Dynamic Programming and Optimal Control
4th Edition, Volume II

by

Dimitri P. Bertsekas

Massachusetts Institute of Technology

Chapter 4
Noncontractive Total Cost Problems

UPDATED

January 1, 2016

This is an updated version of Chapter 4 of the author's Dynamic Programming and Optimal Control, Vol. II, 4th Edition, Athena Scientific, 2012. It includes new material, and it is substantially revised and expanded (it has more than doubled in size).

The new material aims to provide a unified treatment of several models, all of which lack the contractive structure that is characteristic of the discounted problems of Chapters 1 and 2: positive and negative cost models, deterministic optimal control (including adaptive DP), stochastic shortest path models, and risk-sensitive models. Here is a summary of the new material:

(a) Stochastic shortest path problems under weak conditions and their relation to positive cost problems (Sections 4.1.4 and 4.4).

(b) Deterministic optimal control and adaptive DP (Sections 4.2 and 4.3).

(c) Affine monotonic and multiplicative cost models (Section 4.5).

The chapter will be periodically updated, and represents "work in progress." It may contain errors (hopefully not serious ones). Furthermore, its references to the literature are somewhat incomplete at present. Your comments and suggestions to the author at dimitrib@mit.edu are welcome.

4

Noncontractive Total Cost Problems

Contents

In the preceding chapters we dealt with total cost infinite horizon DP problems with rather favorable structure, where contraction properties played a fundamental role. In particular, in the discounted problems of Chapters 1 and 2, the Bellman equations for both stationary policy costs and optimal costs involve (unweighted) contraction mappings, with important analytical and computational benefits resulting. In summary, our main results were the following:

(a) The Bellman equations for stationary policy costs and optimal costs have unique solutions.

(b) There is a strong characterization of optimal stationary policies in terms of attainment of the minimum in the right side of Bellman's equation.

(c) Value iteration (VI) is convergent starting from any bounded initial condition.

(d) Policy iteration (PI) has strong convergence properties, particularly for finite-state problems.

The structure of the SSP problems discussed in Chapter 3 is not quite as favorable, but still has a strong contraction character, thanks to the dominant role of the proper policies, which have a weighted sup-norm contraction property (cf. Section 3.3). A key fact is that the noncontractive improper policies were assumed to produce infinite cost from some initial states (cf. Assumptions 3.1.1 and 3.1.2), and were effectively ruled out from being optimal. Thus strong versions of the results (a)-(d) above were obtained (with some modifications in the case of PI, in order to get around the potential unavailability of an initial proper policy).

In this chapter we consider total cost infinite horizon DP problems without making any kind of contraction assumption, relying only on the fundamental monotonicity property of DP. As a result none of the results (a)-(d) above hold in any generality, and their validity hinges on additional special structure of the problem at hand. Important special structures in his regard are:

(1) *Uniform positivity or uniform negativity of the cost per stage* $g(x,u,w)$. Among others, this ensures that J_π, the cost function of a policy π, is well-defined as a limit of the corresponding finite horizon cost functions. Moreover, J^*, the optimal cost function, satisfies Bellman's equation, as we will see (although it may not be the unique solution).

(2) *A deterministic problem structure*. Among others, we will see that this guarantees that J^* satisfies Bellman's equation. (This need not be true even for finite state SSP problems when the cost per stage can take both positive and negative values, as we will see.)

(3) *The presence of a cost-free and absorbing termination state*, as in

the SSP problems of Chapter 3 (but without requiring the strong assumptions of that chapter).

One of our aims is to look for connecting threads and to integrate the analysis of these special structures.

We start in Section 4.1 with a discussion of two major classes of problems: those with positive cost per stage, and those with negative cost per stage. None of the favorable results (a)-(d) above hold for these problems, and yet some weaker substitutes hold, which may be enhanced through the exploitation of additional problem structure. Interestingly, the results for the positive and the negative cost problems are quite different. In some ways negative cost problems exhibit more benign behavior, while positive cost problems exhibit more interesting and multifaceted behavior, particularly with respect to VI and PI. Positive cost problems with a *finite number of states* have an additional special characteristic, which greatly facilitates their solution: they are strongly related to the SSP problems of Chapter 3 and can be solved using the methodology of that chapter, as we will discuss in Section 4.1.5.

In Section 4.2, we will consider some special cases of positive cost problems that are deterministic, and are central in control system design. While in general, there may be multiple solutions of Bellman's equation, and the VI and PI algorithms may fail, we will delineate reasonable conditions under which these difficulties do not occur.

In Section 4.3, we consider infinite horizon versions of the linear system-quadratic cost problem we considered in Vol. I. This is a special case of the problem of Section 4.2, so the results proved there apply. We also discuss ways to apply simulation-based PI to the adaptive control of linear systems with unknown model parameters.

In Section 4.4, we return to SSP problems, which have neither the positive or negative cost structure of Section 4.1, nor the favorable proper/improper policy structure of Chapter 3. We will encounter here, quite anomalous behavior, including situations where J^*, the optimal cost function over all policies, may not be a solution of Bellman's equation. Instead, Bellman's equation may be solved by another cost function, *the optimal over the restricted set of proper policies*, which are well-behaved with respect to VI.

In Section 4.5, we consider affine monotonic problems, a generalization of the positive cost problems of Chapter 6, where the DP mapping associated with stationary policies has a linear structure. These problems include multiplicative cost problems, which have a strong connection with the SSP problems of Section 4.4. In particular, the analog of a proper policy in SSP is a *stable policy* in affine monotonic models. There are common threads to the analysis of Sections 4.1-4.5, based on an important notion from abstract DP, called *regularity*, which relates to restricted classes of policies that are well-behaved with respect to VI, and to the cost that may be achieved by optimization over this class. This notion is explored in more detail in Appendix B and in the abstract DP monograph [Ber13].

Finally, in Section 4.6, we consider a variety of interesting classes of problems, such as stopping, inventory control, continuous-time models, and nonstationary and periodic problems.

4.1 POSITIVE AND NEGATIVE COST MODELS

We consider the total cost infinite horizon problem of Section 1.1, whereby we want to find a policy $\pi = \{\mu_0, \mu_1, \ldots\}$, where $\mu_k : X \mapsto U$,

$$\mu_k(x_k) \in U(x_k), \qquad \forall\ x_k \in X,\ k = 0, 1, \ldots,$$

that minimizes the cost function

$$J_\pi(x_0) = \lim_{N\to\infty} \mathop{E}_{\substack{w_k \\ k=0,1,\ldots}} \left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\}, \tag{4.1}$$

subject to the system equation constraint

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

In this section we will assume throughout one of the following two conditions, both of which guarantee that the limit in the cost definition (4.1) is well-defined.

Assumption P: (Positivity) The cost per stage g satisfies

$$0 \le g(x, u, w), \qquad \text{for all } (x, u, w) \in X \times U \times W. \tag{4.2}$$

Assumption N: (Negativity) The cost per stage g satisfies

$$g(x, u, w) \le 0, \qquad \text{for all } (x, u, w) \in X \times U \times W. \tag{4.3}$$

Somewhat paradoxically, problems corresponding to Assumption P are sometimes referred to in the research literature as *negative DP problems*. This choice of name is due to historical reasons. It was introduced in the paper [Str66], where the problem of maximizing the infinite sum of negative rewards per stage was considered. Similarly, problems corresponding to Assumption N are sometimes referred to as *positive DP problems* (see [Bla65], [Str66]). Assumption N arises in problems where there is a nonnegative reward per stage and the total expected reward is to be *maximized.*

Note that when $\alpha < 1$ and g is either bounded above or below, we may add a suitable scalar to g in order to satisfy Eq. (4.2) or Eq. (4.3), respectively. Because of the discount factor, an optimal policy will not be affected by this change: the addition of a constant r to g merely adds $(1-\alpha)^{-1}r$ to the cost of every policy.

One complication arising from unbounded costs per stage is that, for some initial states x_0 and some genuinely interesting admissible policies $\pi = \{\mu_0, \mu_1, \ldots\}$, the cost $J_\pi(x_0)$ may be ∞ (in the case of Assumption P) or $-\infty$ (in the case of Assumption N). This is true even if there is discounting. Here is an example:

Example 4.1.1

Consider the scalar system

$$x_{k+1} = \beta x_k + u_k, \qquad k = 0, 1, \ldots,$$

where $x_k \in \Re$ and $u_k \in \Re$, for all k, and β is a positive scalar. The control constraint is

$$|u_k| \leq 1,$$

and the cost is

$$J_\pi(x_0) = \lim_{N\to\infty} \sum_{k=0}^{N-1} \alpha^k |x_k|.$$

Consider the policy $\tilde{\pi} = \{\tilde{\mu}, \tilde{\mu}, \ldots\}$, where $\tilde{\mu}(x) = 0$ for all $x \in \Re$. Then

$$J_{\tilde{\pi}}(x_0) = \lim_{N\to\infty} \sum_{k=0}^{N-1} \alpha^k \beta^k |x_0|,$$

and hence

$$J_{\tilde{\pi}}(x_0) = \begin{cases} 0 & \text{if } x_0 = 0 \\ \infty & \text{if } x_0 \neq 0 \end{cases} \qquad \text{if} \qquad \alpha\beta \geq 1,$$

while

$$J_{\tilde{\pi}}(x_0) = \frac{|x_0|}{1-\alpha\beta} \qquad \text{if} \qquad \alpha\beta < 1.$$

Note a peculiarity here: if $\beta > 1$ the state x_k diverges to ∞ or to $-\infty$, but if the discount factor is sufficiently small ($\alpha < 1/\beta$), the cost $J_{\tilde{\pi}}(x_0)$ is finite.

