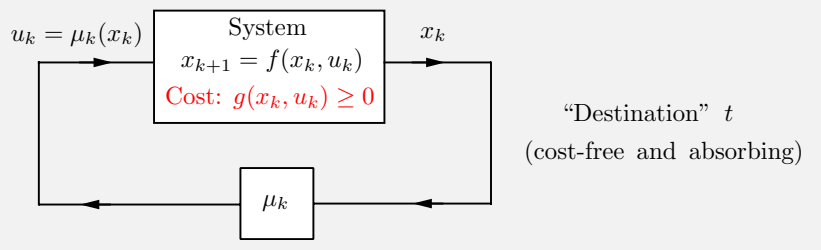


Figure 4.5.1 A deterministic optimal control problem with nonnegative cost per stage, and a cost-free and absorbing destination t.

increasing models. These are the discrete-time infinite horizon deterministic optimal control problems with nonnegative cost per stage, and a destination that is cost-free and absorbing.† Except for the cost nonnegativity, our assumptions are very general, and allow the possibility that the optimal policy may not be stabilizing the system, e.g., may not reach the destination either asymptotically or in a finite number of steps. This situation is illustrated by the one-dimensional linear-quadratic example of Section 3.1.4, where we saw that the Riccati equation may have multiple nonnegative solutions, with the largest solution corresponding to the restricted optimal cost over just the stable policies.

Our approach is similar to the one of the preceding section. We use forcing functions and a perturbation line of analysis like the one of Section 3.4 to delineate collections $\mathcal{C}$ of regular policy-state pairs such that the corresponding restricted optimal cost function $J^*_{\mathcal{C}}$ is a fixed point of T, as required by Prop. 4.4.2.

To this end, we introduce a new unifying notion of p-stability, which in addition to implying convergence of the generated states to the destination, quantifies the speed of convergence. Here is an outline of our analysis:

(a) We consider the properties of several distinct cost functions: J^*, the overall optimal, and $\hat{J}_p$, the restricted optimal over just the p-stable policies. Different choices of p may yield different classes of p-stable policies, with different speeds of convergence.

(b) We show that for any p and associated class of p-stable policies, $\hat{J}_p$ is a solution of Bellman's equation, and we will characterize the smallest and the largest solutions: they are J^*, the optimal cost function, and $\hat{J}^+$, the restricted optimal cost function over the class of (finitely) terminating policies.

(c) We discuss modified versions of the VI and PI algorithms, as substitutes for the standard algorithms, which may not work in general.

† A related line of analysis for deterministic problems with both positive and negative costs per stage is developed in Exercise 4.9.

Consider a deterministic discrete-time infinite horizon optimal control problem involving the system

$$x_{k+1} = f(x_k, u_k), \qquad k = 0, 1, \ldots, \tag{4.45}$$

where x_k and u_k are the state and control at stage k, which belong to sets X and U, referred to as the state and control spaces, respectively, and $f : X \times U \mapsto X$ is a given function. The control u_k must be chosen from a constraint set $U(x_k) \subset U$ that may depend on the current state x_k. The cost per stage $g(x, u)$ is assumed nonnegative and possibly extended real-valued:

$$0 \leq g(x, u) \leq \infty, \qquad \forall\ x \in X,\ u \in U(x),\ k = 0, 1, \ldots. \tag{4.46}$$

We assume that X contains a special state, denoted t, which is referred to as the *destination*, and is cost-free and absorbing:

$$f(t, u) = t, \qquad g(t, u) = 0, \qquad \forall\ u \in U(t).$$

Except for the cost nonnegativity assumption (4.46), this problem is similar to the one of Section 3.5.5. It arises in many classical control applications involving regulation around a set point, and in finite-state and infinite-state versions of shortest path applications; see Fig. 4.5.1.

As earlier, we denote policies by π and stationary policies by μ. Given an initial state x_0, a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ when applied to the system (4.45), generates a unique sequence of state-control pairs $\big(x_k, \mu_k(x_k)\big)$, $k = 0, 1, \ldots$. The cost of π starting from x_0 is

$$J_\pi(x_0) = \sum_{k=0}^{\infty} g\big(x_k, \mu_k(x_k)\big), \qquad x_0 \in X,$$

[the series converges to some number in $[0, \infty]$ thanks to the nonnegativity assumption (4.46)]. The optimal cost function over the set of all policies Π is

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

We denote by $\mathcal{E}^+(X)$ the set of functions $J : X \mapsto [0, \infty]$. In our analysis, we will use the set of functions

$$\mathcal{J} = \big\{J \in \mathcal{E}^+(X) \mid J(t) = 0\big\}.$$

Since t is cost-free and absorbing, this set contains the cost function J_π of every $\pi \in \Pi$, as well as J^*.

Under the cost nonnegativity assumption (4.46), the problem can be cast as a special case of the monotone increasing model with

$$H(x, u, J) = g(x, u) + J\big(f(x, u)\big),$$

and the initial function $\bar{J}$ being identically zero. Thus Prop. 4.4.4 applies and in particular J^* satisfies Bellman's equation:

$$J^*(x) = \inf_{u \in U(x)} \big\{ g(x,u) + J^*\big(f(x,u)\big) \big\}, \quad x \in X.$$

Moreover, an optimal stationary policy (if it exists) may be obtained through the minimization in the right side of this equation, cf. Prop. 4.4.4(c).

The VI method starts from some function $J_0 \in \mathcal{J}$, and generates a sequence of functions $\{J_k\} \subset \mathcal{J}$ according to

$$J_{k+1}(x) = \inf_{u \in U(x)} \big\{ g(x,u) + J_k\big(f(x,u)\big) \big\}, \qquad x \in X,\ k = 0, 1, \ldots. \tag{4.47}$$

From Prop. 4.4.6, we have that the VI sequence $\{J_k\}$ converges to J^* starting from any function $J_0 \in \mathcal{E}^+(X)$ that satisfies

$$J^* \leq J_0 \leq cJ^*,$$

for some scalar $c > 0$. We also have that VI converges to J^* starting from any J_0 with

$$0 \leq J_0 \leq J^*$$

under the compactness condition of Prop. 4.4.4(d). However, $\{J_k\}$ may not always converge to J^* because, among other reasons, Bellman's equation may have multiple solutions within $\mathcal{J}$.

The PI method starts from a stationary policy μ^0, and generates a sequence of stationary policies $\{\mu^k\}$ via a sequence of policy evaluations to obtain J_{μ^k} from the equation

$$J_{\mu^k}(x) = g\big(x, \mu^k(x)\big) + J_{\mu^k}\big(f\big(x, \mu^k(x)\big)\big), \qquad x \in X, \tag{4.48}$$

interleaved with policy improvements to obtain μ^{k+1} from J_{μ^k} according to

$$\mu^{k+1}(x) \in \arg\min_{u \in U(x)} \big\{ g(x,u) + J_{\mu^k}\big(f(x,u)\big) \big\}, \qquad x \in X. \tag{4.49}$$

Here, we implicitly assume that the minimum in Eq. (4.49) is attained for each $x \in X$, which is true under some compactness condition on either $U(x)$ or the level sets of the function $g(x,\cdot) + J_k\big(f(x,\cdot)\big)$, or both. However, as noted in Section 4.3.3, PI may not produce a strict improvement of the cost function of a nonoptimal policy, a fact that was demonstrated with the simple deterministic shortest path example of Section 3.1.1.

The uniqueness of solution of Bellman's equation within $\mathcal{J}$, and the convergence of VI to J^* have been investigated as part of the analysis of Section 3.5.5. There we introduced conditions guaranteeing that J^* is the unique solution of Bellman's equation within a large set of functions

[the near-optimal termination Assumption 3.5.10, but not the cost nonnegativity assumption (4.46)]. Our approach here will make use of the cost nonnegativity but will address the problem under otherwise weaker conditions.

Our analytical approach will also be different than the approach of Section 3.5.5. Here, we will implicitly rely on the regularity ideas for nonstationary policies that we introduced in Section 4.4, and we will make a connection with traditional notions of feedback control system stability. Using nonstationary policies may be important in undiscounted optimal control problems with nonnegative cost per stage because it is not generally true that there exists a stationary ϵ-optimal policy [cf. the ϵ-optimality result of Prop. 4.4.4(e)].

4.5.1 Forcing Functions and p-Stable Policies

We will introduce a notion of stability that involves a function $p : X \mapsto [0, \infty)$ such that

$$p(t) = 0, \qquad p(x) > 0, \qquad \forall\ x \neq t.$$

As in Section 3.4, we refer to p as the *forcing function*, and we associate with it the *p-δ-perturbed optimal control problem*, where $\delta > 0$ is a given scalar. This is the same problem as the original, except that the cost per stage is changed to

$$g(x, u) + \delta p(x).$$

We denote by $J_{\pi,p,\delta}$ the cost function of a policy $\pi \in \Pi$ in the p-δ-perturbed problem:

$$J_{\pi,p,\delta}(x_0) = J_\pi(x_0) + \delta \sum_{k=0}^{\infty} p(x_k), \tag{4.50}$$

where $\{x_k\}$ is the sequence generated starting from x_0 and using π. We also denote by $\hat{J}_{p,\delta}$, the corresponding optimal cost function,

$$\hat{J}_{p,\delta}(x) = \inf_{\pi \in \Pi} J_{\pi,p,\delta}(x), \qquad x \in X.$$

Definition 4.5.1: Let p be a given forcing function. For a state $x_0 \in X$, we say that a policy π is *p-stable from x_0* if for the sequence $\{x_k\}$ generated starting from x_0 and using π we have

$$J_\pi(x_0) < \infty \quad \text{and} \quad \sum_{k=0}^{\infty} p(x_k) < \infty, \tag{4.51}$$

or equivalently [using Eq. (4.50)]

$$J_{\pi,p,\delta}(x_0) < \infty, \qquad \forall\ \delta > 0.$$

The set of all policies that are p-stable from x_0 is denoted by Π_{p,x_0}. We define the *restricted optimal cost function* $\hat{J}_p$ by

$$\hat{J}_p(x) = \inf_{\pi \in \Pi_{p,x}} J_\pi(x), \qquad x \in X, \tag{4.52}$$

(with the convention that the infimum over the empty set is ∞). We say that π is *p-stable* (without qualification) if $\pi \in \Pi_{p,x}$ for all $x \in X$ such that $\Pi_{p,x} \neq \emptyset$. The set of all p-stable policies is denoted by Π_p.

Note that since Eq. (4.51) does not depend on δ, we see that an equivalent definition of a policy π that is p-stable from x_0 is that $J_{\pi,p,\delta}(x_0) < \infty$ for *some* $\delta > 0$ (rather than all $\delta > 0$). Thus the set $\Pi_{p,x}$ of p-stable policies from x depends on p and x but not on δ. Let us make some observations:

(a) *Rate of convergence to t using p-stable policies*: The relation (4.51) shows that the forcing function p quantifies the rate at which the destination is approached using the p-stable policies. As an example, let $X = \Re^n$ and

$$p(x) = \|x\|^\rho,$$

where $\rho > 0$ is a scalar. Then the policies $\pi \in \Pi_{p,x_0}$ are the ones that force x_k towards 0 at a rate faster than $O(1/k^\rho)$, so slower policies are excluded from Π_{p,x_0}.

(b) *Approximation property of $J_{\pi,p,\delta}(x)$*: Consider a pair (π, x_0) with $\pi \in \Pi_{p,x_0}$. By taking the limit as $\delta \downarrow 0$ in the expression

$$J_{\pi,p,\delta}(x_0) = J_\pi(x_0) + \delta \sum_{k=0}^{\infty} p(x_k),$$

[cf. Eq. (4.50)] and by using Eq. (4.51), it follows that

$$\lim_{\delta \downarrow 0} J_{\pi,p,\delta}(x_0) = J_\pi(x_0), \qquad \forall \text{ pairs } (\pi, x_0) \text{ with } \pi \in \Pi_{p,x_0}. \tag{4.53}$$

From this equation, we have that if $\pi \in \Pi_{p,x}$, then $J_{\pi,p,\delta}(x)$ is finite and differs from $J_\pi(x)$ by $O(\delta)$. By contrast, if $\pi \notin \Pi_{p,x}$, then $J_{\pi,p,\delta}(x) = \infty$ by the definition of p-stability, even though we may have $J_\pi(x) < \infty$.

(c) *Limiting property of $\hat{J}_p(x_k)$*: Consider a pair (π, x_0) with $\pi \in \Pi_{p,x_0}$. By breaking down $J_{\pi,p,\delta}(x_0)$ into the sum of the costs of the first k stages and the remaining stages, we have for all $\delta > 0$ and $k > 0$,

$$J_{\pi,p,\delta}(x_0) = \sum_{m=0}^{k-1} g\big(x_m, \mu_m(x_m)\big) + \delta \sum_{m=0}^{k-1} p(x_m) + J_{\pi_k,p,\delta}(x_k),$$

where $\{x_k\}$ is the sequence generated starting from x_0 and using π, and π_k is the policy $\{\mu_k, \mu_{k+1}, \ldots\}$. By taking the limit as $k \to \infty$ and using Eq. (4.50), it follows that

$$\lim_{k\to\infty} J_{\pi_k,p,\delta}(x_k) = 0, \qquad \forall \text{ pairs } (\pi, x_0) \text{ with } \pi \in \Pi_{p,x_0},\ \delta > 0.$$

Also, since $\hat{J}_p(x_k) \leq \hat{J}_{p,\delta}(x_k) \leq J_{\pi_k,p,\delta}(x_k)$, it follows that

$$\lim_{k\to\infty} J_{p,\delta}(x_k) = 0, \quad \forall\ (\pi, x_0) \text{ with } x_0 \in X \text{ and } \pi \in \Pi_{p,x_0},\ \delta > 0, \tag{4.54}$$

$$\lim_{k\to\infty} \hat{J}_p(x_k) = 0, \qquad \forall\ (\pi, x_0) \text{ with } x_0 \in X \text{ and } \pi \in \Pi_{p,x_0}. \tag{4.55}$$

Terminating Policies and Controllability

An important special case is when p is equal to the function

$$p^+(x) = \begin{cases} 0 & \text{if } x = t, \\ 1 & \text{if } x \neq t. \end{cases} \tag{4.56}$$

For $p = p^+$, a policy π is p^+-stable from x if and only if it is *terminating from* x, i.e., reaches t in a finite number of steps starting from x [cf. Eq. (4.51)]. The set of terminating policies from x is denoted by Π_x^+ and it is contained within every other set of p-stable policies $\Pi_{p,x}$, as can be seen from Eq. (4.51). As a result, the restricted optimal cost function over Π_x^+,

$$\hat{J}^+(x) = \inf_{\pi\in\Pi_x^+} J_\pi(x), \qquad x \in X,$$

satisfies $J^*(x) \leq \hat{J}_p(x) \leq \hat{J}^+(x)$ for all $x \in X$. A policy π is said to be *terminating* if it is simultaneously terminating from all $x \in X$ such that $\Pi_x^+ \neq \emptyset$. The set of all terminating policies is denoted by Π^+.

Note that if the state space X is finite, we have for every forcing function p

$$\underline{\beta}\, p^+(x) \leq p(x) \leq \bar{\beta}\, p^+(x), \qquad \forall\ x \in X,$$

for some scalars $\underline{\beta}, \bar{\beta} > 0$. As a result it can be seen that $\Pi_{p,x} = \Pi_x^+$ and $\hat{J}_p = \hat{J}^+$, so in effect the case where $p = p^+$ is the only case of interest for finite-state problems.

The notion of a terminating policy is related to the notion of *controllability*. In classical control theory terms, the system $x_{k+1} = f(x_k, u_k)$ is said to be completely controllable if for every $x_0 \in X$, there exists a policy that drives the state x_k to the destination in a finite number of steps. This notion of controllability is equivalent to the existence of a terminating policy from each $x \in X$.

One of our main results, to be shown shortly, is that J^*, $\hat{J}_p$, and $\hat{J}^+$ are solutions of Bellman's equation, with J^* being the "smallest" solution and $\hat{J}^+$ being the "largest" solution within $\mathcal{J}$. The most favorable situation arises when $J^* = \hat{J}^+$, in which case J^* is the unique solution of Bellman's equation within $\mathcal{J}$. Moreover, in this case it will be shown that the VI algorithm converges to J^* starting with any $J_0 \in \mathcal{J}$ with $J_0 \geq J^*$, and the PI algorithm converges to J^* as well. Once we prove the fixed point property of $\hat{J}_p$, we will be able to bring to bear the regularity ideas of the preceding section (cf. Prop. 4.4.2).

4.5.2 Restricted Optimization over Stable Policies

For a given forcing function p, we denote by $\widehat{X}_p$ the effective domain of $\hat{J}_p$, i.e., the set of all x where $\hat{J}_p$ is finite,

$$\widehat{X}_p = \big\{x \in X \mid \hat{J}_p(x) < \infty\big\}.$$

Since $\hat{J}_p(x) < \infty$ if and only if $\Pi_{p,x} \neq \emptyset$ [cf. Eqs. (4.51) (4.52)], or equivalently $J_{\pi,p,\delta}(x) < \infty$ for some π and all $\delta > 0$, it follows that $\widehat{X}_p$ is also the effective domain of $\hat{J}_{p,\delta}$,

$$\widehat{X}_p = \big\{x \in X \mid \Pi_{p,x} \neq \emptyset\big\} = \big\{x \in X \mid \hat{J}_{p,\delta}(x) < \infty\big\}, \qquad \forall\ \delta > 0.$$

Note that $\widehat{X}_p$ may depend on p and may be a strict subset of the effective domain of J^*, which is denoted by

$$X^* = \big\{x \in X \mid J^*(x) < \infty\big\};$$

(cf. Section 3.5.5). The reason is that there may exist a policy π such that $J_\pi(x) < \infty$, even when there is no p-stable policy from x (for example, no terminating policy from x).

Our first objective is to show that as $\delta \downarrow 0$, the p-δ-perturbed optimal cost function $\hat{J}_{p,\delta}$ converges to the restricted optimal cost function $\hat{J}_p$.

Proposition 4.5.1 (Approximation Property of $\hat{J}_{p,\delta}$): Let p be a given forcing function and $\delta > 0$.

(a) We have

$$J_{\pi,p,\delta}(x) = J_\pi(x) + w_{\pi,p,\delta}(x), \qquad \forall\ x \in X,\ \pi \in \Pi_{p,x}, \quad (4.57)$$

where $w_{\pi,p,\delta}$ is a function such that $\lim_{\delta\downarrow 0} w_{\pi,p,\delta}(x) = 0$ for all $x \in X$.

(b) We have

$$\lim_{\delta\downarrow 0} \hat{J}_{p,\delta}(x) = \hat{J}_p(x), \qquad \forall\ x \in X.$$

Proof: (a) Follows by using Eq. (4.53) for $x \in \widehat{X}_p$, and by taking $w_{p,\delta}(x) = 0$ for $x \notin \widehat{X}_p$.

(b) By Prop. 4.4.4(e), there exists an ϵ-optimal policy π_ϵ for the p-δ-perturbed problem, i.e., $J_{\pi_\epsilon,p,\delta}(x) \leq \hat{J}_{p,\delta}(x) + \epsilon$ for all $x \in X$. Moreover, for $x \in \widehat{X}_p$ we have $\hat{J}_{p,\delta}(x) < \infty$, so $J_{\pi_\epsilon,p,\delta}(x) < \infty$. Hence π_ϵ is p-stable from all $x \in \widehat{X}_p$, and we have $\hat{J}_p \leq J_{\pi_\epsilon}$. Using also Eq. (4.57), we have for all $\delta > 0$, $\epsilon > 0$, $x \in X$, and $\pi \in \Pi_{p,x}$,

$$\hat{J}_p(x) - \epsilon \leq J_{\pi_\epsilon}(x) - \epsilon \leq J_{\pi_\epsilon,p,\delta}(x) - \epsilon \leq \hat{J}_{p,\delta}(x) \leq J_{\pi,p,\delta}(x) = J_\pi(x) + w_{\pi,p,\delta}(x),$$

where $\lim_{\delta \downarrow 0} w_{\pi,p,\delta}(x) = 0$ for all $x \in X$. By taking the limit as $\epsilon \downarrow 0$, we obtain for all $\delta > 0$ and $\pi \in \Pi_{p,x}$,

$$\hat{J}_p(x) \leq \hat{J}_{p,\delta}(x) \leq J_\pi(x) + w_{\pi,p,\delta}(x), \qquad \forall\ x \in X.$$

By taking the limit as $\delta \downarrow 0$ and then the infimum over all $\pi \in \Pi_{p,x}$, we have

$$\hat{J}_p(x) \leq \lim_{\delta \downarrow 0} \hat{J}_{p,\delta}(x) \leq \inf_{\pi \in \Pi_{p,x}} J_\pi(x) = \hat{J}_p(x), \qquad \forall\ x \in X,$$

from which the result follows. **Q.E.D.**

We now consider ϵ-optimal policies, setting the stage for our main proof argument. We know that given any $\epsilon > 0$, by Prop. 4.4.4(e), there exists an ϵ-optimal policy for the p-δ-perturbed problem, i.e., a policy π such that $J_\pi(x) \leq J_{\pi,p,\delta}(x) \leq \hat{J}_{p,\delta}(x) + \epsilon$ for all $x \in X$. We address the question whether there exists a p-stable policy π that is ϵ-optimal for the restricted optimization over p-stable policies, i.e., a policy π that is p-stable simultaneously from all $x \in X_p$, (i.e., $\pi \in \Pi_p$) and satisfies

$$J_\pi(x) \leq \hat{J}_p(x) + \epsilon, \qquad \forall\ x \in X.$$

We refer to such a policy as a *p-ϵ-optimal policy*.

Proposition 4.5.2 (Existence of p-ϵ-Optimal Policy): Let p be a given forcing function and $\delta > 0$. For every $\epsilon > 0$, a policy π that is ϵ-optimal for the p-δ-perturbed problem is p-ϵ-optimal, and hence belongs to Π_p.

Proof: For any ϵ-optimal policy π_ϵ for the p-δ-perturbed problem, we have

$$J_{\pi_\epsilon,p,\delta}(x) \leq \hat{J}_{p,\delta}(x) + \epsilon < \infty, \qquad \forall\ x \in \widehat{X}_p.$$

This implies that $\pi_\epsilon \in \Pi_p$. Moreover, for all sequences $\{x_k\}$ generated from initial state-policy pairs (π, x_0) with $x_0 \in \widehat{X}_p$ and $\pi \in \Pi_{p,x_0}$, we have

$$J_{\pi_\epsilon}(x_0) \leq J_{\pi_\epsilon,p,\delta}(x_0) \leq \hat{J}_{p,\delta}(x_0) + \epsilon \leq J_\pi(x_0) + \delta \sum_{k=0}^{\infty} p(x_k) + \epsilon.$$

Taking the limit as $\delta \downarrow 0$ and using the fact $\sum_{k=0}^{\infty} p(x_k) < \infty$ (since $\pi \in \Pi_{p,x_0}$), we obtain

$$J_{\pi_\epsilon}(x_0) \leq J_\pi(x_0) + \epsilon, \qquad \forall\ x_0 \in \widehat{X}_p,\ \pi \in \Pi_{p,x_0}.$$

By taking infimum over $\pi \in \Pi_{p,x_0}$, it follows that

$$J_{\pi_\epsilon}(x_0) \leq \hat{J}_p(x_0) + \epsilon, \qquad \forall\ x_0 \in \widehat{X}_p,$$

which in view of the fact $J_{\pi_\epsilon}(x_0) = \hat{J}_p(x_0) = \infty$ for $x_0 \notin \widehat{X}_p$, implies that π_ϵ is p-ϵ-optimal. **Q.E.D.**

Note that the preceding proposition implies that

$$\hat{J}_p(x) = \inf_{\pi \in \Pi_p} J_\pi(x), \qquad \forall\ x \in X, \tag{4.58}$$

which is a stronger statement than the definition $\hat{J}_p(x) = \inf_{\pi \in \Pi_{p,x}} J_\pi(x)$ for all $x \in X$. However, it can be shown through examples that there may not exist a restricted-optimal p-stable policy, i.e., a $\pi \in \Pi_p$ such that $J_\pi = \hat{J}_p$, even if there exists an optimal policy for the original problem. One such example is the one-dimensional linear-quadratic problem of Section 3.1.4 for the case where $p = p^+$. Then, there exists a unique linear stable policy that attains the restricted optimal cost $\hat{J}^+(x)$ for all x, but this policy is not terminating. Note also that there may not exist a *stationary* p-ϵ-optimal policy, since generally in undiscounted nonnegative cost optimal control problems there may not exist a stationary ϵ-optimal policy (an example is given following Prop. 4.4.8).

We now take the first steps for bringing regularity ideas into the analysis. We introduce the set of functions S_p given by

$$S_p = \Big\{ J \in \mathcal{J} \;\Big|\; J(x_k) \to 0 \text{ for all sequences } \{x_k\} \text{ generated from initial state-policy pairs } (\pi, x_0) \text{ with } x_0 \in X \text{ and } \pi \in \Pi_{p,x_0} \Big\}. \tag{4.59}$$

In words, S_p consists of the functions in $\mathcal{J}$ whose value is asymptotically driven to 0 by all the policies that are p-stable starting from some $x_0 \in X$. Similar to the analysis of Section 4.4.2, we can prove that the collection

$\mathcal{C}_p = \{(\pi, x_0) \mid \pi \in \Pi_{p,x_0}\}$ is S_p-regular. Moreover, S_p is the largest set S for which $\mathcal{C}_p$ is S-regular.

Note that S_p contains $\hat{J}_p$ and $\hat{J}_{p,\delta}$ for all $\delta > 0$ [cf. Eq. (4.54), (4.55)]. Moreover, S_p contains all functions J such that

$$0 \leq J \leq c\hat{J}_{p,\delta}$$

for some $c > 0$ and $\delta > 0$.

We summarize the preceding discussion in the following proposition, which also shows that $\hat{J}_{p,\delta}$ is the unique solution (within S_p) of Bellman's equation for the p-δ-perturbed problem. This will be needed to prove that $\hat{J}_p$ solves the Bellman equation of the unperturbed problem, but also shows that the p-δ-perturbed problem can be solved more reliably than the original problem (including by VI methods), and yields a close approximation to $\hat{J}_p$ [cf. Prop. 4.5.1(b)].

Proposition 4.5.3: Let p be a forcing function and $\delta > 0$. The function $\hat{J}_{p,\delta}$ belongs to the set S_p, and is the unique solution within S_p of Bellman's equation for the p-δ-perturbed problem,

$$\hat{J}_{p,\delta}(x) = \inf_{u \in U(x)} \Big\{ g(x,u) + \delta p(x) + \hat{J}_{p,\delta}\big(f(x,u)\big) \Big\}, \qquad x \in X. \quad (4.60)$$

Moreover, S_p contains $\hat{J}_p$ and all functions J satisfying

$$0 \leq J \leq c\hat{J}_{p,\delta}$$

for some scalar $c > 0$.

Proof: We have $\hat{J}_{p,\delta} \in S_p$ and $\hat{J}_p \in S_p$ by Eq. (4.54), as noted earlier. We also have that $\hat{J}_{p,\delta}$ is a solution of Bellman's equation (4.60) by Prop. 4.4.4(a). To show that $\hat{J}_{p,\delta}$ is the unique solution within S_p, let $\tilde{J} \in S_p$ be another solution, so that using also Prop. 4.4.4(a), we have

$$\hat{J}_{p,\delta}(x) \leq \tilde{J}(x) \leq g(x,u) + \delta p(x) + \tilde{J}\big(f(x,u)\big), \qquad \forall\, x \in X,\ u \in U(x). \quad (4.61)$$

Fix $\epsilon > 0$, and let $\pi = \{\mu_0, \mu_1, \ldots\}$ be an ϵ-optimal policy for the p-δ-perturbed problem. By repeatedly applying the preceding relation, we have for any $x_0 \in \widehat{X}_p$,

$$\hat{J}_{p,\delta}(x_0) \leq \tilde{J}(x_0) \leq \tilde{J}(x_k) + \delta \sum_{m=0}^{k-1} p(x_m) + \sum_{m=0}^{k-1} g\big(x_m, \mu_m(x_m)\big), \qquad \forall\, k \geq 1, \quad (4.62)$$

where $\{x_k\}$ is the state sequence generated starting from x_0 and using π. We have $\tilde{J}(x_k) \to 0$ (since $\tilde{J} \in S_p$ and $\pi \in \Pi_p$ by Prop. 4.5.2), so that

$$\lim_{k\to\infty} \left\{ \tilde{J}(x_k) + \delta \sum_{m=0}^{k-1} p(x_m) + \sum_{m=0}^{k-1} g\big(x_m, \mu_m(x_m)\big) \right\} = J_{\pi,\delta}(x_0) \leq \hat{J}_{p,\delta}(x_0) + \epsilon. \tag{4.63}$$

By combining Eqs. (4.62) and (4.63), we obtain

$$\hat{J}_{p,\delta}(x_0) \leq \tilde{J}(x_0) \leq \hat{J}_{p,\delta}(x_0) + \epsilon, \qquad \forall\ x_0 \in \widehat{X}_p.$$

By letting $\epsilon \to 0$, it follows that $\hat{J}_{p,\delta}(x_0) = \tilde{J}(x_0)$ for all $x_0 \in \widehat{X}_p$. Also for $x_0 \notin \widehat{X}_p$, we have $\hat{J}_{p,\delta}(x_0) = \tilde{J}(x_0) = \infty$ [since $\hat{J}_{p,\delta}(x_0) = \infty$ for $x_0 \notin \widehat{X}_p$ and $\hat{J}_{p,\delta} \leq \tilde{J}$, cf. Eq. (4.61)]. Thus $\hat{J}_{p,\delta} = \tilde{J}$, proving that $\hat{J}_{p,\delta}$ is the unique solution of the Bellman Eq. (4.60) within S_p. **Q.E.D.**

We next show our main result in this section, namely that $\hat{J}_p$ is the unique solution of Bellman's equation within the set of functions

$$\mathcal{W}_p = \{J \in S_p \mid \hat{J}_p \leq J\}. \tag{4.64}$$

Moreover, we show that the VI algorithm yields $\hat{J}_p$ in the limit for any initial $J_0 \in \mathcal{W}_p$. This result is intimately connected with the regularity ideas of Section 4.4. The idea is that the collection $\mathcal{C}_p = \big\{(\pi, x_0) \mid \pi \in \Pi_{p,x_0}\big\}$ is S_p-regular, as noted earlier. In view of this and the fact that $J^*_{\mathcal{C}_p} = \hat{J}_p$, the result will follow from Prop. 4.4.2 once $\hat{J}_p$ is shown to be a solution of Bellman's equation. This latter property is shown essentially by taking the limit as $\delta \downarrow 0$ in Eq. (4.60).

Proposition 4.5.4: Let p be a given forcing function. Then:

(a) $\hat{J}_p$ is the unique solution of Bellman's equation

$$J(x) = \inf_{u\in U(x)} \Big\{ g(x,u) + J\big(f(x,u)\big) \Big\}, \qquad x \in X, \tag{4.65}$$

within the set $\mathcal{W}_p$ of Eq. (4.64).

(b) (*VI Convergence*) If $\{J_k\}$ is the sequence generated by the VI algorithm (4.47) starting with some $J_0 \in \mathcal{W}_p$, then $J_k \to \hat{J}_p$.

(c) (*Optimality Condition*) If $\hat{\mu}$ is a p-stable stationary policy and

$$\hat{\mu}(x) \in \arg\min_{u\in U(x)} \big\{ g(x,u) + \hat{J}_p\big(f(x,u)\big) \big\}, \qquad \forall\ x \in X, \tag{4.66}$$

then $\hat{\mu}$ is optimal over the set of p-stable policies. Conversely, if $\hat{\mu}$ is optimal within the set of p-stable policies, then it satisfies the preceding condition (4.66).

Proof: (a), (b) We first show that $\hat{J}_p$ is a solution of Bellman's equation. Since $\hat{J}_{p,\delta}$ is a solution of Bellman's equation for the p-δ-perturbed problem (cf. Prop. 4.5.3) and $\hat{J}_{p,\delta} \geq \hat{J}_p$ [cf. Prop. 4.5.1(b)], we have for all $\delta > 0$,

$$\begin{aligned}\hat{J}_{p,\delta}(x) &= \inf_{u\in U(x)} \Big\{g(x,u) + \delta p(x) + \hat{J}_{p,\delta}\big(f(x,u)\big)\Big\} \\ &\geq \inf_{u\in U(x)} \Big\{g(x,u) + \hat{J}_{p,\delta}\big(f(x,u)\big)\Big\} \\ &\geq \inf_{u\in U(x)} \Big\{g(x,u) + \hat{J}_{p}\big(f(x,u)\big)\Big\}.\end{aligned}$$

By taking the limit as $\delta \downarrow 0$ and using the fact $\lim_{\delta\downarrow 0} \hat{J}_{p,\delta} = \hat{J}_p$ [cf. Prop. 4.5.1(b)], we obtain

$$\hat{J}_p(x) \geq \inf_{u\in U(x)} \Big\{g(x,u) + \hat{J}_p\big(f(x,u)\big)\Big\}, \qquad \forall\ x \in X. \tag{4.67}$$

For the reverse inequality, let $\{\delta_m\}$ be a sequence with $\delta_m \downarrow 0$. From Prop. 4.5.3, we have for all m, $x \in X$, and $u \in U(x)$,

$$\begin{aligned}g(x,u) + \delta_m p(x) + \hat{J}_{p,\delta_m}\big(f(x,u)\big) &\geq \inf_{v\in U(x)} \Big\{g(x,v) + \delta_m p(x) \\ &\qquad + \hat{J}_{p,\delta_m}\big(f(x,v)\big)\Big\} \\ &= \hat{J}_{p,\delta_m}(x).\end{aligned}$$

Taking the limit as $m \to \infty$, and using the fact $\lim_{\delta_m\downarrow 0} \hat{J}_{p,\delta_m} = \hat{J}_p$ [cf. Prop. 4.5.1(b)], we have

$$g(x,u) + \hat{J}_p\big(f(x,u)\big) \geq \hat{J}_p(x), \qquad \forall\ x \in X,\ u \in U(x),$$

so that

$$\inf_{u\in U(x)} \Big\{g(x,u) + \hat{J}_p\big(f(x,u)\big)\Big\} \geq \hat{J}_p(x), \qquad \forall\ x \in X. \tag{4.68}$$

By combining Eqs. (4.67) and (4.68), we see that $\hat{J}_p$ is a solution of Bellman's equation. We also have $\hat{J}_p \in S_p$ by Prop. 4.5.3, implying that $\hat{J}_p \in \mathcal{W}_p$ and proving part (a) except for the uniqueness assertion. Part

(b) and the uniqueness part of part (a) follow from Prop. 4.4.2; see the discussion preceding the proposition.

(c) If μ is p-stable and Eq. (4.66) holds, then

$$\hat{J}_p(x) = g\big(x, \mu(x)\big) + \hat{J}_p\big(f(x, \mu(x))\big), \qquad x \in X.$$

By Prop. 4.4.4(b), this implies that $J_\mu \leq \hat{J}_p$, so μ is optimal over the set of p-stable policies. Conversely, assume that μ is p-stable and $J_\mu = \hat{J}_p$. Then by Prop. 4.4.4(b), we have

$$\hat{J}_p(x) = g\big(x, \mu(x)\big) + \hat{J}_p\big(f(x, \mu(x))\big), \qquad x \in X,$$

and since [by part (a)] $\hat{J}_p$ is a solution of Bellman's equation,

$$\hat{J}_p(x) = \inf_{u \in U(x)} \big\{g(x,u) + \hat{J}_p\big(f(x,u)\big)\big\}, \qquad x \in X.$$

Combining the last two relations, we obtain Eq. (4.66). **Q.E.D.**

As a supplement to the preceding proposition, we note the specialization of Prop. 4.4.5 that relates to the optimal cost function J^*.

Proposition 4.5.5: Let S^* be the set

$$S^* = \Big\{J \in \mathcal{J} \;\Big|\; J(x_k) \to 0 \text{ for all sequences } \{x_k\} \text{ generated from initial state-policy pairs } (\pi, x_0) \text{ with } J_\pi(x_0) < \infty\Big\},$$

and $\mathcal{W}^*$ be the set

$$\mathcal{W}^* = \big\{J \in S^* \mid J^* \leq J\big\}.$$

Then J^* belongs to S^* and is the unique solution of Bellman's equation within S^*. Moreover, we have $T^k J \to J^*$ for all $J \in \mathcal{W}^*$.

Proof: Follows from Prop. 4.4.5 in the deterministic special case where w_k takes a single value. **Q.E.D.**

We now consider the special case where p is equal to the function $p^+(x) = 1$ for all $x \neq t$ [cf. Eq. (4.56)]. Then the set of p^+-stable policies from x is Π_x^+, the set of terminating policies from x, and the corresponding restricted optimal cost is $\hat{J}^+(x)$:

$$\hat{J}^+(x) = \hat{J}_{p^+}(x) = \inf_{\pi \in \Pi_x^+} J_\pi(x) = \inf_{\pi \in \Pi^+} J_\pi(x), \qquad x \in X,$$

[the last equality follows from Eq. (4.58)]. In this case, the set S_{p^+} of Eq. (4.59) is the entire set $\mathcal{J}$,

$$S_{p^+} = \mathcal{J},$$

since for all $J \in \mathcal{J}$ and all sequences $\{x_k\}$ generated from initial state-policy pairs (π, x_0) with $x_0 \in X$ and π terminating from x_0, we have $J(x_k) = 0$ for k sufficiently large. Thus, the corresponding set of Eq. (4.64) is

$$\mathcal{W}^+ = \{J \in \mathcal{J} \mid \hat{J}^+ \leq J\}.$$

By specializing to the case $p = p^+$ the result of Prop. 4.5.4, we obtain the following proposition, which makes a stronger assertion than Prop. 4.5.4(a), namely that $\hat{J}^+$ is the largest solution of Bellman's equation within $\mathcal{J}$ (rather than the smallest solution within $\mathcal{W}^+$).

Proposition 4.5.6:

(a) $\hat{J}^+$ is the largest solution of the Bellman equation (4.65) within $\mathcal{J}$, i.e., $\hat{J}^+$ is a solution and if $J' \in \mathcal{J}$ is another solution, then $J' \leq \hat{J}^+$.

(b) (*VI Convergence*) If $\{J_k\}$ is the sequence generated by the VI algorithm (4.47) starting with some $J_0 \in \mathcal{J}$ with $J_0 \geq \hat{J}^+$, then $J_k \to \hat{J}^+$.

(c) (*Optimality Condition*) If μ^+ is a terminating stationary policy and

$$\mu^+(x) \in \arg\min_{u \in U(x)} \big\{g(x,u) + \hat{J}^+\big(f(x,u)\big)\big\}, \qquad \forall\, x \in X, \quad (4.69)$$

then μ^+ is optimal over the set of terminating policies. Conversely, if μ^+ is optimal within the set of terminating policies, then it satisfies the preceding condition (4.69).

Proof: In view of Prop. 4.5.4, we only need to show that $\hat{J}^+$ is the largest solution of the Bellman equation. From Prop. 4.5.4(a), $\hat{J}^+$ is a solution that belongs to $\mathcal{J}$. If $J' \in \mathcal{J}$ is another solution, we have $J' \leq \tilde{J}$ for some $\tilde{J} \in W^+$, so $J' = T^k J' \leq T^k \tilde{J}$ for all k. Since $T^k \tilde{J} \to \hat{J}^+$, it follows that $J' \leq \hat{J}^+$. **Q.E.D.**

We illustrate Props. 4.5.4 and 4.5.6 in Fig. 4.5.2. In particular, each forcing function p delineates the set of initial functions $\mathcal{W}_p$ from which VI converges to $\hat{J}_p$. The function $\hat{J}_p$ is the minimal element of $\mathcal{W}_p$. Moreover, we have $\mathcal{W}_p \cap \mathcal{W}_{p'} = \emptyset$ if $\hat{J}_p \neq \hat{J}_{p'}$, in view of the VI convergence result of Prop. 4.5.4(b).

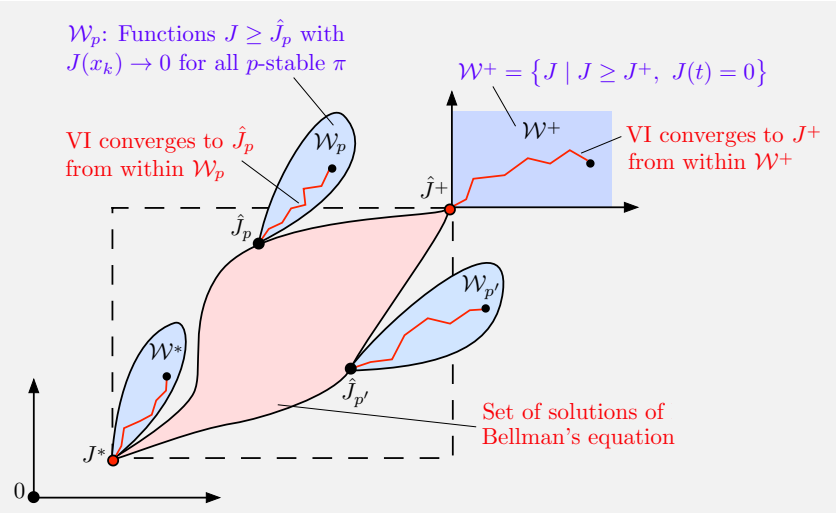


Figure 4.5.2 Schematic two-dimensional illustration of the results of Prop. 4.5.4 and 4.5.6. The functions J^*, $\hat{J}^+$, and $\hat{J}_p$ are all solutions of Bellman's equation. Moreover J^* and $\hat{J}^+$ are the smallest and largest solutions, respectively. Each p defines the set of initial functions $\mathcal{W}_p$ from which VI converges to $\hat{J}_p$ from above. For two forcing functions p and p', we have $\mathcal{W}_p \cap \mathcal{W}_{p'} = \emptyset$ if $\hat{J}_p \neq \hat{J}_{p'}$. Moreover, $\mathcal{W}_p$ contains no solutions of Bellman's equation other than $\hat{J}_p$. It is also possible that $\mathcal{W}_p$ consists of just $\hat{J}_p$.

Note a significant fact: Proposition 4.5.6(b) implies that VI converges to $\hat{J}^+$ starting from the readily available initial condition

$$J_0(x) = \begin{cases} 0 & \text{if } x = t, \\ \infty & \text{if } x \neq t. \end{cases}$$

For this choice of J_0, the value $J_k(x)$ generated by VI is the optimal cost that can be achieved starting from x subject to the constraint that t is reached in k steps or less. As we have noted earlier, in shortest-path type problems VI tends to converge faster when started from above.

Consider now the favorable case where terminating policies are sufficient, in the sense that $\hat{J}^+ = J^*$; cf. Fig. 4.5.3. Then, from Prop. 4.5.6, it follows that J^* is the unique solution of Bellman's equation within $\mathcal{J}$, and the VI algorithm converges to J^* from above, i.e., starting from any $J_0 \in \mathcal{J}$ with $J_0 \geq J^*$. Under additional conditions, such as finiteness of $U(x)$ for all $x \in X$ [cf. Prop. 4.4.4(d)], VI converges to J^* starting from any $J_0 \in \mathcal{E}^+(X)$ with $J_0(t) = 0$. These results are consistent with our analysis of Section 3.5.5.

Examples of problems where terminating policies are sufficient include linear-quadratic problems under the classical conditions of controllability and observability, and finite-node deterministic shortest path prob-

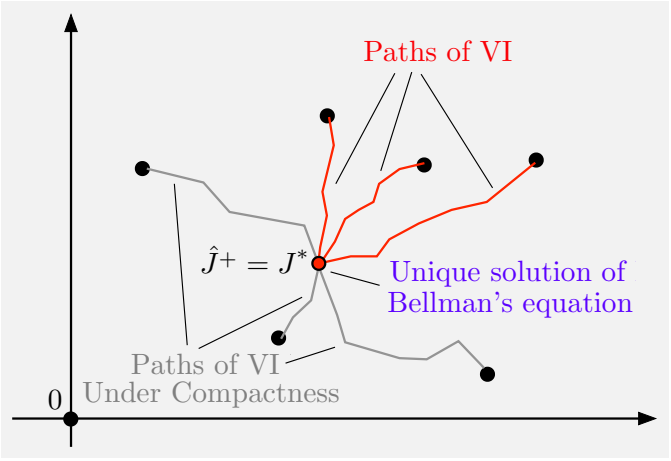


Figure 4.5.3 Schematic two-dimensional illustration of the favorable case where $\hat{J}^+ = J^*$. Then J^* is the unique solution of Bellman's equation within $\mathcal{J}$, and the VI algorithm converges to J^* from above [and also starting from any $J_0 \geq 0$ under a suitable compactness condition; cf. Prop. 4.4.4(d)].

lems with all cycles having positive length. Note that in the former case, despite the fact $\hat{J}^+ = J^*$, there is no optimal terminating policy, since the only optimal policy is a linear policy that drives the system to the origin asymptotically, but not in finite time.

Let us illustrate the results of this section with two examples.

Example 4.5.1 (Minimum Energy Stable Control of Linear Systems)

Consider the linear-quadratic problem of Section 3.5.4. We assume that there exists at least one linear stable policy, so that J^* is real-valued. However, we are making no assumptions on the state weighting matrix Q other than positive semidefiniteness. This includes the case $Q = 0$, when $J^*(x) \equiv 0$. In this case an optimal policy is $\mu^*(x) \equiv 0$, which may not be stable, yet the problem of finding a stable policy that minimizes the "control energy" (a cost that is positive definite quadratic on the control with no penalty on the state) among all stable policies is meaningful.

We consider the forcing function

$$p(x) = \|x\|^2,$$

so the p-δ-perturbed problem includes a positive definite state penalty and from the classical linear-quadratic results, $\hat{J}_{p,\delta}$ is a positive definite quadratic function $x'P_\delta x$, where P_δ is the unique solution of the δ-perturbed Riccati equation

$$P_\delta = A'\big(P_\delta - P_\delta B(B'P_\delta B + R)^{-1}B'P_\delta\big)A + Q + \delta I, \tag{4.70}$$

within the class of positive semidefinite matrices. By Prop. 4.5.1, we have $\hat{J}_p(x) = x'\hat{P}x$, where $\hat{P} = \lim_{\delta \downarrow 0} P_\delta$ is positive semidefinite, and solves the (unperturbed) Riccati equation

$$P = A'\big(P - PB(B'PB + R)^{-1}B'P\big)A + Q.$$

Moreover, by Prop. 4.5.4(a), $\hat{P}$ is the largest solution among positive semidefinite matrices, since all positive semidefinite quadratic functions belong to the set S_p of Eq. (4.59). By Prop. 4.5.4(c), any stable stationary policy $\hat{\mu}$ that is optimal among the set of stable policies must satisfy the optimality condition

$$\hat{\mu}(x) \in \arg\min_{u \in \Re^m} \big\{u'Ru + (Ax + Bu)'\hat{P}(Ax + Bu)\big\}, \qquad \forall\, x \in \Re^n,$$

[cf. Eq. (4.66)], or equivalently, by setting the gradient of the minimized expression to 0,

$$(R + B'\hat{P}B)\hat{\mu}(x) = -B'\hat{P}Ax. \tag{4.71}$$

We may solve Eq. (4.71), and check if any of its solutions $\hat{\mu}$ is p-stable; if this is so, $\hat{\mu}$ is optimal within the class of p-stable policies. Note, however, that in the absence of additional conditions, it is possible that some policies $\hat{\mu}$ that solve Eq. (4.71) are p-unstable.

In the case where there is no linear stable policy, the p-δ-perturbed cost function $\hat{J}_{p,\delta}$ need not be real-valued, and the δ-perturbed Riccati equation (4.70) may not have any solution (consider for example the case where $n = 1$, $m = 1$, $A = 2$, $B = 0$, and $Q = R = 1$). Then, Prop. 4.5.6 still applies, but the preceding analytical approach needs to be modified.

As noted earlier, the Bellman equation may have multiple solutions corresponding to different forcing functions p, with each solution being unique within the corresponding set $\mathcal{W}_p$ of Eq. (4.64), consistently with Prop. 4.5.4(a). The following is an illustrative example.

Example 4.5.2 (An Optimal Stopping Problem)

Consider an optimal stopping problem where the state space X is $\Re^n$. We identify the destination with the origin of $\Re^n$, i.e., $t = 0$. At each $x \neq 0$, we may either stop (move to the origin) at a cost $c > 0$, or move to state γx at cost $\|x\|$, where γ is a scalar with $0 < \gamma < 1$; see Fig. 4.5.4.† Thus the Bellman equation has the form

$$J(x) = \begin{cases} \min\big\{c,\, \|x\| + J(\gamma x)\big\} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

† In this example, the salient feature of the policy that never stops at an $x \neq 0$ is that it drives the system asymptotically to the destination according to an equation of the form $x_{k+1} = f(x_k)$, where f is a contraction mapping. The example admits generalization to the broader class of optimal stopping problems that have this property. For simplicity in illustrating our main point, we consider here the special case where $f(x) = \gamma x$ with $\gamma \in (0, 1)$.

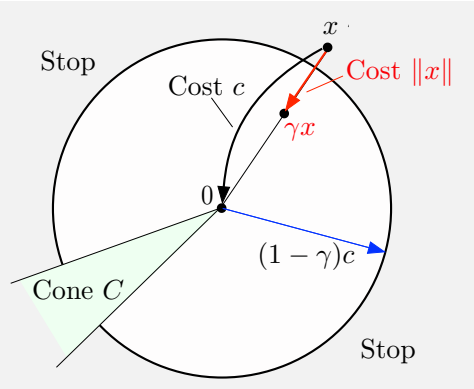


Figure 4.5.4 Illustration of the stopping problem of Example 4.5.2. The optimal policy is to stop outside the sphere of radius $(1-\gamma)c$ and to continue otherwise. Each cone C of the state space defines a different solution $\hat{J}_p$ of Bellman's equation, with $\hat{J}_p(x) = c$ for all nonzero $x \in C$, and a corresponding region of convergence of the VI algorithm.

Let us consider first the forcing function

$$p(x) = \|x\|.$$

Then it can be verified that all policies are p-stable. We have

$$J^*(x) = \hat{J}_p(x) = \min\left\{c,\ \frac{1}{1-\gamma}\|x\|\right\}, \qquad \forall\ x \in \Re^n,$$

and the optimal cost function of the corresponding p-δ-perturbed problem is

$$\hat{J}_{p,\delta}(x) = \min\left\{c + \delta\|x\|,\ \frac{1+\delta}{1-\gamma}\|x\|\right\}, \qquad \forall\ x \in \Re^n.$$

Here the set S_p of Eq. (4.59) is given by

$$S_p = \Big\{J \in \mathcal{J} \mid \lim_{x\to 0} J(x) = 0\Big\},$$

and the corresponding set $\mathcal{W}_p$ of Eq. (4.64) is given by

$$\mathcal{W}_p = \Big\{J \in \mathcal{J} \mid J^* \leq J,\ \lim_{x\to 0} J(x) = 0\Big\}.$$

Let us consider next the forcing function

$$p^+(x) = \begin{cases} 1 & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

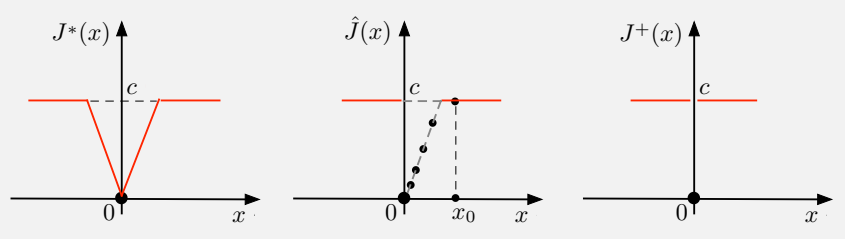


Figure 4.5.5 Illustration of three solutions of Bellman's equation in the one-dimensional case ($n = 1$) of the stopping problem of Example 4.5.2. The solution in the middle is specified by a scalar $x_0 > 0$, and has the form

$$\hat{J}(x) = \begin{cases} 0 & \text{if } x = 0, \\ \frac{1}{1-\gamma}|x| & \text{if } 0 < x < (1-\gamma)c \text{ and } x = \gamma^k x_0 \text{ for some } k \geq 0, \\ c & \text{otherwise.} \end{cases}$$

Then the p^+-stable policies are the terminating policies. Since stopping at some time and incurring the cost c is a requirement for a p^+-stable policy, it follows that the optimal p^+-stable policy is to stop as soon as possible, i.e., stop at every state. The corresponding restricted optimal cost function is

$$\hat{J}^+(x) = \begin{cases} c & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

The optimal cost function of the corresponding p^+-δ-perturbed problem is

$$\hat{J}_{p^+,\delta}(x) = \begin{cases} c + \delta & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases}$$

since in the p^+-δ-perturbed problem it is again optimal to stop as soon as possible, at cost $c + \delta$. Here the set S_{p^+} is equal to $\mathcal{J}$, and the corresponding set $\mathcal{W}^+$ is equal to $\big\{J \in \mathcal{J} \mid \hat{J}^+ \leq J\big\}$.

However, there are infinitely many additional solutions of Bellman's equation between the largest and smallest solutions J^* and $\hat{J}^+$. For example, when $n > 1$, functions $J \in \mathcal{J}$ such that $J(x) = J^*(x)$ for x in some cone and $J(x) = \hat{J}^+(x)$ for x in the complementary cone are solutions; see Fig. 4.5.4. There is also a corresponding infinite number of regions of convergence $\mathcal{W}_p$ of VI [cf. Eq. (4.64)]. Also VI converges to J^* starting from any J_0 with $0 \leq J_0 \leq J^*$ [cf. Prop. 4.4.4(d)]. Figure 4.5.5 illustrates additional solutions of Bellman's equation of a different character.

4.5.3 Policy Iteration Methods

Generally, the standard PI algorithm [cf. Eqs. (4.48), (4.49)] produces unclear results under our assumptions. The following example provides an instance where the PI algorithm may converge to either an optimal or a strictly suboptimal policy.

Example 4.5.3 (Counterexample for PI)

Consider the case $X = \{0, 1\}$, $U(0) = U(1) = \{0, 1\}$, and the destination is $t = 0$. Let also

$$f(x,u) = \begin{cases} 0 & \text{if } u = 0, \\ x & \text{if } u = 1, \end{cases} \qquad g(x,u) = \begin{cases} 1 & \text{if } u = 0,\ x = 1, \\ 0 & \text{if } u = 1 \text{ or } x = 0. \end{cases}$$

This is a one-state-plus-destination shortest path problem where the control $u = 0$ moves the state from $x = 1$ to $x = 0$ (the destination) at cost 1, while the control $u = 1$ keeps the state unchanged at cost 0 (cf. the problem of Section 3.1.1). The policy μ^* that keeps the state unchanged is the only optimal policy, with $J_{\mu^*}(x) = J^*(x) = 0$ for both states x. However, under any forcing function p with $p(1) > 0$, the policy $\hat{\mu}$, which moves from state 1 to 0, is the only p-stable policy, and we have $J_{\hat{\mu}}(1) = \hat{J}_p(1) = 1$. The standard PI algorithm (4.48), (4.49) when started with μ^*, it will repeat μ^*. If this algorithm is started with $\hat{\mu}$, it may generate μ^* or it may repeat $\hat{\mu}$, depending on how the policy improvement iteration is implemented. The reason is that for both x we have

$$\hat{\mu}(x) \in \arg\min_{u \in \{0,1\}} \Big\{ g(x,u) + \hat{J}_p\big(f(x,u)\big) \Big\},$$

as can be verified with a straightforward calculation. Thus a rule for breaking a tie in the policy improvement operation is needed, but such a rule may not be obvious in general.

For another illustration, consider the stopping problem of Example 4.5.2. There if PI is started with the policy that stops at every state, it repeats that policy, and this policy is not optimal even within the class of stable policies with respect to the forcing function $p(x) = \|x\|$.

Motivated by the preceding examples, we consider several types of PI methods that bypass the difficulty above either through assumptions or through modifications. We first consider a favorable case where the standard PI algorithm is reliable. This is the case where the terminating policies are sufficient, in the sense that $J^* = \hat{J}^+$, as in Section 3.5.5.

Policy Iteration for the Case $J^* = \hat{J}^+$

The standard PI algorithm starts with a stationary policy μ^0, and generates a sequence of stationary policies $\{\mu^k\}$ via a sequence of policy evaluations to obtain J_{μ^k} from the equation

$$J_{\mu^k}(x) = g\big(x, \mu^k(x)\big) + J_{\mu^k}\big(f\big(x, \mu^k(x)\big)\big), \qquad x \in X, \tag{4.72}$$

interleaved with policy improvements to obtain μ^{k+1} from J_{μ^k} according to

$$\mu^{k+1}(x) \in \arg\min_{u \in U(x)} \big\{ g(x,u) + J_{\mu^k}\big(f(x,u)\big) \big\}, \qquad x \in X. \tag{4.73}$$

We implicitly assume here that Eq. (4.72) can be solved for J_{μ^k}, and that the minimum in Eq. (4.73) is attained for each $x \in X$, which is true under some compactness condition on either $U(x)$ or the level sets of the function $g(x,\cdot) + J_k\big(f(x,\cdot)\big)$, or both.

Proposition 4.5.7: (Convergence of PI) Assume that $J^* = \hat{J}^+$. Then the sequence $\{J_{\mu^k}\}$ generated by the PI algorithm (4.72), (4.73), satisfies $J_{\mu^k}(x) \downarrow J^*(x)$ for all $x \in X$.

Proof: For a stationary policy μ, let $\bar{\mu}$ satisfy the policy improvement equation

$$\bar{\mu}(x) \in \arg\min_{u \in U(x)} \big\{g(x,u) + J_\mu\big(f(x,u)\big)\big\}, \qquad x \in X.$$

We have shown that

$$J_\mu(x) \geq \inf_{u \in U(x)} \big\{g(x,u) + J_\mu\big(f(x,u)\big)\big\} \geq J_{\bar{\mu}}(x), \qquad x \in X; \tag{4.74}$$

cf. Eq. (4.29). Using μ^k and μ^{k+1} in place of μ and $\bar{\mu}$, we see that the sequence $\{J_{\mu^k}\}$ generated by PI converges monotonically to some function $J_\infty \in E^+(X)$, i.e., $J_{\mu^k} \downarrow J_\infty$. Moreover, from Eq. (4.74) we have

$$J_\infty(x) \geq \inf_{u \in U(x)} \big\{g(x,u) + J_\infty\big(f(x,u)\big)\big\}, \qquad x \in X,$$

as well as

$$g(x,u) + J_{\mu_k}\big(f(x,u)\big) \geq J_\infty(x), \qquad x \in X,\ u \in U(x).$$

We now take the limit in the second relation as $k \to \infty$, then take the infimum over $u \in U(x)$, and then combine with the first relation, to obtain

$$J_\infty(x) \geq \inf_{u \in U(x)} \big\{g(x,u) + J_\infty\big(f(x,u)\big)\big\} \geq J_\infty(x), \qquad x \in X.$$

Thus J_∞ is a solution of Bellman's equation, satisfying $J_\infty \geq J^*$ (since $J_{\mu^k} \geq J^*$ for all k) and $J_\infty \in \mathcal{J}$ (since $J_{\mu^k} \in \mathcal{J}$), so by Prop. 4.5.6(a), it must satisfy $J_\infty = J^*$. **Q.E.D.**

A Perturbed Version of Policy Iteration for the Case $J^* \neq \hat{J}^+$

We now consider PI algorithms without the condition $J^* = \hat{J}^+$. We provide a version of the PI algorithm, which uses a given forcing function p that is fixed, and generates a sequence $\{\mu^k\}$ of p-stable policies such that

$J_{\mu^k} \to \hat{J}_p$. Related algorithms were given in Sections 3.4 and 3.5.1. The following assumption requires that the algorithm generates p-stable policies exclusively, which can be quite restrictive. For instance it is not satisfied for the problem of Example 4.5.3.

Assumption 4.5.1: For each $\delta > 0$ there exists at least one p-stable stationary policy μ such that $J_{\mu,p,\delta} \in S_p$. Moreover, given a p-stable stationary policy μ and a scalar $\delta > 0$, every stationary policy $\overline{\mu}$ such that

$$\overline{\mu}(x) \in \arg\min_{u \in U(x)} \Big\{ g(x,u) + J_{\mu,p,\delta}\big(f(x,u)\big) \Big\}, \qquad \forall\, x \in X,$$

is p-stable, and at least one such policy exists.

The perturbed version of the PI algorithm is defined as follows. Let $\{\delta_k\}$ be a positive sequence with $\delta_k \downarrow 0$, and let μ^0 be a p-stable policy that satisfies $J_{\mu^0,p,\delta_0} \in S_p$. One possibility is that μ^0 is an optimal policy for the δ_0-perturbed problem (cf. the discussion preceding Prop. 4.5.3). At iteration k, we have a p-stable policy μ^k, and we generate a p-stable policy μ^{k+1} according to

$$\mu^{k+1}(x) \in \arg\min_{u \in U(x)} \big\{ g(x,u) + J_{\mu^k,p,\delta_k}\big(f(x,u)\big) \big\}, \qquad x \in X. \tag{4.75}$$

Note that by Assumption 4.5.1 the algorithm is well-defined, and is guaranteed to generate a sequence of p-stable stationary policies. We have the following proposition.

Proposition 4.5.8: Let Assumption 4.5.1 hold. Then for a sequence of p-stable policies $\{\mu^k\}$ generated by the perturbed PI algorithm (4.75), we have $J_{\mu^k,p,\delta_k} \downarrow \hat{J}_p$ and $J_{\mu^k} \to \hat{J}_p$.

Proof: Since the forcing function p is kept fixed, to simplify notation, we abbreviate $J_{\mu,p,\delta}$ with $J_{\mu,\delta}$ for all policies μ and scalars $\delta > 0$. Also, we will use the mappings $T_\mu : \mathcal{E}^+(X) \mapsto \mathcal{E}^+(X)$ and $T_{\mu,\delta} : \mathcal{E}^+(X) \mapsto \mathcal{E}^+(X)$ given by

$$(T_\mu J)(x) = g\big(x,\mu(x)\big) + J\big(f(x,\mu(x))\big), \qquad x \in X,$$

$$(T_{\mu,\delta} J)(x) = g\big(x,\mu(x)\big) + \delta p(x) + J\big(f(x,\mu(x))\big), \qquad x \in X.$$

Moreover, we will use the mapping $T : \mathcal{E}^+(X) \mapsto \mathcal{E}^+(X)$ given by

$$(TJ)(x) = \inf_{u \in U(x)} \big\{ g(x,u) + J\big(f(x,u)\big) \big\}, \qquad x \in X.$$

The algorithm definition (4.75) implies that for all integer $m \geq 1$ we have for all $x_0 \in X$,

$$\begin{aligned} J_{\mu^k,\delta_k}(x_0) &\geq (TJ_{\mu^k,\delta_k})(x_0) + \delta_k p(x_0) \\ &= (T_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k})(x_0) \\ &\geq (T^m_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k})(x_0) \\ &\geq (T^m_{\mu^{k+1},\delta_k} \bar{J})(x_0), \end{aligned}$$

where $\bar{J}$ is the identically zero function [$\bar{J}(x) \equiv 0$]. From this relation we obtain

$$\begin{aligned} J_{\mu^k,\delta_k}(x_0) &\geq \lim_{m\to\infty} (T^m_{\mu^{k+1},\delta_k} \bar{J})(x_0) \\ &= \lim_{m\to\infty} \left\{ \sum_{\ell=0}^{m-1} \big(g\big(x_\ell, \mu^{k+1}(x_\ell)\big) + \delta_k p(x_\ell)\big) \right\} \\ &\geq J_{\mu^{k+1},\delta_{k+1}}(x_0), \end{aligned}$$

as well as

$$J_{\mu^k,\delta_k}(x_0) \geq (TJ_{\mu^k,\delta_k})(x_0) + \delta_k p(x_0) \geq J_{\mu^{k+1},\delta_{k+1}}(x_0).$$

It follows that $\{J_{\mu^k,\delta_k}\}$ is monotonically nonincreasing, so that $J_{\mu^k,\delta_k} \downarrow J_\infty$ for some J_∞, and

$$\lim_{k\to\infty} TJ_{\mu^k,\delta_k} = J_\infty. \tag{4.76}$$

We also have, using the fact $J_\infty \leq J_{\mu^k,\delta_k}$,

$$\begin{aligned} \inf_{u\in U(x)} \big\{g(x,u) + J_\infty\big(f(x,u)\big)\big\} &\leq \lim_{k\to\infty} \inf_{u\in U(x)} \big\{g(x,u) + J_{\mu^k,\delta_k}\big(f(x,u)\big)\big\} \\ &\leq \inf_{u\in U(x)} \lim_{k\to\infty} \big\{g(x,u) + J_{\mu^k,\delta_k}\big(f(x,u)\big)\big\} \\ &= \inf_{u\in U(x)} \left\{g(x,u) + \lim_{k\to\infty} J_{\mu^k,\delta_k}\big(f(x,u)\big)\right\} \\ &= \inf_{u\in U(x)} \big\{g(x,u) + J_\infty\big(f(x,u)\big)\big\}. \end{aligned}$$

Thus equality holds throughout above, so that

$$\lim_{k\to\infty} TJ_{\mu^k,\delta_k} = TJ_\infty.$$

Combining this with Eq. (4.76), we obtain $J_\infty = TJ_\infty$, i.e., J_∞ solves Bellman's equation. We also note that $J_\infty \leq J_{\mu^0,\delta_0}$ and that $J_{\mu^0,\delta_0} \in S_p$ by assumption, so that $J_\infty \in S_p$. By Prop. 4.5.4(a), it follows that $J_\infty = \hat{J}_p$. **Q.E.D.**

Note that despite the fact $J_{\mu^k} \to \hat{J}_p$, the generated sequence $\{\mu^k\}$ may exhibit some serious pathologies in the limit. In particular, if U is a metric space and $\{\mu^k\}_{\mathcal{K}}$ is a subsequence of policies that converges to some $\bar{\mu}$, in the sense that

$$\lim_{k\to\infty,\, k\in\mathcal{K}} \mu^k(x) = \bar{\mu}(x), \qquad \forall\, x \in X,$$

it does not follow that $\bar{\mu}$ is p-stable. In fact it is possible to construct examples where the generated sequence of p-stable policies $\{\mu^k\}$ satisfies $\lim_{k\to\infty} J_{\mu^k} = \hat{J}_p = J^*$, yet $\{\mu^k\}$ may converge to a p-unstable policy whose cost function is strictly larger than $\hat{J}_p$.

An Optimistic Policy Iteration Method

Let us consider an optimistic variant of PI, where policies are evaluated inexactly, with a finite number of VIs. We use a fixed forcing function p. The algorithm aims to compute $\hat{J}_p$, the restricted optimal cost function over the p-stable policies, and generates a sequence $\{J_k, \mu^k\}$ according to

$$T_{\mu^k} J_k = T J_k, \qquad J_{k+1} = T_{\mu^k}^{m_k} J_k, \qquad k = 0, 1, \ldots, \tag{4.77}$$

where m_k is a positive integer for each k. We assume that a policy μ^k satisfying $T_{\mu^k} J_k = T J_k$ can be found for all k, but it need not be p-stable. However, the algorithm requires that

$$J_0 \in \mathcal{W}_p, \qquad J_0 \geq T J_0. \tag{4.78}$$

This may be a restrictive assumption. We have the following proposition.

Proposition 4.5.9: (Convergence of Optimistic PI) Assume that there exists at least one p-stable policy $\pi \in \Pi_p$, and that J_0 satisfies Eq. (4.78). Then a sequence $\{J_k\}$ generated by the optimistic PI algorithm (4.77) belongs to $\mathcal{W}_p$ and satisfies $J_k \downarrow \hat{J}_p$.

Proof: Since $J_0 \geq \hat{J}_p$ and $\hat{J}_p = T\hat{J}_p$ [cf. Prop. 4.5.6(a)], all operations on any of the functions J_k with T_{μ^k} or T maintain the inequality $J_k \geq \hat{J}_p$ for all k, so that $J_k \in \mathcal{W}_p$ for all k. Also the conditions $J_0 \geq T J_0$ and $T_{\mu^k} J_k = T J_k$ imply that

$$J_0 = J_1 \geq T_{\mu^0}^{m_0+1} J_0 = T_{\mu^0} J_1 \geq T J_1 = T_{\mu^1} J_1 \geq \cdots \geq J_2,$$

and continuing similarly,

$$J_k \geq T J_k \geq J_{k+1}, \qquad k = 0, 1, \ldots. \tag{4.79}$$

Thus $J_k \downarrow J_\infty$ for some J_∞, which must satisfy $J_\infty \geq \hat{J}_p$, and hence belong to $\mathcal{W}_p$. By taking limit as $k \to \infty$ in Eq. (4.79) and using an argument similar to the one in the proof of Prop. 4.5.8, it follows that $J_\infty = TJ_\infty$. By Prop. 4.5.6(a), this implies that $J_\infty \leq \hat{J}_p$. Together with the inequality $J_\infty \geq \hat{J}_p$ shown earlier, this proves that $J_\infty = \hat{J}_p$. **Q.E.D.**

As an example, for the shortest path problem of Example 4.5.3, the reader may verify that in the case where $p(x) = 1$ for $x = 1$, the optimistic PI algorithm converges in a single iteration to

$$\hat{J}_p(x) = \begin{cases} 1 & \text{if } x = 1, \\ 0 & \text{if } x = 0, \end{cases}$$

provided that $J_0 \in \mathcal{W}_p = \big\{J \mid J(1) \geq 1,\ J(0) = 0\big\}$. For other starting functions J_0, the algorithm converges in a single iteration to the function

$$J_\infty(1) = \min\big\{1,\, J_0(1)\big\}, \qquad J_\infty(0) = 0.$$

All functions J_∞ of the form above are solutions of Bellman's equation, but only $\hat{J}_p$ is restricted optimal.

4.6 INFINITE-SPACES STOCHASTIC SHORTEST PATH PROBLEMS

In this section we consider a stochastic discrete-time infinite horizon optimal control problem involving the system

$$x_{k+1} = f(x_k, u_k, w_k), \qquad k = 0, 1, \ldots, \tag{4.80}$$

where x_k and u_k are the state and control at stage k, which belong to sets X and U, w_k is a random disturbance that takes values in a countable set W with given probability distribution $P(w_k \mid x_k, u_k)$, and $f : X \times U \times W \mapsto X$ is a given function (cf. Example 1.2.1 in Chapter 1). The state and control spaces X and U are arbitrary, but we assume that W is countable to bypass complex measurability issues in the choice of control (see [BeS78]).

The control u must be chosen from a constraint set $U(x) \subset U$ that may depend on x. The expected cost per stage, $E\big\{g(x, u, w)\big\}$, is assumed nonnegative:

$$0 \leq E\big\{g(x, u, w)\big\} < \infty, \qquad \forall\ x \in X,\ u \in U(x),\ w \in W.$$

We assume that X contains a special cost-free and absorbing state t, referred to as the *destination*:

$$f(t, u, w) = t, \qquad g(t, u, w) = 0, \qquad \forall\ u \in U(t),\ w \in W.$$

This is a special case of an SSP problem, where the cost per stage is nonnegative, but the state and control spaces are arbitrary. It is also a special case of the nonnegative cost stochastic optimal control problem of Section 4.4.2. We adopt the notation and terminology of that section, but we review it here briefly for convenience.

Given an initial state x_0, a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ when applied to the system (4.80), generates a random sequence of state-control pairs $\big(x_k, \mu_k(x_k)\big)$, $k = 0, 1, \ldots$, with cost

$$J_\pi(x_0) = E^\pi_{x_0}\left\{\sum_{k=0}^{\infty} g\big(x_k, \mu_k(x_k), w_k\big)\right\},$$

where $E^\pi_{x_0}\{\cdot\}$ denotes expectation with respect to the probability measure corresponding to initial state x_0 and policy π. For a stationary policy μ, the corresponding cost function is denoted by J_μ. The optimal cost function is

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

and its effective domain is denoted X^*, i.e.,

$$X^* = \big\{x \in X \mid J^*(x) < \infty\big\}.$$

A policy π^* is said to be optimal if $J_{\pi^*}(x) = J^*(x)$ for all $x \in X$.

We denote by $\mathcal{E}^+(X)$ the set of functions $J : X \mapsto [0, \infty]$. In our analysis, we will use the set of functions

$$\mathcal{J} = \big\{J \in \mathcal{E}^+(X) \mid J(t) = 0\big\}.$$

Since t is cost-free and absorbing, this set contains the cost functions J_π of all $\pi \in \Pi$, as well as J^*.

Here the results of Section 4.3 under Assumption I apply, and the optimal cost function J^* is a solution of the Bellman equation

$$J(x) = \inf_{u\in U(x)} E\Big\{g(x, u, w) + J\big(f(x, u, w)\big)\Big\}, \qquad x \in X,$$

where the expected value is with respect to the distribution $P(w \mid x, u)$. Moreover, an optimal stationary policy (if it exists) may be obtained through the minimization in the right side of this equation (with J replaced by J^*, cf. Prop. 4.4.4). The VI algorithm starts from some function $J_0 \in \mathcal{J}$, and generates a sequence $\{J_k\} \subset \mathcal{J}$ according to

$$J_{k+1}(x) = \inf_{u\in U(x)} E\Big\{g(x, u, w) + J_k\big(f(x, u, w)\big)\Big\}, \qquad x \in X,\ k = 0, 1, \ldots.$$

Proper Policies and the δ-Perturbed Problem

We will now introduce a notion of proper policy with a definition that extends the one used for finite-state SSP in Section 3.5.1. For a given state $x \in X$, a policy π is said to be *proper at* x if

$$J_\pi(x) < \infty, \qquad \sum_{k=0}^{\infty} r_k(\pi, x) < \infty, \tag{4.81}$$

where $r_k(\pi, x)$ is the probability that $x_k \neq t$ when using π and starting from $x_0 = x$. We denote by $\widehat{\Pi}_x$ the set of all policies that are proper at x, and we denote by $\hat{J}$ the corresponding restricted optimal cost function,

$$\hat{J}(x) = \inf_{\pi \in \widehat{\Pi}_x} J_\pi(x), \qquad x \in X,$$

(with the convention that the infimum over the empty set is ∞). Finally we denote by $\widehat{X}$ the effective domain of $\hat{J}$, i.e.,

$$\widehat{X} = \big\{x \in X \mid \hat{J}(x) < \infty\big\}. \tag{4.82}$$

Note that $\widehat{X}$ is the set of x such that $\widehat{\Pi}_x$ is nonempty and that $t \in \widehat{X}$.

For any $\delta > 0$, let us consider the *δ-perturbed optimal control problem*. This is the same problem as the original, except that the cost per stage is changed to

$$g(x, u, w) + \delta, \qquad \forall\ x \neq t,$$

while $g(x, u, w)$ is left unchanged at 0 when $x = t$. Thus t is still cost-free as well as absorbing in the δ-perturbed problem. The δ-perturbed cost function of a policy π starting from x is denoted by $J_{\pi,\delta}(x)$ and involves an additional expected cost $\delta r_k(\pi, x)$ for each stage k, so that

$$J_{\pi,\delta}(x) = J_\pi(x) + \delta \sum_{k=0}^{\infty} r_k(\pi, x).$$

Clearly, the sum $\sum_{k=0}^{\infty} r_k(\pi, x)$ is the expected number of steps to reach the destination starting from x and using π, if the sum is finite. We denote by $\hat{J}_\delta$ the optimal cost function of the δ-perturbed problem, i.e., $\hat{J}_\delta(x) = \inf_{\pi \in \Pi} J_{\pi,\delta}(x)$. The following proposition provides some characterizations of proper policies in relation to the δ-perturbed problem.

Proposition 4.6.1:

(a) A policy π is proper at a state $x \in X$ if and only if $J_{\pi,\delta}(x) < \infty$ for all $\delta > 0$.

(b) We have $\hat{J}_\delta(x) < \infty$ for all $\delta > 0$ if and only if $x \in \widehat{X}$.

(c) For every $\epsilon > 0$ and $\delta > 0$, a policy π_ϵ that is ϵ-optimal for the δ-perturbed problem is proper at all $x \in \widehat{X}$, and such a policy exists.

Proof: (a) Follows from Eq. (4.50) and the definition (4.81) of a proper policy.

(b) If $x \in \widehat{X}$ there exists a policy π that is proper at x, and by part (a), $\hat{J}_\delta(x) \leq J_{\pi,\delta}(x) < \infty$ for all $\delta > 0$. Conversely, if $\hat{J}_\delta(x) < \infty$, there exists π such that $J_{\pi,\delta}(x) < \infty$, implying [by part (a)] that $\pi \in \widehat{\Pi}_x$, so that $x \in \widehat{X}$.

(c) An ϵ-optimal π_ϵ exists by Prop. 4.4.4(e). We have $J_{\pi_\epsilon,\delta}(x) \leq \hat{J}_\delta(x) + \epsilon$ for all $x \in X$. Hence $J_{\pi_\epsilon,\delta}(x) < \infty$ for all $x \in \widehat{X}$, implying by part (a) that π_ϵ is proper at all $x \in \widehat{X}$. **Q.E.D.**

The next proposition shows that the cost function $\hat{J}_\delta$ of the δ-perturbed problem can be used to approximate $\hat{J}$.

Proposition 4.6.2: We have $\lim_{\delta \downarrow 0} \hat{J}_\delta(x) = \hat{J}(x)$ for all $x \in X$. Moreover, for any $\epsilon > 0$ and $\delta > 0$, a policy π_ϵ that is ϵ-optimal for the δ-perturbed problem is ϵ-optimal within the class of proper policies, i.e., satisfies

$$J_{\pi_\epsilon}(x) \leq \hat{J}(x) + \epsilon, \qquad \forall\ x \in X.$$

Proof: Let us fix $\delta > 0$, and for a given $\epsilon > 0$, let π_ϵ be a policy that is proper at all $x \in \widehat{X}$ and is ϵ-optimal for the δ-perturbed problem [cf. Prop. 4.6.1(c)]. By using Eq. (4.50), we have for all $\epsilon > 0$, $x \in \widehat{X}$, and $\pi \in \widehat{\Pi}_x$,

$$\hat{J}(x) - \epsilon \leq J_{\pi_\epsilon}(x) - \epsilon \leq J_{\pi_\epsilon,\delta}(x) - \epsilon \leq \hat{J}_\delta(x) \leq J_{\pi,\delta}(x) = J_\pi(x) + w_{\pi,\delta}(x),$$

where

$$w_{\pi,\delta}(x) = \delta \sum_{k=0}^{\infty} r_k(\pi, x) < \infty, \qquad \forall\ x \in \widehat{X},\ \pi \in \widehat{\Pi}_x.$$

By taking the limit as $\epsilon \downarrow 0$, we obtain for all $\delta > 0$ and $\pi \in \widehat{\Pi}_x$,

$$\hat{J}(x) \leq \hat{J}_\delta(x) \leq J_\pi(x) + w_{\pi,\delta}(x), \qquad \forall\ x \in \widehat{X},\ \pi \in \widehat{\Pi}_x.$$

We have $\lim_{\delta \downarrow 0} w_{\pi,\delta}(x) = 0$ for all $x \in \widehat{X}$ and $\pi \in \widehat{\Pi}_x$, so by taking the limit as $\delta \downarrow 0$ and then the infimum over all $\pi \in \widehat{\Pi}_x$,

$$\hat{J}(x) \leq \lim_{\delta \downarrow 0} \hat{J}_\delta(x) \leq \inf_{\pi \in \widehat{\Pi}_x} J_\pi(x) = \hat{J}(x), \qquad \forall\ x \in \widehat{X},$$

from which $\hat{J}(x) = \lim_{\delta \downarrow 0} \hat{J}_\delta(x)$ for all $x \in \widehat{X}$. Moreover, by Prop. 4.6.1(b), $\hat{J}_\delta(x) = \hat{J}(x) = \infty$ for all $x \notin \widehat{X}$, so that $\hat{J}(x) = \lim_{\delta \downarrow 0} \hat{J}_\delta(x)$ for all $x \in X$.