It is also possible to verify that when $\beta > 1$ and $\alpha\beta \geq 1$ the optimal cost satisfies

$$J^*(x_0) = \infty, \qquad \text{if } |x_0| \geq \tfrac{1}{\beta-1},$$

and

$$J^*(x_0) < \infty, \qquad \text{if } |x_0| < \tfrac{1}{\beta-1}.$$

What happens here is that when $\beta > 1$ the system is unstable, and in view of the restriction $|u_k| \leq 1$ on the control, it may not be possible to force the state near zero once it has reached sufficiently large magnitude.

The preceding example shows that there is not much that can be done about the possibility of the cost function being infinite for some policies. To cope with this situation, we conduct our analysis with the notational understanding that the costs $J_\pi(x_0)$ and $J^*(x_0)$ may be ∞ (or $-\infty$) under Assumption P (or N, respectively) for some initial states x_0 and policies π. In other words, we consider $J_\pi(\cdot)$ and $J^*(\cdot)$ to be extended real-valued functions. In fact, the entire subsequent analysis is valid even if the cost per stage $g(x, u, w)$ is ∞ or $-\infty$ for some (x, u, w), as long as Assumption P or Assumption N holds, respectively, although for simplicity, we assume that g is real-valued.

The line of analysis of this section is fundamentally different from the one of the discounted problem of Section 1.2. For the latter problem, the analysis was based on ignoring the "tails" of the cost sequences, which is consistent with a contractive structure. In this section, the tails of the cost sequences may not be small, and for this reason, the control is much more focused on affecting the long-term behavior of the state. For example, let $\alpha = 1$, and assume that the stage cost at all states is nonzero except for a cost-free and absorbing termination state. Then, a primary task of control under Assumption P (or Assumption N) is roughly to bring the state of the system to the termination state or to a region where the cost per stage is nearly zero as *quickly* as possible (as *late* as possible, respectively). Note the difference in control objective between Assumptions P and N. It accounts to some extent for some strikingly different results under the two assumptions.

In what follows in this section, we present results that characterize the optimal cost function J^*, as well as optimal stationary policies. We also discuss VI and give conditions under which it converges to the optimal cost function J^*. In the proofs we will often need to interchange expectation and limit in various relations. This interchange is valid under the assumptions of the following theorem.

Monotone Convergence Theorem: Let $P = (p_1, p_2, \ldots)$ be a probability distribution over $X = \{1, 2, \ldots\}$. Let $\{h_N\}$ be a sequence of extended real-valued functions on X such that for all $i \in X$ and $N = 1, 2, \ldots,$

$$0 \leq h_N(i) \leq h_{N+1}(i).$$

Let $h : X \mapsto [0, \infty]$ be the limit function $h(i) = \lim_{N\to\infty} h_N(i)$. Then

$$\lim_{N\to\infty}\sum_{i=1}^{\infty} p_i h_N(i) = \sum_{i=1}^{\infty} p_i \lim_{N\to\infty} h_N(i) = \sum_{i=1}^{\infty} p_i h(i).$$

Proof: We have

$$\sum_{i=1}^{\infty} p_i h_N(i) \le \sum_{i=1}^{\infty} p_i h(i).$$

By taking the limit, we obtain

$$\lim_{N\to\infty}\sum_{i=1}^{\infty} p_i h_N(i) \le \sum_{i=1}^{\infty} p_i h(i),$$

so there remains to prove the reverse inequality. For every integer $M \ge 1$, we have

$$\lim_{N\to\infty}\sum_{i=1}^{\infty} p_i h_N(i) \ge \lim_{N\to\infty}\sum_{i=1}^{M} p_i h_N(i) = \sum_{i=1}^{M} p_i h(i),$$

and by taking the limit as $M \to \infty$ the reverse inequality follows. **Q.E.D.**

Note that the conclusion of the proposition also holds if instead of monotonically increasing, the sequence $\{h_N\}$ is monotonically decreasing (we simply use the sequence $\{-h_N\}$ in the preceding proof).

4.1.1 Bellman's Equation

Similar to all the infinite horizon problems considered so far, the optimal cost function satisfies Bellman's equation.

Proposition 4.1.1: (Bellman's Equation) Under either Assumption P or N the optimal cost function J^* satisfies

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

or, equivalently,

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

Proof: For any admissible policy $\pi = \{\mu_0, \mu_1, \ldots\}$, consider the cost $J_\pi(x)$ corresponding to π when the initial state is x. We have

$$J_\pi(x) = \mathop{E}_{w}\big\{g\big(x,\mu_0(x),w\big) + V_\pi\big(f(x,\mu_0(x),w)\big)\big\}, \tag{4.4}$$

where, for all $x_1 \in X$,

$$V_\pi(x_1) = \lim_{N\to\infty} \mathop{E}_{\substack{w_k\\ k=1,2,\ldots}} \left\{ \sum_{k=1}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\}.$$

Thus, $V_\pi(x_1)$ is the cost from stage 1 to infinity using π when the initial state is x_1. We clearly have

$$V_\pi(x_1) \geq \alpha J^*(x_1), \qquad \text{for all } x_1 \in X.$$

Hence, from Eq. (4.4),

$$\begin{aligned} J_\pi(x) &\geq \mathop{E}_{w}\big\{g\big(x, \mu_0(x), w\big) + \alpha J^*\big(f(x, \mu_0(x), w)\big)\big\} \\ &\geq \min_{u\in U(x)} \mathop{E}_{w}\big\{g(x,u,w) + \alpha J^*\big(f(x,u,w)\big)\big\}. \end{aligned}$$

Taking the minimum over all admissible policies, we obtain

$$\begin{aligned} \min_\pi J_\pi(x) &= J^*(x) \\ &\geq \min_{u\in U(x)} \mathop{E}_{w}\big\{g(x,u,w) + \alpha J^*\big(f(x,u,w)\big)\big\} \\ &= (TJ^*)(x). \end{aligned}$$

Thus there remains to prove that the reverse inequality also holds. We prove this separately for Assumption N and for Assumption P.

Assume P. The following proof of $J^* \leq TJ^*$ under this assumption would be considerably simplified if we knew that there exists a μ such that $T_\mu J^* = TJ^*$. Since in general such a μ need not exist, we introduce a positive sequence $\{\epsilon_k\}$, and we choose an admissible policy $\pi = \{\mu_0, \mu_1, \ldots\}$ such that

$$(T_{\mu_k} J^*)(x) \leq (TJ^*)(x) + \epsilon_k, \qquad x \in X, \quad k = 0, 1, \ldots$$

Such a choice is possible because under P, we have $0 \leq J^*(x)$ for all x. By using the inequality $TJ^* \leq J^*$ shown earlier, we obtain

$$(T_{\mu_k} J^*)(x) \leq J^*(x) + \epsilon_k, \qquad x \in X, \quad k = 0, 1, \ldots$$

Applying $T_{\mu_{k-1}}$ to both sides of this relation, we have

$$\begin{aligned} (T_{\mu_{k-1}} T_{\mu_k} J^*)(x) &\leq (T_{\mu_{k-1}} J^*)(x) + \alpha\epsilon_k \\ &\leq (TJ^*)(x) + \epsilon_{k-1} + \alpha\epsilon_k \\ &\leq J^*(x) + \epsilon_{k-1} + \alpha\epsilon_k. \end{aligned}$$

Continuing this process, we obtain

$$(T_{\mu_0}T_{\mu_1}\cdots T_{\mu_k}J^*)(x) \le (TJ^*)(x) + \sum_{i=0}^{k} \alpha^i \epsilon_i.$$

By taking the limit as $k \to \infty$ and noting that

$$J^*(x) \le J_\pi(x) = \lim_{k\to\infty} (T_{\mu_0}T_{\mu_1}\cdots T_{\mu_k}J_0)(x) \le \lim_{k\to\infty} (T_{\mu_0}T_{\mu_1}\cdots T_{\mu_k}J^*)(x),$$

where J_0 is the zero function, it follows that

$$J^*(x) \le J_\pi(x) \le (TJ^*)(x) + \sum_{i=0}^{\infty} \alpha^i \epsilon_i, \qquad x \in X.$$

Since the sequence $\{\epsilon_k\}$ is arbitrary, we can take $\sum_{i=0}^{\infty} \alpha^i \epsilon_i$ as small as desired, and we obtain $J^*(x) \le (TJ^*)(x)$ for all $x \in X$. Combining this with the inequality $J^*(x) \ge (TJ^*)(x)$ shown earlier, the result follows (under Assumption P).

Assume N and let J_N be the optimal cost function for the corresponding N-stage problem

$$J_N(x_0) = \min_\pi E\left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\}.$$

We first show that

$$J^*(x) = \lim_{N\to\infty} J_N(x), \qquad x \in X. \tag{4.5}$$

Indeed, in view of Assumption N, we have $J^* \le J_N$ for all N, so

$$J^*(x) \le \lim_{N\to\infty} J_N(x), \qquad x \in X. \tag{4.6}$$

Also, for all $\pi = \{\mu_0, \mu_1, \ldots\}$, we have

$$E\left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\} \ge J_N(x_0),$$

and by taking the limit as $N \to \infty$,

$$J_\pi(x) \ge \lim_{N\to\infty} J_N(x), \qquad x \in X.$$

Taking the minimum over π, we obtain $J^*(x) \ge \lim_{N\to\infty} J_N(x)$, and combining this relation with Eq. (4.6), we obtain Eq. (4.5).

For every admissible μ, we have

$$T_\mu J_N \geq J_{N+1},$$

and by taking the limit as $N \to \infty$, and using the monotone convergence theorem and Eq. (4.5), we obtain

$$T_\mu J^* \geq J^*.$$

Taking the minimum over μ, we obtain $TJ^* \geq J^*$, which combined with the inequality $J^* \geq TJ^*$ shown earlier, proves the result under Assumption N. **Q.E.D.**

Similar to the case of the discounted and SSP problems of the preceding chapters, we also have a Bellman equation for each stationary policy.

Proposition 4.1.2: Let μ be a stationary policy. Then under Assumption P or N, we have

$$J_\mu(x) = \mathop{E}_{w}\bigl\{g\bigl(x,\mu(x),w\bigr) + \alpha J_\mu\bigl(f(x,\mu(x),w)\bigr)\bigr\}, \qquad x \in X,$$

or, equivalently,

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

Uniqueness of Solution of Bellman's Equation

Contrary to discounted problems with bounded cost per stage, the optimal cost function J^* under Assumption P or N need not be the unique solution of Bellman's equation. This is certainly true when $\alpha = 1$, since in this case if $J(\cdot)$ is any solution, then for any scalar r, $J(\cdot)+r$ is also a solution. The following is an example where $\alpha < 1$.

Example 4.1.2

Let $X = [0,\infty)$ (or $X = (-\infty, 0]$) and

$$g(x,u,w) = 0, \qquad f(x,u,w) = \frac{x}{\alpha}.$$

Then for every β, the function J given by $J(x) = \beta x$ for all $x \in X$, is a solution of Bellman's equation, so there are infinitely many solutions. Note, however, that there is a unique solution within the class of bounded functions, the zero function $J_0(x) \equiv 0$, which is the optimal cost function for this problem. More generally, it can be shown by using the following Prop. 4.1.3 that if $\alpha < 1$ and there exists a bounded function that is a solution of Bellman's equation, then that function must be equal to the optimal cost function J^* (see Exercise 4.7).

Later in this chapter we will also encounter finite-state deterministic shortest path examples where $\alpha = 1$ and the set of solutions of Bellman's equations is infinite (see Example 4.2.1 in Section 4.2), and a linear-quadratic infinite horizon problem where Bellman's equation has exactly two solutions within the class of quadratic functions (see Example 4.2.2 in Section 4.2, and also the discussion on the Riccati equation of Section 4.1 in Vol. I). The optimal cost function J^*, however, has the property that it is the smallest (under Assumption P) or largest (under Assumption N) fixed point of T in the sense described in the following proposition.

Proposition 4.1.3:

(a) Under Assumption P, if $\tilde{J} : X \mapsto (-\infty, \infty]$ satisfies $\tilde{J} \geq T\tilde{J}$ and either $\tilde{J}$ is bounded below and $\alpha < 1$, or $\tilde{J} \geq 0$, then $\tilde{J} \geq J^*$.

(b) Under Assumption N, if $\tilde{J} : X \mapsto [-\infty, \infty)$ satisfies $\tilde{J} \leq T\tilde{J}$ and either $\tilde{J}$ is bounded above and $\alpha < 1$, or $\tilde{J} \leq 0$, then $\tilde{J} \leq J^*$.

Proof: (a) Under Assumption P, let r be a scalar such that $\tilde{J}(x) + r \geq 0$ for all $x \in X$ and if $\alpha \geq 1$ let $r = 0$. For any sequence $\{\epsilon_k\}$ with $\epsilon_k > 0$, let $\tilde{\pi} = \{\tilde{\mu}_0, \tilde{\mu}_1, \ldots\}$ be an admissible policy such that, for every $x \in X$ and k,

$$\mathop{E}_{w}\big\{g\big(x, \mu_k(x), w\big) + \alpha\tilde{J}\big(f\big(x, \mu_k(x), w\big)\big)\big\} \leq (T\tilde{J})(x) + \epsilon_k. \tag{4.7}$$

Such a policy exists since $(T\tilde{J})(x) > -\infty$ for all $x \in X$. We have for any initial state $x_0 \in X$,

$$\begin{aligned}
J^*(x_0) &= \min_{\pi} \lim_{N\to\infty} E\left\{\sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\
&\leq \min_{\pi} \liminf_{N\to\infty} E\left\{\alpha^N\big(\tilde{J}(x_N) + r\big) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\
&\leq \liminf_{N\to\infty} E\left\{\alpha^N\big(\tilde{J}(x_N) + r\big) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \tilde{\mu}_k(x_k), w_k\big)\right\}.
\end{aligned}$$

Using Eq. (4.7) and the assumption $\tilde{J} \geq T\tilde{J}$, we obtain

$$E\left\{\alpha^N \tilde{J}(x_N) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \tilde{\mu}(x_k), w_k\big)\right\}$$

$$
\begin{aligned}
&= E\left\{\alpha^N \tilde{J}\big(f\big(x_{N-1}, \tilde{\mu}_{N-1}(x_{N-1}), w_{N-1}\big)\big) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \tilde{\mu}_k(x_k), w_k\big)\right\} \\
&\le E\left\{\alpha^{N-1} \tilde{J}(x_{N-1}) + \sum_{k=0}^{N-2} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} + \alpha^{N-1}\epsilon_{N-1} \\
&\le E\left\{\alpha^{N-2} \tilde{J}(x_{N-2}) + \sum_{k=0}^{N-3} \alpha^k g\big(x_k, \tilde{\mu}_k(x_k), w_k\big)\right\} + \alpha^{N-2}\epsilon_{N-2} \\
&\qquad + \alpha^{N-1}\epsilon_{N-1} \\
&\qquad\qquad\vdots \\
&\le \tilde{J}(x_0) + \sum_{k=0}^{N-1} \alpha^k \epsilon_k .
\end{aligned}
$$

Combining these inequalities, we obtain

$$
J^*(x_0) \le \tilde{J}(x_0) + \lim_{N\to\infty} \left(\alpha^N r + \sum_{k=0}^{N-1} \alpha^k \epsilon_k \right).
$$

Since $\{\epsilon_k\}$ is an arbitrary positive sequence, we may select $\{\epsilon_k\}$ so that $\lim_{N\to\infty} \sum_{k=0}^{N-1} \alpha^k \epsilon_k$ is arbitrarily close to zero, and the result follows.

(b) Under Assumption N, let r be a scalar such that $\tilde{J}(x) + r \le 0$ for all $x \in X$, and if $\alpha \ge 1$, let $r = 0$. We have for every initial state $x_0 \in X$,

$$
\begin{aligned}
J^*(x_0) &= \min_{\pi} \lim_{N\to\infty} E\left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\} \\
&\ge \min_{\pi} \limsup_{N\to\infty} E\left\{ \alpha^N \big(\tilde{J}(x_N) + r\big) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\} \\
&\ge \limsup_{N\to\infty} \min_{\pi} E\left\{ \alpha^N \big(\tilde{J}(x_N) + r\big) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big) \right\},
\end{aligned} \tag{4.8}
$$

where the last inequality follows from the fact that for any sequence $\{h_N(\xi)\}$ of functions of a parameter ξ we have

$$
\min_{\xi} \limsup_{N\to\infty} h_N(\xi) \ge \limsup_{N\to\infty} \min_{\xi} h_N(\xi).
$$

This inequality follows by writing

$$
h_N(\xi) \ge \min_{\xi} h_N(\xi)
$$

and by subsequently taking the lim sup of both sides and the minimum over ξ of the left-hand side.

Now we have, by using the assumption $\tilde{J} \le T\tilde{J}$,

$$\begin{aligned}
\min_\pi E&\left\{\alpha^N \tilde{J}(x_N) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\
&= \min_\pi E\left\{\alpha^{N-1} \min_{u_{N-1}\in U(x_{N-1})} \mathop{E}_{w_{N-1}} \big\{g(x_{N-1}, u_{N-1}, w_{N-1}) \right. \\
&\qquad\qquad + \alpha\tilde{J}\big(f(x_{N-1}, u_{N-1}, w_{N-1})\big)\big\} \\
&\qquad\left. + \sum_{k=0}^{N-2} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\
&\ge \min_\pi E\left\{\alpha^{N-1}\tilde{J}(x_{N-1}) + \sum_{k=0}^{N-2} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\} \\
&\qquad \vdots \\
&\ge \tilde{J}(x_0).
\end{aligned}$$

Using this relation in Eq. (4.8), we obtain

$$J^*(x_0) \ge \tilde{J}(x_0) + \lim_{N\to\infty} \alpha^N r = \tilde{J}(x_0).$$

Q.E.D.