We also have

$$J_{\pi_\epsilon}(x) \leq J_{\pi_\epsilon,\delta}(x) \leq \hat{J}_\delta(x)+\epsilon \leq J_\pi(x)+\delta\sum_{k=0}^{\infty} r(\pi,x)+\epsilon, \quad \forall\, x \in \widehat{X},\ \pi \in \widehat{\Pi}_x.$$

By taking the limit as $\delta \downarrow 0$, we obtain

$$J_{\pi_\epsilon}(x) \leq J_\pi(x) + \epsilon, \qquad \forall\, x \in \widehat{X},\ \pi \in \widehat{\Pi}_x.$$

By taking the infimum over $\pi \in \widehat{\Pi}_x$, it follows that $J_{\pi_\epsilon}(x) \leq \hat{J}(x) + \epsilon$ for all $x \in \widehat{X}$, which combined with the fact $J_{\pi_\epsilon}(x) = \hat{J}(x) = \infty$ for all $x \notin \widehat{X}$, yields the result. **Q.E.D.**

Main Results

By Prop. 4.4.4(a), $\hat{J}_\delta$ solves Bellman's equation for the δ-perturbed problem, while by Prop. 4.6.2, $\lim_{\delta\downarrow 0} \hat{J}_\delta(x) = \hat{J}(x)$. This suggests that $\hat{J}$ solves the unperturbed Bellman equation, which is the "limit" as $\delta \downarrow 0$ of the δ-perturbed version. Indeed we will show a stronger result, namely that $\hat{J}$ is the unique solution of Bellman's equation within the set of functions

$$\widehat{\mathcal{W}} = \{J \in S \mid \hat{J} \leq J\}, \tag{4.83}$$

where

$$S = \Big\{J \in \mathcal{J} \mid E_{x_0}^\pi\big\{J(x_k)\big\} \to 0,\ \forall\ (\pi, x_0) \text{ with } \pi \in \widehat{\Pi}_{x_0}\Big\}. \tag{4.84}$$

Here $E_{x_0}^\pi\big\{J(x_k)\big\}$ denotes the expected value of the function J along the sequence $\{x_k\}$ generated starting from x_0 and using π. Similar to earlier proofs in Sections 4.4 and 4.5, we have that the collection

$$\mathcal{C} = \big\{(\pi, x) \mid \pi \in \widehat{\Pi}_x\big\} \tag{4.85}$$

is S-regular.

We first show a preliminary result. Given a policy $\pi = \{\mu_0, \mu_1, \ldots\}$, we denote by π_k the policy

$$\pi_k = \{\mu_k, \mu_{k+1}, \ldots\}. \tag{4.86}$$

Proposition 4.6.3:

(a) For all pairs $(\pi, x_0) \in \mathcal{C}$ and $k = 0, 1, \ldots$, we have

$$0 \leq E_{x_0}^\pi\big\{\hat{J}(x_k)\big\} \leq E_{x_0}^\pi\big\{J_{\pi_k}(x_k)\big\} < \infty,$$

where π_k is the policy given by Eq. (4.86).

(b) The set $\widehat{\mathcal{W}}$ of Eq. (4.83) contains $\hat{J}$, as well as all functions $J \in S$ satisfying $\hat{J} \leq J \leq c\hat{J}$ for some $c \geq 1$.

Proof: (a) For any pair $(\pi, x_0) \in \mathcal{C}$ and $\delta > 0$, we have

$$J_{\pi,\delta}(x_0) = E^{\pi}_{x_0}\left\{J_{\pi_k,\delta}(x_k) + \sum_{m=0}^{k-1} g\big(x_m, \mu_m(x_m), w_m\big)\right\} + \delta \sum_{m=0}^{k-1} r_m(\pi, x_0).$$

Since $J_{\pi,\delta}(x_0) < \infty$ [cf. Prop. 4.6.1(a)], it follows that $E^{\pi}_{x_0}\big\{J_{\pi_k,\delta}(x_k)\big\} < \infty$. Hence for all x_k that can be reached with positive probability using π and starting from x_0, we have $J_{\pi_k,\delta}(x_k) < \infty$, implying [by Prop. 4.6.1(a)] that $(\pi_k, x_k) \in \mathcal{C}$. Hence $\hat{J}(x_k) \leq J_{\pi_k}(x_k)$ and by applying $E^{\pi}_{x_0}\{\cdot\}$, the result follows.

(b) We have for all $(\pi, x_0) \in \mathcal{C}$,

$$J_{\pi}(x_0) = E^{\pi}_{x_0}\Big\{g\big(x_0, \mu_0(x_0), w_0\big)\Big\} + E^{\pi}_{x_0}\big\{J_{\pi_1}(x_1)\big\}, \tag{4.87}$$

and for $m = 1, 2, \ldots,$

$$E^{\pi}_{x_0}\big\{J_{\pi_m}(x_m)\big\} = E^{\pi}_{x_0}\Big\{g\big(x_m, \mu_m(x_m), w_m\big)\Big\} + E^{\pi}_{x_0}\big\{J_{\pi_{m+1}}(x_{m+1})\big\}, \tag{4.88}$$

where $\{x_m\}$ is the sequence generated starting from x_0 and using π. By using repeatedly the expression (4.88) for $m = 1, \ldots, k-1$, and combining it with Eq. (4.87), we obtain for all $k = 1, 2, \ldots,$

$$J_{\pi}(x_0) = E^{\pi}_{x_0}\big\{J_{\pi_k}(x_k)\big\} + \sum_{m=0}^{k-1} E^{\pi}_{x_0}\Big\{g\big(x_m, \mu_m(x_m), w_m\big)\Big\}, \quad \forall\ (\pi, x_0) \in \mathcal{C}.$$

The rightmost term above tends to $J_{\pi}(x_0)$ as $k \to \infty$, so by using the fact $J_{\pi}(x_0) < \infty$, we obtain

$$E^{\pi}_{x_0}\big\{J_{\pi_k}(x_k)\big\} \to 0, \qquad \forall\ (\pi, x_0) \in \mathcal{C}.$$

By part (a), it follows that

$$E^{\pi}_{x_0}\big\{\hat{J}(x_k)\big\} \to 0, \qquad \forall\ (\pi, x_0) \in \mathcal{C},$$

so that $\hat{J} \in \widehat{\mathcal{W}}$. This also implies that

$$E^{\pi}_{x_0}\big\{J(x_k)\big\} \to 0, \qquad \forall\ (\pi, x_0) \in \mathcal{C},$$

if $\hat{J} \leq J \leq c\hat{J}$ for some $c \geq 1$. **Q.E.D.**

We can now prove our main result.

Proposition 4.6.4: Assume that either W is finite or there exists a $\delta > 0$ such that

$$E\Big\{g(x,u,w) + \hat{J}_\delta\big(f(x,u,w)\big)\Big\} < \infty, \qquad \forall\ x \in X^*,\ u \in U(x).$$

(a) $\hat{J}$ is the unique solution of the Bellman Eq. (4.65) within the set $\widehat{\mathcal{W}}$ of Eq. (4.83).

(b) (*VI Convergence*) If $\{J_k\}$ is the sequence generated by the VI algorithm (4.47) starting with some $J_0 \in \widehat{\mathcal{W}}$, then $J_k \to \hat{J}$.

(c) (*Optimality Condition*) If μ is a stationary policy that is proper at all $x \in \widehat{X}$ and

$$\mu(x) \in \arg\min_{u \in U(x)} E\Big\{g(x,u,w) + \hat{J}\big(f(x,u,w)\big)\Big\}, \qquad \forall\ x \in X, \tag{4.89}$$

then μ is optimal over the set of proper policies, i.e., $J_\mu = \hat{J}$. Conversely, if μ is proper at all $x \in \widehat{X}$ and $J_\mu = \hat{J}$, then μ satisfies the preceding condition (4.89).

Proof: (a), (b) By Prop. 4.6.3(b), $\hat{J} \in \widehat{\mathcal{W}}$. We will first show that $\hat{J}$ is a solution of Bellman's equation. Since $\hat{J}_\delta$ solves the Bellman equation for the δ-perturbed problem, and $\hat{J}_\delta \geq \hat{J}$ (cf. Prop. 4.6.2), we have for all $\delta > 0$ and $x \neq t$,

$$\begin{aligned}
\hat{J}_\delta(x) &= \inf_{u \in U(x)} E\Big\{g(x,u,w) + \delta + \hat{J}_\delta\big(f(x,u,w)\big)\Big\} \\
&\geq \inf_{u \in U(x)} E\Big\{g(x,u,w) + \hat{J}_\delta\big(f(x,u,w)\big)\Big\} \\
&\geq \inf_{u \in U(x)} E\Big\{g(x,u,w) + \hat{J}\big(f(x,u,w)\big)\Big\}.
\end{aligned}$$

By taking the limit as $\delta \downarrow 0$ and using Prop. 4.6.2, we obtain

$$\hat{J}(x) \geq \inf_{u \in U(x)} E\Big\{g(x,u,w) + \hat{J}\big(f(x,u,w)\big)\Big\}, \qquad \forall\ x \in X. \tag{4.90}$$

For the reverse inequality, let $\{\delta_m\}$ be a sequence with $\delta_m \downarrow 0$. We have for all m, $x \neq t$, and $u \in U(x)$,

$$\begin{aligned}
E\Big\{g(x,u,w) + \delta_m + \hat{J}_{\delta_m}\big(f(x,u,w)\big)\Big\} &\geq \inf_{v \in U(x)} E\Big\{g(x,v,w) + \delta_m \\
&\qquad + \hat{J}_{\delta_m}\big(f(x,v,w)\big)\Big\} \\
&= \hat{J}_{\delta_m}(x).
\end{aligned}$$

We now take limit as $m \to \infty$ in the preceding relation, and we interchange limit and expectation (our assumptions allow the use of the monotone convergence theorem for this purpose; Exercise 4.11 illustrates the need for these assumptions). Using also the fact $\lim_{\delta_m \downarrow 0} \hat{J}_{\delta_m} = \hat{J}$ (cf. Prop. 4.6.2), we have

$$E\Big\{g(x,u,w) + \hat{J}\big(f(x,u,w)\big)\Big\} \geq \hat{J}(x), \qquad \forall\ x \in X,\ u \in U(x),$$

so that

$$\inf_{u \in U(x)} E\Big\{g(x,u,w) + \hat{J}\big(f(x,u,w)\big)\Big\} \geq \hat{J}(x), \qquad \forall\ x \in X. \tag{4.91}$$

By combining Eqs. (4.90) and (4.91), we see that $\hat{J}$ is a solution of Bellman's equation.

Part (b) follows by using the S-regularity of the collection (4.85) and Prop. 4.4.2(b). Finally, since $\hat{J} \in \widehat{\mathcal{W}}$ and $\hat{J}$ is a solution of Bellman's equation, part (b) implies the uniqueness assertion of part (a).

(c) If μ is proper at all $x \in \widehat{X}$ and Eq. (4.89) holds, then

$$\hat{J}(x) = E\Big\{g\big(x,\mu(x),w\big) + \hat{J}\big(f(x,\mu(x),w)\big)\Big\}, \qquad x \in X.$$

By Prop. 4.4.4(b), this implies that $J_\mu \leq \hat{J}$, so μ is optimal over the set of proper policies. Conversely, assume that μ is proper at all $x \in \widehat{X}$ and $J_\mu = \hat{J}$. Then by Prop. 4.4.4(b), we have

$$\hat{J}(x) = E\Big\{g\big(x,\mu(x),w\big) + \hat{J}\big(f(x,\mu(x),w)\big)\Big\}, \qquad x \in X,$$

while [by part (a)] $\hat{J}$ is a solution of Bellman's equation,

$$\hat{J}(x) = \inf_{u \in U(x)} E\Big\{g(x,u,w) + \hat{J}\big(f(x,u,w)\big)\Big\}, \qquad x \in X.$$

Combining the last two relations, we obtain Eq. (4.89). **Q.E.D.**

We illustrate Prop. 4.6.4 in Fig. 4.6.1. Let us consider now the favorable case where the set of proper policies is sufficient in the sense that it can achieve the same optimal cost as the set of all policies, i.e., $\hat{J} = J^*$. This is true for example if all policies are proper at all x such that $J^*(x) < \infty$. Moreover it is true in some of the finite-state formulations of SSP that we discussed in Chapter 3; see also the subsequent Prop. 4.6.5. When $\hat{J} = J^*$, it follows from Prop. 4.6.4 that J^* is the unique solution of Bellman's equation within $\widehat{\mathcal{W}}$, and that the VI algorithm converges to J^* starting from any $J_0 \in \widehat{\mathcal{W}}$. Under an additional compactness condition, such as finiteness

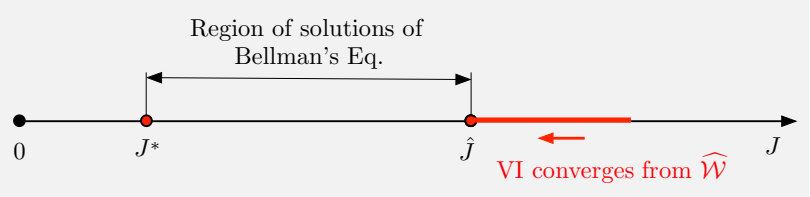


Figure 4.6.1 Illustration of the solutions of Bellman's equation. All solutions either lie between J^* and $\hat{J}$, or they lie outside the set $\widehat{\mathcal{W}}$. The VI algorithm converges to $\hat{J}$ starting from any $J_0 \in \widehat{\mathcal{W}}$.

of $U(x)$ for all $x \in X$ [cf. Prop. 4.4.4(e)], VI converges to J^* starting from any J_0 in the set S of Eq. (4.84).

Proposition 4.6.4 does not say anything about the existence of a proper policy that is optimal within the class of proper policies. For a simple example where $J^* = \hat{J}$ but the only optimal policy is improper, consider a deterministic shortest path problem with a single state 1 plus the destination t. At state 1 we may choose $u \in [0,1]$ with cost u, and move to t if $u \neq 0$ and stay at 1 if $u = 0$. Note that here we have $J^*(1) = \hat{J}(1) = 0$, and the minimum over $u \in [0,1]$ is attained in Bellman's equation, which has the form

$$J^*(1) = \min\left\{\inf_{u\in(0,1]} u,\ J^*(1)\right\}.$$

However, the only optimal policy (staying at 1) is improper.

4.6.1 The Multiplicity of Solutions of Bellman's Equation

Let us now discuss the issue of multiplicity of solutions of Bellman's equation within the set of functions

$$\mathcal{J} = \big\{J \in \mathcal{E}^+(X) \mid J(t) = 0\big\}.$$

We know from Props. 4.4.4(a) and 4.6.4(a) that J^* and $\hat{J}$ are solutions, and that all other solutions J must satisfy either $J^* \leq J \leq \hat{J}$ or $J \notin \widehat{\mathcal{W}}$.

In the special case of a deterministic problem (one where the disturbance w_k takes a single value), it was shown in Section 4.5 that $\hat{J}$ is the largest solution of Bellman's equation within $\mathcal{J}$, so all solutions $J' \in \mathcal{J}$ satisfy $J^* \leq J' \leq \hat{J}$. It was also shown through examples that there can be any number of solutions that lie between J^* and $\hat{J}$: a finite number, an infinite number, or none at all.

In stochastic problems, however, the situation is strikingly different in the following sense: there can be an infinite number of solutions that do not lie below $\hat{J}$, i.e., solutions $J' \in \mathcal{J}$ that do not satisfy $J' \leq \hat{J}$. Of course, by Prop. 4.6.4(a), these solutions must lie outside $\widehat{\mathcal{W}}$. The following example, which involves a finite set W, is an illustration.

Example 4.6.1

Let $X = \Re$, $t = 0$, and assume that there is only one control at each state, and hence a single policy π. The disturbance w_k takes two values: 1 and 0 with probabilities $\alpha \in (0,1)$ and $1-\alpha$, respectively. The system equation is

$$x_{k+1} = \frac{w_k x_k}{\alpha},$$

and there is no cost at each state and stage:

$$g(x,u,w) \equiv 0.$$

Thus from state x_k we move to state x_k/α with probability α and to the termination state $t = 0$ with probability $1-\alpha$.

Here, the unique policy is stationary and proper at all $x \in X$, and we have

$$J^*(x) = \hat{J}(x) = 0, \qquad \forall\ x \in X.$$

Bellman's equation has the form

$$J(x) = (1-\alpha)J(0) + \alpha J\left(\frac{x}{a}\right),$$

which within $\mathcal{J}$ reduces to

$$J(x) = \alpha J\left(\frac{x}{\alpha}\right), \qquad \forall\ J \in \mathcal{J},\ x \in X. \tag{4.92}$$

It can be seen that Bellman's equation has an infinite number of solutions within $\mathcal{J}$ in addition to J^* and $\hat{J}$: any positively homogeneous function, such as, for example,

$$J(x) = \gamma|x|, \qquad \gamma > 0,$$

is a solution. Consistently with Prop. 4.6.4(a), none of these solutions belongs to $\widehat{\mathcal{W}}$, since x_k is either equal to x_0/α^k (with probability α^k) or equal to 0 (with probability $1-\alpha^k$). For example, in the case of $J(x) = \gamma|x|$, we have

$$E^{\pi}_{x_0}\left\{J(x_k)\right\} = \alpha^k \gamma \left|\frac{x_0}{\alpha^k}\right| = \gamma|x_0|, \qquad \forall\ k \geq 0,$$

so $J(x_k)$ does not converge to 0, unless $x_0 = 0$. Moreover, none of these additional solutions seems to be significant in some discernible way.

The preceding example illustrates an important structural difference between deterministic and stochastic shortest path problems with infinite state space. For a terminating policy μ in the context of the deterministic problem of Section 4.5, the corresponding Bellman equation $J = T_\mu J$ has a unique solution within $\mathcal{J}$ [to see this, consider the restricted problem for which μ is the only policy, and apply Prop. 4.5.6(a)]. By contrast, for a proper policy in the stochastic context of the present section, the corresponding Bellman equation may have an infinite number of solutions within $\mathcal{J}$, as Example 4.6.1 shows. This discrepancy does not occur when the state space is finite, as we have seen in Section 3.5.1. We will next elaborate on the preceding observations and refine our analysis regarding multiplicity of solutions of Bellman's equation for problems where the cost per stage is bounded.

4.6.2 The Case of Bounded Cost per Stage

Let us consider the special case where the cost per stage g is bounded over $X \times U \times W$, i.e.,

$$\sup_{(x,u,w)\in X\times U\times W} g(x,u,w) < \infty. \tag{4.93}$$

We will show that $\hat{J}$ is the largest solution of Bellman's equation within the class of functions that are bounded over the effective domain $\widehat{X}$ of $\hat{J}$ [cf. Eq. (4.82)].

We say that a policy π is *uniformly proper* if there is a uniform bound on the expected number of steps to reach the destination from states $x \in \widehat{X}$ using π:

$$\sup_{x\in\widehat{X}} \sum_{k=0}^{\infty} r_k(\pi, x) < \infty.$$

Since we have

$$J_\pi(x_0) \leq \left(\sup_{(x,u,w)\in X\times U\times W} g(x,u,w) \right) \cdot \sum_{k=0}^{\infty} r_k(\pi, x_0) < \infty, \quad \forall\ \pi \in \widehat{\Pi}_{x_0},$$

it follows that the cost function J_π of a uniformly proper π belongs to the set $\mathcal{B}$, defined by

$$\mathcal{B} = \left\{ J \in \mathcal{J} \;\middle|\; \sup_{x\in\widehat{X}} J(x) < \infty \right\}. \tag{4.94}$$

When $\widehat{X} = X$, the notion of a uniformly proper policy coincides with the notion of a transient policy used in [Pli78] and [JaC06], which itself descends from earlier works. However, our definition is somewhat more general, since it also applies to the case where $\widehat{X}$ is a strict subset of X.

Let us denote by $\widehat{\mathcal{W}}_b$ the set of functions

$$\widehat{\mathcal{W}}_b = \{J \in \mathcal{B} \mid \hat{J} \leq J\}.$$

The following proposition, illustrated in Fig. 4.6.2, provides conditions for $\hat{J}$ to be the largest fixed point of T within $\mathcal{B}$. Its assumptions include the existence of a uniformly proper policy, which implies that $\hat{J}$ belongs to $\mathcal{B}$. The proposition also uses the earlier Prop. 4.4.6 in order to provide conditions for $J^* = \hat{J}$, in which case J^* is the unique fixed point of T within $\mathcal{B}$.

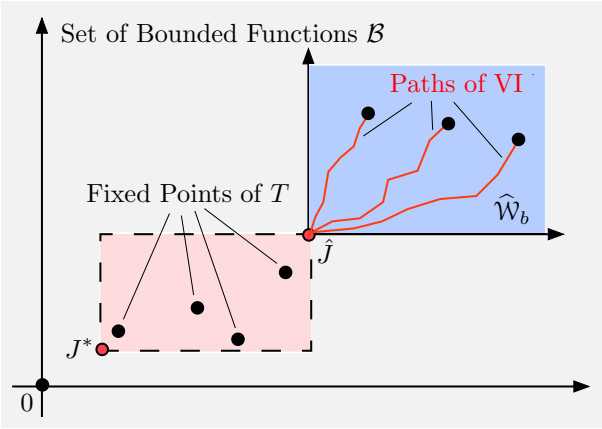


Figure 4.6.2. Schematic illustration of Prop. 4.6.5 for a nonnegative cost SSP problem. The functions J^* and $\hat{J}$ are the smallest and largest solutions, respectively, of Bellman's equation within the set $\mathcal{B}$. Moreover, the VI algorithm converges to $\hat{J}$ starting from $J_0 \in \widehat{\mathcal{W}}_b = \{J \in \mathcal{B} \mid \hat{J} \leq J\}$.

Proposition 4.6.5: Let the assumptions of Prop. 4.6.4 hold, and assume further that the cost per stage g is bounded over $X \times U \times W$ [cf. Eq. (4.93)], and that there exists a uniformly proper policy. Then:

(a) $\hat{J}$ is the largest solution of the Bellman Eq. (4.65) within the set $\mathcal{B}$ of Eq. (4.94), i.e., $\hat{J}$ is a solution that belongs to $\mathcal{B}$ and if $J' \in \mathcal{B}$ is another solution, then $J' \leq \hat{J}$. Moreover, if $\hat{J} = J^*$, then J^* is the unique solution of Bellman's equation within $\mathcal{B}$.

(b) If $\{J_k\}$ is the sequence generated by the VI algorithm (4.47) starting with some $J_0 \in \mathcal{B}$ with $J_0 \geq \hat{J}$, then $J_k \to \hat{J}$.

(c) Assume in addition that X is finite, that $J^*(x) > 0$ for all $x \neq t$, and that $X^* = \widehat{X}$. Then $\hat{J} = J^*$.

Proof: (a) Since the cost function of a uniformly proper policy belongs to $\mathcal{B}$, we have $\hat{J} \in \mathcal{B}$. On the other hand, for all $J \in \mathcal{B}$, we have

$$E_{x_0}^{\pi}\{J(x_k)\} \leq \left(\sup_{x \in \widehat{X}} J(x)\right) \cdot r_k(\pi, x_0) \to 0, \qquad \forall\, \pi \in \widehat{\Pi}_{x_0}.$$

It follows that the set $\widehat{\mathcal{W}}_b$ is contained in $\widehat{\mathcal{W}}$, while the function $\hat{J}$ belongs to $\widehat{\mathcal{W}}_b$. Since $\widehat{\mathcal{W}}_b$ is unbounded above within the set $\mathcal{B}$, for every solution $J' \in \mathcal{B}$ of Bellman's equation we have $J' \leq J$ for some $J \in \widehat{\mathcal{W}}_b$, and hence also $J' \leq \tilde{J}$ for some $\tilde{J}$ in the set S of Eq. (4.84). It follows from Prop. 4.4.2(a) and the S-regularity of the collection (4.85) that $J' \leq \hat{J}$.

If in addition $\hat{J} = J^*$, from Prop. 4.4.4(a), $\hat{J}$ is also the smallest solution of Bellman's equation within $\mathcal{J}$. Hence J^* is the unique solution of Bellman's equation within $\mathcal{B}$.

(b) Follows from Prop. 4.6.4(b), since $\widehat{\mathcal{W}}_b \subset \widehat{\mathcal{W}}$, as shown in the proof of part (a).

(c) We have by assumption

$$0 < J^*(x) \leq \hat{J}(x), \qquad \forall\ x \neq t,$$

while $\hat{J}(x) < \infty$ for all $x \in X^*$ since $X^* = \widehat{X}$. In view of the finiteness of X, we can find a sufficiently large c such that $\hat{J} \leq cJ^*$, so by Prop. 4.4.6, it follows that $\hat{J} = J^*$. **Q.E.D.**

The uniqueness of solution of Bellman's equation within $\mathcal{B}$ when $\hat{J} = J^*$ [cf. part (a) of the preceding proposition] is consistent with Example 4.6.1. In that example, J^* and $\hat{J}$ are equal and bounded, and all the additional solutions of Bellman's equation are unbounded, as can be verified by using Eq. (4.92).

Note that without the assumption of existence of a uniformly proper π, $\hat{J}$ and J^* need not belong to $\mathcal{B}$. As an example, let X be the set of nonnegative integers, let $t = 0$, and let there be a single policy that moves the system deterministically from a state $x \geq 1$ to the state $x - 1$ at cost $g(x, x-1) = 1$. Then

$$\hat{J}(x) = J^*(x) = x, \qquad \forall\ x \in X,$$

so $\hat{J}$ and J^* do not belong to $\mathcal{B}$, even though g is bounded. Here the unique policy is proper at all x, but is not uniformly proper.

In a given practical application, we may be interested in computing either J^* or $\hat{J}$. If the cost per stage is bounded, we may compute $\hat{J}$ with the VI algorithm, assuming that an initial function in the set $\widehat{\mathcal{W}}_b$ can be found. The computation of J^* is also possible by using the VI algorithm and starting from the zero initial condition, assuming that the conditions of Prop. 4.4.4(d) are satisfied.

An alternative possibility for the case of a finite spaces SSP is to approximate the problem with a sequence of α_k-discounted problems where the discount factors α_k tend to 1. This approach, developed in some detail in Exercise 5.28 of the book [Ber17a], has the advantage that the discounted problems can be solved more reliably and with a broader variety of methods than the original undiscounted SSP.

Another technique, developed in the paper [BeY16], is to transform a finite-state SSP problem such that $J^*(x) = 0$ for some $x \neq t$ into an equivalent SSP problem that satisfies the conditions of Prop. 4.6.5(c), and thus allow the computation of J^* by a VI or PI algorithm. The idea is to lump t together with the states x for which $J^*(x) = 0$ into a single

state, which is the termination state for the equivalent SSP problem. This technique is strictly limited to finite-state problems, since in general the conditions $J^*(x) > 0$ for all $x \neq t$ and $X^* = \widehat{X}$ do not imply that $\hat{J} = J^*$, even under the bounded cost and uniform properness assumptions of this section (see the deterministic stopping Example 4.5.2).

4.7 NOTES, SOURCES, AND EXERCISES

Sections 4.1: The use of monotonicity as the foundational property of abstract DP models was initiated in the author's papers [Ber75], [Ber77].

Section 4.2: The finite horizon analysis of Section 4.2 was given in Chapter 3 of the monograph by Bertsekas and Shreve [BeS78].

Section 4.3: The monotone increasing and decreasing abstract DP models of Section 4.3 were introduced in the author's papers [Ber75], [Ber77]. Their analysis was also given in Chapter 5 of the monograph [BeS78].

Important examples of noncontractive infinite horizon models are the classical negative cost DP problems, analyzed by Blackwell [Bla65], and by Dubins and Savage [DuS65], and the positive cost DP problems analyzed in Strauch [Str66] (and also in Strauch's Ph.D. thesis, written under the supervision of Blackwell). The monograph by Bertsekas and Shreve [BeS78] provides a detailed treatment of these two models, which also resolves the associated measurability questions using the notion of universally measurable policies. The paper by Yu and Bertsekas [YuB15] provides a more recent analysis that addresses some issues regarding the convergence of the VI and PI algorithms that were left unresolved in the monograph [BeS78]. A simpler textbook treatment, which bypasses the measurability questions, is given in the author's [Ber12a], Chapter 4.

The compactness condition that guarantees convergence of VI to J^* starting with the initial condition $J_0 = \bar{J}$ under Assumption I (cf. Prop. 4.3.14) was obtained by the author in [Ber72] for reachability problems (see Exercise 4.5), and in [Ber75], [Ber77] for positive cost DP models; see also Schal [Sch75] and Whittle [Whi80]. A more refined analysis of the question of convergence of VI to J^* is possible. This analysis provides a necessary and sufficient condition for convergence, and improves over the compactness condition of Prop. 4.3.14. In particular, the following characterization is shown in [Ber77], Prop. 11 (see also [BeS78], Prop. 5.9):

For a set $C \subset X \times U \times \Re$, let $\Pi(C)$ be the projection of C onto $X \times \Re$:

$$\Pi(C) = \big\{(x, \lambda) \mid (x, u, \lambda) \in C \text{ for some } u \in U(x)\big\},$$

and denote also

$$\overline{\Pi(C)} = \big\{(x, \lambda) \mid \lambda_m \to \lambda \text{ for some sequence } \{\lambda_m\} \text{ with } \big\{(x, \lambda_m)\big\} \subset C\big\}.$$

Consider the sets $C_k \subset X \times U \times \Re$ given by

$$C_k = \big\{(x,u,\lambda) \mid H(x,u,T^k\bar{J}) \leq \lambda,\, x \in X,\, u \in U(x)\big\}, \qquad k = 0,1,\ldots.$$

Then under Assumption I we have $T^k\bar{J} \to J^*$ if and only if

$$\overline{\Pi\big(\cap_{k=0}^{\infty} C_k\big)} = \cap_{k=0}^{\infty}\overline{\Pi(C_k)}.$$

Moreover we have $T^k\bar{J} \to J^*$ and in addition there exists an optimal stationary policy if and only if

$$\Pi\big(\cap_{k=0}^{\infty} C_k\big) = \cap_{k=0}^{\infty}\overline{\Pi(C_k)}. \tag{4.95}$$

For a connection with Prop. 4.3.14, it can be shown that compactness of

$$U_k(x,\lambda) = \big\{u \in U(x) \big|\ H(x,u,T^k\bar{J}) \leq \lambda\big\}$$

implies Eq. (4.95) (see [Ber77], Prop. 12, or [BeS78], Prop. 5.10).

The analysis of convergence of VI to J^* under Assumption I and starting with an initial condition $J_0 \geq J^*$ is far more complicated than for the initial condition $J_0 = \bar{J}$. A principal reason for this is the multiplicity of solutions of Bellman's equation within the set $\big\{J \in \mathcal{E}^+(X) \mid J \geq \bar{J}\big\}$. We know that J^* is the smallest solution (cf. Prop. 4.4.9), and an interesting issue is the characterization of the largest solution and other solutions within some restricted class of functions of interest. We substantially resolved this question in Sections 4.5 and 4.6 for infinite-spaces deterministic and stochastic shortest path problems, respectively (as well in Sections 3.5.1 and 3.52 for finite-state stochastic shortest path and affine monotonic problems). Generally, optimal control problems with nonnegative cost per stage can typically be reduced to problems with a cost-free and absorbing termination state (see [BeY16] for an analysis of the finite-state case). However, the fuller characterization of the set of solutions of Bellman's equation for general abstract DP models under Assumption I requires further investigation.

Optimistic PI and λ-PI under Assumption D have not been considered prior to the 2013 edition of this book, and the corresponding analysis of Section 4.3.3 is new. See [BeI96], [ThS10a], [ThS10b], [Ber11b], [Sch11], [Ber16b] for analyses of λ-PI for discounted and SSP problems.

Section 4.4: The definition and analysis of regularity for nonstationary policies was introduced in the author's paper [Ber15]. We have primarily used regularity in this book to analyze the structure of the solution set of Bellman's equation, and to identify the region of attraction of value and policy iteration algorithms. This analysis is multifaceted, so it is worth summarizing here:

(a) We have characterized the fixed point properties of the optimal cost function J^* and the restricted optimal cost function $J_{\mathcal{C}}^*$ over S-regular

collections $\mathcal{C}$, for various sets S. While J^* and $J^*_{\mathcal{C}}$ need not be fixed points of T, they are fixed points in a large variety of interesting contexts (Sections 3.3-3.5 and 4.4-4.6).

(b) We have shown that when $J^* = J^*_{\mathcal{C}}$, then J^* is the unique solution of Bellman's equation in several interesting noncontractive contexts. In particular, Section 3.3 deals with an important case that covers among others, the most common type of stochastic shortest path problems. However, even when $J^* \neq J^*_{\mathcal{C}}$, the functions J^* and $J^*_{\mathcal{C}}$ often bound the set of solutions from below and/or from above (see Sections 3.5.1, 3.5.2, 4.5, 4.6).

(c) Simultaneously with the analysis of the fixed point properties of J^* and $J^*_{\mathcal{C}}$, we have used regularity to identify the region of convergence of value iteration. Often convergence to $J^*_{\mathcal{C}}$ can be shown from starting functions $J \geq J^*_{\mathcal{C}}$, assuming that $J^*_{\mathcal{C}}$ is a fixed point of T. In the favorable case where $J^* = J^*_{\mathcal{C}}$, convergence to J^* can often be shown from every starting function of interest. In addition regularity has been used to guarantee the validity of policy iteration algorithms that generate exclusively regular policies, and are guaranteed to converge to J^* or $J^*_{\mathcal{C}}$.

(d) We have been able to characterize some of the solutions of Bellman's equation, but not the entire set. Generally, there may exist an infinite number of solutions, and some of them may not be associated with an S-regular collection for any set S, unless we change the starting function $\bar{J}$ that is part of the definition of the cost function J_π of the policies. There is a fundamental difficulty here: the solutions of the Bellman equation $J = TJ$ do not depend on $\bar{J}$, but S-regularity of a collection of policy-state pairs depends strongly on $\bar{J}$. A sharper characterization of the solution set of Bellman's equation remains an open interesting question, in both specific problem contexts as well as in generality.

The use of regularity in the analysis of undiscounted and discounted stochastic optimal control in Sections 4.4.2 and 4.4.3 is new, and was presented in the author's paper [Ber15]. The analysis of convergent models in Section 4.4.4, under the condition

$$J^*(x) \geq \bar{J}(x) > -\infty, \qquad \forall\, x \in X,$$

is also new. A survey of stochastic optimal control problems under convergence conditions that are more general than the ones considered here is given by Feinberg [Fei02]. An analysis of convergent models for stochastic optimal control, which illustrates the broad range of pathological behaviors that can occur without the condition $J^* \geq \bar{J}$, is given in the paper by Yu [Yu15].

Section 4.5: This section follows the author's paper [Ber17a]. The issue of the connection of optimality with stability (and also with controllability and observability) was raised in the classic paper by Kalman [Kal60] in the context of linear-quadratic problems.

The set of solutions of the Riccati equation has been extensively investigated starting with the papers by Willems [Wil71] and Kucera [Kuc72], [Kuc73], which were followed up by several other works; see the book by Lancaster and Rodman [LaR95] for a comprehensive treatment. In these works, the "largest" solution of the Riccati equation is referred to as the "stabilizing" solution, and the stability of the corresponding policy is shown, although the author could not find an explicit statement in the literature regarding the optimality of this policy within the class of all linear stable policies. Also the lines of analysis of these works are tied to the structure of the linear-quadratic problem and are unrelated to our analysis of Section 4.5, which is based on semicontractive ideas.

Section 4.6: Proper policies for infinite-state SSP problems have been considered earlier in the works of Pliska [Pli78], and James and Collins [JaC06], where they are called "transient." There are a few differences between the frameworks of [Pli78], [JaC06] and Section 4.6, which impact on the results obtained. In particular, the papers [Pli78] and [JaC06] use a related (but not identical) definition of properness to the one of Section 4.6, while the notion of a transient policy used in [JaC06] coincides with the notion of a uniformly proper policy of Section 4.6.2 when $\widehat{X} = X$. Furthermore, [Pli78] and [JaC06] do not consider the notion of policy that is "proper at a state." The paper [Pli78] assumes that all policies are transient, that g is bounded, and that J^* is real-valued. The paper [JaC06] allows for notransient policies that have infinite cost from some initial states, and extends the analysis of Bertsekas and Tsitsiklis [BeT91] from finite state space to infinite state space (addressing also measurability issues). Also, [JaC06] allows the cost per stage g to take both positive and negative values, and uses assumptions that guarantee that $J^* = \hat{J}$, that J^* is real-valued, and that improper policies cannot be optimal. Instead, in Section 4.6 we allow that $J^* \neq \hat{J}$ and that J^* can take the value ∞, while requiring that g is nonnegative and that the disturbance space W is countable.

The analysis of Section 4.6 comes from the author's paper [Ber17b], and is most closely related to the SSP analysis under the weak conditions of Section 3.5.1, where we assumed that the state space is finite, but allowed g to take both positive and negative values. The extension of some of our results of Section 4.6 to SSP problems where g takes both positive and negative values may be possible; Exercises 4.8 and 4.9 suggest some research directions. However, our analysis of infinite-spaces SSP problems in this chapter relies strongly on the nonnegativity of g and cannot be extended without major modifications. In this connection, it is worth mentioning the example of Section 3.1.2, which shows that J^* may not be a solution

of Bellman's equation when g can take negative values.

E X E R C I S E S

4.1 (Example of Nonexistence of an Optimal Policy Under D)

This is an example of a deterministic stopping problem where Assumption D holds, and an optimal policy does not exist, even though only two controls are available at each state (stop and continue). The state space is $X = \{1, 2, \ldots\}$. Continuation from state x leads to state $x + 1$ with certainty and no cost, while the stopping cost is $-1 + (1/x)$, so that there is an incentive to delay stopping at every state. Here for all x, $\bar{J}(x) = 0$, and

$$H(x, u, J) = \begin{cases} J(x+1) & \text{if } u = \text{continue}, \\ -1 + (1/x) & \text{if } u = \text{stop}. \end{cases}$$

Show that $J^*(x) = -1$ for all x, but there is no policy (stationary or not) that attains the optimal cost starting from x.

Solution: Since a cost is incurred only upon stopping, and the stopping cost is greater than -1, we have $J_\mu(x) > -1$ for all x and μ. On the other hand, starting from any state x and stopping at $x + n$ yields a cost $-1 + \frac{1}{x+n}$, so by taking n sufficiently large, we can attain a cost arbitrarily close to -1. Thus $J^*(x) = -1$ for all x, but no policy can attain this optimal cost.

4.2 (Counterexample for Optimality Condition Under D)

For the problem of Exercise 4.1, show that the policy μ that never stops is not optimal but satisfies $T_\mu J^* = TJ^*$.

Solution: We have $J^*(x) = -1$ and $J_\mu(x) = 0$ for all $x \in X$. Thus μ is nonoptimal, yet attains the minimum in Bellman's equation

$$J^*(x) = \min \left\{ J^*(x+1),\ -1 + \frac{1}{x} \right\}$$

for all x.

4.3 (Counterexample for Optimality Condition Under I)

Let

$$X = \Re, \qquad U(x) \equiv (0, 1], \qquad \bar{J}(x) \equiv 0,$$

$$H(x, u, J) = |x| + J(ux), \qquad \forall\ x \in X,\ u \in U(x).$$

Let $\mu(x) = 1$ for all $x \in X$. Then $J_\mu(x) = \infty$ if $x \neq 0$ and $J_\mu(0) = 0$. Verify that $T_\mu J_\mu = T J_\mu$. Verify also that $J^*(x) = |x|$, and hence μ is not optimal.

Solution: The verification of $T_\mu J_\mu = T J_\mu$ is straightforward. To show that $J^*(x) = |x|$, we first note that $|x|$ is a fixed point of T, so by Prop. 4.3.2, $J^*(x) \leq |x|$. Also $(T\bar{J})(x) = |x|$ for all x, while under Assumption I, we have $J^* \geq T\bar{J}$, so $J^*(x) \geq |x|$. Hence $J^*(x) = |x|$.