As before, we have a version of Prop. 4.1.3 for stationary policies.

Proposition 4.1.4: Let μ be a stationary policy.

(a) Under Assumption P, if $\tilde{J} : X \mapsto (-\infty, \infty]$ satisfies $\tilde{J} \ge T_\mu \tilde{J}$ and either $\tilde{J}$ is bounded below and $\alpha < 1$, or $\tilde{J} \ge 0$, then $\tilde{J} \ge J_\mu$.

(b) Under Assumption N, if $\tilde{J} : X \mapsto [-\infty, \infty)$ satisfies $\tilde{J} \le T_\mu \tilde{J}$ and either $\tilde{J}$ is bounded above and $\alpha < 1$, or $\tilde{J} \le 0$, then $\tilde{J} \le J_\mu$.

4.1.2 Optimality Conditions

Under Assumption P, we have the same optimality condition as for discounted problems with bounded cost per stage.

Proposition 4.1.5: (Necessary and Sufficient Condition for Optimality under P) Let Assumption P hold. A stationary policy μ is optimal if and only if $TJ^* = T_\mu J^*$.

Proof: If $TJ^* = T_\mu J^*$, Bellman's equation ($J^* = TJ^*$) implies that $J^* = T_\mu J^*$. From Prop. 4.1.4(a) we then obtain $J^* \geq J_\mu$, showing that μ is optimal. Conversely, if $J^* = J_\mu$, we have using Prop. 4.1.2, $TJ^* = J^* = J_\mu = T_\mu J_\mu = T_\mu J^*$. **Q.E.D.**

Note that when $U(x)$ is a finite set for every $x \in X$, the above proposition implies the existence of an optimal stationary policy under Assumption P. This may not be true under Assumption N (see Exercise 4.3). Instead, we have a different characterization of an optimal stationary policy.

Proposition 4.1.6: (Necessary and Sufficient Condition for Optimality under N) Let Assumption N hold. A stationary policy μ is optimal if and only if $TJ_\mu = T_\mu J_\mu$.

Proof: If $TJ_\mu = T_\mu J_\mu$, then from Prop. 4.1.2 we have $J_\mu = T_\mu J_\mu$, so that J_μ is a fixed point of T. Then by Prop. 4.1.3, we have $J_\mu \leq J^*$, which implies that μ is optimal. Conversely, if $J_\mu = J^*$, then

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

Q.E.D.

The following deterministic shortest path example illustrates the limitations of the preceding two propositions. It shows that under Assumption P, we may have $TJ_\mu = T_\mu J_\mu = J_\mu$, while μ is not optimal (which incidentally shows that VI may get stuck at the cost function of a suboptimal policy). Moreover, under Assumption N, we may have $TJ^* = T_\mu J^*$ while μ is not optimal (which incidentally shows that when starting PI with an optimal policy, the next policy may be suboptimal).

Example 4.1.3

Let $X = \{1, t\}$, where t is a cost-free and absorbing state (cf. Fig. 4.1.1). At state 1 there are two choices: u which leads to t at cost b, and u' that self-transitions to 1 at cost 0. We have for all $J = \big(J(1), J(t)\big)$,

$$(TJ)(1) = \min\big\{J(1),\, b + J(t)\big\}, \qquad (TJ)(t) = J(t).$$

Bellman's equation takes the form

$$J(1) = \min\big\{J(1),\, b + J(t)\big\}, \qquad J(t) = J(t),$$

and the set of its solutions is shown in Fig. 4.1.1. Consider two cases:

(a) $b > 0$ in which case Assumption P holds. Then applying u' at state 1 is optimal and we have $J^*(1) = J^*(t) = 0$. However, for the suboptimal

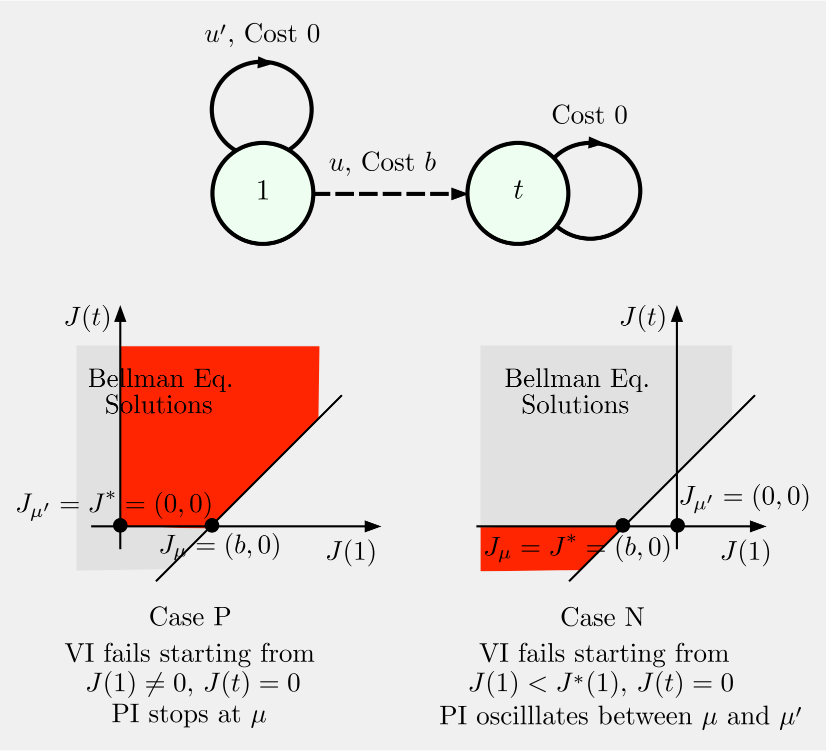


Figure 4.1.1. A deterministic two-state problem with termination node t (cf. Example 4.1.3). At state 1 there are two choices; u which leads to t at cost b (policy μ), and u' that self-transitions to 1 at cost 0 (policy μ'). The figure shows the set of solutions of Bellman's equation for $b > 0$ (case P) and $b < 0$ (case N).

In case P, J^* is the smallest solution within the positive orthant [cf. Prop. 4.1.3(a)]. In case N, J^* is the largest solution within the negative orthant [cf. Prop. 4.1.3(b)].

policy μ that applies $\mu(1) = u$, we have $J_\mu(1) = b$, $J_\mu(t) = 0$, which satisfy $TJ_\mu = T_\mu J_\mu$.

(b) $b < 0$ in which case Assumption N holds. Then applying u at state 1 is optimal and we have $J^*(1) = b$, $J^*(t) = 0$. However, for the suboptimal policy μ' that applies $\mu'(1) = u'$, we have $TJ^* = T_{\mu'}J^*$.

4.1.3 Computational Methods

We now turn to computational methods such as VI and PI. As Example 4.1.3 suggests, even for very simple problems, these methods may have serious difficulties. We will first discuss the validity of the methods under just Assumptions P or N. In subsequent sections, we will introduce additional conditions and special classes of problems for which VI and PI can be applied more reliably.

We first consider the VI algorithm, which generates a sequence of functions T^kJ, $k = 1, 2, \ldots$, starting from some initial function J. We will see that, contrary to the case of the discounted problems of Chapter 2 and the SSP problems of Chapter 3, convergence of $\{T^kJ\}$ to J^* depends on the initial condition J, and may require additional assumptions, beyond P or N. We have the following proposition.

Proposition 4.1.7: (Convergence of VI)

(a) Let Assumption P hold and assume that the control constraint set $U(x)$ is finite for all $x \in X$. Then if $J : X \mapsto \Re$ is any bounded function and $\alpha < 1$, or otherwise if $0 \le J \le J^*$, we have

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

(b) Let Assumption N hold. Then if $J : X \mapsto \Re$ is any bounded function and $\alpha < 1$, or otherwise if $J^* \le J \le 0$, we have

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

Proof: (a) We will first assume that $\alpha = 1$ and show convergence of VI starting from the zero function, which we will denote by J_0. Under Assumption P, we have

$$J_0 \le TJ_0 \le \cdots \le T^kJ_0 \le \cdots \le J^*,$$

where the inequality on the right follows from the relation $J_0 \le J^*$, which in view of the monotonicity of T and Bellman's equation ($J^* = TJ^*$), implies that $T^kJ_0 \le T^kJ^* = J^*$. Thus $\{T^kJ\}$ converges to some $J_\infty \le J^*$.