4.4 (Solution by Mathematical Programming)

This exercise shows that under Assumptions I and D, it is possible to use a computational method based on mathematical programming when $X = \{1, \ldots, n\}$.

(a) Under Assumption I, show that J^* is the unique solution of the following optimization problem in $z = (z_1, \ldots, z_n)$:

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{n} z_i \\ &\text{subject to} \quad z_i \geq \bar{J}(i), \quad z_i \geq \inf_{u \in U(i)} H(i,u,z), \qquad i = 1, \ldots, n. \end{aligned}$$

(b) Under Assumption D, show that J^* is the unique solution of the following optimization problem in $z = (z_1, \ldots, z_n)$:

$$\begin{aligned} &\text{maximize} \quad \sum_{i=1}^{n} z_i \\ &\text{subject to} \quad z_i \leq \bar{J}(i), \quad z_i \leq H(i,u,z), \qquad i = 1, \ldots, n, \quad u \in U(i). \end{aligned}$$

Note: Generally, these programs may not be linear or even convex.

Solution: (a) Any feasible solution z of the given optimization problem satisfies $z \geq \bar{J}$ as well as $z_i \geq \inf_{u \in U(i)} H(i,u,z)$ for all $i = 1, \ldots, n$, so that $z \geq Tz$. It follows from Prop. 4.4.9 that $z \geq J^*$, which implies that J^* is an optimal solution of the given optimization problem. Also J^* is the unique optimal solution since if z is feasible and $z \neq J^*$, the inequality $z \geq J^*$ implies that $\sum_i z_i > \sum_i J^*(i)$, so z cannot be optimal.

(b) Any feasible solution z of the given optimization problem satisfies $z \leq \bar{J}$ as well as $z_i \leq H(i,u,z)$ for all $i = 1, \ldots, n$ and $u \in U(i)$, so that $z \leq Tz$. It follows from Prop. 4.3.6 that $z \leq J^*$, which implies that J^* is an optimal solution of the given optimization problem. Similar to part (a), J^* is the unique optimal solution.

4.5 (Infinite Time Reachability [Ber71], [Ber72])

This exercise provides an instance of an interesting problem where the mapping H is naturally extended real-valued. Consider a dynamic system

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

where w_k is viewed as an uncertain disturbance that may be any point in a set $W(x_k, u_k)$ (this is known in the literature as an "unknown but bounded" disturbance, and is the basis for a worst case/minimax treatment of uncertainty in the control of uncertain dynamic systems). We introduce an abstract DP model where the objective is to find a policy that keeps the state x_k of the system within a given set X at all times, for all possible values of the sequence $\{w_k\}$. This is a common objective, which arises in a variety of control theory contexts, including model predictive control (see [Ber17a], Section 6.4.3).

Let

$$\bar{J}(x) = \begin{cases} 0 & \text{if } x \in X, \\ \infty & \text{otherwise,} \end{cases}$$

and

$$H(x,u,J) = \begin{cases} 0 & \text{if } J(x) = 0,\ u \in U(x), \text{ and } J\big(f(x,u,w)\big) = 0,\ \forall\ w \in W(x,u), \\ \infty & \text{otherwise.} \end{cases}$$

(a) Show that Assumption I holds, and that the optimal cost function has the form

$$J^*(x) = \begin{cases} 0 & \text{if } x \in X^*, \\ \infty & \text{otherwise,} \end{cases}$$

where X^* is some subset of X.

(b) Consider the sequence of sets $\{X_k\}$, where

$$X_k = \big\{x \in X \mid (T^k \bar{J})(x) = 0\big\}.$$

Show that $X_{k+1} \subset X_k$ for all k, and that $X^* \subset \cap_{k=0}^{\infty} X_k$. Show also that convergence of VI (i.e., $T^k \bar{J} \to J^*$) is equivalent to $X^* = \cap_{k=0}^{\infty} X_k$.

(c) Show that $X^* = \cap_{k=0}^{\infty} X_k$ and there exists an optimal stationary policy if the sets

$$\hat{U}_k(x) = \big\{u \in U(x) \mid f(x,u,w) \in X_k,\ \forall\ w \in W(x,u)\big\}$$

are compact for all k greater than some index $\bar{k}$. *Hint*: Use Prop. 4.3.14.

Solution: Let $\hat{\mathcal{E}}(X)$ be the subset of $\mathcal{E}(X)$ that consists of functions that take only the two values 0 and ∞, and for all $J \in \hat{\mathcal{E}}(X)$ denote

$$D(J) = \big\{x \in X \mid J(x) = 0\big\}.$$

Note that for all $J \in \hat{\mathcal{E}}(X)$ we have $T_\mu J \in \hat{\mathcal{E}}(X)$, $TJ \in \hat{\mathcal{E}}(X)$, and that

$$D(T_\mu J) = \big\{x \in X \mid x \in D(J),\ f\big(x, \mu(x), w\big) \in D(J),\ \forall\ w \in W(x, \mu(x))\big\},$$

$$D(TJ) = \cup_{\mu \in \mathcal{M}} D(T_\mu J).$$

(a) For all $J \in \hat{\mathcal{E}}(X)$, we have $D(T_\mu J) \subset D(J)$ and $T_\mu J \geq J$, so condition (1) of Assumption I holds, and it is easily verified that the remaining two conditions of

Assumption I also hold. We have $\bar{J} \in \hat{\mathcal{E}}(X)$, so for any policy $\pi = \{\mu_0, \mu_1, \ldots\}$, we have $T_{\mu_0} \cdots T_{\mu_k} \bar{J} \in \hat{\mathcal{E}}(X)$. It follows that J_π, given by

$$J_\pi = \lim_{k\to\infty} T_{\mu_0} \cdots T_{\mu_k} \bar{J},$$

also belongs to $\hat{\mathcal{E}}(X)$, and the same is true for $J^* = \inf_{\pi\in\Pi} J_\pi$. Thus J^* has the given form with $D(J^*) = X^*$.

(b) Since $\{T^k \bar{J}\}$ is monotonically nondecreasing we have $D(T^{k+1}\bar{J}) \subset D(T^k \bar{J})$, or equivalently $X_{k+1} \subset X_k$ for all k. Generally for a sequence $\{J_k\} \subset \hat{\mathcal{E}}(X)$, if $J_k \uparrow J$, we have $J \in \hat{\mathcal{E}}(X)$ and $D(J) = \cap_{k=0}^\infty D(J_k)$. Thus convergence of VI (i.e., $T^k \bar{J} \uparrow J^*$) is equivalent to $D(J^*) = \cap_{k=0}^\infty D(J_k)$ or $X^* = \cap_{k=0}^\infty X_k$.

(c) The compactness condition of Prop. 4.3.14 guarantees that $T^k \bar{J} \uparrow J^*$, or equivalently by part (b), $X^* = \cap_{k=0}^\infty X_k$. This condition requires that the sets

$$U_k(x,\lambda) = \big\{ u \in U(x) \;\big|\; H(x,u,T^k\bar{J}) \le \lambda \big\}$$

are compact for every $x \in X$, $\lambda \in \Re$, and for all k greater than some integer $\overline{k}$. It can be seen that $U_k(x,\lambda)$ is equal to the set

$$\hat{U}_k(x) = \big\{ u \in U(x) \;\big|\; f(x,u,w) \in X_k,\ \forall\ w \in W(x,u) \big\}$$

given in the statement of the exercise.

4.6 (Exceptional Linear-Quadratic Problems)

Consider the deterministic linear-quadratic problem of Section 3.5.4 and Example 4.5.1. Assume that there is a single control variable u_k, and two state variables, x_k^1 and x_k^2, which evolve according to

$$x_{k+1}^1 = \gamma x_k^1 + b u_k, \qquad x_{k+1}^2 = x_k^1 + x_k^2 + u_k,$$

where $\gamma > 1$. The cost of stage k is quadratic of the form

$$q\big((x_k^1)^2 + (x_k^2)^2\big) + (u_k)^2.$$

Consider the four cases of pairs of values (b,q) where $b \in \{0,1\}$ and $q \in \{0,1\}$. For each case, use the theory of Section 4.5 to find the optimal cost function J^* and the optimal cost function over stable policies $\hat{J}^+$, and to describe the convergence behavior of VI.

Solution: When $b = 1$ and $q = 1$, the classical controllability and observability conditions are satisfied, and we have $J^* = \hat{J}^+$, while there exists an optimal policy that is linear and stable (so J^* and $\hat{J}^+$ are real-valued and positive definite quadratic). Moreover, the VI algorithm converges to J^* starting from any $J_0 \ge 0$ (even extended real-valued J_0) with $J_0(0) = 0$.

When $b = 0$ and $q = 0$, we clearly have $J^*(x) \equiv 0$. Also $\hat{J}^+(x^1,x^2) = \infty$ for $x^1 \ne 0$, while $\hat{J}^+(0,x^2)$ is finite for all x^2, but positive for $x^2 \ne 0$ (since for

$x^1 = 0$, the problem becomes essentially one-dimensional, and similar to the one of Section 3.5.4). The VI algorithm converges to $\hat{J}^+$ starting from any positive semidefinite quadratic initial condition J_0 with $J_0(0, x^2) = 0$ and $J_0 \neq J^*$.

When $b = 0$ and $q = 1$, we have $J^* = \hat{J}^+$, but J^* and $\hat{J}^+$ are not real-valued. In particular, since x_k^1 stays constant under all policies when $b = 0$, we have $J^*(x^1, x^2) = \hat{J}^+(x^1, x^2) = \infty$ for $x^1 \neq 0$. Moreover, for an initial state with $x_0^1 = 0$, the problem becomes essentially a one-dimensional problem that satisfies the classical controllability and observability conditions, and we have $J^*(0, x^2) = \hat{J}^+(0, x^2)$ for all x^2. The VI algorithm takes the form

$$J_{k+1}(0, x^2) = \min_u \Big\{ (x^2)^2 + (u)^2 + J_k(0, x^2 + u) \Big\},$$

$$J_{k+1}(x^1, x^2) = \min_u \Big\{ (x^1)^2 + (x^2)^2 + (u)^2 + J_k(\gamma x^1, x^1 + x^2 + u) \Big\}, \qquad \text{if } x^1 \neq 0.$$

It can be seen that the VI iterates $J_k(0, x^2)$ evolve as in the case of a single state variable problem, where x^1 is fixed at 0. For $x^1 \neq 0$, the VI iterates $J_k(x^1, x^2)$ diverge to ∞.

When $b = 1$ and $q = 0$, we have $J^*(x) \equiv 0$, while $0 < \hat{J}^+(x) < \infty$ for all $x \neq 0$. Similar to Example 4.5.1, the VI algorithm converges to $\hat{J}^+$ starting from any initial condition $J_0 \geq \hat{J}^+$. The functions J^* and $\hat{J}^+$ are real-valued and satisfy Bellman's equation, which has the form

$$J(x^1, x^2) = \min_u \Big\{ (u)^2 + J(\gamma x^1 + u, x^1 + x^2 + u) \Big\}.$$

However, Bellman's equation has additional solutions, other than J^* and $\hat{J}^+$. One of these is

$$\hat{J}(x^1, x^2) = P(x^1)^2,$$

where $P = \gamma^2 - 1$ (cf. the example of Section 3.5.4).

4.7 (Discontinuities in Infinite-State Shortest Path Problems)

The purpose of this exercise is to show that different types of perturbations in infinite-state shortest path problems, may yield different solutions of Bellman's equation. Consider the optimal stopping problem of Example 4.5.2, and introduce a perturbed version by modifying the effect of the action that moves the state from $x \neq 0$ to γx. Instead, this action stops the system with probability $\delta > 0$ at cost $\beta \geq 0$, and moves the state from x to γx with probability $1 - \delta$ at cost $\|x\|$. Note that with this modification, all policies become uniformly proper. Show that:

(a) The optimal cost function of the (δ, β)-perturbed version of the problem, denoted $\hat{J}_{\delta,\beta}$, is the unique solution of the corresponding Bellman equation within the class of bounded functions $\mathcal{B}$ of Eq. (4.94).

(b) For $\beta = 0$, we have $\lim_{\delta \downarrow 0} \hat{J}_{\delta,0} = J^*$, where J^* is the optimal cost function of the deterministic problem of Example 4.5.2.

(c) For $\beta = c$, we have $\hat{J}_{\delta,c} = \hat{J}^+$ for all $\delta > 0$, where $\hat{J}^+$ is the largest solution of Bellman's equation in the deterministic problem of Example

4.5.2 [$\hat{J}^+(x) = c$ for all $x \neq 0$, which corresponds to the policy that stops at all states].

Solution: (a) It can be seen that the Bellman equation for the (δ, β)-perturbed version of the problem is

$$J(x) = \begin{cases} \min\big\{c,\, \delta\beta + (1-\delta)\big(\|x\| + J(\gamma x)\big)\big\} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases}$$

and has exactly the same solutions as the equation

$$J(x) = \begin{cases} \min\big\{c,\, \delta\beta + (1-\delta)\big(\min\big\{c/(1-\delta),\, \|x\|\big\} + J(\gamma x)\big)\big\} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

The latter equation involves a bounded cost per stage, and hence according to the theory of Section 4.6, has a unique solution within $\mathcal{B}$, when all policies are proper.

(b) Evident since the effect of δ on the cost of the optimal policy of the problem of Example 4.5.2 diminishes as $\delta \to 0$.

(c) Since termination at cost c is inevitable (with probability 1) under every policy, the optimal policy for the (δ, β)-perturbed version of the problem is to stop as soon as possible.

4.8 (A Perturbation Approach for Semicontractive Models)

The purpose of this exercise is to adapt the perturbation approach of Section 3.4 so that it can be used in conjunction with the regularity notion for nonstationary policies of Definition 4.4.1. Given a set of functions $S \subset \mathcal{E}(X)$ and a collection $\mathcal{C}$ of policy-state pairs (π, x) that is S-regular, let $J^*_{\mathcal{C}}$ be the restricted optimal cost function defined by

$$J^*_{\mathcal{C}}(x) = \inf_{(\pi,x)\in\mathcal{C}} J_\pi(x), \qquad x \in X.$$

Consider also a nonnegative forcing function $p : X \mapsto [0, \infty)$, and for each $\delta > 0$ and stationary policy μ, the mappings $T_{\mu,\delta}$ and T_δ given by

$$(T_{\mu,\delta}J)(x) = H\big(x, \mu(x), J\big) + \delta p(x), \qquad (T_\delta J)(x) = \inf_{\mu\in\mathcal{M}} (T_{\mu,\delta}J)(x), \qquad x \in X.$$

We refer to the problem associated with the mappings $T_{\mu,\delta}$ as the δ-perturbed problem. The cost function of a policy $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi$ for this problem is

$$J_{\pi,\delta} = \limsup_{k\to\infty} T_{\mu_0,\delta} \cdots T_{\mu_k,\delta} \bar{J},$$

and the optimal cost function is $\hat{J}_\delta = \inf_{\pi\in\Pi} J_{\pi,\delta}$. Assume that for every $\delta > 0$:

(1) $\hat{J}_\delta$ satisfies the Bellman equation of the δ-perturbed problem, $\hat{J}_\delta = T_\delta \hat{J}_\delta$.

(2) For every $x \in X$, we have $\inf_{(\pi,x)\in\mathcal{C}} J_{\pi,\delta}(x) = \hat{J}_\delta(x)$.

(3) For all $x \in X$ and $(\pi, x) \in \mathcal{C}$, we have

$$J_{\pi,\delta}(x) \leq J_\pi(x) + w_{\pi,\delta}(x),$$

where $w_{\pi,\delta}$ is a function such that $\lim_{\delta\downarrow 0} w_{\pi,\delta} = 0$.

(4) For every sequence $\{J_m\} \subset S$ with $J_m \downarrow J$, we have

$$\lim_{m\to\infty} H(x, u, J_m) = H(x, u, J), \qquad \forall\ x \in X,\ u \in U(x).$$

Then $J^*_{\mathcal{C}}$ is a fixed point of T and the conclusions of Prop. 4.4.2 hold. Moreover, we have

$$J^*_{\mathcal{C}} = \lim_{\delta\downarrow 0} \hat{J}_\delta.$$

Solution: The proof is very similar to the one of Prop. 3.4.1. Condition (2) implies that for every $x \in X$ and $\epsilon > 0$, there exists a policy $\pi_{x,\epsilon}$ such that $(\pi_{x,\epsilon}, x) \in \mathcal{C}$ and $J_{\pi_{x,\epsilon},\delta}(x) \leq \hat{J}_\delta(x) + \epsilon$. Thus, using conditions (2) and (3), we have for all $x \in X$, $\delta > 0$, $\epsilon > 0$, and π with $(\pi, x) \in \mathcal{C}$,

$$J^*_{\mathcal{C}}(x) - \epsilon \leq J_{\pi_{x,\epsilon}}(x) - \epsilon \leq J_{\pi_{x,\epsilon},\delta}(x) - \epsilon \leq \hat{J}_\delta(x) \leq J_{\pi,\delta}(x) \leq J_\pi(x) + w_{\pi,\delta}(x).$$

By taking the limit as $\epsilon \downarrow 0$, we obtain for all $x \in X$, $\delta > 0$, and π with $(\pi, x) \in \mathcal{C}$,

$$J^*_{\mathcal{C}}(x) \leq \hat{J}_\delta(x) \leq J_{\pi,\delta}(x) \leq J_\pi(x) + w_{\pi,\delta}(x).$$

By taking the limit as $\delta \downarrow 0$ and then the infimum over all π with $(\pi, x) \in \mathcal{C}$, it follows [using also condition (3)] that for all $x \in X$,

$$J^*_{\mathcal{C}}(x) \leq \lim_{\delta\downarrow 0} \hat{J}_\delta(x) \leq \inf_{\{\pi|(\pi,x)\in\mathcal{C}\}} \lim_{\delta\downarrow 0} J_{\pi,\delta}(x) \leq \inf_{\{\pi|(\pi,x)\in\mathcal{C}\}} J_\pi(x) = J^*_{\mathcal{C}}(x),$$

so that $J^*_{\mathcal{C}} = \lim_{\delta\downarrow 0} \hat{J}_\delta$.

To prove that $J^*_{\mathcal{C}}$ is a fixed point of T, we prove that both $J^*_{\mathcal{C}} \geq TJ^*_{\mathcal{C}}$ and $J^*_{\mathcal{C}} \leq TJ^*_{\mathcal{C}}$ hold. Indeed, from condition (1) and the fact $\hat{J}_\delta \geq J^*_{\mathcal{C}}$ shown earlier, we have for all $\delta > 0$,

$$\hat{J}_\delta = T_\delta \hat{J}_\delta \geq T\hat{J}_\delta \geq TJ^*_{\mathcal{C}},$$

and by taking the limit as $\delta \downarrow 0$ and using the fact $J^*_{\mathcal{C}} = \lim_{\delta\downarrow 0} \hat{J}_\delta$ shown earlier, we obtain $J^*_{\mathcal{C}} \geq TJ^*_{\mathcal{C}}$. For the reverse inequality, let $\{\delta_m\}$ be a sequence with $\delta_m \downarrow 0$. Using condition (1) we have for all m,

$$H(x, u, \hat{J}_{\delta_m}) + \delta_m p(x) \geq (T_{\delta_m}\hat{J}_{\delta_m})(x) = \hat{J}_{\delta_m}(x), \qquad \forall\ x \in X,\ u \in U(x).$$

Taking the limit as $m \to \infty$, and using condition (4) and the fact $\hat{J}_{\delta_m} \downarrow J^*_{\mathcal{C}}$ shown earlier, we have

$$H(x, u, J^*_{\mathcal{C}}) \geq J^*_{\mathcal{C}}(x), \qquad \forall\ x \in X,\ u \in U(x),$$

so that by minimizing over $u \in U(x)$, we obtain $TJ^*_{\mathcal{C}} \geq J^*_{\mathcal{C}}$.

4.9 (Deterministic Optimal Control with Positive and Negative Costs per Stage)

In this exercise, we consider the infinite-spaces optimal control problem of Section 4.5 and its notation, but without the assumption $g \geq 0$ [cf. Eq. (4.46)]. Instead, we assume that

$$-\infty < g(x,u) \leq \infty, \qquad \forall\ x \in X,\ u \in U(x),\ k = 0, 1, \ldots,$$

and that $J^*(x) > -\infty$ for all $x \in X$. The latter assumption was also made in Section 3.5.5, but in the present exercise, we will not assume the additional near-optimal termination Assumption 3.5.9 of that section, and we will use instead the perturbation framework of Exercise 4.8. Note that J^* is a fixed point of T because the problem is deterministic (cf. Exercise 3.1).

We say that a policy π is *terminating from state* $x_0 \in X$ if the sequence $\{x_k\}$ generated by π starting from x_0 terminates finitely (i.e., satisfies $x_{\bar{k}} = t$ for some index $\bar{k}$). We denote by Π_x the set of all policies that are terminating from x, and we consider the collection

$$\mathcal{C} = \big\{(\pi, x) \mid \pi \in \Pi_x\big\}.$$

Let $J^*_{\mathcal{C}}$ be the corresponding restricted optimal cost function,

$$J^*_{\mathcal{C}}(x) = \inf_{(\pi,x)\in\mathcal{C}} J_\pi(x) = \inf_{\pi\in\Pi_x} J_\pi(x), \qquad x \in X,$$

and let S be the set of functions

$$S = \big\{J \in \mathcal{E}(X) \mid J(t) = 0,\ J(x) > -\infty,\ x \in X\big\}.$$

Clearly $\mathcal{C}$ is S-regular, so we may consider the perturbation framework of Exercise 4.8 with $p(x) = 1$ for all $x \neq t$ and $p(t) = 0$. Apply the results of that exercise to show that:

(a) We have
$$J^*_{\mathcal{C}} = \lim_{\delta\downarrow 0} \hat{J}_\delta.$$

(b) $J^*_{\mathcal{C}}$ is the only fixed point of T within the set
$$\mathcal{W} = \big\{J \in \mathcal{E}(X) \mid J(t) = 0,\ J \geq J^*_{\mathcal{C}}\big\}.$$

(c) We have $T^k J \to J^*_{\mathcal{C}}$ for all $J \in \mathcal{W}$.

Solution: Part (a) follows from Exercise 4.8, and parts (b), (c) follow from Exercise 4.8 and Prop. 4.4.2.

4.10 (On Proper Policies for Stochastic Shortest Paths)

Consider the infinite-spaces SSP problem of Section 4.6 under the assumptions of Prop. 4.6.4, and assume that g is bounded over $X \times U \times W$.

(a) Show that if μ is a uniformly proper policy, then J_μ is the unique solution of the equation $J = T_\mu J$ within $\mathcal{B}$ and that $T_\mu^k J \to J_\mu$ for all $J \in \mathcal{B}$.

(b) Let J' be a fixed point of T such that $J' \in \mathcal{B}$ and $J' \neq \hat{J}$. Show that a policy μ satisfying $T_\mu J' = TJ'$ cannot be uniformly proper.

Solution: (a) Consider the problem where the only policy is μ, i.e., with control constraint set $\tilde{U}(x) = \{\mu(x)\}$, $x \in X$, and apply Props. 4.6.5 and 4.4.4.

(b) Assume to come to a contradiction that μ is uniformly proper. We have $T_\mu J' = TJ' = J'$, so by part (a) we have $J' = J_\mu$, while $J_\mu \geq \hat{J}$ since μ is uniformly proper. Thus $J' \geq \hat{J}$ while $J' \neq \hat{J}$ by assumption. This contradicts the largest fixed point property of $\hat{J}$ [cf. Prop. 4.6.5(a)].

4.11 (Example where $\hat{J}$ is not a Fixed Point of T in Infinite Spaces SSP)

We noted in Section 4.6 that some additional assumption, like

$$E\Big\{g(x,u,w) + \hat{J}_\delta\big(f(x,u,w)\big)\Big\} < \infty, \qquad \forall\ x \in X^*,\ u \in U(x), \tag{4.96}$$

or the finiteness of W, is necessary to prove that $\hat{J}$ is a fixed point for SSP problems (cf. Prop. 4.6.4). [The condition (4.96) is satisfied for example if there exists a policy π (necessarily proper at all $x \in X^*$) such that $J_{\pi,\delta}$ is bounded over X^*.] To see what can happen without such an assumption, consider the following example, which was constructed by Yi Zhang (private communication).

Let $X = \{t, 0, 1, 2, \ldots\}$, where t is the termination state, and let $g(x,u,w) \equiv 0$, so that $J^*(x) \equiv 0$. There is only one control at each state, and hence only one policy. The transitions are as follows:

From each state $x = 2, 3, \ldots$, we move deterministically to state $x-1$, from state 1 we move deterministically to state t, and from state 0 we move to state $x = 1, 2, \ldots$, with probability p_x such that $\sum_{x=1}^\infty x p_x = \infty$.

Verify that the unique policy is proper at all $x = 1, 2, \ldots$, and we have $\hat{J}(x) = J^*(x) = 0$. However, the policy is not proper at $x = 0$, since the expected number of transitions from $x = 0$ to termination is $\sum_{x=1}^\infty x p_x = \infty$. As a result the set $\widehat{\Pi}_0$ is empty and we have $\hat{J}(0) = \infty$. Thus $\hat{J}$ does not satisfy the Bellman equation for $x = 0$, since

$$\infty = \hat{J}(0) \neq E\Big\{g(0,u,w) + \hat{J}\big(f(0,u,w)\big)\Big\} = \sum_{x=1}^\infty p_x \hat{J}(x) = 0.$$

4.12 (Convergence of Nonexpansive Monotone Fixed Point Iterations with a Unique Fixed Point)

Consider the mapping H of Section 2.1 under the monotonicity Assumption 2.1.1. Assume that instead of the contraction Assumption 2.1.2, the following hold:

(1) For every $J \in \mathcal{B}(X)$, the function TJ belongs to $\mathcal{B}(X)$, the space of functions on X that are bounded with respect to the weighted sup-norm corresponding to a positive weighting function v.

(2) T is nonexpansive, i.e., $\|TJ - TJ'\| \leq \|J - J'\|$ for all $J, J' \in \mathcal{B}(X)$.

(3) T has a unique fixed point within $\mathcal{B}(X)$, denoted J^*.

(4) If X is infinite the following continuity property holds: For each $J \in \mathcal{B}(X)$ and $\{J_m\} \subset \mathcal{B}(X)$ with either $J_m \downarrow J$ or $J_m \uparrow J$,

$$H(x,u,J) = \lim_{m\to\infty} H(x,u,J_m), \qquad \forall\ x \in X,\ u \in U(x).$$

Show the following:

(a) For every $J \in \mathcal{B}(X)$, we have $\|T^k J - J^*\| \to 0$ if X is finite, and $T^k J \to J^*$ if X is infinite.

(b) Part (a) holds if $\mathcal{B}(X)$ is replaced by $\big\{J \in \mathcal{B}(X) \mid J \geq 0\big\}$, or by $\big\{J \in \mathcal{B}(X) \mid J(t) = 0\big\}$, or by $\big\{J \in \mathcal{B}(X) \mid J(t) = 0,\ J \geq 0\big\}$, where t is a special cost-free and absorbing destination state t.

(Unpublished joint work of the author with H. Yu.)

Solution: (a) Assume first that X is finite. For any $c > 0$, let $V_0 = J^* + c\,v$ and consider the sequence $\{V_k\}$ defined by $V_{k+1} = TV_k$ for $k \geq 0$. Note that $\{V_k\} \subset \mathcal{B}(X)$, since $\|V_0\| \leq \|J^*\| + c$ so that $V_0 \in \mathcal{B}(X)$, and we have $V_{k+1} = TV_k$, so that property (1) applies. From the nonexpansiveness property (2), we have

$$H(x,u,J^* + c\,v) \leq H(x,u,J^*) + c\,v(x), \qquad x \in X,\ u \in U(x),$$

and by taking the infimum over $u \in U(x)$, we obtain $J^* \leq T(J^* + c\,v) \leq J^* + c\,v$, i.e., $J^* \leq V_1 \leq V_0$. From this and the monotonicity of T it follows that $J^* \leq V_{k+1} \leq V_k$ for all k, so that for each $x \in X$, $V_k(x) \downarrow \overline{V}(x)$ where $\overline{V}(x) \geq J^*(x)$. Moreover, $\overline{V}$ lies in $\mathcal{B}(X)$ (since $J^* \leq \overline{V} \leq V_k$), and also satisfies $\|V_k - \overline{V}\| \to 0$ (since X is finite). From property (2), we have $\|TV_k - T\overline{V}\| \leq \|V_k - \overline{V}\|$, so that $\|TV_k - T\overline{V}\| \to 0$, which together with the fact $TV_k = V_{k+1} \to \overline{V}$, implies that $\overline{V} = T\overline{V}$. Thus $\overline{V} = J^*$ by the uniqueness property (3), and it follows that $V_k \downarrow J^*$.

Similarly, define $W_k = T^k(J^* - c\,v)$, and by an argument symmetric to the above, $W_k \uparrow J^*$. Now for any $J \in \mathcal{B}(X)$, let $c = \|J - J^*\|$ in the definition of V_k and W_k. Then $J^* - c\,v \leq J \leq J^* + c\,v$, so by the monotonicity of T, we have $W_k \leq T^k J \leq V_k$ as well as $W_k \leq J^* \leq V_k$ for all k. Therefore $\|T^k J - J^*\| \leq \|W_k - V_k\|$ for all $k \geq 0$. Since $\|W_k - V_k\| \leq \|W_k - J^*\| + \|V_k - J^*\| \to 0$, the conclusion follows.

If X is infinite and property (4) holds, the preceding proof goes through, except for the part that shows that $\|V_k - \overline{V}\| \to 0$. Instead we use a different

argument to prove that $\overline{V} = T\overline{V}$. Indeed, since $V_k \geq V_{k+1} = TV_k \geq T\overline{V}$, it follows that $\overline{V} \geq T\overline{V}$. For the reverse inequality we write

$$T\overline{V} = \inf_{u\in U(x)} \lim_{k\to\infty} H(x,u,V_k) \geq \lim_{k\to\infty} \inf_{u\in U(x)} H(x,u,V_k) = \lim_{k\to\infty} TV_k = \overline{V},$$

where the first equality follows from the continuity property (4), and the inequality follows from the generic relation $\inf \lim H \geq \lim \inf H$. Thus we have $\overline{V} = T\overline{V}$, which by the uniqueness property (3), implies that $\overline{V} = J^*$ and $V_k \downarrow J^*$. With a similar argument we obtain $W_k \uparrow J^*$, implying that $T^k J \to J^*$.

(b) The proof of part (a) applies with simple modifications.

4.13 (Convergence of Nonexpansive Monotone Fixed Point Iterations with Multiple Fixed Points)

Consider the mapping H of Section 2.1 under the monotonicity Assumption 2.1.1. Assume that instead of the contraction Assumption 2.1.2, the following hold:

(1) For every $J \in \mathcal{B}(X)$, the function TJ belongs to $\mathcal{B}(X)$, the space of functions on X that are bounded with respect to the weighted sup-norm corresponding to a positive weighting function v.

(2) T is nonexpansive, i.e., $\|TJ - TJ'\| \leq \|J - J'\|$ for all $J, J' \in \mathcal{B}(X)$.

(3) T has a largest fixed point within $\mathcal{B}(X)$, denoted $\hat{J}$, i.e., $\hat{J} \in \mathcal{B}(X)$, $\hat{J}$ is a fixed point of T, and for every other fixed point $J' \in \mathcal{B}(X)$ we have $J' \leq \hat{J}$.

(4) If X is infinite the following continuity property holds: For each $J \in \mathcal{B}(X)$ and $\{J_m\} \subset \mathcal{B}(X)$ with either $J_m \downarrow J$ or $J_m \uparrow J$,

$$H(x,u,J) = \lim_{m\to\infty} H(x,u,J_m), \qquad \forall\ x \in X,\ u \in U(x).$$

Show the following:

(a) For every $J \in \mathcal{B}(X)$ such that $\hat{J} \leq J \leq \hat{J} + c\,v$ for some $c > 0$, we have $\|T^k J - \hat{J}\| \to 0$ if X is finite, and $T^k J \to \hat{J}$ if X is infinite.

(b) Part (a) holds if $\mathcal{B}(X)$ is replaced by $\big\{J \in \mathcal{B}(X) \mid J \geq 0\big\}$, or by $\big\{J \in \mathcal{B}(X) \mid J(t) = 0\big\}$, or by $\big\{J \in \mathcal{B}(X) \mid J(t) = 0,\ J \geq 0\big\}$, where t is a special cost-free and absorbing destination state t.

(Note the similarity with the preceding exercise.)

Solution: (a) The proof follows the line of proof of the preceding exercise. Assume first that X is finite. For any $c > 0$, let $V_0 = \hat{J} + c\,v$ and consider the sequence $\{V_k\}$ defined by $V_{k+1} = TV_k$ for $k \geq 0$. Note that $\{V_k\} \subset \mathcal{B}(X)$, since $\|V_0\| \leq \|\hat{J}\| + c$ so that $V_0 \in \mathcal{B}(X)$, and we have $V_{k+1} = TV_k$, so that property (1) applies. From the nonexpansiveness property (2), we have

$$H(x,u,\hat{J} + c\,v) \leq H(x,u,\hat{J}) + c\,v(x), \qquad x \in X,\ u \in U(x),$$

and by taking the infimum over $u \in U(x)$, we obtain $\hat{J} \leq T(\hat{J} + c\,v) \leq \hat{J} + c\,v$, i.e., $\hat{J} \leq V_1 \leq V_0$. From this and the monotonicity of T it follows that $\hat{J} \leq V_{k+1} \leq V_k$

for all k, so that for each $x \in X$, $V_k(x) \downarrow \overline{V}(x)$ where $\overline{V}(x) \geq \hat{J}(x)$. Moreover, $\overline{V}$ lies in $\mathcal{B}(X)$ (since $\hat{J} \leq \overline{V} \leq V_k$), and also satisfies $\|V_k - \overline{V}\| \to 0$ (since X is finite). From property (2), we have $\|TV_k - T\overline{V}\| \leq \|V_k - \overline{V}\|$, so that $\|TV_k - T\overline{V}\| \to 0$, which together with the fact $TV_k = V_{k+1} \to \overline{V}$, implies that $\overline{V} = T\overline{V}$. Thus $\overline{V} = \hat{J}$ by property (3), and it follows that $V_k \downarrow \hat{J}$.

If X is infinite and property (4) holds, the preceding proof goes through, except for the part that shows that $\|V_k - \overline{V}\| \to 0$. Instead we use a different argument to prove that $\overline{V} = T\overline{V}$. Indeed, since $V_k \geq V_{k+1} = TV_k \geq T\overline{V}$, it follows that $\overline{V} \geq T\overline{V}$. For the reverse inequality we write

$$T\overline{V} = \inf_{u \in U(x)} \lim_{k \to \infty} H(x, u, V_k) \geq \lim_{k \to \infty} \inf_{u \in U(x)} H(x, u, V_k) = \lim_{k \to \infty} TV_k = \overline{V},$$

where the first equality follows from the continuity property (4). Thus we have $\overline{V} = T\overline{V}$, which by property (3), implies that $\overline{V} = \hat{J}$ and $V_k \downarrow \hat{J}$.

(b) The proof of part (a) applies with simple modifications.

4.14 (Necessary and Sufficient Condition for an Interpolated Nonexpansive Mapping to be a Contraction)

This exercise (due to unpublished joint work with H. Yu) considers a nonexpansive mapping $G : \Re^n \mapsto \Re^n$, and derives conditions under which the interpolated mapping G_γ defined by

$$G_\gamma(x) = (1 - \gamma)x + \gamma G(x), \qquad x \in \Re^n,$$

is a contraction for all $\gamma \in (0, 1)$. Consider $\Re^n$ equipped with a strictly convex norm $\|\cdot\|$, and the set

$$C = \left\{ \left(\frac{x - y}{\|x - y\|}, \frac{G(x) - G(y)}{\|x - y\|} \right) \;\middle|\; x, y \in \Re^n,\ x \neq y \right\},$$

which can be viewed as a set of "slopes" of G along all directions. Show that the mapping G_γ defined by

$$G_\gamma(x) = (1 - \gamma)x + \gamma G(x), \qquad x \in \Re^n,$$

is a contraction for all $\gamma \in (0, 1)$ if and only if there is no closure point (z, w) of C such that $z = w$. *Note*: To illustrate with some one-dimensional examples what can happen if this closure condition is violated, let $G : \Re \mapsto \Re$ be continuously differentiable, monotonically nondecreasing, and satisfying $0 \leq \frac{dG(x)}{dx} \leq 1$. Note that G is nonexpansive. We consider two cases.

(1) $G(0) = 0$, $\frac{dG(0)}{dx} = 1$, $0 \leq \frac{dG(x)}{dx} < 1$ for $x \neq 0$, $\lim_{x \to \infty} \frac{dG(x)}{dx} < 1$ and $\lim_{x \to -\infty} \frac{dG(x)}{dx} < 1$. Here $(z, w) = (1, 1)$ is a closure point of C and satisfies $z = w$. Note that G_γ is not a contraction for any $\gamma \in (0, 1)$, although it has 0 as its unique fixed point.

(2) $\lim_{x\to\infty} \frac{dG(x)}{dx} = 1$. Here we have $\lim_{x\to\infty}\big(G(x) - G(y)\big) = x - y$ for $x = y + 1$, so $(1, 1)$ is a closure point of C. It can also be seen that because $\lim_{x\to\infty} \frac{dG_\gamma(x)}{dx} = 1$, G_γ is not a contraction for any $\gamma \in (0, 1)$, and may have one, more than one, or no fixed points.

Solution: Assume there is no closure point (z, w) of C such that $z = w$, and for $\gamma \in (0, 1)$, let

$$\rho = \sup_{(z,w)\in C} \big\|(1-\gamma)z + \gamma w\big\|.$$

The set C is bounded since for all $(z, w) \in C$, we have $\|z\| = 1$, and $\|w\| \le 1$ by the nonexpansiveness of G. Hence, there exists a sequence $\big\{(z_k, w_k)\big\} \subset C$ that converges to some $(\overline{z}, \overline{w})$, and is such that

$$\big\|(1-\gamma)z_k + \gamma w_k\big\| \to \rho.$$

Since $(\overline{z}, \overline{w})$ is a closure point of C, we have $\overline{z} \ne \overline{w}$. Using the continuity of the norm, we have

$$\rho = \big\|(1-\gamma)\overline{z} + \gamma\overline{w}\big\| < (1-\gamma)\|\overline{z}\| + \gamma\|\overline{w}\| \le 1,$$

where for the strict inequality we use the strict convexity of the norm, and for the last inequality we use the fact $\|\overline{z}\| = 1$ and $\|\overline{w}\| \le 1$. Thus $\rho < 1$, and since

$$\begin{aligned}\left\|(1-\gamma)\frac{x-y}{\|x-y\|} + \gamma\frac{G(x)-G(y)}{\|x-y\|}\right\| &= \frac{\big\|G_\gamma(x) - G_\gamma(y)\big\|}{\|x-y\|}\\ &\le \sup_{(z,w)\in C}\big\|(1-\gamma)z + \gamma w\big\|\\ &= \rho, \qquad \forall\ x \ne y,\end{aligned}$$

it follows that G_γ is a contraction of modulus ρ.

Conversely, if G_γ is a contraction, we have

$$\begin{aligned}\sup_{(z,w)\in C}\big\|(1-\gamma)z + \gamma w\big\| &= \sup_{x\ne y}\left\|(1-\gamma)\frac{x-y}{\|x-y\|} + \gamma\frac{G(x)-G(y)}{\|x-y\|}\right\|\\ &\le \sup_{x\ne y}\frac{\big\|G_\gamma(x) - G_\gamma(y)\big\|}{\|x-y\|}\\ &< 1.\end{aligned}$$

Thus for every closure point (z, w) of C,

$$\big\|(1-\gamma)z + \gamma w\big\| < 1,$$

which implies that we cannot have $z = w$.

5

Sequential Zero-Sum Games and Minimax Control

Contents

In this chapter, we introduce a contractive abstract DP framework and related policy iteration (PI) algorithms, specifically designed for sequential zero-sum games and minimax problems with a general structure. Aside from greater generality, the advantage of our algorithms over alternatives is that they resolve some long-standing convergence difficulties of the "natural" PI algorithm, which have been known since the Pollatschek and Avi-Itzhak method [PoA69] for finite-state Markov games. Mathematically, this "natural" algorithm is a form of Newton's method for solving Bellman's equation, but Newton's method, contrary to the case of single-player DP problems, is not globally convergent in the case of a minimax problem, because of an additional difficulty: the Bellman operator may have components that are neither convex nor concave.

Our algorithms address this difficulty by introducing alternating player choices, and by using a policy-dependent mapping with a uniform sup-norm contraction property, similar to earlier works by Bertsekas and Yu [BeY10], [BeY12], [YuB13a], which has been described in part in Section 2.6.3. Moreover, our algorithms allow a convergent and highly parallelizable implementation, which is based on state space partitioning, and distributed asynchronous policy evaluation and policy improvement operations within each set of the partition. They are also suitable for approximations based on an aggregation approach.

5.1 INTRODUCTION

We will discuss abstract DP frameworks and PI methods for sequential minimax problems. In addition to being more efficient and reliable than alternatives, our methods are well suited for distributed asynchronous implementation. In Sections 5.1 and 5.2, we will discuss an abstract DP framework, which can be derived from the contractive framework of Chapter 2. We will revisit abstract PI algorithms within this framework and show how they relate to known algorithms for minimax control. We will also discuss how these algorithms when applied to discounted and terminating zero-sum Markov games, lead to methods such as the ones by Hoffman and Karp [HoK66], and by Pollatschek and Avi-Itzhak [PoA69]. We will note some of the drawbacks of these methods, particularly the need to solve a substantial optimization problem as part of the policy evaluation phase. These drawbacks motivate new PI algorithms and a different abstract framework, based on an alternating player choices format, which we will introduce in Section 5.3.

In our initial problem formulation, the focus of Sections 5.1 and 5.2, we consider abstract sequential infinite horizon zero-sum game and minimax problems, which involve two players that choose controls at each state x of some state space X, from within some state-dependent constraint sets: a *minimizer*, who selects a control u from within a subset $U(x)$ of a control

space U, and a *maximizer*, who selects a control v from within a subset $V(x)$ of a control space V. The spaces X, U, and V are arbitrary. Functions $\mu : X \mapsto U$ and $\nu : X \mapsto V$ such that $\mu(x) \in U(x)$ and $\nu(x) \in V(x)$ for all $x \in X$, are called *policies* for the minimizer and the maximizer, respectively. The set of policies for the minimizer and the maximizer are denoted by $\mathcal{M}$ and $\mathcal{N}$, respectively.

As in earlier chapters, the main idea is to start with a general mapping that defines the Bellman equation of the problem. In particular, we introduce a real-valued mapping that is suitable for minimax problems, and has the form

$$H(x, u, v, J), \qquad x \in X,\ u \in U(x),\ v \in V(x),\ J \in B(X); \tag{5.1}$$

cf. Example 2.6.4. In Eq. (5.1), $B(X)$ is the space of real-valued functions on X that are bounded with respect to a weighted sup-norm

$$\|J\| = \sup_{x \in X} \frac{\big|J(x)\big|}{\xi(x)}, \qquad J \in B(X), \tag{5.2}$$

where ξ is a function taking a positive value $\xi(x)$ for each $x \in X$. Our main assumption is the following:

Assumption 5.1.1: (Contraction for Minimax Problems) For every $\mu \in \mathcal{M}$, $\nu \in \mathcal{N}$, consider the operator $T_{\mu,\nu}$ that maps a function $J \in B(X)$ to the function $T_{\mu,\nu}J$ defined by

$$(T_{\mu,\nu}J)(x) = H\big(x, \mu(x), \nu(x), J\big), \qquad x \in X, \tag{5.3}$$

and assume the following:

(a) $T_{\mu,\nu}J$ belongs to $B(X)$ for all $J \in B(X)$.

(b) There exists an $\alpha \in (0, 1)$ such that for all $\mu \in \mathcal{M}$, $\nu \in \mathcal{N}$, the operator $T_{\mu,\nu}$ is a contraction mapping of modulus α with respect to the weighted sup-norm (5.2), i.e., for all $J, J' \in B(X)$, $\mu \in \mathcal{M}$, and $\nu \in \mathcal{N}$,

$$\|T_{\mu,\nu}J - T_{\mu,\nu}J'\| = \sup_{x \in X} \frac{\big|(T_{\mu,\nu}J)(x) - (T_{\mu,\nu}J')(x)\big|}{\xi(x)} \leq \alpha \|J - J'\|.$$

Since $T_{\mu,\nu}$ is a contraction within the complete space $B(X)$, under the preceding assumption, it has a unique fixed point $J_{\mu,\nu} \in B(X)$. We are interested in the operator $T : B(X) \mapsto B(X)$, defined by

$$(TJ)(x) = \inf_{u \in U(x)} \sup_{v \in V(x)} H(x, u, v, J), \qquad x \in X, \tag{5.4}$$

or equivalently,

$$(TJ)(x) = \inf_{\mu \in \mathcal{M}} \sup_{\nu \in \mathcal{N}} (T_{\mu,\nu} J)(x), \qquad x \in X. \tag{5.5}$$

An important fact is that T *is a contraction mapping from* $B(X)$ *to* $B(X)$. Indeed from Assumption 1.1(b), we have for all $x \in X$, $\mu \in \mathcal{M}$, and $\nu \in \mathcal{N}$,

$$(T_{\mu,\nu} J)(x) \leq (T_{\mu,\nu} J')(x) + \alpha \|J - J'\| \, \xi(x).$$

Taking the supremum over $\nu \in \mathcal{N}$ of both sides above, and then the infimum over $\mu \in \mathcal{M}$, and using Eq. (5.5), we obtain

$$(TJ)(x) \leq (TJ')(x) + \alpha \|J - J'\| \, \xi(x), \qquad \text{for all } x \in X.$$

Similarly, by reversing the roles of J and J', we obtain

$$(TJ')(x) \leq (TJ')(x) + \alpha \|J - J'\| \, \xi(x), \qquad \text{for all } x \in X.$$

Combining the preceding two relations, we have

$$\big|(TJ)(x) - (TJ')(x)\big| \leq \alpha \|J - J'\| \, \xi(x), \qquad \text{for all } x \in X,$$

and by dividing with $\xi(x)$, and taking supremum over $x \in X$, it follows that

$$\|TJ - TJ'\| \leq \alpha \|J - J'\|.$$

Thus T is a contraction mapping from $B(X)$ to $B(X)$, with respect to the sup-norm (5.2), with modulus α, and has a unique fixed point within $B(X)$, which we denote by J^*.

Bellman's Equation and Minimax Optimal Policies

Given a mapping H of the form (5.1) that satisfies Assumption 1.1, we are interested in computing the fixed point J^* of T, i.e., a function J^* such that

$$J^*(x) = \inf_{u \in U(x)} \sup_{v \in V(x)} H(x, u, v, J^*), \qquad \text{for all } x \in X. \tag{5.6}$$

Moreover, we are interested in finding a policy $\mu^* \in \mathcal{M}$ (if it exists) that attains the infimum for all $x \in X$ as in the following equation

$$\mu^*(x) \in \arg \min_{u \in U(x)} \overline{H}(x, u, J^*), \qquad \text{for all } x \in X,$$

where for all $x \in X$, $u \in U(x)$, and $J \in B(X)$, the mapping $\overline{H}$ is defined by

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

We are also interested in finding a policy $\nu^* \in \mathcal{N}$ (if it exists) that attains the supremum for all $x \in X$ as in the following equation

$$\nu^*(x) \in \arg\max_{v \in V(x)} H\big(x, \mu^*(x), v, J^*\big), \qquad \text{for all } x \in X.$$

In the context of a sequential minimax problem that is addressed by DP, the fixed point equation $J^* = TJ^*$ is viewed as a form of Bellman's equation. In this case, $J^*(x)$ is the minimax cost starting from state x. Moreover μ^* *is an optimal policy for the minimizer in a minimax sense, while* ν^* *is a corresponding worst case response of the maximizer*. Under suitable assumptions on H (such as convexity in u and concavity in v) the order of minimization and maximization can be interchanged in the preceding relations, in which case it can be shown that (μ^*, ν^*) is a saddle point (within the space $\mathcal{M} \times \mathcal{N}$) of the minimax value $J_{\mu,\nu}(x)$, for every $x \in X$.

Markov Games

The simplest special case of a sequential stochastic game problem, which relates to our abstract framework, was introduced in the paper by Shapley [Sha53] for undiscounted finite-state problems, with a termination state, where the Bellman operator $T_{\mu,\nu}$ is contractive with respect to the (unweighted) sup-norm for all $\mu \in \mathcal{M}$, and $\nu \in \mathcal{N}$. Shapley's work brought the contraction mapping approach to prominence in DP and sequential game analysis, and was subsequently extended by several authors in both undiscounted and discounted settings; see e.g., the book by Filar and Vrieze [FiV97], the lecture notes by Kallenberg [Kal20], and the works referenced there. Let us now describe a class of finite-state zero-sum game problems that descend from Shapley's work, and are often called "Markov games" (the name was introduced by Zachrisson [Zac64]).

Example 5.1.1 (Discounted Finite-State Markov Games)

Consider two players that play repeated matrix games at each of an infinite number of stages, using mixed strategies. The game played at a given stage is defined by a state x that takes values in a finite set X, and changes from one stage to the next according to a Markov chain whose transition probabilities are influenced by the players' choices. At each stage and state $x \in X$, the minimizer selects a probability distribution $u = (u_1, \ldots, u_n)$ over n possible choices $i = 1, \ldots, n$, and the maximizer selects a probability distribution $v = (v_1, \ldots, v_m)$ over m possible choices $j = 1, \ldots, m$. If the minimizer chooses i and the maximizer chooses j, the payoff of the stage is $a_{ij}(x)$ and depends on the state x. Thus the expected payoff of the stage is $\sum_{i,j} a_{ij}(x) u_i v_j$ or $u' A(x) v$, where $A(x)$ is the $n \times m$ matrix with components $a_{ij}(x)$ (u and v are viewed as column vectors, and a prime denotes transposition).

The state evolves according to transition probabilities $q_{xy}(i,j)$, where i and j are the moves selected by the minimizer and the maximizer, respectively (here y represents the next state and game to be played after moves i and j are chosen at the game represented by x). When the state is x, under u and v, the state transition probabilities are

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

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

As shown by Shapley [Sha53], the problem can be formulated as a fixed point problem involving the mapping H given by

$$\begin{aligned} H(x,u,v,J) &= u' A(x) v + \alpha \sum_{y \in X} p_{xy}(u,v) J(y) \\ &= u' \left(A(x) + \alpha \sum_{y \in X} Q_{xy} J(y) \right) v. \end{aligned} \tag{5.7}$$

It can be verified that H satisfies the contraction Assumption 1.1 [with $\xi(x) \equiv 1$]. Thus the corresponding operator T is an unweighted sup-norm contraction, and its unique fixed point J^* satisfies the Bellman equation

$$J^*(x) = (TJ^*)(x) = \min_{u \in U} \max_{v \in V} H(x,u,v,J^*), \qquad \text{for all } x \in X, \tag{5.8}$$

where U and V denote the sets of probability distributions $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_m)$, respectively.

Since the matrix defining the mapping H of Eq. (5.7),

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

is independent of u and v, we may view $J^*(x)$ as the value of a static (nonsequential) matrix game that depends on x. In particular, from a fundamental saddle point theorem for matrix games, we have

$$\min_{u \in U} \max_{v \in V} H(x,u,v,J^*) = \max_{v \in V} \min_{u \in U} H(x,u,v,J^*), \qquad \text{for all } x \in X. \tag{5.9}$$

It was shown by Shapley [Sha53] that the strategies obtained by solving the static saddle point problem (5.9) correspond to a saddle point of the sequential game in the space of strategies. Thus once we find J^* as the fixed point of the mapping T [cf. Eq. (5.8)], we can obtain equilibrium policies for the minimizer and maximizer by solving the matrix game (5.9).

Example 5.1.2 (Undiscounted Finite-State Markov Games with a Termination State)

Here the problem is the same as in the preceding example, except that there is no discount factor ($\alpha = 1$), and in addition to the states in X, there is a termination state t that is cost-free and absorbing. In this case the mapping H is given by

$$H(x,u,v,J) = u' \left(A(x) + \sum_{y \in X} Q_{xy} J(y) \right) v, \tag{5.10}$$

cf. Eq. (5.7), where the matrix of transition probabilities Q_{xy} may be substochastic, while T has the form

$$(TJ)(x) = \min_{u \in U} \max_{v \in V} H(x,u,v,J). \tag{5.11}$$

Assuming that the termination state t is reachable with probability one under all policy pairs, it can be shown that the mapping H satisfies the contraction Assumption 1.1, so results and algorithms that are similar to the ones for the preceding example apply. This reachability assumption, however, is restrictive and is not satisfied when the problem has a semicontractive character, whereby $T_{\mu,\nu}$ is a contraction under some policy pairs but not for others. In this case the analysis is more complicated and requires the notion of proper and improper policies from single-player stochastic shortest path problems; see the papers [BeT91], [PaB99], [YuB13a], [Yu14].

In the next section, we will view our abstract minimax problem, involving the Bellman equation (5.6), as an optimization by a single player who minimizes against a worst-case response by an antagonistic opponent/maximizer, and we will describe the corresponding PI algorithm. This algorithm has been known for the case of Markov games since the 1960s. We will highlight the main weakness of this algorithm: the computational cost of the policy evaluation operation, which involves the solution of the maximizer's problem for a fixed policy of the minimizer. We will then discuss an attractive proposal by Pollatschek and Avi-Itzhak [PoA69] that overcomes this difficulty, albeit with an algorithm that requires restrictive assumptions for its validity. Then, in Section 5.3, we will introduce and analyze a new algorithm, which maintains the attractive structure of the Pollatschek and Avi-Itzhak algorithm without requiring restrictive assumptions. We will also show the validity of our algorithm in the context of a distributed asynchronous implementation, as well as in an on-line context, which involves one-state-at-a-time policy improvement, with the states generated by an underlying dynamic system or Markov chain.

5.2 RELATIONS TO SINGLE-PLAYER ABSTRACT DP FORMULATIONS

In this section, we will reformulate our minimax problem in a way that will bring to bear the theory of Chapter 2. In particular, we will view the problem of finding a fixed point of the minimax operator T of Eq. (5.4) [cf. the Bellman equation (5.6)] as a single-player optimization problem by redefining T in terms of the mapping $\overline{H}$ given by

$$\overline{H}(x,u,J) = \sup_{v\in V(x)} H(x,u,v,J), \qquad x\in X,\ u\in U(x),\ J\in B(X). \tag{5.12}$$

In particular, we write T as

$$(TJ)(x) = \inf_{u\in U(x)} \overline{H}(x,u,J), \qquad x\in X, \tag{5.13}$$

or equivalently, by introducing for each $\mu\in\mathcal{M}$ the operator $\overline{T}_\mu$ given by

$$(\overline{T}_\mu J)(x) = \overline{H}\big(x,\mu(x),J\big) = \sup_{v\in V(x)} H\big(x,\mu(x),v,J\big), \qquad x\in X, \tag{5.14}$$

we write T as

$$(TJ)(x) = \inf_{\mu\in\mathcal{M}} (\overline{T}_\mu J)(x), \qquad x\in X. \tag{5.15}$$

Our contraction assumption implies that all the operators $\overline{T}_\mu$, $\mu\in\mathcal{M}$, as well as the operator T are weighted sup-norm contractions from $B(X)$ to $B(X)$, with modulus α.

Thus the single-player weighted sup-norm contractive DP framework of Chapter 2 applies directly to the operator T as defined by Eq. (5.15). In particular, to apply this framework to a minimax problem, we start from the mapping $\overline{H}$ of Eq. (5.12), which defines $\overline{T}_\mu$ via Eq. (5.14), and then T, using Eq. (5.15).

PI Algorithms

In view of the preceding transformation of our minimax problem to the single-player abstract DP formalism, the PI algorithms developed for the latter apply, and in fact these algorithms have been known for a long time for the special case of finite-state Markov games, cf. Examples 5.1.1 and 5.1.2.

In particular, the standard form of PI generates iteratively a sequence of policies $\{\mu^t\}$. The typical iteration starts with μ^t and computes μ^{t+1} with a minimization that involves the optimal cost function of a maxi-

mizer's abstract DP problem with the minimizer's policy fixed at μ^t, as follows:†

Iteration $(t+1)$ of Abstract PI Algorithm from the Minimizer's Point of View

Given μ^t, generate μ^{t+1} with a two-step process:

(a) **Policy evaluation**, which computes J_{μ^t} as the unique fixed point of the mapping $\overline{T}_{\mu^t}$ given by Eq. (5.14), i.e.,

$$J_{\mu^t} = \overline{T}_{\mu^t} J_{\mu^t}, \tag{5.16}$$

or equivalently

$$J_{\mu^t}(x) = \max_{v \in V(x)} H\big(x, \mu^t(x), v, J_{\mu^t}\big), \qquad x \in X. \tag{5.17}$$

(b) **Policy improvement**, which computes μ^{t+1} as a policy that satisfies

$$\overline{T}_{\mu^{t+1}} J_{\mu^t} = T J_{\mu^t}, \tag{5.18}$$

or equivalently

$$\mu^{t+1}(x) \in \arg\min_{u \in U(x)} \left(\max_{v \in V(x)} H(x, u, v, J_{\mu^t}) \right), \qquad x \in X. \tag{5.19}$$

There are also *optimistic forms of PI*, which starting with a function $J^0 \in B(X)$, generate a sequence of function-policy pairs $\{J^t, \mu^t\}$ with the algorithm

$$\overline{T}_{\mu^t} J^t = T J^t, \qquad J^{t+1} = \overline{T}_{\mu^t}^{m_t} J^t, \qquad k = 0, 1, \ldots, \tag{5.20}$$

where $\{m_t\}$ is a sequence of positive integers; see Section 2.5. Here the policy evaluation operation (5.16) that finds the fixed point of the mapping $\overline{T}_{\mu^t}$ is approximated by m_t value iterations using $\overline{T}_{\mu^t}$, and starting from J^t, as in the second equation of (5.20). The convergence of the abstract forms of these PI algorithms has been established under the additional

† Policy improvement involves an optimization operation that defines the new/improved policy. Throughout this chapter, and in the context of PI algorithms, we implicitly assume that this optimization can be carried out, i.e., that the optimum is attained, and write accordingly "min" and "max" in place of "inf" and "sup," respectively.

monotonicity assumption

$$\overline{T}_\mu J \leq \overline{T}_\mu J' \qquad \text{for all } J, J' \in B(X) \text{ with } J \leq J', \tag{5.21}$$

which is typically satisfied in DP-type single-player and two-player problem formulations.

The drawback of the preceding PI algorithms is that the policy evaluation operation of Eq. (5.16) and its optimistic counterpart of Eq. (5.20) aim to find or approximate the fixed point of $\overline{T}_{\mu^t}$, which involves a potentially time-consuming maximization over $v \in V(x)$; cf. the definition (5.14) and Eq. (5.17). This can be seen from the fact that Eq. (5.17) is Bellman's equation for a maximizer's abstract DP problem, where the minimizer is known to use the policy μ^t. There is a PI algorithm for finite-state Markov games, due to Pollatschek and Avi-Itzhak [PoA69], which was specifically designed to avoid the use of maximization over $v \in V(x)$ in the policy evaluation operation. We present this algorithm next, together with a predecessor PI algorithm, due to Hoffman and Karp [HoK66], which is in fact the algorithm (5.16)-(5.19) applied to the Markov game Example 5.1.1.

The Hoffman-Karp, and Pollatschek and Avi-Itzhak Algorithms for Finite-State Markov Games

The PI algorithm (5.16)-(5.19) for the special case of finite-state Markov games (cf. Example 5.1.1), has been proposed by Hoffman and Karp [HoK66]. It takes the form

$$J_{\mu^t}(x) = \max_{v \in V} H\big(x, \mu^t(x), v, J_{\mu^t}\big), \qquad x \in X, \tag{5.22}$$

where H is the Markov game mapping (5.7) (this is the policy evaluation step), followed by solving the static minimax problem

$$\min_{u \in U} \max_{v \in V} H(x, u, v, J_{\mu^t}), \qquad x \in X, \tag{5.23}$$

and letting μ^{t+1} be a policy that attains the minimum above (this is the policy improvement step). The policy improvement subproblem (5.23) is a matrix saddle point problem, involving the matrix

$$A(x) + \sum_{y \in X} Q_{xy} J_{\mu^t}(y),$$

[cf. Eq. (5.10)], which is easily solvable by linear programming for each x (this is well-known in the theory of matrix games).

However, the policy evaluation step (5.22) involves the solution of the maximizer's Markov decision problem, for the fixed policy μ^t of the minimizer. This can be a quite difficult problem that requires an expensive

computation. The same is true for a modified version of the Hoffman-Karp algorithm proposed by van der Wal [Van78], which involves an approximate policy evaluation, based on a limited number of value iterations, as in the optimistic PI algorithm (5.20). The computational difficulty of the policy evaluation phase of the Hoffman-Karp algorithm is also shared by other PI algorithms for sequential games that have been suggested in the literature in subsequent works (e.g., Patek and Bertsekas [PaB99], and Yu [Yu14]).

Following the publication of the Hoffman-Karp algorithm, another PI algorithm for finite-state Markov games was proposed by Pollatschek and Avi-Itzhak [PoA69], and has attracted considerable attention because it is more computationally expedient. It generates a sequence of minimizer-maximizer policy pairs $\{\mu^t, \nu^t\}$ and corresponding game value functions $J_{\mu^t,\nu^t}(x)$, starting from each state x. In particular, the standard form of PI generates iteratively a sequence of policies $\{\mu^t\}$. We give this algorithm in an abstract form, which parallels the PI algorithm (5.16)-(5.19). The typical iteration starts with a pair (μ^t, ν^t) and computes a pair (μ^{t+1}, ν^{t+1}) as follows:

Iteration $(t+1)$ of the Pollatschek and Avi-Itzhak PI Algorithm in Abstract Form

Given (μ^t, ν^t), generate (μ^{t+1}, ν^{t+1}) with a two-step process:

(a) **Policy evaluation**, which computes J_{μ^t,ν^t} by solving the fixed point equation

$$J_{\mu^t,\nu^t}(x) = H\big(x, \mu^t(x), \nu^t(x), J_{\mu^t,\nu^t}\big), \qquad x \in X. \tag{5.24}$$

(b) **Policy improvement**, which computes (μ^{t+1}, ν^{t+1}) by solving the saddle point problem

$$\min_{u\in U} \max_{v\in V} H(x, u, v, J_{\mu^t,\nu^t}), \qquad x \in X. \tag{5.25}$$

The Pollatschek and Avi-Itzhak algorithm [PoA69] is the algorithm (5.24)-(5.25), specialized to the Markov game case of the mapping H that involves the matrix

$$A(x) + \sum_{y\in X} Q_{xy} J_{\mu^t,\nu^t}(y),$$

similar to the Hoffman-Karp algorithm, cf. Eq. (5.10). A key observation is that the policy evaluation operation (5.24) is computationally comparable to policy evaluation in a single-player Markov decision problem, i.e., solving a linear system of equations. In particular, it does not involve solution of the Markov decision problem of the maximizer like the Hoffman-Karp

PI algorithm [cf. Eq. (5.22)], or its approximate solution by multiple value iterations, as in the van der Wal optimistic version (5.20) for Markov games.

Computational studies have shown that the Pollatschek and Avi-Itzhak algorithm converges much faster than its competitors, *when it converges* (see Breton et al. [BFH86], and also Filar and Tolwinski [FiT91], who proposed a modification of the algorithm). Moreover, the number of iterations required for convergence is fairly small. This is consistent with an interpretation given by Pollatschek and Avi-Itzhak in their paper [PoA69], where they have shown that their algorithm coincides with a form of Newton's method for solving the fixed point/Bellman equation $J = TJ$ (see Fig. 5.2.1).† The close connection of PI with Newton's method is well-known in control theory and operations research, through several works, including Kleinman [Kle68] for linear-quadratic optimal control problems, and Puterman and Brumelle [PuB78], [PuB79] for more abstract settings. Its significance in reinforcement learning contexts has been discussed at length in the author's recent books [Ber20] and [Ber22]; see also Section 1.3.

Unfortunately, however, the Pollatschek and Avi-Itzhak algorithm is valid only under restrictive assumptions (given in their paper [PoA69]). The difficulty is that Newton's method applied to the Bellman equation $J = TJ$ need not be globally convergent when the operator T corresponds to a minimax problem. This is illustrated in Fig. 5.2.1, which also illustrates why Newton's method (equivalently, the PI algorithm) is globally

† Newton's method for solving a general fixed point problem of the form $z = F(z)$, where z is an n-dimensional vector, operates as follows: At the current iterate z_k, we linearize F and find the solution z_{k+1} of the corresponding linear fixed point problem, obtained using a first order Taylor expansion:

$$z_{k+1} = F(z_k) + \frac{\partial F(z_k)}{\partial z}(z_{k+1} - z_k),$$

where $\partial F(z_k)/\partial z$ is the $n \times n$ Jacobian matrix of F evaluated at the n-dimensional vector z_k. The most commonly given convergence rate property of Newton's method is *quadratic convergence*. It states that near the solution z^*, we have

$$\|z_{k+1} - z^*\| = O\big(\|z_k - z^*\|^2\big),$$

where $\|\cdot\|$ is the Euclidean norm, and holds assuming the Jacobian matrix exists and is Lipschitz continuous (see [Ber16c], Section 1.4). Qualitatively similar results hold under other assumptions. In particular a superlinear convergence statement (suitably modified to account for lack of differentiability of F) can be proved for the case where $F(z)$ has components that are either monotonically increasing or monotonically decreasing, and either concave or convex. In the case of the Pollatschek and Avi-Itzhak algorithm, the main difficulty is that the concavity/convexity condition is violated; see Fig. 5.2.1.

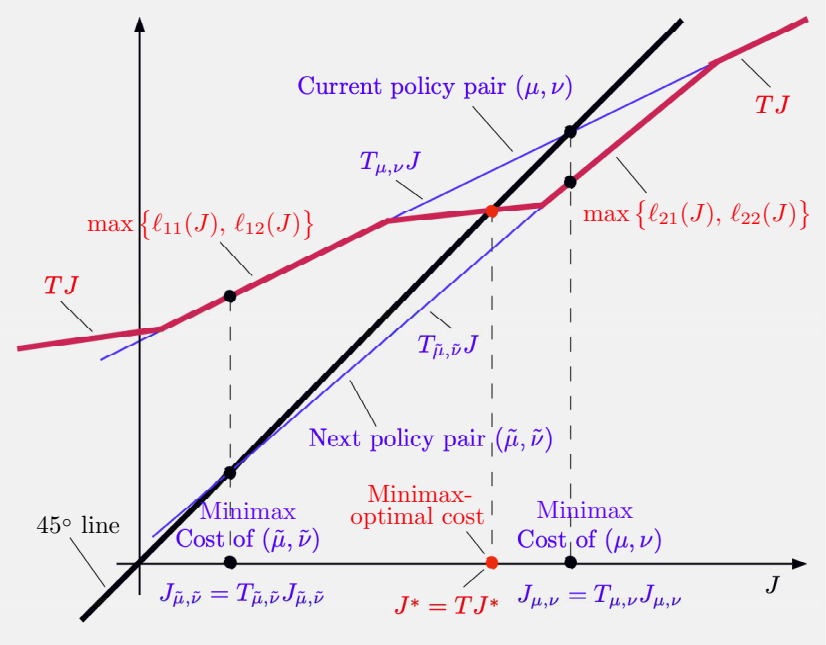


Figure 5.2.1 Schematic illustration of the abstract minimax PI algorithm (5.24)-(5.25) in the case of a minimax problem involving a single state, in addition to a termination state t; cf. Example 5.1.2. We have $J^*(t) = 0$ and $(TJ)(t) = 0$ for all J with $J(t) = 0$, so that the operator T can be graphically represented in just one dimension (denoted by J) that corresponds to the nontermination state. This makes it easy to visualize T and geometrically interpret why Newton's method does not converge. Because the operator T may be neither convex nor concave for a minimax problem, the algorithm may cycle between pairs (μ, ν) and $(\tilde{\mu}, \tilde{\nu})$, as shown in the figure. By contrast in a (single-player) finite-state Markovian decision problem, T has piecewise linear and concave components, and the PI algorithm converges in a finite number of iterations. The figure illustrates an operator T of the form

$$TJ = \min\Big\{\max\big\{\ell_{11}(J), \ell_{12}(J)\big\}, \max\big\{\ell_{21}(J), \ell_{22}(J)\big\}\Big\},$$

where $\ell_{ij}(J)$, are linear functions of J, corresponding to the choices $i = 1, 2$ of the minimizer and $j = 1, 2$ of the maximizer. Thus TJ is the minimum of the convex functions

$$\max\big\{\ell_{11}(J), \ell_{12}(J)\big\} \qquad \text{and} \qquad \max\big\{\ell_{21}(J), \ell_{22}(J)\big\},$$

as shown in the figure. Newton's method linearizes TJ at the current iterate [i.e., replaces TJ with one of the four linear functions $\ell_{ij}(J)$, $i = 1, 2$, $j = 1, 2$ (the one attaining the min-max at the current iterate)] and solves the corresponding linear fixed point problem to obtain the next iterate.

convergent in the case of a single-player finite-state Markov decision problem, as is well known. In this case each component $(TJ)(x)$ of the function TJ is concave and piecewise linear, thereby guaranteeing the finite termination of the PI algorithm. This is not true in the case of finite-state minimax problems and Markov games. The difficulty is that *the functions* $(TJ)(x)$ *may be neither convex nor concave in* J, even though they are piecewise linear and have a monotonicity property (cf. Fig. 5.2.1). In fact a two-state example where the Pollatschek and Avi-Itzhak algorithm does not converge to J^* was given by van der Wal [Van78]. This example involves a single state in addition to a termination state, and the algorithm oscillates similar to Fig. 5.2.1. Note that the Hoffman-Karp algorithm does not admit an interpretation as Newton's method, and is not subject to the convergence difficulties of the Pollatschek and Avi-Itzhak algorithm.

5.3 A NEW PI ALGORITHM FOR ABSTRACT MINIMAX DP PROBLEMS

In this section, we will introduce modifications to the Pollatschek and Avi-Itzhak algorithm, and its abstract version (5.24)-(5.25), given in the preceding section, with the aim to enhance its convergence properties, while maintaining its favorable structure. These modifications will apply to a general minimax problem of finding a fixed point of a suitable contractive operator, and offer the additional benefit that they allow asynchronous, distributed, and on-line implementations. They are also suitable for approximations based on an aggregation approach, which will be discussed in Section 5.5.

Our PI algorithm is motivated by a line of analysis and corresponding algorithms introduced by Bertsekas and Yu [BeY10], [BeY12] for discounted infinite horizon DP problems, and by Yu and Bertsekas [YuB13a] for stochastic shortest path problems (with both proper and improper policies). These algorithms were also presented in general abstract form in the author's book [Ber12a], as well as in Section 2.6.3. The PI algorithm of this section uses a similar abstract formulation, but replaces the single mapping that is minimized in these works with two mappings, one of which is minimized while the other is maximized. Mathematically, the difficulty of the Pollatschek and Avi-Itzhak algorithm is that the policies (μ^{t+1}, ν^{t+1}) obtained from the policy improvement/static game (5.25) are not "improved" in a clear sense, such as

$$J_{\mu^{t+1},\nu^{t+1}}(x) \leq J_{\mu^t,\nu^t}(x), \qquad \text{for all } x \in X,$$

as they are in the case of single-player DP, where a policy improvement property is central in the standard convergence proof of single-player PI. Our algorithm, however, does not rely on policy improvement, but rather derives its validity from a *uniform contraction property of an underlying*

operator, to be given in Section 5.4 (cf. Prop. 5.4.2). In fact, *our algorithm does not require the monotonicity assumption (5.21) for its convergence*, and thus it can be used in minimax problems that are beyond the scope of DP.†

As an aid to understanding intuitively the abstract framework of this section, we note that it is patterned after a multistage process, whereby at each stage, the following sequence of events is envisioned (cf. Fig. 5.3.1):

(1) We start at some state x_1 from a space X_1.

(2) The minimizer, knowing x_1, chooses a control $u \in U(x_1)$. Then a new state x_2 from a space X_2 is generated as a function of (x_1, u). (It is possible that $X_1 = X_2$, but for greater generality, we do not assume so. Also the transition from x_1 to x_2 may involve a random disturbance; see the subsequent Example 3.3.)

(3) The maximizer, knowing x_2, chooses a control $v \in V(x_2)$. Then a new state $\overline{x}_1 \in X_1$ is generated.

(4) The next stage is started at $\overline{x}_1$ and the process is repeated.

If we start with $x_1 \in X_1$, this sequence of events corresponds to finding the optimal minimizer policy against a worst case choice of the maximizer, and the corresponding min-max value is denoted by $J_1^*(x_1)$. Symmetrically, if we start with $x_2 \in X_2$, this sequence of events corresponds to finding the optimal maximizer policy against a worst case choice of the minimizer, and the corresponding max-min value is denoted by $J_2^*(x_2)$.

This type of framework can be viewed within the context of the theory of zero-sum games in extensive form, a methodology with a long history [Kuh53]. Games in extensive form involve sequential/alternating choices by the players with knowledge of prior choices. By contrast, for games in simultaneous form, such as the Markov games of the preceding section, the players make their choices without being sure of the other player's choices.

Fixed Point Formulation

We consider the space of bounded functions of $x_1 \in X_1$, denoted by $B(X_1)$, and the space of bounded functions of $x_2 \in X_2$, denoted by $B(X_2)$, with respect to the norms $\|J_1\|_1$ and $\|J_2\|_2$ defined by

$$\|J_1\|_1 = \sup_{x_1 \in X_1} \frac{\big|J_1(x_1)\big|}{\xi_1(x_1)}, \qquad \|J_2\|_2 = \sup_{x_2 \in X_2} \frac{\big|J_2(x_2)\big|}{\xi_2(x_2)}, \tag{5.26}$$

† For example, our algorithm can be used for the asynchronous distributed computation of fixed points of concave operators, arising in fields like economics and polulation dynamics. The key fact here is that a concave function can be described as the minimum of a collection of linear functions through the classical conjugacy operation.

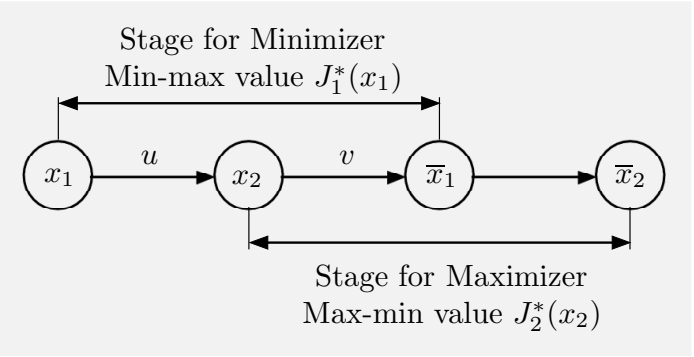


Figure 5.3.1 Schematic illustration of the sequence of events at each stage of the minimax problem. We start at $x_1 \in X_1$. The minimizer chooses a control $u \in U(x_1)$, a new state $x_2 \in X_2$ is generated, the maximizer chooses a $v \in V(x_2)$, and a new state $\overline{x}_1 \in X_1$ is generated, etc. If the stage begins at x_2 rather than x_1, this corresponds to the max-min problem. The corresponding min-max and max-min values are $J_1^*(x_1)$ and $J_2^*(x_2)$, respectively.

where ξ_1 and ξ_2 are positive weighting functions, respectively. We also consider the space $B(X_1) \times B(X_2)$ with the norm

$$\big\|(J_1, J_2)\big\| = \max\big\{\|J_1\|_1,\ \|J_2\|_2\big\}. \tag{5.27}$$

We will be interested in finding a pair of functions (J_1^*, J_2^*) that are the fixed point of mappings

$$H_1 : X_1 \times U \times B(X_2) \mapsto B(X_1), \qquad H_2 : X_2 \times V \times B(X_1) \mapsto B(X_2),$$

in the following sense: for all $x_1 \in X_1$ and $x_2 \in X_2$,

$$J_1^*(x_1) = \inf_{u \in U(x_1)} H_1(x_1, u, J_2^*), \qquad J_2^*(x_2) = \sup_{v \in V(x_2)} H_2(x_2, v, J_1^*). \tag{5.28}$$

These two equations form an abstract version of Bellman's equation for the infinite horizon sequential min-max problem described by the sequence of events (1)-(4) given earlier. We will assume later (see Section 5.4) that H_1 and H_2 have a contraction property like Assumption 5.1.1, which will guarantee that (J_1^*, J_2^*) is the unique fixed point within $B(X_1) \times B(X_2)$.

Note that the fixed point problem (5.28) involves both min-max and max-min values, without assuming that they are equal. By contrast the algorithms of Section 5.2 aim to compute only the min-max value. In the case of a Markov game (cf. Examples 5.1.1 and 5.1.2), the min-max value is equal to the max-min value, but in general min-max may not be equal to max-min, and the algorithms of Section 5.2 will only find min-max explicitly. We will next provide an example to interpret J_1^* and J_2^* as the min-max and max-min value functions of a sequential infinite horizon problem involving the sequence of events (1)-(4) given earlier.

Example 5.3.1 (Discounted Minimax Control - Explicit Separation of the Two Players)

In this formulation of a discounted minimax control problem, the states of the minimizer and the maximizer, respectively, at time k are denoted by $x_{1,k} \in X_1$ and $x_{2,k} \in X_2$, and they evolve according to

$$x_{2,k+1} = f_1(x_{1,k}, u_k), \qquad x_{1,k+1} = f_2(x_{2,k+1}, v_k), \qquad k = 0, 1, \ldots. \tag{5.29}$$

The mappings H_1 and H_2 are given by

$$H_1(x_1, u, J_2) = g_1(x_1, u) + \alpha J_2\big(f_1(x_1, u)\big), \tag{5.30}$$

$$H_2(x_2, v, J_1) = g_2(x_2, v) + \alpha J_1\big(f_2(x_2, v)\big), \tag{5.31}$$

where g_1 and g_2 are stage cost functions for the minimizer and the maximizer, respectively. The corresponding fixed point problem of Eq. (5.28) has the form

$$J_1^*(x_1) = \inf_{u \in U(x_1)} \Big[g_1(x_1, u) + \alpha J_2^*\big(f_1(x_1, u)\big)\Big], \tag{5.32}$$

$$J_2^*(x_2) = \sup_{v \in V(x_2)} \Big[g_2(x_2, v) + \alpha J_1^*\big(f_2(x_2, v)\big)\Big]. \tag{5.33}$$

Example 5.3.2 (Markov Games)

We will show that the discounted Markov game of Example 5.1.1 can be reformulated within our fixed point framework of Eq. (5.28) by letting $X_1 = X$, $X_2 = X \times U$, and by redefining the minimizer's control to be a probability distribution $(u_1, \ldots, u_n)$, and the maximizer's control to be one of the m possible choices $j = 1, \ldots, m$.