We will now show that J_∞ is a fixed point of T, so from Prop. 4.1.3(a), it will follow that $J^* \le J_\infty$, implying that $J_\infty = J^*$. Indeed, by applying T to the relation $T^kJ_0 \le J_\infty \le J^*$, we obtain

$$(T^{k+1}J_0)(x) = \min_{u\in U(x)} \mathop{E}_{w}\Big\{g(x,u,w) + \alpha(T^kJ_0)\big(f(x,u,w)\big)\Big\} \le (TJ_\infty)(x), \tag{4.9}$$

and by taking the limit as $k \to \infty$, it follows that

$$J_\infty \le TJ_\infty.$$

Assume to arrive at a contradiction that there exists a state $\tilde{x} \in X$ such that

$$J_\infty(\tilde{x}) < (TJ_\infty)(\tilde{x}). \tag{4.10}$$

Let u_k minimize in Eq. (4.9) when $x = \tilde{x}$. Since $U(\tilde{x})$ is finite, there must exist some $\tilde{u} \in U(\tilde{x})$ such that $u_k = \tilde{u}$ for all k in some infinite subset K of the positive integers. By Eq. (4.9) we have for all $k \in K$

$$(T^{k+1}J_0)(\tilde{x}) = \mathop{E}_{w}\Big\{g(\tilde{x},\tilde{u},w) + \alpha(T^kJ_0)\big(f(\tilde{x},\tilde{u},w)\big)\Big\} \le (TJ_\infty)(\tilde{x}).$$

Taking the limit as $k \to \infty$, $k \in K$, we obtain

$$\begin{aligned} J_\infty(\tilde{x}) &= \mathop{E}_{w}\Big\{g\big(\tilde{x},\tilde{u},w\big) + \alpha J_\infty\big(f(\tilde{x},\tilde{u},w)\big)\Big\} \\ &\ge (TJ_\infty)(\tilde{x}) \\ &= \min_{u \in U(\tilde{x})} \mathop{E}_{w}\Big\{g(\tilde{x},u,w) + \alpha J_\infty\big(f(\tilde{x},u,w)\big)\Big\}. \end{aligned}$$

This contradicts Eq. (4.10), so we must have $J_\infty = TJ_\infty$, and as noted earlier, this implies that T^kJ converges to J^* starting from the initial condition $J = J_0$.

In the case where $\alpha = 1$ and the initial condition J satisfies $J_0 \le J \le J^*$, we have using the monotonicity of T,

$$T^kJ_0 \le T^kJ \le J^*, \qquad k = 0, 1, \ldots$$

Since T^kJ_0 converges to J^*, so does T^kJ.

Finally, in the case where $\alpha < 1$ and J is bounded, let r be a scalar such that

$$J_0 - re \le J \le J_0 + re.$$

Applying T^k to this relation, we obtain

$$T^kJ_0 - \alpha^k re \le T^kJ \le T^kJ_0 + \alpha^k re.$$

Since T^kJ_0 converges to J^*, as shown earlier, this relation implies that T^kJ converges also to J^*.

(b) It was shown earlier [cf. Eq. (4.5)] that under Assumption N, we have

$$\lim_{k\to\infty} (T^kJ_0)(x) = J^*(x), \qquad x \in X.$$

The proof from this point is identical to that for part (a). **Q.E.D.**

We will now strengthen part (a) of the preceding proposition, by replacing the finiteness assumption on the control constraint set with a weaker compactness assumption. Let us recall that a subset X of a metric space is said to be *compact* if every sequence $\{x_k\}$ with $x_k \in X$ contains a subsequence $\{x_k\}_{k\in K}$ that converges to a point $x \in X$. Equivalently, X is compact if and only if it is closed and bounded. The empty set is (trivially) considered compact. Given any collection of compact sets, their intersection is a compact set (possibly empty). For a sequence of nonempty compact sets $X_1, X_2 \ldots, X_k, \ldots$ such that

$$X_1 \supset X_2 \supset \cdots \supset X_k \supset X_{k+1} \supset \cdots$$

the intersection $\cap_{k=1}^{\infty} X_k$ is both nonempty and compact. In view of this fact, it follows that if $f : \Re^n \mapsto [-\infty, \infty]$ is a function such that the set

$$F_\lambda = \bigl\{ x \in \Re^n \mid f(x) \le \lambda \bigr\} \tag{4.11}$$

is compact for every $\lambda \in R$, then there exists a vector x^* minimizing f; i.e., there exists an $x^* \in \Re^n$ such that

$$f(x^*) = \min_{x \in \Re^n} f(x).$$

To see this, take a decreasing sequence $\{\lambda_k\}$ such that $\lambda_k \downarrow \min_{x \in \Re^n} f(x)$. If $\min_{x \in \Re^n} f(x) < \infty$, such a sequence exists and the sets

$$F_{\lambda_k} = \{ x \in \Re^n \mid f(x) \le \lambda_k \}$$

are nonempty and compact. Furthermore, $F_{\lambda_k} \supset F_{\lambda_{k+1}}$ for all k, and hence the intersection $\cap_{k=1}^{\infty} F_{\lambda_k}$ is also nonempty and compact. Let x^* be any vector in $\cap_{k=1}^{\infty} F_{\lambda_k}$. Then

$$f(x^*) \le \lambda_k, \qquad k = 1, 2, \ldots,$$

and taking the limit as $k \to \infty$, we obtain $f(x^*) \le \min_{x \in \Re^n} f(x)$, proving that x^* minimizes $f(x)$. The most common case where we can guarantee that the set F_λ of Eq. (4.11) is compact for all λ is when f is continuous and $f(x) \to \infty$ as $\|x\| \to \infty$.

Proposition 4.1.8: (Convergence of VI Under P) Let Assumption P hold and let J_0 be the identically zero function. Assume that the sets

$$U_k(x, \lambda) = \left\{ u \in U(x) \,\middle|\, \mathop{E}_{w} \bigl\{ g(x, u, w) + \alpha (T^k J_0)\bigl(f(x, u, w)\bigr) \bigr\} \le \lambda \right\} \tag{4.12}$$

are compact subsets of a metric space for every $x \in X$, $\lambda \in \Re$, and for all k greater than some integer $\overline{k}$. Then the conclusion of Prop. 4.1.7(a) holds. Furthermore, there exists a stationary optimal policy.

Proof: We follow the line of proof and the notation of Prop. 4.1.7(a). We have $J_\infty \le T J_\infty$. Suppose that there existed a state $\tilde{x} \in X$ such that

$$J_\infty(\tilde{x}) < (T J_\infty)(\tilde{x}). \tag{4.13}$$

Clearly, we must have $J_\infty(\tilde{x}) < \infty$. Consider the sets

$$\begin{aligned} &U_k\big(\tilde{x}, J_\infty(\tilde{x})\big) \\ &\qquad = \Big\{ u \in U(\tilde{x}) \;\Big|\; \mathop{E}_{w}\big\{g(\tilde{x}, u, w) + \alpha (T^k J_0)\big(f(\tilde{x}, u, w)\big)\big\} \le J_\infty(\tilde{x}) \Big\} \end{aligned}$$

for $k \ge \overline{k}$. Let also u_k be a point attaining the minimum in

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

i.e., u_k is such that

$$(T^{k+1} J_0)(\tilde{x}) = \mathop{E}_{w}\big\{g(\tilde{x}, u_k, w) + \alpha (T^k J_0)\big(f(\tilde{x}, u_k, w)\big)\big\}.$$

Such minimizing points u_k exist by our compactness assumption. For every $k \ge \overline{k}$, consider the sequence $\{u_i\}_{i=k}^\infty$. Since $T^k J_0 \le T^{k+1} J_0 \le \cdots \le J_\infty$, it follows that

$$\begin{aligned} &\mathop{E}_{w}\big\{g(\tilde{x}, u_i, w) + \alpha (T^k J_0)\big(f(\tilde{x}, u_i, w)\big)\big\} \\ &\qquad\qquad \le \mathop{E}_{w}\big\{g(\tilde{x}, u_i, w) + \alpha (T^i J_0)\big(f(\tilde{x}, u_i, w)\big)\big\} \\ &\qquad\qquad \le J_\infty(\tilde{x}), \qquad i \ge k. \end{aligned}$$

Therefore $\{u_i\}_{i=k}^\infty \subset U_k\big(\tilde{x}, J_\infty(\tilde{x})\big)$, and since $U_k\big(\tilde{x}, J_\infty(\tilde{x})\big)$ is compact, all the limit points of $\{u_i\}_{i=k}^\infty$ belong to $U_k\big(\tilde{x}, J_\infty(\tilde{x})\big)$ and at least one such limit point exists. Hence the same is true of the limit points of the whole sequence $\{u_i\}_{i=\overline{k}}^\infty$. It follows that if $\tilde{u}$ is a limit point of $\{u_i\}_{i=\overline{k}}^\infty$ then

$$\tilde{u} \in \cap_{k=\overline{k}}^\infty U_k\big(\tilde{x}, J_\infty(\tilde{x})\big).$$

This implies by Eq. (4.12) that for all $k \ge \overline{k}$

$$J_\infty(\tilde{x}) \ge \mathop{E}_{w}\big\{g(\tilde{x}, \tilde{u}, w) + \alpha (T^k J_0)\big(f(\tilde{x}, \tilde{u}, w)\big)\big\} \ge (T^{k+1} J_0)(\tilde{x}). \qquad (4.14)$$

Taking the limit as $k \to \infty$, we obtain

$$J_\infty(\tilde{x}) = \mathop{E}_{w}\big\{g(\tilde{x}, \tilde{u}, w) + \alpha J_\infty\big(f(\tilde{x}, \tilde{u}, w)\big)\big\}.$$

Since the right-hand side is greater than or equal to $(TJ_\infty)(\tilde{x})$, Eq. (4.13) is contradicted. Hence $J_\infty = TJ_\infty$ and the result follows similar to the proof of Prop. 4.1.7(a).