To introduce into our problem formulation an appropriate contraction structure that we will need in the next section, we use a scaling parameter β such that

$$\beta > 1, \qquad \alpha\beta < 1. \tag{5.34}$$

The idea behind the use of the scaling parameter β is to introduce discounting into the stages of both the minimizer and the maximizer. We consider functions $J_1^*(x)$ and $J_2^*(x, u)$ that solve the equations

$$J_1^*(x) = \frac{1}{\beta} \min_{u \in U} J_2^*(x, u), \tag{5.35}$$

$$\begin{aligned} J_2^*(x, u) = \max \Bigg\{ & u' \left(A(x) + \alpha\beta \sum_{y \in X} Q_{xy} J_1^*(y) \right) (1), \ldots, \\ & u' \left(A(x) + \alpha\beta \sum_{y \in X} Q_{xy} J_1^*(y) \right) (m) \Bigg\}, \end{aligned} \tag{5.36}$$

where

$$\left(A(x) + \alpha\beta \sum_{y \in X} Q_{xy} J_1^*(y) \right)(j), \qquad j = 1, \ldots, m, \tag{5.37}$$

denotes the jth column of the matrix

$$A(x) + \alpha\beta \sum_{y \in X} Q_{xy} J_1^*(y). \tag{5.38}$$

It can be seen from these equations that

$$J_2^*(x,u) = \max_{v \in V} u' \left(A(x) + \alpha\beta \sum_{y \in X} Q_{xy} J_1^*(y) \right) v, \tag{5.39}$$

since the maximization over $v \in V$ above is equivalent to the maximization over the m alternatives in Eq. (5.36), which correspond to the extreme points of the unit simplex V. Thus from Eqs. (5.35) and (5.39), it follows that the function βJ_1^* satisfies

$$(\beta J_1^*)(x) = \min_{u \in U} \max_{v \in V} u' \left(A(x) + \alpha \sum_{y \in X} Q_{xy} (\beta J_1^*)(y) \right) v,$$

so it coincides with the vector of equilibrium values J^* of the Markov game formulation of Example 5.1.1 [cf. Eq. (5.7)-(5.8)].

Note that $J_2^*(x, \cdot)$ is a piecewise linear function of u with at most m pieces, defined by the columns (5.37). Thus the fixed point (J_1^*, J_2^*) can be stored and be computed as a finite set of numbers: the real numbers $J_1^*(x)$, $x \in X$, which can also be used to compute the $n \times m$ matrices

$$A(x) + \alpha\beta \sum_{y \in X} Q_{xy} J_1^*(y), \qquad x \in X,$$

whose columns define $J_2^*(x,u)$, cf. Eq. (5.36).

We finally observe that the two equations (5.35) and (5.39) can be written in the form (5.28), with $x_1 = x$, $x_2 = (x,u)$, and H_1, H_2 defined by

$$H_1(x,u,J_2) = \frac{1}{\beta} J_2(x,u),$$

$$H_2(x,u,v,J_1) = u' \left(A(x) + \alpha\beta \sum_{y \in X} Q_{xy} J_1^*(y) \right) v.$$

An important area of application of our two-player framework is control under set-membership uncertainty within a game-against-nature formulation, whereby nature is modeled as an antagonistic opponent choosing $v \in V(x_2)$. Here only the min-max value is of practical interest, but our subsequent PI methodology will find the max-min value as well. We provide two examples of this type of formulation.

Example 5.3.3 (Discounted Minimax Control Over an Infinite Horizon)

Consider a dynamic system whose state evolves at each time k according to a discrete time equation of the form

$$x_{k+1} = f(x_k, u_k, v_k), \qquad k = 0, 1, \ldots, \tag{5.40}$$

where x_k is the state, u_k is the control to be selected from some given set $U(x_k)$ (with perfect knowledge of x_k), and v_k is a disturbance that is selected by an antagonistic nature from a set $V(x_k, u_k)$ [with perfect knowledge of (x_k, u_k)]. A cost $g(x_k, u_k, v_k)$ is incurred at time k, it is accumulated over an infinite horizon, and it is discounted by $\alpha \in (0, 1)$. The Bellman equation for this problem is

$$J^*(x) = \inf_{u \in U(x)} \sup_{v \in V(x,u)} \Big[g(x, u, v) + \alpha J^*\big(f(x, u, v)\big) \Big], \tag{5.41}$$

and the optimal cost function J^* is the unique fixed point of this equation, assuming that the cost per stage g is a bounded function.

To reformulate this problem into the fixed point format (5.28), we identify the minimizer's state x_1 with the state x of the system (5.40), and the maximizer's state x_2 with the state-control pair (x, u). We also introduce a scaling parameter β that satisfies $\beta > 1$ and $\alpha\beta < 1$; cf. Eq. (5.34). We define H_1 and H_2 as follows:

$$H_1(x, u, J_2) \text{ maps } (x, u, J_2) \text{ to the real value } \frac{1}{\beta} J_2(x, u),$$

$$H_2(x, u, v, J_1) \text{ maps } (x, u, v, J_1) \text{ to the real value } g(x, u, v) + \alpha\beta J_1\big(f(x, u, v)\big).$$

Then the resulting fixed point problem (5.28) takes the form

$$(\beta J_1^*)(x) = \inf_{u \in U(x)} J_2^*(x, u),$$

$$J_2^*(x, u) = \sup_{v \in V(x,u)} \Big[g(x, u, v) + \alpha(\beta J_1^*)\big(f(x, u, v)\big) \Big],$$

which is equivalent to the Bellman equation (5.41) with $J^* = \beta J_1^*$.

Example 5.3.4 (Discounted Minimax Control with Partially Stochastic Disturbances)

Consider a dynamic system such as the one of Eq. (5.40) in the preceding example, except that there is an additional stochastic disturbance w with

known conditional probability distribution given (x, u, v). Thus the state evolves at each time k according to

$$x_{k+1} = f(x_k, u_k, v_k, w_k), \qquad k = 0, 1, \ldots, \tag{5.42}$$

and the cost per stage is $g(x_k, u_k, v_k, w_k)$. The Bellman equation now is

$$J^*(x) = \inf_{u \in U(x)} \sup_{v \in V(x,u)} E_w \Big\{ g(x, u, v, w) + \alpha J^*\big(f(x, u, v, w)\big) \;\Big|\; x, u, v \Big\}, \tag{5.43}$$

and J^* is the unique fixed point of this equation, assuming that g is a bounded function.

Similar to Example 5.3.3, we let the minimizer's state be x, and the maximizer's state be (x, u), we introduce a scaling parameter β that satisfies $\beta > 1$ and $\alpha\beta < 1$; cf. Eq. (5.34), and we define H_1 and H_2 as follows:

$$H_1(x, u, J_2) \text{ maps } (x, u, J_2) \text{ to the real value } \frac{1}{\beta} J_2(x, u),$$

$$H_2(x, u, v, J_1) \text{ maps } (x, u, v, J_1)$$

$$\text{to the real value } E_w \Big\{ g(x, u, v, w) + \alpha\beta J_1\big(f(x, u, v, w)\big) \;\Big|\; x, u, v \Big\}.$$

The resulting fixed point problem (5.28) takes the form

$$(\beta J_1^*)(x) = \inf_{u \in U(x)} J_2^*(x, u),$$

$$J_2^*(x, u) = \sup_{v \in V(x,u)} E_w \Big\{ g(x, u, v, w) + \alpha(\beta J_1^*)\big(f(x, u, v, w)\big) \;\Big|\; x, u, v \Big\}.$$

which is equivalent to the Bellman equation (5.43) with $J^* = \beta J_1^*$.

Other examples of application of our abstract fixed point framework (5.28) include two-player versions of multiplicative and exponential cost problems. One-player cases of these problems have a long tradition in DP; see e.g., Jacobson [Jac73], Denardo and Rothblum [DeR79], Whittle [Whi81], Rothblum [Rot84], Patek [Pat01]. Abstract versions of these problems come under the general framework of *affine monotonic problems*, for which we refer to Section 3.5.2 and the author's paper [Ber19a] for further discussion. Two-player versions of affine monotonic problems involve a state space $X = \{1, \ldots, n\}$, and the mapping

$$H(x, u, v, J) = g(x, u, v) + \sum_{y=1}^{n} A_{xy}(u, v) J(y), \qquad x = 1, \ldots, n,$$

where g and A_{xy} satisfy for all $x, y = 1, \ldots, n,\ u \in U(x),\ v \in V(x)$,

$$g(x, u, v) \geq 0, \qquad A_{xy}(u, v) \geq 0.$$

Our PI algorithms can be suitably adapted to address these problems, along the lines of the preceding examples. Of course, the corresponding convergence analysis may pose special challenges, depending on whether our assumptions of the next section are satisfied.

"Naive" PI Algorithms

A PI algorithm for the fixed point problem (5.28), which is patterned after the Pollatschek and Avi-Itzhak algorithm, generates a sequence of policy pairs $\{\mu^t, \nu^t\} \subset \mathcal{M} \times \mathcal{N}$ and corresponding sequence of cost function pairs $\{J_{1,\mu^t,\nu^t}, J_{2,\mu^t,\nu^t}\} \subset B(X_1) \times B(X_2)$. We use the term "naive" to indicate that the algorithm does not address adequately the convergence issue of the underlying Newton's method.† Given $\{\mu^t, \nu^t\}$ it generates $\{\mu^{t+1}, \nu^{t+1}\}$ with a two-step process as follows:

(a) **Policy evaluation**, which computes the functions $\{J_{1,\mu^t,\nu^t}, J_{2,\mu^t,\nu^t}\}$ by solving the fixed point equations

$$J_{1,\mu^t,\nu^t}(x_1) = H_1\big(x_1, \mu^t(x), J_{2,\mu^t,\nu^t}\big), \qquad x_1 \in X_1, \tag{5.44}$$

$$J_{2,\mu^t,\nu^t}(x_2) = H_2\big(x_2, \nu^t(x), J_{1,\mu^t,\nu^t}\big), \qquad x_2 \in X_2. \tag{5.45}$$

(b) **Policy improvement**, which computes (μ^{t+1}, ν^{t+1}) with the minimizations

$$\mu^{t+1}(x_1) \in \arg\min_{u \in U(x_1)} H_1\big(x_1, u, J_{2,\mu^t,\nu^t}\big), \qquad x_1 \in X_1, \tag{5.46}$$

$$\nu^{t+1}(x_2) \in \arg\max_{v \in V(x_2)} H_2\big(x_2, v, J_{1,\mu^t,\nu^t}\big), \qquad x_2 \in X_2. \tag{5.47}$$

This algorithm resembles the abstract version of the Pollatschek and Avi-Itzhak algorithm (5.24)-(5.25) in that it involves simple policy evaluations, which do not require the solution of a multistage DP problem for either the minimizer or the maximizer. Unfortunately, however, the algorithm (5.44)-(5.47) cannot be proved to be convergent, as it does not deal effectively with the oscillatory behavior illustrated in Fig. 5.2.1.

An optimistic version of the PI algorithm (5.44)-(5.47) evaluates the fixed point pair $(J_{1,\mu^t,\nu^t}, J_{2,\mu^t,\nu^t})$ approximately, by using some number, say $\bar{k} \geq 1$, of value iterations. It has the form

$$J_{1,k+1}(x_1) = H_1\big(x_1, \mu^t(x_1), J_{2,k}\big), \qquad x_1 \in X_1, \qquad k = 0, 1, \ldots, \bar{k} - 1, \tag{5.48}$$

$$J_{2,k+1}(x_2) = H_2\big(x_2, \nu^t(x_2), J_{1,k}\big), \qquad x_2 \in X_2, \qquad k = 0, 1, \ldots, \bar{k} - 1, \tag{5.49}$$

starting from an initial approximation $(J_{1,0}, J_{2,0})$, instead of solving the fixed point equations (5.44)-(5.45). As $\bar{k}$ (i.e., the number of value iterations used for policy evaluation) increases, the pair $(J_{1,\bar{k}}, J_{2,\bar{k}})$ converges to

† We do not mean the term in a pejorative sense. In fact the Pollatschek and Avi-Itzhak paper [PoA69] embodies original ideas, includes sophisticated and insightful analysis, and has stimulated considerable followup work.

$(J_{1,\mu^t,\nu^t}, J_{2,\mu^t,\nu^t})$, and the optimistic and nonoptimistic policy evaluations coincide in the limit (under suitable contraction assumptions to be introduced in the next section). Still the PI algorithm that uses this optimistic policy evaluation, followed by a policy improvement operation similar to Eqs. (5.46)-(5.47), i.e.,

$$\mu^{t+1}(x_1) \in \arg\min_{u\in U(x_1)} H_1\big(x_1, u, J_{2,\bar{k}+1}\big), \qquad x_1 \in X_1, \tag{5.50}$$

$$\nu^{t+1}(x_2) \in \arg\max_{v\in V(x_2)} H_2\big(x_2, v, J_{1,\bar{k}+1}\big), \qquad x_2 \in X_2, \tag{5.51}$$

cannot be proved convergent and is subject to oscillatory behavior. However, this optimistic algorithm can be made convergent through modifications that we describe next.

Our Distributed Optimistic Abstract PI Algorithm

Our PI algorithm for finding the solution (J_1^*, J_2^*) of the Bellman equation (5.28) has structural similarity with the "naive" PI algorithm that uses optimistic policy evaluations of the form (5.48)-(5.49) and policy improvements of the form (5.50)-(5.51). It differs from the PI algorithms of the preceding section, such as the Hoffman-Karp and van der Wal algorithms, in two ways:

(a) *It treats symmetrically the minimizer and the maximizer*, in that it aims to find both the min-max and the max-min cost functions, which are J_1^* and J_2^*, respectively, and it ignores the possibility that we may have $J_1^* = J_2^*$.

(b) *It separates the policy evaluations and policy improvements of the minimizer and the maximizer, in asynchronous fashion.* In particular, in the algorithm that we will present shortly, each iteration will consist of only one of four operations: (1) an approximate policy evaluation (consisting of a single value iteration) by the minimizer, (2) a policy improvement by the minimizer, (3) an approximate policy evaluation (consisting of a single value iteration) by the maximizer, (4) a policy improvement by the maximizer.

The order and frequency by which these four operations are performed does not affect the convergence of the algorithm, as long as all of these operations are performed infinitely often. Thus the algorithm is well suited for distributed implementation. Moreover, by executing the policy evaluation steps (1) and (3) much more frequently than the policy improvement operations (2) and (4), we obtain an algorithm involving nearly exact policy evaluation.

Our algorithm generates *two* sequences of function pairs,

$$\{J_1^t, J_2^t\} \subset B(X_1) \times B(X_2), \qquad \{V_1^t, V_2^t\} \subset B(X_1) \times B(X_2),$$

and a sequence of policy pairs:

$$\{\mu^t, \nu^t\} \subset \mathcal{M} \times \mathcal{N}.$$

The algorithm involves pointwise minimization and maximization operations on pairs of functions, which we treat notationally as follows: For any pair of functions (V, J) from within $B(X_1)$ or $B(X_2)$, we denote by $\min[V, J]$ and by $\max[V, J]$ the functions defined on $B(X_1)$ or $B(X_2)$, respectively, that take values

$$\min[V, J](x) = \min\big\{V(x), J(x)\big\}, \qquad \max[V, J](x) = \max\big\{V(x), J(x)\big\},$$

for every x in X_1 or X_2, respectively.

At iteration t, our algorithm starts with

$$J_1^t, V_1^t, J_2^t, V_2^t, \mu^t, \nu^t,$$

and generates

$$J_1^{t+1}, V_1^{t+1}, J_2^{t+1}, V_2^{t+1}, \mu^{t+1}, \nu^{t+1},$$

by executing *one* of the following four operations.†

Iteration $(t+1)$ of Distributed Optimistic Abstract PI Algorithm

Given $(J_1^t, V_1^t, J_2^t, V_2^t, \mu^t, \nu^t)$, do one of the following four operations (a)-(d):

(a) **Single value iteration for policy evaluation of the minimizer**: For all $x_1 \in X_1$, set

$$J_1^{t+1}(x_1) = H_1\big(x_1, \mu^t(x_1), \max[V_2^t, J_2^t]\big), \tag{5.52}$$

and leave $J_2^t, V_1^t, V_2^t, \mu^t, \nu^t$ unchanged, i.e., the corresponding $(t+1)$-iterates are set to the t-iterates: $J_2^{t+1} = J_2^t$, $V_1^{t+1} = V_1^t$, $V_2^{t+1} = V_2^t$, $\mu^{t+1} = \mu^t$, $\nu^{t+1} = \nu^t$.

† The choice of operation is arbitrary at iteration t, as long as each type of operation is executed for infinitely many t. It can be extended by introducing "communication delays," and state space partitioning, whereby the operations are carried out in just a subset of the corresponding state space. This is a type of asynchronous operation that was also used in the earlier works [BeY10], [BeY12], [YuB13a]. It is supported by an asynchronous convergence analysis originated in the author's papers [Ber82], [Ber83]; see also Section 2.6.1 of the present book, the book [BeT89], and the book [Ber12a], Section 2.6. This asynchronous convergence analysis applies because the mapping underlying our algorithm is a contraction with respect to a sup-norm (rather than some other norm such as an L_2 norm).

(b) **Policy improvement for the minimizer**: For all $x_1 \in X_1$, set

$$J_1^{t+1}(x_1) = V_1^{t+1}(x_1) = \min_{u \in U(x_1)} H_1\big(x_1, u, \max[V_2^t, J_2^t]\big), \quad (5.53)$$

set $\mu^{t+1}(x_1)$ to a control $u \in U(x_1)$ that attains the above minimum, and leave J_2^t, V_2^t, ν^t unchanged.

(c) **Single value iteration for policy evaluation of the maximizer**: For all $x_2 \in X_2$, set

$$J_2^{t+1}(x_2) = H_2\big(x_2, \nu^t(x_2), \min[V_1^t, J_1^t]\big), \quad (5.54)$$

and leave $J_1^t, V_1^t, V_2^t, \mu^t, \nu^t$ unchanged.

(d) **Policy improvement for the maximizer**: For all $x_2 \in X_2$, set

$$J_2^{t+1}(x_2) = V_2^{t+1}(x_2) = \max_{v \in V(x_2)} H_2\big(x_2, v, \min[V_1^t, J_1^t]\big), \quad (5.55)$$

set $\nu^{t+1}(x_2)$ to a control $v \in V(x_2)$ that attains the above maximum, and leave J_1^t, V_1^t, μ^t unchanged.

Example 5.3.5 (Our PI Algorithm for Minimax Control - Explicit Separation of the Players)

Consider the minimax control problem with explicit separation of the two players of Example 5.3.1, which involves the dynamic system $x_{1,k} \in X_1$ and $x_{2,k} \in X_2$, and they evolve according to

$$x_{2,k+1} = f_1(x_{1,k}, u_k), \qquad x_{1,k+1} = f_2(x_{2,k+1}, v_k), \qquad k = 0, 1, \ldots,$$

[cf. Eq. (5.29)]. The Bellman equation for this problem can be broken down into the two equations (5.32), (5.33):

$$J_1^*(x_1) = \inf_{u \in U(x_1)} \Big[g_1(x_1, u) + \alpha J_2^*\big(f_1(x_1, u)\big) \Big],$$

$$J_2^*(x_2) = \sup_{v \in V(x_2)} \Big[g_2(x_2, v) + \alpha J_1^*\big(f_2(x_2, v)\big) \Big].$$

In the context of this problem, the four operations (5.52)-(5.55) of our PI algorithm take the following form:

(a) **Single value iteration for policy evaluation for the minimizer**: For all $x_1 \in X_1$, set

$$J_1^{t+1}(x_1) = g_1\big(x_1, \mu^t(x_1)\big) + \alpha \max\Big[V_2^t\big(f_1(x_1, \mu^t(x_1))\big), J_2^t\big(f_1(x_1, \mu^t(x_1))\big)\Big], \tag{5.56}$$

and leave $J_2^t, V_1^t, V_2^t, \mu^t, \nu^t$ unchanged.

(b) **Policy improvement for the minimizer**: For all $x_1 \in X_1$, set

$$J_1^{t+1}(x_1) = V_1^{t+1}(x_1) = \min_{u \in U(x_1)} \Big[g_1(x_1, u) + \alpha \max\big[V_2^t\big(f_1(x_1, u)\big), J_2^t\big(f_1(x_1, u)\big)\big]\Big], \tag{5.57}$$

set $\mu^{t+1}(x_1)$ to a control $u \in U(x_1)$ that attains the above minimum, and leave J_2^t, V_2^t, ν^t unchanged.

(c) **Single value iteration for policy evaluation of the maximizer**: For all $x_2 \in X_2$ and $v \in V(x_2)$, set

$$J_2^{t+1}(x_2) = g_2\big(x_2, \nu^t(x_2)\big) + \alpha \min\Big[V_1^t\big(f_2(x_2, \nu^t(x_2))\big), J_1^t\big(f_2(x_2, \nu^t(x_2))\big)\Big], \tag{5.58}$$

and leave $J_1^t, V_1^t, V_2^t, \mu^t, \nu^t$ unchanged.

(d) **Policy improvement for the maximizer**: For all $x_2 \in X_2$, set

$$J_2^{t+1}(x_2) = V_2^{t+1}(x_2) = \max_{v \in V(x_2)} \Big[g_2(x_2, v) + \alpha \min\big[V_1^t\big(f_2(x_2, v)\big), J_1^t\big(f_2(x_2, v)\big)\big]\Big], \tag{5.59}$$

set $\nu^{t+1}(x_2)$ to a control $u \in V(x_2)$ that attains the above maximum, and leave J_1^t, V_1^t, μ^t unchanged.

Example 5.3.6 (Our PI Algorithm for Markov Games)

Let us consider the Markov game formulation of Example 5.3.2. Our PI algorithm with x_1, x_2, H_1, and H_2 defined earlier, can be implemented by storing J_1^t, V_1^t as the real numbers $J_1^t(x)$ and $V_1^t(x)$, $x \in X$, and by storing and representing the piecewise linear functions J_2^t, V_2^t using the m columns of the $n \times m$ matrices

$$A(x) + \alpha\beta \sum_{y \in X} Q_{xy} \min\big[V_1^t(y), J_1^t(y)\big], \qquad x \in X; \tag{5.60}$$

cf. Eq. (5.36). None of the operations (5.52)-(5.55) require the solution of a Markovian decision problem as in the Hoffman-Karp algorithm. This is similar to the Pollatschek and Avi-Itzhak algorithm.

More specifically, the policy evaluation (5.52) for the minimizer takes the form

$$J_1^{t+1}(x) = \frac{1}{\beta} \max \Big[V_2^{t+1}\big(x, \mu^t(x)\big), J_2^{t+1}\big(x, \mu^t(x)\big) \Big], \qquad \text{for all } x \in X, \quad (5.61)$$

while the policy improvement (5.53) for the minimizer takes the form

$$J_1^{t+1}(x) = V_1^{t+1}(x) = \frac{1}{\beta} \min_{u \in U} \max \big[V_2^{t+1}(x,u), J_2^{t+1}(x,u) \big], \qquad \text{for all } x \in X. \quad (5.62)$$

The policy evaluation (5.54) for the maximizer takes the form

$$J_2^{t+1}(x,u) = u' \left(A(x) + \alpha\beta \sum_{y \in X} Q_{xy} \min \big[V_1^t(y), J_1^t(y) \big] \right) \big(\nu^t(x) \big), \quad (5.63)$$

for all $x \in X$ and $u \in U$, while the policy improvement (5.55) for the maximizer takes the form

$$\begin{aligned} J_2^{t+1}(x,u) &= V_2^{t+1}(x,u) \\ &= \max \left\{ u' \left(A(x) + \alpha\beta \sum_{y \in X} Q_{xy} \min \big[V_1^t(y), J_1^t(y) \big] \right) (1), \ldots, \right. \\ &\qquad \left. u' \left(A(x) + \alpha\beta \sum_{y \in X} Q_{xy} \min \big[V_1^t(y), J_1^t(y) \big] \right) (m) \right\}, \end{aligned} \quad (5.64)$$

for all $x \in X$ and $u \in U$, where

$$\left(A(x) + \alpha\beta \sum_{y \in X} Q_{xy} \min \big[V_1^t(y), J_1^t(y) \big] \right) (j)$$

is the jth column of the $n \times m$ matrix (5.60).

Again it can be seen that except for the extra memory storage to maintain V_1^t and V_2^t, the preceding PI algorithm (5.61)-(5.64) requires roughly similar/comparable computations to the ones of the "naive" optimistic PI algorithm (5.48)-(5.51), when applied to the Markov game model.

Discussion of our Algorithm

Let us now provide a discussion of some of the properties of our PI algorithm (5.52)-(5.55). We first note that except for the extra memory storage to maintain V_1^t and V_2^t, the algorithm requires roughly similar/comparable computations to the ones of the "naive" optimistic PI algorithm (5.48)-(5.51). Note also that by performing a large number of value iterations of the form (5.52) or (5.54) we obtain an algorithm that involves nearly

exact policy evaluation, similar to the "naive" nonoptimistic PI algorithm (5.44)-(5.47).

Mathematically, under the contraction assumption to be introduced in the next section, our algorithm (5.52)-(5.55) avoids the oscillatory behavior illustrated in Fig. 5.2.1because it embodies a policy-dependent sup-norm contraction, which has a *uniform fixed point*, the pair (J_1^*, J_2^*), regardless of the policies. This is the essence of the key Prop. 5.4.2, which will be shown in the next section.

Aside from this mathematical insight, one may gain intuition into the mechanism of our algorithm (5.52)-(5.55), by comparing it with the optimistic version of the "naive" optimistic PI algorithm (5.48)-(5.51). Our algorithm (5.52)-(5.55) involves additionally the functions V_1^t and V_2^t, which are changed only during the policy improvement operations, and tend to provide a guarantee against oscillatory behavior. In particular, since

$$\max[V_2^t, J_2^t] \geq J_2^t,$$

the iterations of the minimizer in our algorithm, (5.52) and (5.53), are more "pessimistic" about the choices of the maximizer than the iterates of the minimizer in the "naive" PI iterates (5.48) and (5.49). Similarly, since

$$\min[V_1^t, J_1^t] \leq J_1^t,$$

the iterations of the maximizer in our algorithm, (5.54) and (5.55), are more "pessimistic" than the iterates of the maximizer in the naive PI iterates (5.48) and (5.49). As a result *the use of V_1^t and V_2^t in our PI algorithm makes it more conservative*, and mitigates the oscillatory swings that are illustrated in Fig. 5.2.1.

Let us also note that the use of the functions V_1 and V_2 in our algorithm (5.52)-(5.55) may slow down the algorithmic progress relative to the (nonconvergent) "naive" algorithm (5.44)-(5.47). To remedy this situation an interpolation device has been suggested in the paper [BeY10] (Section V), which roughly speaking interpolates between the two algorithms, while still guaranteeing the algorithm's convergence; see also Section 2.6.3. Basically, such a device makes the algorithm less "pessimistic," as it guards against nonconvergence, and it can similarly be used in our algorithm (5.52)-(5.55).

In the next section, we will show convergence of our PI algorithm (5.52)-(5.55) with a line of proof that can be summarized as follows. Using a contraction argument, based on an assumption to be introduced shortly, we show that the sequences $\{V_1^t\}$ and $\{V_2^t\}$ converge to some functions $V_1^* \in B(X_1)$ and $V_2^* \in B(X_2)$, respectively. From the policy improvement operations (5.53) and (5.55) it will then follow that the sequences $\{J_1^t\}$ and $\{J_2^t\}$ converge to the same functions V_1^* and V_2^*, respectively, so that $\min[V_1^t, J_1^t]$ and $\max[V_2^t, J_2^t]$ converge to V_1^* and V_2^*, respectively, as well.

Using the continuity of H_1 and H_2 (a consequence of our contraction assumption), it follows from Eqs. (5.53) and (5.55) that (V_1^*, V_2^*) is the fixed point of H_1 and H_2 [in the sense of Eq. (5.28)], and hence is also equal to (J_1^*, J_2^*) [cf. Eq. (5.28)]. Thus we finally obtain convergence:

$$V_1^t \to J_1^*, \qquad J_1^t \to J_1^*, \qquad V_2^t \to J_2^*, \qquad J_2^t \to J_2^*.$$

5.4 CONVERGENCE ANALYSIS

For each $\mu \in \mathcal{M}$, we consider the operator $T_{1,\mu}$ that maps a function $J_2 \in B(X_2)$ into the function of x_1 given by

$$(T_{1,\mu}J_2)(x_1) = H_1\big(x_1, \mu(x_1), J_2\big), \qquad x_1 \in X_1. \tag{5.65}$$

Also for each $\nu \in \mathcal{N}$, we consider the operator $T_{2,\nu}$ that maps a function $J_1 \in B(X_1)$ into the function of x_2 given by

$$(T_{2,\nu}J_1)(x_2) = H_2\big(x_2, \nu(x_2), J_1\big), \qquad x_2 \in X_2. \tag{5.66}$$

We will also consider the operator $T_{\mu,\nu}$ that maps a function $(J_1, J_2) \in B(X_1) \times B(X_2)$ into the function of $(x_1, x_2) \in X_1 \times X_2$, given by

$$\big(T_{\mu,\nu}(J_1, J_2)\big)(x_1, x_2) = \big((T_{1,\mu}J_2)(x_1),\, (T_{1,\nu}J_1)(x_2)\big). \tag{5.67}$$

[Recall here that the norms on $B(X_1)$, $B(X_2)$, and $B(X_1) \times B(X_2)$ are given by Eqs. (5.26) and (5.27).]

We will show convergence of our algorithm assuming the following.

Assumption 5.4.1: (Contraction Assumption) Consider the operator $T_{\mu,\nu}$ given by Eq. (5.67).

(a) For all $(\mu, \nu) \in \mathcal{M} \times \mathcal{N}$, and $(J_1, J_2) \in B(X_1) \times B(X_2)$, the function $T_{\mu,\nu}(J_1, J_2)$ belongs to $B(X_1) \times B(X_2)$.

(b) There exists an $\alpha \in (0, 1)$ such that for all $(\mu, \nu) \in \mathcal{M} \times \mathcal{N}$, $T_{\mu,\nu}$ is a contraction mapping of modulus α within $B(X_1) \times B(X_2)$.

By writing the contraction property as

$$\begin{aligned} \max\big\{\|T_{1,\mu}J_2 - T_{1,\mu}J_2'\|_1, \|T_{2,\nu}J_1 - T_{2,\nu}J_1'\|_2\big\} \\ \leq \alpha \max\big\{\|J_1 - J_1'\|_1,\, \|J_2 - J_2'\|_2\big\}, \end{aligned} \tag{5.68}$$

for all $J_1, J_1' \in B(X_1)$ and $J_2, J_2' \in B(X_2)$ [cf. the norm definition (5.27)], we have

$$\|T_{1,\mu}J_2 - T_{1,\mu}J_2'\|_1 \leq \alpha\|J_2 - J_2'\|_2, \qquad \text{for all } J_2, J_2' \in B(X_2), \tag{5.69}$$

and

$$\|T_{2,\nu}J_1 - T_{2,\nu}J_1'\|_2 \leq \alpha\|J_1 - J_1'\|_1, \qquad \text{for all } J_1, J_1' \in B(X_1); \tag{5.70}$$

[set $J_1 = J_1'$ or $J_2 = J_2'$, respectively, in Eq. (5.68)]. From these relations, we obtain†

$$\|T_1J_2 - T_1J_2'\|_1 \leq \alpha\|J_2 - J_2'\|_2, \qquad \text{for all } J_2, J_2' \in B(X_2), \tag{5.71}$$

and

$$\|T_2J_1 - T_2J_1'\|_2 \leq \alpha\|J_1 - J_1'\|_1, \qquad \text{for all } J_1, J_1' \in B(X_1), \tag{5.72}$$

where

$$(T_1J_2)(x_1) = \inf_{\mu\in\mathcal{M}}(T_{1,\mu}J_2)(x_1) = \inf_{u\in U(x_1)} H_1(x_1, u, J_2), \qquad x_1 \in X_1,$$

$$(T_2J_1)(x_2) = \sup_{\nu\in\mathcal{N}}(T_{2,\nu}J_1)(x_2) = \sup_{v\in V(x_2)} H_2(x_2, v, J_1), \qquad x_2 \in X_2.$$

The relations (5.71)-(5.72) also imply that *the operator*

$$T : B(X_1) \times B(X_2) \mapsto B(X_1) \times B(X_2),$$

defined by

$$T(J_1, J_2) = (T_1J_2, T_2J_1), \tag{5.73}$$

is a contraction mapping from $B(X_1) \times B(X_2)$ *to* $B(X_1) \times B(X_2)$ *with modulus* α. It follows that T *has a unique fixed point* $(J_1^*, J_2^*) \in B(X_1) \times B(X_2)$. We will show that our algorithm yields in the limit this fixed point.

---

† For a proof, we write Eq. (5.69) as

$$(T_{1,\mu}J_2)(x_1) \leq (T_{1,\mu}J_2')(x_1) + \alpha\|J_2 - J_2'\|_2\,\xi_1(x_1),$$

$$(T_{2,\mu}J_2)(x_1) \leq (T_{2,\mu}J_2')(x_1) + \alpha\|J_2 - J_2'\|_2\,\xi_1(x_1),$$

for all $x_1 \in X_1$. By taking infimum of both sides over $\mu \in \mathcal{M}$, we obtain

$$\frac{\big|(T_1J_2)(x_1) - T_1J_2')(x_1)\big|}{\xi_1(x_1)} \leq \alpha\|J_2 - J_2'\|_2,$$

and by taking supremum over $x_1 \in X_1$, the desired relation

$$\|T_1J_2 - T_1J_2'\|_1 \leq \alpha\|J_2 - J_2'\|_2$$

follows. The proof of the other relation, $\|T_2J_1 - T_2J_1'\|_2 \leq \alpha\|J_1 - J_1'\|_1$, is similar.

The following is our main convergence result [convergence here is meant in the sense of the norm (5.27) on $B(X_1) \times B(X_2)$]. Note that this result applies to any order and frequency of policy evaluations and policy improvements of the two players.

Proposition 5.4.1: (Convergence) Let Assumption 5.4.1 hold, and assume that each of the four operations of the PI algorithm (5.52)-(5.55) is performed infinitely often. Then the sequences $\{(J_1^t, J_2^t)\}$ and $\{(V_1^t, V_2^t)\}$ generated by the algorithm converge to (J_1^*, J_2^*).

The proof is long but follows closely the steps of the proof for the single-player abstract DP case in Section 2.6.3.

An Extended Algorithm and its Convergence Proof

We first show the following lemma.

Lemma 5.4.1: For all $(V_1, V_2), (J_1, J_2), (V_1', V_2'), (J_1', J_2') \in B(X_1) \times B(X_2)$, we have

$$\big\| \min[V_1, J_1] - \min[V_1', J_1'] \big\|_1 \leq \max\big\{ \|V_1 - V_1'\|_1,\ \|J_1 - J_1'\|_1 \big\}, \quad (5.74)$$

$$\big\| \max[V_2, J_2] - \min[V_2', J_2'] \big\|_2 \leq \max\big\{ \|V_2 - V_2'\|_2,\ \|J_2 - J_2'\|_2 \big\}. \quad (5.75)$$

Proof: For every $x_1 \in X_1$, we write

$$\frac{V_1(x_1)}{\xi_1(x_1)} \leq \frac{V_1'(x_1)}{\xi_1(x_1)} + \max\big\{ \|V_1 - V_1'\|_1,\ \|J_1 - J_1'\|_1 \big\},$$

$$\frac{J_1(x_1)}{\xi_1(x_1)} \leq \frac{J_1'(x_1)}{\xi_1(x_1)} + \max\big\{ \|V_1 - V_1'\|_1,\ \|J_1 - J_1'\|_1 \big\},$$

from which we obtain

$$\frac{\min\big\{V_1(x_1), J_1(x_1)\big\}}{\xi_1(x_1)} \leq \frac{\min\big\{V_1'(x_1), J_1'(x_1)\big\}}{\xi_1(x_1)} + \max\big\{ \|V_1 - V_1'\|_1,\ \|J_1 - J_1'\|_1 \big\},$$

so that

$$\frac{\min\big\{V_1(x_1), J_1(x_1)\big\} - \min\big\{V_1'(x_1), J_1'(x_1)\big\}}{\xi_1(x_1)} \leq \max\big\{ \|V_1 - V_1'\|_1,\ \|J_1 - J_1'\|_1 \big\}.$$

By exchanging the roles of (V_1, J_1) and (V_1', J_1'), and combining the two inequalities, we have

$$\frac{\Big|\min\big\{V_1(x_1), J_1(x_1)\big\} - \min\big\{V_1'(x_1), J_1'(x_1)\big\}\Big|}{\xi_1(x_1)} \leq \max\big\{\|V_1 - V_1'\|_1,\, \|J_1 - J_1'\|_1\big\},$$

and by taking the supremum over $x_1 \in X_1$, we obtain Eq. (5.74). We similarly prove Eq. (5.75). **Q.E.D.**

We consider the spaces of bounded functions $Q_1(x_1, u)$ of $(x_1, u) \in X_1 \times U$ and $Q_1(x_2, v)$ of $(x_2, v) \in X_2 \times V$, with norms

$$\|Q_1\|_1 = \sup_{x_1 \in X_1,\, u \in U} \frac{\big|Q_1(x_1, u)\big|}{\xi_1(x_1)}, \qquad \|Q_2\|_2 = \sup_{x_2 \in X_2,\, v \in V} \frac{\big|Q_2(x_2, v)\big|}{\xi_2(x_2)}, \tag{5.76}$$

respectively, where ξ_1 and ξ_2 are the weighting functions that define the norm of $B(X_1)$ and $B(X_2)$ [cf. Eq. (5.26)]. We denote these spaces by $B(X_1 \times U)$ and $B(X_2 \times V)$, respectively. Functions in these spaces have the meaning of *Q-factors for the minimizer and the maximizer*.

We next introduce a new operator, denoted by $G_{\mu,\nu}$, which is parametrized by the policy pair (μ, ν), and will be shown to have a common fixed point for all $(\mu, \nu) \in \mathcal{M} \times \mathcal{N}$, from which (J_1^*, J_2^*) can be readily obtained. The operator $G_{\mu,\nu}$ involves operations on Q-factor pairs (Q_1, Q_2) for the minimizer and the maximizer, in addition to functions of state (V_1, V_2), and is used define an "extended" PI algorithm that operates over a larger function space than the one of Section 5.3. Once the convergence of this "extended" PI algorithm is shown, the convergence of our algorithm of Section 5.3 will readily follow.

To define the operator $G_{\mu,\nu}$, we note that it consists of four components, maps $B(X_1) \times B(X_2) \times B(X_1 \times U) \times B(X_2 \times V)$ into itself. It is given by

$$\begin{aligned} G_{\mu,\nu}(V_1, V_2, Q_1, Q_2) = \big(&M_{1,\nu}(V_2, Q_2), M_{2,\mu}(V_1, Q_1), \\ &F_{1,\nu}(V_2, Q_2), F_{2,\mu}(V_1, Q_1)\big), \end{aligned} \tag{5.77}$$

with the functions $M_{1,\nu}(V_2, Q_2)$, $M_{2,\mu}(V_1, Q_1)$, $F_{1,\nu}(V_2, Q_2)$, $F_{2,\mu}(V_1, Q_1)$, defined as follows:

- $M_{1,\nu}(V_2, Q_2)$: This is the function of x_1 given by

$$\begin{aligned} \big(M_{1,\nu}(V_2, Q_2)\big)(x_1) &= \big(T_1 \max[V_2, \hat{Q}_{2,\nu}]\big)(x_1) \\ &= \inf_{u \in U(x_1)} H_1\big(x_1, u, \max[V_2, \hat{Q}_{2,\nu}]\big), \end{aligned} \tag{5.78}$$

  where $\hat{Q}_{2,\nu}$ is the function of x_2 given by

$$\hat{Q}_{2,\nu}(x_2) = Q_2\big(x_2, \nu(x_2)\big). \tag{5.79}$$

- $M_{2,\mu}(V_1, Q_1)$: This is the function of x_2 given by

$$\begin{aligned}\big(M_{2,\mu}(V_1, Q_1)\big)(x_2) &= \big(T_2 \min[V_1, \hat{Q}_{1,\mu}]\big)(x_2) \\ &= \sup_{v \in V(x_2)} H_2\big(x_2, v, \min[V_1, \hat{Q}_{1,\mu}]\big),\end{aligned} \tag{5.80}$$

  where $\hat{Q}_{1,\mu}$ is the function of x_1 given by

$$\hat{Q}_{1,\mu}(x_1) = Q_1\big(x_1, \mu(x_1)\big). \tag{5.81}$$

- $F_{1,\nu}(V_2, Q_2)$: This is the function of (x_1, u), given by

$$F_{1,\nu}(V_2, Q_2)(x_1, u) = H_1\big(x_1, u, \max[V_2, \hat{Q}_{2,\nu}]\big). \tag{5.82}$$

- $F_{2,\mu}(V_1, Q_1)$: This is the function of (x_2, v), given by

$$F_{2,\mu}(V_1, Q_1)(x_2, v) = H_2\big(x_2, v, \min[V_1, \hat{Q}_{1,\mu}]\big). \tag{5.83}$$

Note that the four components of $G_{\mu,\nu}$ correspond to the four operations of our algorithm (5.52)-(5.55). In particular,

- $M_{1,\nu}(V_2, Q_2)$ corresponds to policy improvement of the minimizer.
- $M_{2,\mu}(V_1, Q_1)$ corresponds to policy improvement of the maximizer.
- $F_{1,\nu}(V_2, Q_2)$ corresponds to policy evaluation of the minimizer.
- $F_{2,\mu}(V_1, Q_1)$ corresponds to policy evaluation of the maximizer.