To show that there exists an optimal stationary policy, observe that Eq. (4.14) and the fact $J_\infty = J^*$ imply that $\tilde{u}$ attains the minimum in

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

for any state $\tilde{x} \in X$ with $J^*(\tilde{x}) < \infty$. For states $\tilde{x} \in X$ such that $J^*(\tilde{x}) = \infty$, every $u \in U(\tilde{x})$ attains the preceding minimum. Hence by Prop. 4.1.5(a) an optimal stationary policy exists. **Q.E.D.**

The reader may verify by inspection of the preceding proof that if $\mu_k(\tilde{x})$, $k = 0, 1, \ldots$, attains the minimum in the relation

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

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 *VI under the assumptions of either Prop. 4.1.7(a) or Prop. 4.1.8 yields in the limit not only the optimal cost function J^* but also an optimal stationary policy*.

Example 4.2.3, to be given later, shows that VI may not converge to J^* starting from the identically zero function, in the absence of the compactness hypothesis of the preceding proposition (see also Exercise 4.1). Generally, under Assumption P, we are not guaranteed that $T^k J$ converges to J^* starting from initial conditions $J \geq J^*$, even if $\alpha < 1$. For the case where $\alpha = 1$, case (a) of Example 4.1.3 shows that convergence of VI starting from $J \geq J^*$ is not guaranteed even for finite-state/finite-control problems. However, if we restrict the initial condition for VI within a suitable class of functions, convergence to J^* may be obtained under certain conditions (see Sections 4.2-4.4). We note here that experience with various types of problems, including deterministic shortest path problems, suggests that generally if VI works at all, it works faster starting from “large” initial conditions (those satisfying $J \geq J^*$) than starting from “small” initial conditions (those satisfying $J \leq J^*$).

Let us also the following condition for VI convergence from above, first derived and proved in [YuB13], and given in Appendix B as Prop. B.4.3.

Proposition 4.1.9: (Convergence of VI from Above Under P) Let Assumption P hold. If a function $J : X \mapsto [0, \infty]$ satisfies

$$J^* \leq J \leq cJ^* \qquad \text{for some } c > 0, \tag{4.15}$$

then we have $T^k J \to J^*$.

The condition (4.15) highlights a requirement for the reliable implementation of VI: it is important to know the sets

$$\big\{x \in X \mid J^*(x) = 0\big\}, \qquad \big\{x \in X \mid J^*(x) = \infty\big\}$$

in order to obtain a suitable initial condition. For finite-state problems, the set of x such that $J^*(x) = 0$ can be computed in polynomial time as will be shown in Section 4.1.4, which also provides a method for dealing with cases where the set X_∞ is nonempty.

Asynchronous Value Iteration

The concepts of asynchronous VI that we developed in Section 2.6.1 apply also under the Assumptions P and N of this section. Under Assumption P, if J^* is real-valued, we may apply Prop. 2.6.1 with the sets $S(k)$ defined by

$$S(k) = \{J \mid T^k J_0 \leq J \leq J^*\}, \qquad k = 0, 1, \ldots,$$

where J_0 is the zero function on X. Assuming that $T^k J_0 \to J^*$ (cf. Props. 4.1.7-4.1.8), 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 $R(X)$ is replaced by the set of all nonnegative extended real-valued functions.

Under Assumption N similar conclusions hold for the asynchronous version of VI that starts with a function J satisfying $J^* \leq J \leq 0$. Asynchronous pointwise convergence to J^* can be shown, based on an extension of the asynchronous convergence theorem (Prop. 2.6.1), where $R(X)$ is replaced by the set of all extended real-valued functions $J \leq 0$.

Policy Iteration

Let us now discuss PI. Unfortunately, PI is not a valid algorithm under Assumption P in the absence of further conditions. This is true despite the fact that the policy improvement property

$$J_{\overline{\mu}}(x) \leq J_\mu(x), \qquad \forall\, x \in X, \tag{4.16}$$

holds for any policies μ and $\overline{\mu}$ such that $T_{\overline{\mu}} J_\mu = T J_\mu$. To see this, note that

$$T_{\overline{\mu}} J_\mu = T J_\mu \leq T_\mu J_\mu = J_\mu,$$

from which we obtain $\lim_{N\to\infty} T_{\overline{\mu}}^N J_\mu \leq J_\mu$. Since $J_{\overline{\mu}} = \lim_{N\to\infty} T_{\overline{\mu}}^N J_0$ and $J_0 \leq J_\mu$, we obtain $J_{\overline{\mu}} \leq J_\mu$.

However, the inequality $J_{\overline{\mu}} \leq J_\mu$ by itself is not sufficient to guarantee the validity of PI. In particular, it is not clear that strict inequality holds in Eq. (4.16) for at least one state $x \in X$ when μ is not optimal. This occurs in the shortest path problem of Example 4.1.3 for the case $b > 0$, where it can be verified that PI can stop with the suboptimal policy that moves from node 1 to t. The difficulty here is that the equality $J_\mu = T J_\mu$ does not imply that μ is optimal, and additional conditions are needed to guarantee the validity of PI. However, for special cases such conditions can be verified, and various forms of PI may be reliably used for some important classes of problems (see Sections 4.2-4.4).

Note that without conditions that guarantee that PI will not stop with a suboptimal policy, also has an implication in the context of a rollout algorithm. It implies that under Assumption P, the rollout algorithm may not strictly improve the cost of a suboptimal base policy. The reason is that rollout can be viewed as a single step of PI starting from the base policy.

Under Assumption N, the policy improvement property (4.16) may fail. In particular, under Assumption N, 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 [cf. case (b) of Example 4.1.3]. As a result, there may be an oscillation between nonoptimal policies even when the state and control spaces are finite.

On the other hand, under Assumption N, optimistic PI (cf. Section 2.3.3) has much better convergence properties, because it embodies the mechanism of VI, which is convergent to J^* as we saw in Prop. 4.1.7(b). Indeed, let us consider an optimistic PI algorithm that generates a sequence $\{J_k, \mu^k\}$ according to †

$$T_{\mu^k} J_k = TJ_k, \qquad J_{k+1} = T_{\mu^k}^{m_k} J_k, \tag{4.17}$$

where m_k is a positive integer. We assume that the algorithm starts with a function J_0 that satisfies $0 \geq J_0 \geq TJ_0$ and $J_0 \geq J^*$. For example, we may choose J_0 to be the identically zero function. We have the following proposition.

Proposition 4.1.10: Let Assumption N hold and let $\{J_k, \mu^k\}$ be a sequence generated by the optimistic PI algorithm (4.17), assuming that $0 \geq J_0 \geq J^*$ and $J_0 \geq TJ_0$. Then $J_k \downarrow J^*$.

Proof: We have

$$J_0 \geq T_{\mu^0} J_0 \geq T_{\mu^0}^{m_0} J_0 = J_1 \geq T_{\mu^0}^{m_0+1} J_0 = T_{\mu^0} J_1 \geq TJ_1 = T_{\mu^1} J_1 \geq \cdots \geq J_2,$$

where the first, second, and third inequalities hold because the assumption $J_0 \geq TJ_0 = T_{\mu^0} J_0$ implies that $T_{\mu^0}^{\ell} J_0 \geq T_{\mu^0}^{\ell+1} J_0$ for all $\ell \geq 0$. Continuing similarly we obtain

$$J_k \geq TJ_k \geq J_{k+1}, \qquad \forall\ k \geq 0. \tag{4.18}$$

† 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.

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

$$J_{k+1} = T_{\mu^k}^{m_k} J_k \geq T^{m_k} J_k \geq T^{m_k} J^* = J^*, \tag{4.19}$$

where the last equality follows from Bellman's equation, $TJ^* = J^*$, thus completing the induction. Thus, by combining the preceding two relations,

$$J_k \geq TJ_k \geq J_{k+1} \geq J^*, \qquad \forall\ k \geq 0. \tag{4.20}$$

We will now show by induction that

$$T^k J_0 \geq J_k \geq J^*, \qquad \forall\ k \geq 0. \tag{4.21}$$

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. (4.20),

$$T^{k+1} J_0 \geq TJ_k \geq J_{k+1} \geq J^*,$$

thus completing the induction. Since $T^k J_0 \to J^*$ [cf. Prop. 4.1.7(b)], from Eq. (4.21), we obtain $J_k \downarrow J^*$. **Q.E.D.**

Note that in the preceding proposition, we have $J_k \to J^*$, even if $J^*(x) = -\infty$ for some x; for an example, see the blackmailer's problem of Example 3.2.1. The reason why optimistic PI can deal with the absence of an optimal policy is that it acts as a form of VI, which is convergent to J^* under Assumption N [cf. Prop. 4.1.7(b)]. However, optimistic PI tends to be more computationally efficient than VI, as experience has shown, so it is usually preferable in practice.

Linear Programming

Finally, let us note that it is possible to devise a computational method based on mathematical programming when X and U are finite sets by making use of Prop. 4.1.3. Under Assumption N and $\alpha = 1$, $J^*(1), \ldots, J^*(n)$ solve the following linear programming problem (in $z_1, \ldots, z_n$):

$$\begin{aligned} &\text{maximize} \quad \sum_{i=1}^n z_i \\ &\text{subject to} \quad z_i \leq \sum_{j=1}^n p_{ij}(u)\big(g(i,u,j) + z_j\big), \qquad i = 1, \ldots, n, \quad u \in U(i). \end{aligned}$$

When $\alpha = 1$ and Assumption P holds, the corresponding optimization problem takes the form

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^n z_i \\ &\text{subject to} \quad z_i \geq \min_{u \in U(i)} \left[\sum_{j=1}^n p_{ij}(u)\big(g(i,u,j) + z_j\big) \right], \qquad i = 1, \ldots, n, \end{aligned}$$

but unfortunately this problem is not linear or even convex.

4.1.4 Finite-State Positive Cost Models: Equivalence to a Stochastic Shortest Path Problem

We will now consider finite-state, finite-control problems under Assumption P. We will show that such problems have a special property that greatly facilitates their solution: *they can be transformed to equivalent SSP problems for which the powerful analysis and computational methodology of Chapter 3 can be applied.*† In particular, we will define an SSP problem, which will be shown to be "equivalent" to the given problem. In this new SSP problem, all the states x in the set

$$X_0 = \big\{x \in X \mid J^*(x) = 0\big\},$$

including the termination state t, are merged into a new termination state $\bar{t}$. We assume that the set X_0 is nonempty, and this does not involve any loss of generality, since if needed we may include in X an artificial cost-free and absorbing termination state that is not reachable from any of the other states with a feasible transition. To facilitate the exposition, we will also assume without essential loss of generality that $X_0 \neq X$, or equivalently that the set X_+ given by

$$X_+ = \big\{x \in X \mid J^*(x) > 0\big\},$$

is nonempty.

Note that from the Bellman equation $J^* = TJ^*$, and the finiteness of $U(x)$, we obtain the following useful characterization of X_0:

$$\begin{aligned} x \in X_0 \quad &\text{if and only if there exists } u \in U(x) \text{ such that} \\ &\qquad g(x,u) = 0 \text{ and } p_{xy}(u) = 0 \text{ for all } y \notin X_0. \end{aligned} \tag{4.22}$$

In words, there exists a policy under which X_0 is an "absorbing" set of states where the one-stage cost is equal to 0. Algorithms for constructing X_0 will be given shortly. We introduce a new SSP problem as follows.

† Actually, the finiteness of the control space is not essential, and can be replaced by the compactness assumption that was briefly discussed in Section 3.2 for SSP. The analysis of finite-state positive cost SSP problems with infinite control space and a compactness assumption will be given as a special case of the affine monotonic analysis of Section 4.5. The papers [BeT91] and [BeY15] provide the corresponding analysis for finite-state SSP problems with both positive and negative costs per stage.

Definition of Equivalent SSP Problem:

State space: $\overline{X} = X_+ \cup \{\overline{t}\}$, where $\overline{t}$ is a cost-free and absorbing termination state.

Controls and one-stage costs: For $x \in X_+$, we have $\overline{U}(x) = U(x)$ and $\overline{g}(x,u) = g(x,u)$, for all $u \in \overline{U}(x)$.

Transition probabilities: For $x \in X_+$ and $u \in \overline{U}(x)$, we have

$$\overline{p}_{xy}(u) = \begin{cases} p_{xy}(u) & \text{if } y \in X_+, \\ \sum_{z \in X_0} p_{xz}(u) & \text{if } y = \overline{t}. \end{cases}$$

The optimal cost vector for the equivalent SSP problem is denoted by $\bar{J}$, and is the smallest nonnegative solution of the corresponding Bellman equation $J = \overline{T}J$, where

$$\begin{aligned} (\overline{T}J)(x) &\overset{\text{def}}{=} \min_{u \in \overline{U}(x)} \left[\overline{g}(x,u) + \sum_{y \in X_+} \overline{p}_{xy}(u) J(y) \right] \\ &= \min_{u \in U(x)} \left[g(x,u) + \sum_{y \in X_+} p_{xy}(u) J(y) \right], \qquad x \in X_+, \end{aligned} \tag{4.23}$$

[cf. Prop. 4.1.3(a)].

We will now clarify the relation of the equivalent SSP problem with the given problem (also referred to as the "original" problem). The key fact for our purposes, given in the following proposition, is that $\bar{J}$ coincides with J^* on the set X_+. Moreover if J^* is real-valued, then the equivalent SSP problem satisfies Assumptions 3.1.1 and 3.1.2 of Section 3.1 (we will refer to these as the "standard SSP conditions" in what follows). As a result we may transfer the available analytical results from the equivalent SSP problem to the original problem. We may also apply the VI and PI methods discussed in Chapter 3 to the equivalent SSP problem, after first obtaining the set X_0, in order to compute the solution of the original problem.

Proposition 4.1.11: Assume that X and U are finite sets, and that $g(x,u) \geq 0$ for all $x \in X$ and $u \in U(x)$. Then:

(a) $J^*(x) = \bar{J}(x)$ for all $x \in X_+$.

(b) A policy μ^* is optimal for the original problem if and only if

$$\mu^*(x) = \overline{\mu}(x), \qquad \forall\, x \in X_+,$$

$$g\big(x,\mu^*(x)\big) = 0, \quad p_{xy}\big(\mu^*(x)\big) = 0, \quad \forall\, x \in X_0,\ y \in X_+, \tag{4.24}$$

where $\overline{\mu}$ is an optimal policy for the equivalent SSP problem.

(c) If J^* is real-valued, then in the equivalent SSP problem every improper policy has infinite cost starting from some initial state. Moreover, there exists at least one proper policy, so the equivalent SSP problem satisfies the standard SSP conditions.

Proof: (a) Let us extend $\bar{J}$ to a function $\hat{J}$ that has domain X:

$$\hat{J}(x) = \begin{cases} \bar{J}(x) & \text{if } x \in X_+, \\ 0 & \text{if } x \in X_0. \end{cases}$$

Then from the Bellman equation $\bar{J} = \overline{T}\bar{J}$, and the definition (4.23) of $\overline{T}$, we have $\hat{J}(x) = (T\hat{J})(x)$ for all $x \in X_+$, while from Eq. (4.22), we have $(T\hat{J})(x) = 0 = \hat{J}(x)$ for all $x \in X_0$. Thus $\hat{J}$ is a fixed point of T, so that $\hat{J} \geq J^*$ [since J^* is the smallest nonnegative fixed point of T, cf. Prop. 4.1.3(a)], and hence $\bar{J}(x) \geq J^*(x)$ for all $x \in X_+$. Conversely, the restriction of J^* to X_+ is a solution of the Bellman equation $J = \overline{T}J$, with $\overline{T}$ given by Eq. (4.23), so we have $\bar{J}(x) \leq J^*(x)$ for all $x \in X_+$ [since $\bar{J}$ is the smallest nonnegative fixed point of $\overline{T}$, cf. Prop. 4.1.3(a)].

(b) A policy μ^* is optimal for the original problem if and only if $J^* = TJ^* = T_{\mu^*}J^*$ (cf. Prop. 4.1.5), or

$$\begin{aligned} J^*(x) &= \min_{u \in U(x)} \left[g(x,u) + \sum_{y=1}^{n} p_{xy}(u) J^*(y) \right] \\ &= g\big(x,\mu^*(x)\big) + \sum_{y=1}^{n} p_{xy}\big(\mu^*(x)\big) J^*(y), \qquad \forall\, x \in X. \end{aligned}$$

Equivalently, μ^* is optimal if and only if

$$\begin{aligned} J^*(x) &= \min_{u \in U(x)} \left[g(x,u) + \sum_{y \in X_+} p_{xy}(u) J^*(y) \right] \\ &= g\big(x,\mu^*(x)\big) + \sum_{y \in X_+} p_{xy}\big(\mu^*(x)\big) J^*(y), \qquad \forall\, x \in X_+, \end{aligned} \tag{4.25}$$

and Eq. (4.24) holds. Using part (a), the Bellman equation $\bar{J} = \overline{T}\bar{J}$, and the definition (4.23) of $\overline{T}$, we see that Eq. (4.25) is the necessary and sufficient condition for optimality of the restriction of μ^* to X_+ in the equivalent SSP problem, and the result follows.

(c) Let μ be an improper policy of the equivalent SSP problem. Then the Markov chain induced by μ contains a recurrent class R that consists of states x with $J^*(x) > 0$ [since we have $J^*(x) > 0$, for all $x \in X_+$]. We have $g\big(x, \mu(x)\big) > 0$ for some state $\overline{x} \in R$ [otherwise, $g\big(x, \mu(x)\big) = 0$ for all $x \in R$, implying that $J_\mu(x) = 0$ and hence $J^*(x) = 0$ for all $x \in R$]. From this it follows that $\bar{J}(x) = J^*(x) = \infty$ for all $x \in R$, since under μ, the state $\overline{x}$ is visited infinitely often with probability 1 starting from within R.