The key step in our convergence proof is to show that $G_{\mu,\nu}$ has a contraction property with respect to the norm on $B(X_1) \times B(X_2) \times B(X_1 \times U) \times B(X_2 \times V)$ given by

$$\big\|(V_1, V_2, Q_1, Q_2)\big\| = \max\big\{\|V_1\|_1, \|V_2\|_2, \|Q_1\|_1, \|Q_2\|_2\big\}, \tag{5.84}$$

where $\|V_1\|_1$, $\|V_2\|_2$ are the weighted sup-norms of V_1, V_2, respectively, defined by Eq. (5.26), and $\|Q_1\|_1$, $\|Q_2\|_2$ are the weighted sup-norms of Q_1, Q_2, defined by Eq. (5.76). Moreover, the contraction property is uniform, in the sense that *the fixed point of $G_{\mu,\nu}$ does not depend on (μ, ν)*. This means that *we can carry out iterations with $G_{\mu,\nu}$, while changing μ and ν arbitrarily between iterations, and still aim at the same fixed point*. We have the following proposition.

Proposition 5.4.2: (Uniform Contraction) Let Assumption 5.4.1 hold. Then for all $(\mu,\nu) \in \mathcal{M} \times \mathcal{N}$, the operator $G_{\mu,\nu}$ is a contraction mapping with modulus α with respect to the norm of Eqs. (5.84), (5.26), and (5.76). Moreover, the corresponding fixed point of $G_{\mu,\nu}$ is $(J_1^*, J_2^*, Q_1^*, Q_2^*)$ [independently of the choice of (μ,ν)], where (J_1^*, J_2^*) is the fixed point of the mapping T of Eq. (5.73), and Q_1^*, Q_2^* are the functions defined by

$$Q_1^*(x_1,u) = H_1(x_1,u,J_2^*), \qquad x_1 \in X_1,\ u \in U(x_1), \tag{5.85}$$

$$Q_2^*(x_2,v) = H_2(x_2,v,J_1^*), \qquad x_2 \in X_2,\ v \in V(x_2). \tag{5.86}$$

Proof: We prove the contraction property of $G_{\mu,\nu}$ by breaking it down to four inequalities, which hold for all $(V_1,V_2),(V_1',V_2') \in B(X_1)\times B(X_2)$ and $(Q_1,Q_2),(Q_1',Q_2') \in B(X_1,U)\times B(X_2,V)$. In particular, we have

$$\begin{aligned}
\big\|M_{1,\nu}(V_2,Q_2) - M_{1,\nu}(V_2',Q_2')\big\|_1 &= \Big\|T_1\big(\max[V_2,\hat{Q}_{2,\nu}]\big) - T_1\big(\max[V_2',\hat{Q}_{2,\nu}']\big)\Big\|_1 \\
&\le \alpha\big\|\max[V_2,\hat{Q}_{2,\nu}]_2 - \max[V_2',\hat{Q}_{2,\nu}']\big\|_2 \\
&\le \alpha\max\big\{\|V_2-V_2'\|_2,\ \|\hat{Q}_{2,\nu}-\hat{Q}_{2,\nu}'\|_2\big\} \\
&\le \alpha\max\big\{\|V_2-V_2'\|_2,\ \|Q_2-Q_2'\|_2\big\} \\
&\le \alpha\max\big\{\|V_1-V_1'\|_1,\ \|Q_1-Q_1'\|_1, \\
&\qquad\qquad \|V_2-V_2'\|_2,\ \|Q_2-Q_2'\|_2\big\} \\
&= \alpha\big\|(V_1,V_2,Q_1,Q_2)-(V_1',V_2',Q_1',Q_2')\big\|,
\end{aligned} \tag{5.87}$$

where the first equality uses the definitions of $M_{1,\nu}(V_2,Q_2)$, $M_{1,\nu}(V_2',Q_2')$ [cf. Eqs. (5.78) and (5.80)], the first inequality follows from Eq. (5.69), the second inequality follows using Lemma 5.4.1, the third inequality follows from the definition of $\hat{Q}_{2,\nu}$ and $\hat{Q}_{2,\nu}'$, the last inequality is trivial, and the last equality follows from the norm definition (5.84). Similarly, we prove that

$$\big\|M_{2,\mu}(V_1,Q_1) - M_{2,\mu}(V_1',Q_1')\big\|_2 \le \alpha\big\|(V_1,V_2,Q_1,Q_2)-(V_1',V_2',Q_1',Q_2')\big\|, \tag{5.88}$$

$$\big\|F_{1,\nu}(V_2,Q_2) - F_{1,\nu}(V_2',Q_2')\big\|_1 \le \alpha\big\|(V_1,V_2,Q_1,Q_2)-(V_1',V_2',Q_1',Q_2')\big\|, \tag{5.89}$$

$$\big\|F_{2,\mu}(V_1,Q_1) - F_{2,\mu}(V_1',Q_1')\big\|_2 \le \alpha\big\|(V_1,V_2,Q_1,Q_2)-(V_1',V_2',Q_1',Q_2')\big\|. \tag{5.90}$$

From the preceding relations (5.87)-(5.90), it follows that each of the four components of the maximization that comprises the norm

$$\big\|G_{\mu,\nu}(V_1,V_2,Q_1,Q_2) - G_{\mu,\nu}(V_1',V_2',Q_1',Q_2')\big\|$$

[cf. Eq. (5.77)] is less or equal to

$$\alpha \big\|(V_1, V_2, Q_1, Q_2) - (V_1', V_2', Q_1', Q_2')\big\|.$$

Thus we have

$$\begin{aligned}\big\|G_{\mu,\nu}(V_1, V_2, Q_1, Q_2) - G_{\mu,\nu}(V_1', V_2', Q_1', Q_2')\big\| \\ \leq \alpha \big\|(V_1, V_2, Q_1, Q_2) - (V_1', V_2', Q_1', Q_2')\big\|,\end{aligned}$$

which shows the desired contraction property of $G_{\mu,\nu}$.

In view of the contraction property just shown, the mapping $G_{\mu,\nu}$ has a unique fixed point for each $(\mu,\nu) \in \mathcal{M} \times \mathcal{N}$, which we denote by (V_1, V_2, Q_1, Q_2) [with some notational abuse, we do not show the possible dependence of the fixed point on (μ,ν)]. In view of Eqs. (5.77)-(5.83), this fixed point satisfies for all $x_1 \in X_1$, $x_2 \in X_2$, $(x_1, u) \in X_1 \times U$, $(x_2, v) \in X_2 \times V$,

$$V_1(x_1) = \inf_{u' \in U(x_1)} H_1\big(x_1, u', \max[V_2, \hat{Q}_{2,\nu}]\big), \tag{5.91}$$

$$V_2(x_2) = \sup_{v' \in V(x_2)} H_2\big(x_2, v', \min[V_1, \hat{Q}_{1,\mu}]\big), \tag{5.92}$$

$$Q_1(x_1, u) = H_1\big(x_1, u, \max[V_2, \hat{Q}_{2,\nu}]\big), \quad Q_2(x_2, v) = H_2\big(x_2, v, \min[V_1, \hat{Q}_{1,\mu}]\big). \tag{5.93}$$

By comparing the preceding two relations, it follows that for all $x_1 \in X_1$, $x_2 \in X_2$,

$$V_1(x_1) \leq Q_1(x_1, u), \quad \text{for all } x_1,\ u \in U(x_1),$$

$$V_2(x_2) \geq Q_2(x_2, v), \quad \text{for all } x_2,\ v \in V(x_2),$$

which implies that

$$\min[V_1, \hat{Q}_{1,\mu}] = V_1, \qquad \max[V_2, \hat{Q}_{2,\nu}] = V_2.$$

Using Eqs. (5.91)-(5.92), this in turn shows that

$$V_1(x_1) = \inf_{u \in U(x_1)} H_1(x_1, u, V_2), \qquad V_2(x_2) = \sup_{v \in V(x_2)} H_2(x_2, v, V_1).$$

Thus, independently of (μ,ν), (V_1, V_2) is the unique fixed point of the contraction mapping T of Eq. (5.73), which is (J_1^*, J_2^*). Moreover from Eq. (5.93), we have that (Q_1, Q_2) is precisely (Q_1^*, Q_2^*) as given by Eqs. (5.85) and (5.86). This shows that, independently of (μ,ν), the fixed point of $G_{\mu,\nu}$ is $(J_1^*, J_2^*, Q_1^*, Q_2^*)$, and proves the desired result. **Q.E.D.**

The preceding proposition implies the convergence of the "extended" algorithm, which at each iteration t applies one of the four components of

G_{μ^t,ν^t} evaluated at the current iterate $(V_1^t, V_2^t, Q_1^t, Q_2^t, \mu^t, \nu^t)$, and updates this iterate accordingly. This algorithm is well-suited for the calculation of both (J_1^*, J_2^*) and (Q_1^*, Q_2^*). However, since we are just interested to calculate (J_1^*, J_2^*), a simpler and more efficient algorithm is possible, which is in fact our PI algorithm based on the four operations (5.52)-(5.55). To this end, we observe that the algorithm that updates $(V_1^t, V_2^t, Q_1^t, Q_2^t, \mu^t, \nu^t)$ can be operated so that it does not require the maintenance of the full Q-factor functions (Q_1^t, Q_2^t). The reason is that the values $Q_1^t(x_1, u)$ and $Q_2^t(x_2, v)$ with $u \neq \mu^t(x_1)$ and $v \neq \nu^t(x_2)$, do not appear in the calculations, and hence we need only the values $\hat{Q}^t_{1,\mu^t}(x_1)$ and $\hat{Q}^t_{2,\nu^t}(x_2)$, which we store in functions J_1^t and J_2^t, i.e., we set

$$J_1^t(x_1) = \hat{Q}^t_{1,\mu^t}(x_1) = Q_1^t\big(x_1, \mu^t(x)\big),$$

$$J_2^t(x_2) = \hat{Q}^t_{2,\nu^t}(x_2) = Q_2^t\big(x_2, \nu^t(x_2)\big).$$

Once we do that, the resulting algorithm is precisely our PI algorithm (5.52)-(5.55).

In summary, our PI algorithm that updates $(V_1^t, V_2^t, J_1^t, J_2^t, \mu^t, \nu^t)$ is a reduced space implementation of the asynchronous fixed point algorithm that updates $(V_1^t, V_2^t, Q_1^t, Q_2^t, \mu^t, \nu^t)$ using the uniform contraction mapping G_{μ^t,ν^t}, with the identifications

$$J_1^t = \hat{Q}_{1,\mu^t}, \qquad J_2^t = \hat{Q}_{2,\nu^t}.$$

This proves its convergence as stated in Prop. 5.4.1.

5.5 APPROXIMATION BY AGGREGATION

Our algorithm of Section 5.3 involves exact implementation without function approximations, and thus is not suitable for large state and control spaces. An important research direction is approximate implementations based on our PI algorithmic structure of Section 5.3, whereby we use approximation in value space with cost function approximations obtained through reinforcement learning methods. An interesting algorithmic approach is aggregation with representative states, as described in the book [Ber19b] (Section 6.1).

In particular, let us consider the minimax formulation of Example 5.3.1 and Eqs. (5.29), (5.32), (5.33), which involves separate state spaces X_1 and X_2 for the minimizer and the maximizer, respectively. In the aggregation with representative states formalism, we execute our PI algorithm over reduced versions of the spaces X_1 and X_2. In particular, we discretize X_1 and X_2 by using suitable finite collections of representative

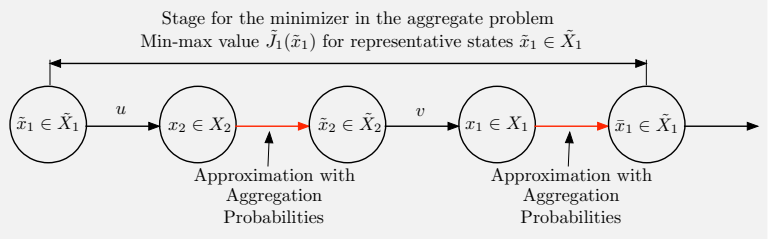


Figure 5.5.1 Schematic illustration of an aggregation framework that is patterned after the sequence of events of the multistage process of Fig. 5.3.1. The aggregate problem is specified by a finite subset of representative states $\tilde{X}_1 \subset X_1$, a finite subset of representative states $\tilde{X}_2 \subset X_2$, and aggregation probabilities for passing from states $x_2 \in X_2$ to representative states $\tilde{x}_2 \in \tilde{X}_2$, and for passing from states $x_1 \in X_1$ to representative states $\tilde{x}_1 \in \tilde{X}_1$. A stage starts at a representative state $\tilde{x}_1 \in \tilde{X}_1$ and ends at some other representative state $\bar{x}_1 \in \tilde{X}_1$, by going successively through a state $x_2 \in X_2$ under the influence of the minimizer's choice $u \in U(\tilde{x}_1)$, then to a representative state $\tilde{x}_2 \in \tilde{X}_2$ using aggregation probabilities $\phi_{x_2\tilde{x}_2}$ (i.e., the transition $x_2 \to \tilde{x}_2$ takes place with probability $\phi_{x_2\tilde{x}_2}$), then to a state $x_1 \in X_1$ under the influence of the maximizer's choice $v \in V(\tilde{x}_2)$, and finally to $\bar{x}_1 \in \tilde{X}_1$ using aggregation probabilities $\phi_{x_1\bar{x}_1}$ (the transition $x_1 \to \bar{x}_1$ takes place with probability $\phi_{x_1\bar{x}_1}$). The transitions $\tilde{x}_1 \to x_2$ and $\tilde{x}_2 \to x_1$ produce costs $g_1(\tilde{x}_1, u)$ and $g_2(\tilde{x}_2, v)$, respectively [cf. Eqs. (5.30), (5.31)]. The aggregation probabilities $\phi_{x_2\tilde{x}_2}$ and $\phi_{x_1\bar{x}_1}$ can be arbitrary. However, their choice affects the min-max and max-min functions of the aggregate problem.

We can solve the aggregate problem by using simulation-based versions of our PI algorithm (5.52)-(5.55) of Section 5.3 to obtain the min-max and max-min functions $\tilde{J}_1(\tilde{x}_1)$ and $\tilde{J}_2(\tilde{x}_2)$ at all the representative states $\tilde{x}_1 \in \tilde{X}_1$ and $\tilde{x}_2 \in \tilde{X}_2$, respectively [cf. [Ber19b] (Chapter 6)]. Then, min-max and max-min function approximations are computed from

$$\tilde{J}_1(x_1) = \sum_{\tilde{x}_1 \in \tilde{X}_1} \phi_{x_1\tilde{x}_1} \tilde{J}_1(\tilde{x}_1), \qquad \tilde{J}_2(x_2) = \sum_{\tilde{x}_2 \in \tilde{X}_2} \phi_{x_2\tilde{x}_2} \tilde{J}_2(\tilde{x}_2).$$

Suboptimal decision choices by the minimizer and the maximizer are then obtained from the one-step lookahead optimizations

$$\min_{u \in U(x_1)} H_1(x_1, u, \tilde{J}_2), \qquad \max_{v \in V(x_2)} H_2(x_2, v, \tilde{J}_1).$$

See the book [Ber19b] (Section 6.1) and the paper [Ber18a] for a detailed accounting of the aggregation approach with representative states for single-player infinite horizon DP.

states $\tilde{X}_1 \subset X_1$ and $\tilde{X}_2 \subset X_2$, and construct a lower-dimensional aggregate problem. The typical stage involves transitions between representative states, with intermediate artificial transitions $x_1 \to \tilde{x}_1$ and $x_2 \to \tilde{x}_2$, which involve randomization with aggregation probabilities $\phi_{x_1\tilde{x}_1}$ and $\phi_{x_2\tilde{x}_2}$, respectively; see Fig. 5.5.1.

The structure of the aggregate problem is amenable to a DP formulation, and as a result, it can be solved by using simulation-based versions

of the PI methods of Section 5.3 [we refer to the book [Ber19b] (Chapter 6) for more details]. The cost function approximations thus obtained, call them $\tilde{J}_1$, $\tilde{J}_2$, are used in the one-step lookahead minimization

$$\min_{u \in U(x_1)} H_1(x_1, u, \tilde{J}_2),$$

to obtain a suboptimal minimizer's policy, and in the one-step lookahead maximization

$$\max_{v \in V(x_2)} H_2(x_2, v, \tilde{J}_1),$$

to obtain a suboptimal maximizer's policy.

The aggregation with representative states approach has the advantage that it maintains the DP structure of the original minimax problem. This allows the use of our PI methods of Section 5.3, with convergence guaranteed by the results of Section 5.4. Another aggregation approach that can be similarly used within our context, is hard aggregation, whereby the state spaces X_1 and X_2 are partitioned into subsets that form aggregate states; see [Ber18a], [Ber18b], [Ber19b]. Other reinforcement learning methods, based for example on the use of neural networks, can also be used for approximate implementation of our PI algorithms. However, their convergence properties are problematic, in the absence of additional assumptions. The papers by Bertsekas and Yu ([BeY12], Sections 6 and 7), and by Yu and Bertsekas [YuB13a] (Section 4), also describe alternative simulation-based approximation possibilities that may serve as a starting point for minimax PI algorithms with function approximation.

5.6 NOTES AND SOURCES

In this chapter, we have discussed PI algorithms that are specifically tailored to sequential zero-sum games and minimax problems with a contractive abstract DP structure. We used as starting point the methods by Hoffman and Karp [HoK66], and by Pollatschek and Avi-Itzhak [PoA69] for discounted and terminating zero-sum Markov games. Related methods have been discussed for Markov games by van der Wal [Van78], Tolwinski [Tol89], Filar and Tolwinski [FiT91], Filar and Vrieze [FiV96], and for stochastic shortest games, by Patek and Bertsekas [PaB99], and Yu [Yu14]; see also Perolat et al. [PPG16], [PSP15], and the survey by Zhang, Yang, and Basar [ZYB21] for related reinforcement learning methods. Our algorithms of Section 5.3 resolve the long-standing convergence difficulties of the Pollatschek and Avi-Itzhak PI algorithm [PoA69], and allow an asynchronous implementation, whereby the policy evaluation and policy improvement operations can be done in any order and with different frequencies. Moreover, our algorithms find simultaneously the min-max and

the max-min values, and they are suitable for Markov zero-sum game problems, as well as for minimax control problems involving set-membership uncertainty.

While we have not addressed in detail the issue of asynchronous distributed implementation in a multiprocessor system, our algorithm admits such an implementation, as has been discussed for its single-player counterparts in the papers by Bertsekas and Yu [BeY10], [BeY12], [YuB13a], and also in a more abstract form in the author's books [Ber12a] and [Ber20]. In particular, there is a highly parallelizable and convergent distributed implementation, which is based on state space partitioning, and asynchronous policy evaluation and policy improvement operations within each set of the partition. The key idea, which forms the core of asynchronous DP algorithms [Ber82], [Ber83] (see also the books [BeT89], [Ber12a], [Ber20]) is that the mapping $G_{\mu,\nu}$ of Eq. (5.77) has two components *for every state* (policy evaluation and policy improvement) for the minimizer and two corresponding components for every state for the maximizer. Because of the uniform sup-norm contraction property of $G_{\mu,\nu}$, iterating with any one of these components, and at any single state, does not impede the progress made by iterations with the other components, while making eventual progress towards the solution.

In view of its asynchronous convergence capability, our framework is also suitable for on-line implementations where policy improvement and evaluations are done at only one state at a time. In such implementations, the algorithm performs a policy improvement at a single state, followed by a number of policy evaluations at other states, with the current policy pair (μ^t, ν^t) evaluated at only one state x at a time, and the cycle is repeated. One may select states cyclically for policy improvement, but there are alternative possibilities, including the case where states are selected on-line as the system operates. An on-line PI algorithm of this type, which may also be operated as a rollout algorithm (a control selected by a policy improvement at each encountered state), was given recently in the author's paper [Ber21a], and can be straightforwardly adapted to the minimax and Markov game cases of this chapter.

Other algorithmic possibilities, also discussed in the works just noted, involve the presence of "communication delays" between processors, which roughly means that the iterates generated at some processors may involve iterates of other processors that are out-of-date. This is possible because the asynchronous convergence line of analysis framework of [Ber83] in combination with the uniform weighted sup-norm contraction property of Prop. 5.4.2 can tolerate the presence of such delays. Implementations that involve forms of stochastic sampling have also been given in the papers [BeY12], [YuB13a].

An important issue for efficient implementation of our algorithm is the relative frequency of policy improvement and policy evaluation operations. If a very large number of contiguous policy evaluation operations, using the

same policy pair (μ^t, ν^t), is done between policy improvement operations, the policy evaluation is nearly exact. Then the algorithm's behavior is essentially the same as the one of the nonoptimistic algorithm where policy evaluation is done according to

$$J_{1,\mu^t,\nu^t}(x_1) = H_1\Big(x_1, \mu^t(x_1), \max\big[V_2^t, J_{2,\mu^t,\nu^t}\big]\Big), \qquad x_1 \in X_1,$$

$$J_{2,\mu^t,\nu^t}(x_2) = H_2\Big(x_2, \nu^t(x_2), \min\big[V_1^t, J_{1,\mu^t,\nu^t}\big]\Big), \qquad x_2 \in X_2,$$

cf. Eqs. (5.44)-(5.45) (in the context of Markovian decision problems, this type of policy evaluation involves the solution of an optimal stopping problem; cf. the paper [BeY12]). Otherwise the policy evaluation is inexact/optimistic, and in the extreme case where only one policy evaluation is done between policy improvements, the algorithm resembles a value iteration method. Based on experience with optimistic PI, it appears that the optimal number of policy evaluations between policy improvements should be substantially larger than one, and should also be problem-dependent.

We mention the possibility of extensions to other related minimax and Markov game problems. In particular, the treatment of undiscounted problems that involve a termination state can be patterned after the distributed asynchronous PI algorithm for stochastic shortest path problems by Yu and Bertsekas [YuB13a], and will be the subject of a separate report. A related area of investigation is on-line algorithms applied to robust shortest path planning problems, where the aim is to reach a termination state at minimum cost and against the actions of an antagonistic opponent. The author's paper [Ber19c] (see also Section 3.5.3) has provided analysis and algorithms, some of the PI type, for these minimax versions of shortest path problems, and has given many references of related works. Still our PI algorithm of Section 5.3, appropriately extended, offers some substantial advantages within the shortest path context, in both a serial and a distributed computing environment.

Note that a sequential minimax problem with a finite horizon may be viewed as a simple special case of an infinite horizon problem with a termination state. The PI algorithms of the present chapter are directly applicable and can be simply modified for such a problem. In conjunction with function approximation methods, such as the aggregation method described earlier, they may provide an attractive alternative to exact, but hopelessly time-consuming solution approaches.

For an interesting class of finite horizon problems, consider a two-stage "robust" version of stochastic programming, patterned after Example 5.3.3 and Eq. (5.42). Here, at an initial state x_0, the decision maker/minimizer applies a decision $u_0 \in U(x_0)$, an antagonistic nature chooses $v_0 \in V(x_0, u_0)$, and a random disturbance w_0 is generated according to a probability distribution than depends on (x_0, u_0, v_0). A cost

$g_0(x_0, u_0, v_0, w_0)$ is then incurred and the next state

$$x_1 = f(x_0, u_0, v_0, w_0)$$

is generated. Then the process is repeated at the second stage, with (x_1, u_1, v_1, w_1) replacing (x_0, u_0, v_0, w_0), and finally a terminal cost $G_2(x_2)$ is incurred where

$$x_2 = f(x_1, u_1, v_1, w_1).$$

Here the decision maker aims to minimize the expected total cost assuming a worst-case selection of (v_0, v_1). The maximizing choices (v_0, v_1) may have a variety of problem-dependent interpretations, including prices affecting the costs g_0, g_1, G_2, and forecasts affecting the probability distributions of the disturbances (w_0, w_1). The distributed asynchronous PI algorithm of Section 5.3 is easily modified for this problem, and similarly can be interpreted as Newton's method for solving a two-stage version of Bellman's equation. Exact solution of the problem may be a daunting computational task, but a satisfactory suboptimal solution, along the lines of Section 5.5, using approximation in value space with function approximation based on aggregation may prove feasible.

Finally, let us note a theoretical use of our line of analysis that is based on uniform contraction properties. It may form the basis for a rigorous mathematical treatment of PI algorithms in stochastic two-player DP models that involve universally measurable policies. We refer to the paper by Yu and Bertsekas [YuB15], where the associated issues of validity and convergence of PI methods for single-player problems have been addressed using algorithmic ideas that are closely related to the ones of the present chapter.

APPENDIX A:

Notation and Mathematical Conventions

In this appendix we collect our notation, and some related mathematical facts and conventions.

A.1 SET NOTATION AND CONVENTIONS

If X is a set and x is an element of X, we write $x \in X$. A set can be specified in the form $X = \{x \mid x \text{ satisfies } P\}$, as the set of all elements satisfying property P. The union of two sets X_1 and X_2 is denoted by $X_1 \cup X_2$, and their intersection by $X_1 \cap X_2$. The empty set is denoted by $\emptyset$. The symbol $\forall$ means "for all."

The set of real numbers (also referred to as scalars) is denoted by $\Re$. The set of extended real numbers is denoted by $\Re^*$:

$$\Re^* = \Re \cup \{\infty, -\infty\}.$$

We write $-\infty < x < \infty$ for all real numbers x, and $-\infty \leq x \leq \infty$ for all extended real numbers x. We denote by $[a, b]$ the set of (possibly extended) real numbers x satisfying $a \leq x \leq b$. A rounded, instead of square, bracket denotes strict inequality in the definition. Thus $(a, b]$, $[a, b)$, and (a, b) denote the set of all x satisfying $a < x \leq b$, $a \leq x < b$, and $a < x < b$, respectively.

Generally, we adopt standard conventions regarding addition and multiplication in $\Re^*$, except that we take

$$\infty - \infty = -\infty + \infty = \infty,$$

and we take the product of 0 and ∞ or $-\infty$ to be 0. In this way the sum and product of two extended real numbers is well-defined. Division by 0 or ∞ does not appear in our analysis. In particular, we adopt the following rules in calculations involving ∞ and $-\infty$:

$$\alpha + \infty = \infty + \alpha = \infty, \qquad \forall\ \alpha \in \Re^*,$$

$$\alpha - \infty = -\infty + \alpha = -\infty, \qquad \forall\ \alpha \in [-\infty, \infty),$$

$$\alpha \cdot \infty = \infty, \qquad \alpha \cdot (-\infty) = -\infty, \qquad \forall\ \alpha \in (0, \infty],$$

$$\alpha \cdot \infty = -\infty, \qquad \alpha \cdot (-\infty) = \infty, \qquad \forall\ \alpha \in [-\infty, 0),$$

$$0 \cdot \infty = \infty \cdot 0 = 0 = 0 \cdot (-\infty) = (-\infty) \cdot 0, \qquad -(-\infty) = \infty.$$

Under these rules, the following laws of arithmetic are still valid within $\Re^*$:

$$\alpha_1 + \alpha_2 = \alpha_2 + \alpha_1, \qquad (\alpha_1 + \alpha_2) + \alpha_3 = \alpha_1 + (\alpha_2 + \alpha_3),$$

$$\alpha_1 \alpha_2 = \alpha_2 \alpha_1, \qquad (\alpha_1 \alpha_2)\alpha_3 = \alpha_1(\alpha_2 \alpha_3).$$

We also have

$$\alpha(\alpha_1 + \alpha_2) = \alpha\alpha_1 + \alpha\alpha_2$$

if either $\alpha \geq 0$ or else $(\alpha_1 + \alpha_2)$ is not of the form $\infty - \infty$.

Inf and Sup Notation

The *supremum* of a nonempty set $X \subset \Re^*$, denoted by $\sup X$, is defined as the smallest $y \in \Re^*$ such that $y \geq x$ for all $x \in X$. Similarly, the *infimum* of X, denoted by $\inf X$, is defined as the largest $y \in \Re^*$ such that $y \leq x$ for all $x \in X$. For the empty set, we use the convention

$$\sup \varnothing = -\infty, \qquad \inf \varnothing = \infty.$$

If $\sup X$ is equal to an $\overline{x} \in \Re^*$ that belongs to the set X, we say that $\overline{x}$ is the *maximum point* of X and we write $\overline{x} = \max X$. Similarly, if $\inf X$ is equal to an $\overline{x} \in \Re^*$ that belongs to the set X, we say that $\overline{x}$ is the *minimum point* of X and we write $\overline{x} = \min X$. Thus, when we write $\max X$ (or $\min X$) in place of $\sup X$ (or $\inf X$, respectively), we do so just for emphasis: we indicate that it is either evident, or it is known through earlier analysis, or it is about to be shown that the maximum (or minimum, respectively) of the set X is attained at one of its points.

A.2 FUNCTIONS

If f is a function, we use the notation $f : X \mapsto Y$ to indicate the fact that f is defined on a nonempty set X (its *domain*) and takes values in a set Y (its *range*). Thus when using the notation $f : X \mapsto Y$, we implicitly assume that X is nonempty. We will often use the *unit function* $e : X \mapsto \Re$, defined by

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

Given a set X, we denote by $\mathcal{R}(X)$ the set of real-valued functions $J : X \mapsto \Re$, and by $\mathcal{E}(X)$ the set of all extended real-valued functions $J : X \mapsto \Re^*$. For any collection $\{J_\gamma \mid \gamma \in \Gamma\} \subset \mathcal{E}(X)$, parameterized by the elements of a set Γ, we denote by $\inf_{\gamma\in\Gamma} J_\gamma$ the function taking the value $\inf_{\gamma\in\Gamma} J_\gamma(x)$ at each $x \in X$.

For two functions $J_1, J_2 \in \mathcal{E}(X)$, we use the shorthand notation $J_1 \leq J_2$ to indicate the pointwise inequality

$$J_1(x) \leq J_2(x), \qquad \forall\, x \in X.$$

We use the shorthand notation $\inf_{i\in I} J_i$ to denote the function obtained by pointwise infimum of a collection $\{J_i \mid i \in I\} \subset \mathcal{E}(X)$, i.e.,

$$\left(\inf_{i\in I} J_i\right)(x) = \inf_{i\in I} J_i(x), \qquad \forall\, x \in X.$$

We use similar notation for sup.

Given subsets $S_1, S_2, S_3 \subset \mathcal{E}(X)$ and mappings $T_1 : S_1 \mapsto S_3$ and $T_2 : S_2 \mapsto S_1$, the *composition* of T_1 and T_2 is the mapping $T_1T_2 : S_2 \mapsto S_3$ defined by

$$(T_1T_2J)(x) = \big(T_1(T_2J)\big)(x), \qquad \forall\, J \in S_2,\ x \in X.$$

In particular, given a subset $S \subset \mathcal{E}(X)$ and mappings $T_1 : S \mapsto S$ and $T_2 : S \mapsto S$, the composition of T_1 and T_2 is the mapping $T_1T_2 : S \mapsto S$ defined by

$$(T_1T_2J)(x) = \big(T_1(T_2J)\big)(x), \qquad \forall\, J \in S,\ x \in X.$$

Similarly, given mappings $T_k : S \mapsto S$, $k = 1, \ldots, N$, their composition is the mapping $(T_1 \cdots T_N) : S \mapsto S$ defined by

$$(T_1T_2 \cdots T_NJ)(x) = \big(T_1(T_2(\cdots(T_NJ)))\big)(x), \qquad \forall\, J \in S,\ x \in X.$$

In our notation involving compositions we minimize the use of parentheses, as long as clarity is not compromised. In particular, we write T_1T_2J instead of (T_1T_2J) or $(T_1T_2)J$ or $T_1(T_2J)$, but we write $(T_1T_2J)(x)$ to indicate the value of T_1T_2J at $x \in X$.

If X and Y are nonempty sets, a mapping $T : S_1 \mapsto S_2$, where $S_1 \subset \mathcal{E}(X)$ and $S_2 \subset \mathcal{E}(Y)$, is said to be *monotone* if for all $J, J' \in S_1$,

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

Sequences of Functions

For a sequence of functions $\{J_k\} \subset \mathcal{E}(X)$ that converges pointwise, we denote by $\lim_{k\to\infty} J_k$ the pointwise limit of $\{J_k\}$. We denote by $\limsup_{k\to\infty} J_k$ (or $\liminf_{k\to\infty} J_k$) the pointwise limit superior (or inferior, respectively) of $\{J_k\}$. If $\{J_k\} \subset \mathcal{E}(X)$ converges pointwise to J, we write $J_k \to J$. Note that we reserve this notation for pointwise convergence. To denote convergence with respect to a norm $\|\cdot\|$, we write $\|J_k - J\| \to 0$.

A sequence of functions $\{J_k\} \subset \mathcal{E}(X)$ is said to be *monotonically nonincreasing* (or *monotonically nondecreasing*) if $J_{k+1} \leq J_k$ for all k (or $J_{k+1} \geq J_k$ for all k, respectively). Such a sequence always has a (pointwise) limit within $\mathcal{E}(X)$. We write $J_k \downarrow J$ (or $J_k \uparrow J$) to indicate that $\{J_k\}$ is monotonically nonincreasing (or monotonically nondecreasing, respectively) and that its limit is J.

Let $\{J_{mn}\} \subset \mathcal{E}(X)$ be a double indexed sequence, which is monotonically nonincreasing separately for each index in the sense that

$$J_{(m+1)n} \leq J_{mn}, \qquad J_{m(n+1)} \leq J_{mn}, \qquad \forall\, m, n = 0, 1, \ldots.$$

For such sequences, a useful fact is that

$$\lim_{m\to\infty} \Big(\lim_{n\to\infty} J_{mn} \Big) = \lim_{m\to\infty} J_{mm}.$$

There is a similar fact for monotonically nondecreasing sequences.

Expected Values

Given a random variable w defined over a probability space Ω, the expected value of w is defined by

$$E\{w\} = E\{w^+\} + E\{w^-\},$$

where w^+ and w^- are the positive and negative parts of w,

$$w^+(\omega) = \max\big\{0, w(\omega)\big\}, \qquad w^-(\omega) = \min\big\{0, w(\omega)\big\}.$$

In this way, taking also into account the rule $\infty - \infty = \infty$, the expected value $E\{w\}$ is well-defined if Ω is finite or countably infinite. In more general cases, $E\{w\}$ is similarly defined by the appropriate form of integration, and more detail will be given at specific points as needed.

APPENDIX B:

Contraction Mappings

B.1 CONTRACTION MAPPING FIXED POINT THEOREMS

The purpose of this appendix is to provide some background on contraction mappings and their properties. Let Y be a real vector space with a norm $\|\cdot\|$, i.e., a real-valued function satisfying for all $y \in Y$, $\|y\| \geq 0$, $\|y\| = 0$ if and only if $y = 0$, and

$$\|ay\| = |a|\|y\|, \quad \forall\ a \in \Re, \qquad \|y + z\| \leq \|y\| + \|z\|, \quad \forall\ y, z \in Y.$$

Let $\overline{Y}$ be a closed subset of Y. A function $F : \overline{Y} \mapsto \overline{Y}$ is said to be a *contraction mapping* if for some $\rho \in (0, 1)$, we have

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

The scalar ρ is called the *modulus of contraction* of F.

Example B.1 (Linear Contraction Mappings in $\Re^n$)

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

$$Fy = b + Ay,$$

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

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

$$\|Ay\|_s \leq \big(\sigma(A) + \epsilon\big)\|y\|_s, \qquad \forall\ y \in \Re^n. \tag{B.1}$$

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

$$\|Fy - Fz\|_s = \big\|A(y-z)\big\|_s \leq \rho \|y - z\|_s, \qquad \forall\ y, z \in \Re^n. \tag{B.2}$$

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

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

A sequence $\{y_k\} \subset Y$ is said to be a *Cauchy sequence* if $\|y_m - y_n\| \to 0$ as $m, n \to \infty$, i.e., given any $\epsilon > 0$, there exists N such that $\|y_m - y_n\| \leq \epsilon$ for all $m, n \geq N$. The space Y is said to be *complete* under the norm $\|\cdot\|$ if every Cauchy sequence $\{y_k\} \subset Y$ is convergent, in the sense that for some $y \in Y$, we have $\|y_k - y\| \to 0$. Note that a Cauchy sequence is always bounded. Also, a Cauchy sequence of real numbers is convergent, implying that the real line is a complete space and so is every real finite-dimensional vector space. On the other hand, an infinite dimensional space may not be complete under some norms, while it may be complete under other norms.

When Y is complete and $\overline{Y}$ is a closed subset of Y, an important property of a contraction mapping $F : \overline{Y} \mapsto \overline{Y}$ is that it has a unique fixed point within $\overline{Y}$, i.e., the equation

$$y = Fy$$

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

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

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

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

so that $\|Ay\|_s \leq \big(\sigma(A) + \beta\delta\big)\|y\|_s$, for all $y \in \Re^n$. For a given $\epsilon > 0$, we choose $\delta = \epsilon/\beta$, so the preceding relation yields Eq. (B.1).

converges to y^*, starting from an arbitrary initial point y_0.

Proposition B.1: (Contraction Mapping Fixed-Point Theorem) Let Y be a complete vector space and let $\overline{Y}$ be a closed subset of Y. Then if $F : \overline{Y} \mapsto \overline{Y}$ is a contraction mapping with modulus $\rho \in (0,1)$, there exists a unique $y^* \in \overline{Y}$ such that

$$y^* = Fy^*.$$

Furthermore, the sequence $\{F^k y\}$ converges to y^* for any $y \in \overline{Y}$, and we have

$$\|F^k y - y^*\| \leq \rho^k \|y - y^*\|, \qquad k = 1, 2, \ldots.$$

Proof: Let $y \in \overline{Y}$ and consider the iteration $y_{k+1} = Fy_k$ starting with $y_0 = y$. By the contraction property of F,

$$\|y_{k+1} - y_k\| \leq \rho \|y_k - y_{k-1}\|, \qquad k = 1, 2, \ldots,$$

which implies that

$$\|y_{k+1} - y_k\| \leq \rho^k \|y_1 - y_0\|, \qquad k = 1, 2, \ldots.$$

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

$$\begin{aligned} \|y_{k+m} - y_k\| &\leq \sum_{i=1}^{m} \|y_{k+i} - y_{k+i-1}\| \\ &\leq \rho^k (1 + \rho + \cdots + \rho^{m-1}) \|y_1 - y_0\| \\ &\leq \frac{\rho^k}{1-\rho} \|y_1 - y_0\|. \end{aligned}$$

Therefore, $\{y_k\}$ is a Cauchy sequence in $\overline{Y}$ and must converge to a limit $y^* \in \overline{Y}$, since Y is complete and $\overline{Y}$ is closed. We have for all $k \geq 1$,

$$\|Fy^* - y^*\| \leq \|Fy^* - y_k\| + \|y_k - y^*\| \leq \rho \|y^* - y_{k-1}\| + \|y_k - y^*\|$$

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

$$\|y^* - \tilde{y}\| = \|Fy^* - F\tilde{y}\| \leq \rho \|y^* - \tilde{y}\|,$$

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

To show the convergence rate bound of the last part, note that

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

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

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

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

We say that a function $F : Y \mapsto Y$ is an *m-stage contraction mapping* if there exists a positive integer m and some $\rho < 1$ such that

$$\left\|F^m y - F^m y'\right\| \leq \rho \|y - y'\|, \qquad \forall\ y, y' \in Y,$$

where F^m denotes the composition of F with itself m times. Thus, F is an m-stage contraction if F^m is a contraction. Again, the scalar ρ is called the modulus of contraction. We have the following generalization of Prop. B.1.

Proposition B.2: (m-Stage Contraction Mapping Fixed-Point Theorem) Let Y be a complete vector space and let $\overline{Y}$ be a closed subset of Y. Then if $F : \overline{Y} \mapsto \overline{Y}$ is an m-stage contraction mapping with modulus $\rho \in (0, 1)$, there exists a unique $y^* \in \overline{Y}$ such that

$$y^* = F y^*.$$

Furthermore, $\{F^k y\}$ converges to y^* for any $y \in \overline{Y}$.

Proof: Since F^m maps $\overline{Y}$ into $\overline{Y}$ and is a contraction mapping, by Prop. B.1, it has a unique fixed point in $\overline{Y}$, denoted y^*. Applying F to both sides of the relation $y^* = F^m y^*$, we see that $F y^*$ is also a fixed point of F^m, so by the uniqueness of the fixed point, we have $y^* = F y^*$. Therefore y^* is a fixed point of F. If F had another fixed point, say $\tilde{y}$, then we would have $\tilde{y} = F^m \tilde{y}$, which by the uniqueness of the fixed point of F^m implies that $\tilde{y} = y^*$. Thus, y^* is the unique fixed point of F.

To show the convergence of $\{F^k y\}$, note that by Prop. B.1, we have for all $y \in \overline{Y}$,

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

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

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

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

B.2 WEIGHTED SUP-NORM CONTRACTIONS

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

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

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

$$\|J\| = \sup_{x\in X} \frac{\big|J(x)\big|}{v(x)}. \tag{B.3}$$

It is easily verified that $\|\cdot\|$ thus defined has the required properties for being a norm. Furthermore, $\mathcal{B}(X)$ *is complete under this norm*. To see this, consider a Cauchy sequence $\{J_k\} \subset \mathcal{B}(X)$, and note that $\|J_m - J_n\| \to 0$ as $m, n \to \infty$ implies that for all $x \in X$, $\{J_k(x)\}$ is a Cauchy sequence of real numbers, so it converges to some $J^*(x)$. We will show that $J^* \in \mathcal{B}(X)$ and that $\|J_k - J^*\| \to 0$. To this end, it will be sufficient to show that given any $\epsilon > 0$, there exists an integer K such that

$$\frac{\big|J_k(x) - J^*(x)\big|}{v(x)} \le \epsilon, \qquad \forall\ x \in X,\ k \ge K.$$

This will imply that

$$\sup_{x\in X} \frac{\big|J^*(x)\big|}{v(x)} \le \epsilon + \|J_k\|, \qquad \forall\ k \ge K,$$

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

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

The right-hand side above is less or equal to

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

The first term in the above sum is less than $\epsilon/2$ for i and n larger than some threshold; fixing i and letting n be sufficiently large, the second term can also be made less than $\epsilon/2$, so the sum is made less than ϵ - a contradiction. In conclusion, the space $\mathcal{B}(X)$ is complete, so the fixed point results of Props. B.1 and B.2 apply.

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

Finite-Dimensional Cases

Let us now focus on the finite-dimensional case $X = \{1, \ldots, n\}$, in which case $\mathcal{R}(X)$ and $\mathcal{B}(X)$ can be identified with $\Re^n$. We first consider a linear mapping (cf. Example B.1). We have the following proposition.

Proposition B.3: Consider a linear mapping $F : \Re^n \mapsto \Re^n$ of the form

$$Fy = b + Ay,$$

where A is an $n \times n$ matrix with components a_{ij}, and b is a vector in $\Re^n$. Denote by $|A|$ the matrix whose components are the absolute values of the components of A and let $\sigma(A)$ and $\sigma\big(|A|\big)$ denote the spectral radii of A and $|A|$, respectively. Then:

(a) $|A|$ has a real eigenvalue λ, which is equal to its spectral radius, and an associated nonnegative eigenvector.

(b) F is a contraction with respect to some weighted sup-norm if and only if $\sigma\big(|A|\big) < 1$. In particular, any substochastic matrix P ($p_{ij} \geq 0$ for all i, j, and $\sum_{j=1}^n p_{ij} \leq 1$, for all i) is a contraction with respect to some weighted sup-norm if and only if $\sigma(P) < 1$.

(c) F is a contraction with respect to the weighted sup-norm

$$\|y\| = \max_{i=1,\ldots,n} \frac{|y_i|}{v(i)}$$

if and only if

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

Proof: (a) This is the Perron-Frobenius Theorem; see e.g., [BeT89], Chapter 2, Prop. 6.6.

(b) This follows from the Perron-Frobenius Theorem; see [BeT89], Chapter 2, Cor. 6.2.

(c) This is proved in more general form in the following Prop. B.4. **Q.E.D.**

Consider next a nonlinear mapping $F : \Re^n \mapsto \Re^n$ that has the property

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

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

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

for some $\alpha \in (0, 1)$, where $\big(|Fy - Fz|\big)(i)$ and $\big(P\,|y - z|\big)(i)$ are the ith components of the vectors $|Fy - Fz|$ and $P\,|y - z|$, respectively. Thus, F is a contraction with respect to $\|\cdot\|$. For additional discussion of linear and nonlinear contraction mapping properties and characterizations such as the one above, see the book [OrR70].

Linear Mappings on Countable Spaces

The case where X is countable (or, as a special case, finite) is frequently encountered in DP. The following proposition provides some useful criteria for verifying the contraction property of mappings that are either linear or are obtained via a parametric minimization of other contraction mappings.

Proposition B.4: Let $X = \{1, 2, \ldots\}$.

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

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

where b_i and a_{ij} are some scalars. Then F is a contraction with modulus ρ with respect to the weighted sup-norm (B.3) if and only if

$$\frac{\sum_{j\in X} |a_{ij}|\, v(j)}{v(i)} \le \rho, \qquad i \in X. \tag{B.4}$$

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

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

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

Proof: (a) Assume that Eq. (B.4) holds. For any $J, J' \in \mathcal{B}(X)$, we have

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

where the last inequality follows from the hypothesis.

Conversely, arguing by contradiction, let's assume that Eq. (B.4) is violated for some $i \in X$. Define $J(j) = v(j)\,\mathrm{sgn}(a_{ij})$ and $J'(j) = 0$ for all $j \in X$. Then we have $\|J - J'\| = \|J\| = 1$, and

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

showing that F is not a contraction of modulus ρ.

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

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

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

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

Reversing the roles of J and J', we obtain

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

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

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

Proposition B.5: Let $X = \{1, 2, \ldots\}$, let M be a parameter set, and for each $\mu \in M$, let F_μ be a linear mapping of the form

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

where we assume that the summation above is well-defined for all $J \in \mathcal{B}(X)$.

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

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

with

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

(b) Consider the mapping F

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

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

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

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

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

$$(F_\mu J)(i) \leq \big|b_i(\mu)\big| + \sum_{j \in X} \big|a_{ij}(\mu)\big| \, \big|J(j)/v(j)\big| \, v(j)$$

$$\begin{aligned} &\leq \big|b_i(\mu)\big| + \|J\| \sum_{j \in X} \big|a_{ij}(\mu)\big|\, v(j) \\ &= \big|b_i(\mu)\big| + \|J\|\, V_i(\mu), \end{aligned}$$

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

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

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

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

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

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

Q.E.D.

References

[ABB02] Abounadi, J., Bertsekas, B. P., and Borkar, V. S., 2002. "Stochastic Approximation for Non-Expansive Maps: Q-Learning Algorithms," SIAM J. on Control and Opt., Vol. 41, pp. 1-22.

[AnM79] Anderson, B. D. O., and Moore, J. B., 1979. Optimal Filtering, Prentice Hall, Englewood Cliffs, N. J.

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

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

[BFH86] Breton, M., Filar, J. A., Haurie, A., and Schultz, T. A., 1986. "On the Computation of Equilibria in Discounted Stochastic Dynamic Games," in Dynamic Games and Applications in Economics, Springer, pp. 64-87.

[Bau78] Baudet, G. M., 1978. "Asynchronous Iterative Methods for Multiprocessors," Journal of the ACM, Vol. 25, pp. 226-244.

[BeI96] Bertsekas, D. P., and Ioffe, S., 1996. "Temporal Differences-Based Policy Iteration and Applications in Neuro-Dynamic Programming," Lab. for Info. and Decision Systems Report LIDS-P-2349, MIT.

[BeK65] Bellman, R., and Kalaba, R. E., 1965. Quasilinearization and Nonlinear Boundary-Value Problems, Elsevier, N.Y.

[BeS78] Bertsekas, D. P., and Shreve, S. E., 1978. Stochastic Optimal Control: The Discrete Time Case, Academic Press, N. Y.; may be downloaded from http://web.mit.edu/dimitrib/www/home.html

[BeT89] Bertsekas, D. P., and Tsitsiklis, J. N., 1989. Parallel and Distributed Computation: Numerical Methods, Prentice-Hall, Engl. Cliffs, N. J.; may be downloaded from http://web.mit.edu/dimitrib/www/home.html

[BeT91] Bertsekas, D. P., and Tsitsiklis, J. N., 1991. "An Analysis of Stochastic Shortest Path Problems," Math. of OR, Vol. 16, pp. 580-595.

[BeT96] Bertsekas, D. P., and Tsitsiklis, J. N., 1996. Neuro-Dynamic Programming, Athena Scientific, Belmont, MA.

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

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

[BeY09] Bertsekas, D. P., and Yu, H., 2009. "Projected Equation Methods for Approximate Solution of Large Linear Systems," J. of Computational and Applied Mathematics, Vol. 227, pp. 27-50.

[BeY10] Bertsekas, D. P., and Yu, H., 2010. "Asynchronous Distributed Policy Iteration in Dynamic Programming," Proc. of Allerton Conf. on Communication, Control and Computing, Allerton Park, Ill, pp. 1368-1374.

[BeY12] Bertsekas, D. P., and Yu, H., 2012. "Q-Learning and Enhanced Policy Iteration in Discounted Dynamic Programming," Math. of OR, Vol. 37, pp. 66-94.

[BeY16] Bertsekas, D. P., and Yu, H., 2016. "Stochastic Shortest Path Problems Under Weak Conditions," Lab. for Information and Decision Systems Report LIDS-2909, January 2016.

[Ber71] Bertsekas, D. P., 1971. "Control of Uncertain Systems With a Set-Membership Description of the Uncertainty," Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA (available from the author's website).

[Ber72] Bertsekas, D. P., 1972. "Infinite Time Reachability of State Space Regions by Using Feedback Control," IEEE Trans. Aut. Control, Vol. AC-17, pp. 604-613.

[Ber75] Bertsekas, D. P., 1975. "Monotone Mappings in Dynamic Programming," 1975 IEEE Conference on Decision and Control, pp. 20-25.

[Ber77] Bertsekas, D. P., 1977. "Monotone Mappings with Application in Dynamic Programming," SIAM J. on Control and Opt., Vol. 15, pp. 438-464.

[Ber82] Bertsekas, D. P., 1982. "Distributed Dynamic Programming," IEEE Trans. Aut. Control, Vol. AC-27, pp. 610-616.

[Ber83] Bertsekas, D. P., 1983. "Asynchronous Distributed Computation of Fixed Points," Math. Programming, Vol. 27, pp. 107-120.

[Ber87] Bertsekas, D. P., 1987. Dynamic Programming: Deterministic and Stochastic Models, Prentice-Hall, Englewood Cliffs, N. J.

[Ber96] Bertsekas, D. P., 1996. Lecture at NSF Workshop on Reinforcement Learning, Hilltop House, Harper's Ferry, N. Y.

[Ber98] Bertsekas, D. P., 1998. Network Optimization: Continuous and Discrete Models, Athena Scientific, Belmont, MA.

[Ber09] Bertsekas, D. P., 2009. Convex Optimization Theory, Athena Scientific, Belmont, MA.

[Ber10] Bertsekas, D. P., 2010. "Williams-Baird Counterexample for Q-Factor Asynchronous Policy Iteration,"
http://web.mit.edu/dimitrib/www/Williams-Baird Counterexample.pdf

[Ber11a] Bertsekas, D. P., 2011. "Temporal Difference Methods for General Projected Equations," IEEE Trans. Aut. Control, Vol. 56, pp. 2128-2139.

[Ber11b] Bertsekas, D. P., 2011. "λ-Policy Iteration: A Review and a New Implementation," Lab. for Info. and Decision Systems Report LIDS-P-2874, MIT; appears in Reinforcement Learning and Approximate Dynamic Programming for Feedback Control, by F. Lewis and D. Liu (eds.), IEEE Press, 2012.

[Ber11c] Bertsekas, D. P., 2011. "Approximate Policy Iteration: A Survey and

Some New Methods," J. of Control Theory and Applications, Vol. 9, pp. 310-335; a somewhat expanded version appears as Lab. for Info. and Decision Systems Report LIDS-2833, MIT, 2011.

[Ber12a] Bertsekas, D. P., 2012. Dynamic Programming and Optimal Control, Vol. II, 4th Edition: Approximate Dynamic Programming, Athena Scientific, Belmont, MA.

[Ber12b] Bertsekas, D. P., 2012. "Weighted Sup-Norm Contractions in Dynamic Programming: A Review and Some New Applications," Lab. for Info. and Decision Systems Report LIDS-P-2884, MIT.

[Ber15] Bertsekas, D. P., 2015. "Regular Policies in Abstract Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-P-3173, MIT, May 2015; arXiv preprint arXiv:1609.03115; SIAM J. on Optimization, Vol. 27, 2017, pp. 1694-1727.

[Ber16a] Bertsekas, D. P., 2016. "Affine Monotonic and Risk-Sensitive Models in Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-3204, MIT, June 2016; arXiv preprint arXiv:1608.01393; IEEE Trans. on Aut. Control, Vol. 64, 2019, pp. 3117-3128.

[Ber16b] Bertsekas, D. P., 2016. "Proximal Algorithms and Temporal Differences for Large Linear Systems: Extrapolation, Approximation, and Simulation," Report LIDS-P-3205, MIT, Oct. 2016; arXiv preprint arXiv:1610.1610.05427.

[Ber16c] Bertsekas, D. P., 2016. Nonlinear Programming, 3rd Edition, Athena Scientific, Belmont, MA.

[Ber17a] Bertsekas, D. P., 2017. Dynamic Programming and Optimal Control, Vol. I, 4th Edition, Athena Scientific, Belmont, MA.

[Ber17b] Bertsekas, D. P., 2017. "Value and Policy Iteration in Deterministic Optimal Control and Adaptive Dynamic Programming," IEEE Transactions on Neural Networks and Learning Systems, Vol. 28, pp. 500-509.

[Ber17c] Bertsekas, D. P., 2017. "Stable Optimal Control and Semicontractive Dynamic Programming," Report LIDS-P-3506, MIT, May 2017; SIAM J. on Control and Optimization, Vol. 56, 2018, pp. 231-252.

[Ber17d] Bertsekas, D. P., 2017. "Proper Policies in Infinite-State Stochastic Shortest Path Problems," Report LIDS-P-3507, MIT, May 2017; arXiv preprint arXiv:1711.10129.

[Ber18a] Bertsekas, D. P., 2018. "Feature-Based Aggregation and Deep Reinforcement Learning: A Survey and Some New Implementations," Lab. for Information and Decision Systems Report, MIT; arXiv preprint arXiv:1804.04577; IEEE/CAA Journal of Automatica Sinica, Vol. 6, 2019, pp. 1-31.

[Ber18b] Bertsekas, D. P., 2018. "Biased Aggregation, Rollout, and Enhanced Policy Improvement for Reinforcement Learning," Lab. for Information and Decision Systems Report, MIT; arXiv preprint arXiv:1910.02426.

[Ber18c] Bertsekas, D. P., 2018. "Proximal Algorithms and Temporal Differences for Solving Fixed Point Problems," Computational Optimization and Applications J., Vol. 70, pp. 709-736.

[Ber19a] Bertsekas, D. P., 2019. "Affine Monotonic and Risk-Sensitive Models in Dynamic Programming," IEEE Transactions on Aut. Control, Vol. 64, pp. 3117-3128.

[Ber19b] Bertsekas, D. P., 2019. Reinforcement Learning and Optimal Control, Athena Scientific, Belmont, MA.

[Ber19c] Bertsekas, D. P., 2019. "Robust Shortest Path Planning and Semicontractive Dynamic Programming," Naval Research Logistics, Vol. 66, pp. 15-37.

[Ber20] Bertsekas, D. P., 2020. Rollout, Policy Iteration, and Distributed Reinforcement Learning, Athena Scientific, Belmont, MA.

[Ber21a] Bertsekas, D. P., 2021. "On-Line Policy Iteration for Infinite Horizon Dynamic Programming," arXiv preprint arXiv:2106.00746.

[Ber21b] Bertsekas, D. P., 2021. "Multiagent Reinforcement Learning: Rollout and Policy Iteration," IEEE/CAA J. of Automatica Sinica, Vol. 8, pp. 249-271.

[Ber21c] Bertsekas, D. P., 2021. "Distributed Asynchronous Policy Iteration for Sequential Zero-Sum Games and Minimax Control," arXiv preprint arXiv:2107.10406, July 2021.

[Ber22] Bertsekas, D. P., 2022. Lessons from AlphaZero for Optimal, Model Predictive, and Stochastic Control, Athena Scientific, Belmont, MA.

[Bla65] Blackwell, D., 1965. "Positive Dynamic Programming," Proc. Fifth Berkeley Symposium Math. Statistics and Probability, pp. 415-418.

[BoM99] Borkar, V. S., Meyn, S. P., 1999. "Risk Sensitive Optimal Control: Existence and Synthesis for Models with Unbounded Cost," SIAM J. Control and Opt., Vol. 27, pp. 192-209.

[BoM00] Borkar, V. S., Meyn, S. P., 2000. "The O.D.E. Method for Convergence of Stochastic Approximation and Reinforcement Learning," SIAM J. Control and Opt., Vol. 38, pp. 447-469.

[BoM02] Borkar, V. S., Meyn, S. P., 2002. "Risk-Sensitive Optimal Control for Markov Decision Processes with Monotone Cost," Math. of OR, Vol. 27, pp. 192-209.

[Bor98] Borkar, V. S., 1998. "Asynchronous Stochastic Approximation," SIAM J. Control Opt., Vol. 36, pp. 840-851.

[Bor08] Borkar, V. S., 2008. Stochastic Approximation: A Dynamical Systems Viewpoint, Cambridge Univ. Press, N. Y.

[CFH07] Chang, H. S., Fu, M. C., Hu, J., Marcus, S. I., 2007. Simulation-Based Algorithms for Markov Decision Processes, Springer, N. Y.

[CaM88] Carraway, R. L., and Morin, T. L., 1988. "Theory and Applications of Generalized Dynamic Programming: An Overview," Computers and Mathematics with Applications, Vol. 16, pp. 779-788.

[CaR13] Canbolat, P. G., and Rothblum, U. G., 2013. "(Approximate) Iterated Successive Approximations Algorithm for Sequential Decision Processes," Annals of Operations Research, Vol. 208, pp. 309-320.

[Cao07] Cao, X. R., 2007. Stochastic Learning and Optimization: A Sensitivity-Based Approach, Springer, N. Y.

[ChM69] Chazan D., and Miranker, W., 1969. "Chaotic Relaxation," Linear Algebra and Applications, Vol. 2, pp. 199-222.

[ChS87] Chung, K.-J., and Sobel, M. J., 1987. "Discounted MDPs: Distribution Functions and Exponential Utility Maximization," SIAM J. Control and Opt., Vol. 25, pp. 49-62.

[CoM99] Coraluppi, S. P., and Marcus, S. I., 1999. "Risk-Sensitive and Minimax Control of Discrete-Time, Finite-State Markov Decision Processes," Automatica, Vol. 35, pp. 301-309.

[DFV00] de Farias, D. P., and Van Roy, B., 2000. "On the Existence of Fixed Points for Approximate Value Iteration and Temporal-Difference Learning," J. of Optimization Theory and Applications, Vol. 105, pp. 589-608.

[DeM67] Denardo, E. V., and Mitten, L. G., 1967. "Elements of Sequential Decision Processes," J. Indust. Engrg., Vol. 18, pp. 106-112.

[DeR79] Denardo, E. V., and Rothblum, U. G., 1979. "Optimal Stopping, Exponential Utility, and Linear Programming," Math. Programming, Vol. 16, pp. 228-244.

[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.

[FeM97] Fernandez-Gaucherand, E., and Marcus, S. I., 1997. "Risk-Sensitive Optimal Control of Hidden Markov Models: Structural Results," IEEE Trans. Aut. Control, Vol. AC-42, pp. 1418-1422.

[Fei02] Feinberg, E. A., 2002. "Total Reward Criteria," in E. A. Feinberg and A. Shwartz, (Eds.), Handbook of Markov Decision Processes, Springer, N. Y.

[FiT91] Filar, J. A., and Tolwinski, B., 1991. "On the Algorithm of Pollatschek and Avi-ltzhak," in Stochastic Games and Related Topics, Theory and Decision Library, Springer, Vol. 7, pp. 59-70.

[FiV96] Filar, J., and Vrieze, K., 1996. Competitive Markov Decision Processes, Springer, N. Y.

[FlM95] Fleming, W. H., and McEneaney, W. M., 1995. "Risk-Sensitive Control on an Infinite Time Horizon," SIAM J. Control and Opt., Vol. 33, pp. 1881-1915.

[Gos03] Gosavi, A., 2003. Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning, Springer, N. Y.

[GuS17] Guillot, M., and Stauffer, G., 2017. "The Stochastic Shortest Path Problem: A Polyhedral Combinatorics Perspective," Univ. of Grenoble Report.

[HCP99] Hernandez-Lerma, O., Carrasco, O., and Perez-Hernandez. 1999. "Markov Control Processes with the Expected Total Cost Criterion: Optimality, Stability, and Transient Models," Acta Appl. Math., Vol. 59, pp. 229-269.

[Hay08] Haykin, S., 2008. Neural Networks and Learning Machines, (3rd Edition), Prentice-Hall, Englewood-Cliffs, N. J.

[HeL99] Hernandez-Lerma, O., and Lasserre, J. B., 1999. Further Topics on Discrete-Time Markov Control Processes, Springer, N. Y.

[HeM96] Hernandez-Hernandez, D., and Marcus, S. I., 1996. "Risk Sensitive Control of Markov Processes in Countable State Space," Systems and Control Letters, Vol. 29, pp. 147-155.

[HiW05] Hinderer, K., and Waldmann, K.-H., 2005. "Algorithms for Countable State Markov Decision Models with an Absorbing Set," SIAM J. of Control and

Opt., Vol. 43, pp. 2109-2131.

[HoK66] Hoffman, A. J., and Karp, R. M., 1966. "On Nonterminating Stochastic Games," Management Science, Vol. 12, pp. 359-370.

[HoM72] Howard, R. S., and Matheson, J. E., 1972. "Risk-Sensitive Markov Decision Processes," Management Science, Vol. 8, pp. 356-369.

[JBE94] James, M. R., Baras, J. S., Elliott, R. J., 1994. "Risk-Sensitive Control and Dynamic Games for Partially Observed Discrete-Time Nonlinear Systems," IEEE Trans. Aut. Control, Vol. AC-39, pp. 780-792.

[JaC06] James, H. W., and Collins, E. J., 2006. "An Analysis of Transient Markov Decision Processes," J. Appl. Prob., Vol. 43, pp. 603-621.

[Jac73] Jacobson, D. H., 1973. "Optimal Stochastic Linear Systems with Exponential Performance Criteria and their Relation to Deterministic Differential Games," IEEE Transactions on Automatic Control, Vol. AC-18, pp. 124-131.

[Kal60] Kalman, R. E., 1960. "Contributions to the Theory of Optimal Control," Bol. Soc. Mat. Mexicana, Vol. 5, pp. 102-119.

[Kal20] Kallenberg, L., 2020. Markov Decision Processes, Lecture Notes, University of Leiden.

[Kle68] Kleinman, D. L., 1968. "On an Iterative Technique for Riccati Equation Computations," IEEE Trans. Automatic Control, Vol. AC-13, pp. 114-115.

[Kuc72] Kucera, V., 1972. "The Discrete Riccati Equation of Optimal Control," Kybernetika, Vol. 8, pp. 430-447.

[Kuc73] Kucera, V., 1973. "A Review of the Matrix Riccati Equation," Kybernetika, Vol. 9, pp. 42-61.

[Kuh53] Kuhn, H. W., 1953. "Extensive Games and the Problem of Information," in Kuhn, H. W., and Tucker, A. W. (eds.), Contributions to the Theory of Games, Vol. II, Annals of Mathematical Studies No. 28, Princeton University Press, pp. 193-216.

[LaR95] Lancaster, P., and Rodman, L., 1995. Algebraic Riccati Equations, Clarendon Press, Oxford, UK.

[Mey07] Meyn, S., 2007. Control Techniques for Complex Networks, Cambridge Univ. Press, N. Y.

[Mit64] Mitten, L. G., 1964. "Composition Principles for Synthesis of Optimal Multistage Processes," Operations Research, Vol. 12, pp. 610-619.

[Mit74] Mitten, L. G., 1964. "Preference Order Dynamic Programming," Management Science, Vol. 21, pp. 43 - 46.

[Mor82] Morin, T. L., 1982. "Monotonicity and the Principle of Optimality," J. of Math. Analysis and Applications, Vol. 88, pp. 665-674.

[NeB03] Nedić, A., and Bertsekas, D. P., 2003. "Least-Squares Policy Evaluation Algorithms with Linear Function Approximation," J. of Discrete Event Systems, Vol. 13, pp. 79-110.

[OrR70] Ortega, J. M., and Rheinboldt, W. C., 1970. Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, N. Y.

[PPG16] Perolat, J., Piot, B., Geist, M., Scherrer, B., and Pietquin, O., 2016. "Softened Approximate Policy Iteration for Markov Games," in Proc. International Conference on Machine Learning, pp. 1860-1868.

[PSP15] Perolat, J., Scherrer, B., Piot, B., and Pietquin, O., 2015. "Approximate Dynamic Programming for Two-Player Zero-Sum Markov Games," in Proc. International Conference on Machine Learning, pp. 1321-1329.

[PaB99] Patek, S. D., and Bertsekas, D. P., 1999. "Stochastic Shortest Path Games," SIAM J. on Control and Opt., Vol. 36, pp. 804-824.

[Pal67] Pallu de la Barriere, R., 1967. Optimal Control Theory, Saunders, Phila; republished by Dover, N. Y., 1980.

[Pat01] Patek, S. D., 2001. "On Terminating Markov Decision Processes with a Risk Averse Objective Function," Automatica, Vol. 37, pp. 1379-1386.

[Pat07] Patek, S. D., 2007. "Partially Observed Stochastic Shortest Path Problems with Approximate Solution by Neuro-Dynamic Programming," IEEE Trans. on Systems, Man, and Cybernetics Part A, Vol. 37, pp. 710-720.

[Pli78] Pliska, S. R., 1978. "On the Transient Case for Markov Decision Chains with General State Spaces," in Dynamic Programming and its Applications, by M. L. Puterman (ed.), Academic Press, N. Y.

[PoA69] Pollatschek, M., and Avi-Itzhak, B., 1969. "Algorithms for Stochastic Games with Geometrical Interpretation," Management Science, Vol. 15, pp. 399-413.

[Pow07] Powell, W. B., 2007. Approximate Dynamic Programming: Solving the Curses of Dimensionality, J. Wiley and Sons, Hoboken, N. J; 2nd ed., 2011.

[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.

[PuB79] Puterman, M. L., and Brumelle, S. L., 1979. "On the Convergence of Policy Iteration in Stationary Dynamic Programming," Math. of Operations Research, Vol. 4, pp. 60-69.

[Put94] Puterman, M. L., 1994. Markovian Decision Problems, J. Wiley, N. Y.

[Rei16] Reissig, G., 2016. "Approximate Value Iteration for a Class of Deterministic Optimal Control Problems with Infinite State and Input Alphabets," Proc. 2016 IEEE Conf. on Decision and Control, pp. 1063-1068.

[Roc70] Rockafellar, R. T., 1970. Convex Analysis, Princeton Univ. Press, Princeton, N. J.

[Ros67] Rosenfeld, J., 1967. "A Case Study on Programming for Parallel Processors," Research Report RC-1864, IBM Res. Center, Yorktown Heights, N. Y.

[Rot79] Rothblum, U. G., 1979. "Iterated Successive Approximation for Sequential Decision Processes," in Stochastic Control and Optimization, by J. W. B. van Overhagen and H. C. Tijms (eds), Vrije University, Amsterdam.

[Rot84] Rothblum, U. G., 1984. "Multiplicative Markov Decision Chains," Math. of OR, Vol. 9, pp. 6-24.

[ScL12] Scherrer, B., and Lesner, B., 2012. "On the Use of Non-Stationary Policies for Stationary Infinite-Horizon Markov Decision Processes," NIPS 2012 - Neural Information Processing Systems, South Lake Tahoe, Ne.

[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.

[Sch11] Scherrer, B., 2011. "Performance Bounds for Lambda Policy Iteration and Application to the Game of Tetris," Report RR-6348, INRIA, France; J. of Machine Learning Research, Vol. 14, 2013, pp. 1181-1227.

[Sch12] Scherrer, B., 2012. "On the Use of Non-Stationary Policies for Infinite-Horizon Discounted Markov Decision Processes," INRIA Lorraine Report, France.

[Sha53] Shapley, L. S., 1953. "Stochastic Games," Proc. Nat. Acad. Sci. U.S.A., Vol. 39.

[Sob75] Sobel, M. J., 1975. "Ordinal Dynamic Programming," Management Science, Vol. 21, pp. 967-975.

[Str66] Strauch, R., 1966. "Negative Dynamic Programming," Ann. Math. Statist., Vol. 37, pp. 871-890.

[SuB98] Sutton, R. S., and Barto, A. G., 1998. Reinforcement Learning, MIT Press, Cambridge, MA.

[Sze98a] Szepesvari, C., 1998. Static and Dynamic Aspects of Optimal Sequential Decision Making, Ph.D. Thesis, Bolyai Institute of Mathematics, Hungary.

[Sze98b] Szepesvari, C., 1998. "Non-Markovian Policies in Sequential Decision Problems," Acta Cybernetica, Vol. 13, pp. 305-318.

[Sze10] Szepesvari, C., 2010. Algorithms for Reinforcement Learning, Morgan and Claypool Publishers, San Franscisco, CA.

[TBA86] Tsitsiklis, J. N., Bertsekas, D. P., and Athans, M., 1986. "Distributed Asynchronous Deterministic and Stochastic Gradient Optimization Algorithms," IEEE Trans. Aut. Control, Vol. AC-31, pp. 803-812.

[ThS10a] Thiery, C., and Scherrer, B., 2010. "Least-Squares λ-Policy Iteration: Bias-Variance Trade-off in Control Problems," in ICML'10: Proc. of the 27th Annual International Conf. on Machine Learning.

[ThS10b] Thiery, C., and Scherrer, B., 2010. "Performance Bound for Approximate Optimistic Policy Iteration," Technical Report, INRIA, France.

[Tol89] Tolwinski, B., 1989. "Newton-Type Methods for Stochastic Games," in Basar T. S., and Bernhard P. (eds), Differential Games and Applications, Lecture Notes in Control and Information Sciences, vol. 119, Springer, pp. 128-144.

[Tsi94] Tsitsiklis, J. N., 1994. "Asynchronous Stochastic Approximation and Q-Learning," Machine Learning, Vol. 16, pp. 185-202.

[VVL13] Vrabie, V., Vamvoudakis, K. G., and Lewis, F. L., 2013. Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles, The Institution of Engineering and Technology, London.

[Van78] van der Wal, J., 1978. "Discounted Markov Games: Generalized Policy Iteration Method," J. of Optimization Theory and Applications, Vol. 25, pp. 125-138.

[VeP87] Verdu, S., and Poor, H. V., 1987. "Abstract Dynamic Programming Models under Commutativity Conditions," SIAM J. on Control and Opt., Vol. 25, pp. 990-1006.

[Wat89] Watkins, C. J. C. H., Learning from Delayed Rewards, Ph.D. Thesis, Cambridge Univ., England.

[Whi80] Whittle, P., 1980. "Stability and Characterization Conditions in Negative Programming," Journal of Applied Probability, Vol. 17, pp. 635-645.

[Whi81] Whittle, P., 1981. "Risk-Sensitive Linear/Quadratic/Gaussian Control," Advances in Applied Probability, Vol. 13, pp. 764-777.

[Whi82] Whittle, P., 1982. Optimization Over Time, Wiley, N. Y., Vol. 1, 1982, Vol. 2, 1983.

[Whi90] Whittle, P., 1990. Risk-Sensitive Optimal Control, Wiley, Chichester.

[WiB93] Williams, R. J., and Baird, L. C., 1993. "Analysis of Some Incremental Variants of Policy Iteration: First Steps Toward Understanding Actor-Critic Learning Systems," Report NU-CCS-93-11, College of Computer Science, Northeastern University, Boston, MA.

[Wil71] Willems, J., 1971. "Least Squares Stationary Optimal Control and the Algebraic Riccati Equation," IEEE Trans. on Automatic Control, Vol. 16, pp. 621-634.

[YuB10] Yu, H., and Bertsekas, D. P., 2010. "Error Bounds for Approximations from Projected Linear Equations," Math. of OR, Vol. 35, pp. 306-329.

[YuB12] Yu, H., and Bertsekas, D. P., 2012. "Weighted Bellman Equations and their Applications in Dynamic Programming," Lab. for Info. and Decision Systems Report LIDS-P-2876, MIT.

[YuB13a] Yu, H., and Bertsekas, D. P., 2013. "Q-Learning and Policy Iteration Algorithms for Stochastic Shortest Path Problems," Annals of Operations Research, Vol. 208, pp. 95-132.

[YuB13b] Yu, H., and Bertsekas, D. P., 2013. "On Boundedness of Q-Learning Iterates for Stochastic Shortest Path Problems," Math. of OR, Vol. 38, pp. 209-227.

[YuB15] Yu, H., and Bertsekas, D. P., 2015. "A Mixed Value and Policy Iteration Method for Stochastic Control with Universally Measurable Policies," Math. of OR, Vol. 40, pp. 926-968.

[Yu14] Yu, H., 2014. "Stochastic Shortest Path Games and Q-Learning," arXiv preprint arXiv:1412.8570.

[Yu15] Yu, H., 2015. "On Convergence of Value Iteration for a Class of Total Cost Markov Decision Processes," SIAM J. on Control and Optimization, Vol. 53, pp. 1982-2016.

[ZYB21] Zhang, K., Yang, Z. and Basar, T., 2021. "Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms," Handbook of Reinforcement Learning and Control, pp. 321-384.

[Zac64] Zachrisson, L. E., 1964. "Markov Games," in Advances in Game Theory, by M. Dresher, L. S. Shapley, and A. W. Tucker, (eds.), Princeton Univ. Press, Princeton, N. J., pp. 211-253.

INDEX

Neuro-Dynamic Programming
Dimitri P. Bertsekas and John N. Tsitsiklis

Athena Scientific, 1996
512 pp., hardcover, ISBN 1-886529-10-8

This is the first textbook that fully explains the neuro-dynamic programming/reinforcement learning methodology, a breakthrough in the practical application of neural networks and dynamic programming to complex problems of planning, optimal decision making, and intelligent control.

From the review by George Cybenko for IEEE Computational Science and Engineering, May 1998:

"Neuro-Dynamic Programming is a remarkable monograph that integrates a sweeping mathematical and computational landscape into a coherent body of rigorous knowledge. The topics are current, the writing is clear and to the point, the examples are comprehensive and the historical notes and comments are scholarly."

"In this monograph, Bertsekas and Tsitsiklis have performed a Herculean task that will be studied and appreciated by generations to come. I strongly recommend it to scientists and engineers eager to seriously understand the mathematics and computations behind modern behavioral machine learning."

Among its special features, the book:

- Describes and unifies a large number of NDP methods, including several that are new
- Describes new approaches to formulation and solution of important problems in stochastic optimal control, sequential decision making, and discrete optimization
- Rigorously explains the mathematical principles behind NDP
- Illustrates through examples and case studies the practical application of NDP to complex problems from optimal resource allocation, optimal feedback control, data communications, game playing, and combinatorial optimization
- Presents extensive background and new research material on dynamic programming and neural network training

Neuro-Dynamic Programming is the winner of the 1997 INFORMS CSTS prize for research excellence in the interface between Operations Research and Computer Science

Reinforcement Learning and Optimal Control
Dimitri P. Bertsekas

Athena Scientific, 2019
388 pp., hardcover, ISBN 978-1-886529-39-7

This book explores the common boundary between optimal control and artificial intelligence, as it relates to reinforcement learning and simulation-based neural network methods. These are popular fields with many applications, which can provide approximate solutions to challenging sequential decision problems and large-scale dynamic programming (DP). The aim of the book is to organize coherently the broad mosaic of methods in these fields, which have a solid analytical and logical foundation, and have also proved successful in practice.

The book discusses both approximation in value space and approximation in policy space. It adopts a gradual expository approach, which proceeds along four directions:

- From exact DP to approximate DP: We first discuss exact DP algorithms, explain why they may be difficult to implement, and then use them as the basis for approximations.
- From finite horizon to infinite horizon problems: We first discuss finite horizon exact and approximate DP methodologies, which are intuitive and mathematically simple, and then progress to infinite horizon problems.
- From model-based to model-free implementations: We first discuss model-based implementations, and then we identify schemes that can be appropriately modified to work with a simulator.

The mathematical style of this book is somewhat different from the one of the author's DP books, and the 1996 neuro-dynamic programming (NDP) research monograph, written jointly with John Tsitsiklis. While we provide a rigorous, albeit short, mathematical account of the theory of finite and infinite horizon DP, and some fundamental approximation methods, we rely more on intuitive explanations and less on proof-based insights. Moreover, our mathematical requirements are quite modest: calculus, a minimal use of matrix-vector algebra, and elementary probability (mathematically complicated arguments involving laws of large numbers and stochastic convergence are bypassed in favor of intuitive explanations).

The book is supported by on-line video lectures and slides, as well as new research material, some of which has been covered in the present monograph.

Rollout, Policy Iteration, and Distributed Reinforcement Learning

Dimitri P. Bertsekas

Athena Scientific, 2020
480 pp., hardcover, ISBN 978-1-886529-07-6

This book develops in greater depth some of the methods from the author's Reinforcement Learning and Optimal Control textbook (Athena Scientific, 2019). It presents new research, relating to rollout algorithms, policy iteration, multiagent systems, partitioned architectures, and distributed asynchronous computation.

The application of the methodology to challenging discrete optimization problems, such as routing, scheduling, assignment, and mixed integer programming, including the use of neural network approximations within these contexts, is also discussed.

Much of the new research is inspired by the remarkable AlphaZero chess program, where policy iteration, value and policy networks, approximate lookahead minimization, and parallel computation all play an important role.

Among its special features, the book:

- Presents new research relating to distributed asynchronous computation, partitioned architectures, and multiagent systems, with application to challenging large scale optimization problems, such as combinatorial/discrete optimization, as well as partially observed Markov decision problems.
- Describes variants of rollout and policy iteration for problems with a multiagent structure, which allow the dramatic reduction of the computational requirements for lookahead minimization.
- Establishes connections of rollout algorithms and model predictive control, one of the most prominent control system design methodology.
- Expands the coverage of some research areas discussed in the author's 2019 textbook Reinforcement Learning and Optimal Control.
- Provides the mathematical analysis that supports the Newton step interpretations and the conclusions of the present book.

The book is supported by on-line video lectures and slides, as well as new research material, some of which has been covered in the present monograph.