To prove the existence of a proper policy, we note that by the finiteness of the control space, the original problem has an optimal policy (cf. Prop. 4.1.5), and since J^* is real-valued this policy cannot be improper (as we have just shown, improper policies have infinite cost starting from at least one initial state). **Q.E.D.**

The following proposition provides analytical and computational results for the original problem, using the equivalent SSP problem.

Proposition 4.1.12: Assume that X and U are finite sets, that $g(x, u) \geq 0$ for all $x \in X$ and $u \in U(x)$, and that J^* is real-valued. Consider the set

$$\mathcal{J} = \big\{ J \geq 0 \mid J(x) = 0, \ \forall\, x \in X_0 \big\}.$$

Then:

(a) J^* is the unique fixed point of T within $\mathcal{J}$.

(b) We have $T^k J \to J^*$ for any $J \in \mathcal{J}$.

Proof: (a) Since the standard SSP conditions hold for the equivalent SSP problem by Prop. 4.1.11(c), $\bar{J}$ is the unique fixed point of $\overline{T}$. From Prop. 4.1.11(a) and the definition of the equivalent SSP problem, it follows that J^* is the unique fixed point of T within the set $\mathcal{J}$.

(b) Similar to the proof of part (a), the VI algorithm for the equivalent SSP problem is convergent to $\bar{J}$ from any initial condition, which implies the result. **Q.E.D.**

To make use of Prop. 4.1.12 we should know the sets X_0 and X_+, and also be able to deal with the case where J^* is not real-valued. We will provide an algorithm to determine X_0 next, and we will consider the case where J^* can take infinite values in the next subsection.

Algorithm for Constructing X_0 and X_+

In practice, the sets X_0 and X_+ can often be determined by a simple analysis that relies on the special structure of the given problem. When this is not so, we may compute these sets with a simple algorithm that requires at most n iterations, where n is the number of states in X. Let

$$\hat{U}(x) = \big\{u \in U(x) \mid g(x,u) = 0\big\}, \qquad x \in X.$$

Denote $X_1 = \big\{x \in X \mid \hat{U}(x) \neq \varnothing\big\}$, and define for $k \geq 1$,

$$X_{k+1} = \big\{x \in X_k \mid \text{there exists } u \in \hat{U}(x) \text{ such that } y \in X_k \\ \text{for all } y \text{ with } p_{xy}(u) > 0\big\}.$$

It can be seen with a straightforward induction that

$$X_k = \big\{x \in X \mid (T^k J_0)(x) = 0\big\},$$

where J_0 is the zero vector. Clearly we have $X_{k+1} \subset X_k$ for all k, and since X is finite, the algorithm terminates at some iteration $\bar{k}$ with $X_{\bar{k}+1} = X_{\bar{k}}$. Moreover the set $X_{\bar{k}}$ is equal to X_0, since we have $T^k J_0 \uparrow J^*$ because of the finiteness of the control space [cf. Prop. 4.1.7(a)]. If m is the number of state-control pairs, each iteration requires $O(m)$ computation, so the complexity of the algorithm for finding X_0 and X_+ is $O(mn)$.

The Case Where J^* is not Real-Valued

In order to use effectively the equivalent SSP problem, J^* must be real-valued, so that Prop. 4.1.12 can apply. It turns out that this restriction can be circumvented by introducing an artificial high-cost stopping action at each state, thereby making J^* real-valued.

In particular, let us assume without loss of generality that the original problem is already in SSP format, so it includes a termination state t. Let us introduce for each scalar $c > 0$, an SSP problem that is identical to the original, except that an additional control is added to each $U(x)$, under which the transition to t occurs with probability 1 and a cost c is incurred. We refer to this problem as the c-SSP problem, and we denote its optimal cost vector by $\hat{J}_c$. Note that

$$\hat{J}_c(x) \leq c, \qquad \hat{J}_c(x) \leq J^*(x), \qquad \forall\, x \in X,\ c > 0,$$

and that $\hat{J}_c$ is the unique fixed point of the corresponding mapping $\hat{T}_c$ given by

$$(\hat{T}_c J)(x) = \min\left[c,\ \min_{u \in U(x)}\left[g(x,u) + \sum_{y=1}^{n} p_{xy}(u) J(y)\right]\right], \qquad x = 1,\ldots,n, \tag{4.26}$$

within the set of $J \in \Re^n$ with $J(x) = 0$ for all $x \in X_0$ [cf. Prop. 4.1.12(a)]. Let

$$X_f = \big\{x \in X \mid J^*(x) < \infty\big\}, \qquad X^\infty = \big\{x \in X \mid J^*(x) = \infty\big\}.$$

We have the following proposition.

Proposition 4.1.13: Assume that X and U are finite sets, and that $g(x, u) \geq 0$ for all $x \in X$ and $u \in U(x)$. Then there exists $\overline{c} > 0$ such that for all $c \geq \overline{c}$, we have

$$\hat{J}_c(x) = J^*(x), \qquad \forall\, x \in X_f,$$

and if $\hat{\mu}$ is an optimal policy for the c-SSP problem, then any policy μ^* such that $\mu^*(x) = \hat{\mu}(x)$ for $x \in X_f$ is optimal for the original problem.

Proof: The result is clearly true if J^* is real-valued, since then for $c \geq \max_{x \in X} J^*(x)$, the VI algorithm starting from J_0 produces identical results for the c-SSP and original SSP problems, so for such c, $\hat{J}_c = J^*$. For the case where X^∞ is nonempty, we will formulate "reduced" versions of these two problems, where the states in X_f do not communicate with the states in X^∞, so that by restricting the reduced problems to X_f, we revert to the case where J^* is real-valued.

Indeed, for both the c-SSP problem and the original problem, let us replace the constraint set $U(x)$ by the set

$$\hat{U}(x) = \begin{cases} U(x) & \text{if } x \in X^\infty, \\ \big\{u \in U(x) \mid p_{xy}(u) = 0,\ \forall\, y \in X^\infty\big\} & \text{if } x \in X_f, \end{cases}$$

so that the infinite cost states in X^∞ are unreachable from the finite cost states in X_f. We refer to the problems thus created as the "reduced" c-SSP problem and the "reduced" original problem.

We now apply Prop. 4.1.5 to both the original and the "reduced" original problems. In the original problem, for each $x \in X_f$, the minimum in the expression

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

is attained for some $u \in \hat{U}(x)$ [controls $u \notin \hat{U}(x)$ are inferior because they lead with positive probability to states $y \in X^\infty$]. Thus an optimal policy for the original problem is feasible for the reduced original problem, and hence also optimal since the optimal cost cannot become smaller at any state when passing from the original to the reduced original problem. Similarly, for each $x \in X_f$, the minimum in the expression

$$\min \left[c,\ \min_{u \in U(x)} \left[g(x,u) + \sum_{y=1}^{n} p_{xy}(u) J_c(y) \right] \right],$$

[cf. Eq. (4.26)] is attained for some $u \in \hat{U}(x)$ once c becomes sufficiently large. The reason is that for $y \in X^\infty$, $\hat{J}_c(y) \uparrow J^*(y) = \infty$, so for sufficiently large c, each control $u \notin \hat{U}(x)$ becomes inferior to the controls $u \in \hat{U}(x)$, for which $p_{xy}(u) = 0$. Thus by taking c large enough, an optimal policy for the original c-SSP problem, becomes feasible and hence optimal for the reduced c-SSP problem [here the size of the "large enough" c depends on x and u, so finiteness of X and $U(x)$ is important for this argument]. We have thus shown that the optimal cost vector of the reduced original SSP problem is also J^*, and the optimal cost vector of the reduced c-SSP problem is also $\hat{J}_c$ for sufficiently large c.

Clearly, starting from any state in X_f it is impossible to transition to a state $x \in X^\infty$ in the reduced original problem and the reduced c-SSP problem. Thus if we restrict these problems to the set of states in X_f, we will not affect their optimal costs for these states. Since J^* is real-valued in X_f, it follows that for sufficiently large c, these optimal cost vectors are equal (as noted in the beginning of the proof), i.e., $\hat{J}_c(x) = J^*(x)$ for all $x \in X_f$. **Q.E.D.**

We note that the finiteness of U is needed for Prop. 4.1.13 to hold, and that a compactness condition such as the one of Assumption 3.2.1 is not sufficient. We demonstrate this with examples.

Example 4.1.4 (Counterexamples)

Consider the SSP problem of Fig. 4.1.2, which involves transition probabilities and costs that depend continuously on u, and the following two cases:

(a) Let $U(2) = (0,1]$, which is infinite but not compact. Then we have $J^*(1) = J^*(2) = \infty$. Let us now calculate $\hat{J}_c(1)$ and $\hat{J}_c(2)$ from the Bellman equation

$$\hat{J}_c(1) = \min \big[c,\ 1 + \hat{J}_c(1) \big],$$

$$\hat{J}_c(2) = \min \left[c,\ \min_{u \in (0,1]} \big[1 - \sqrt{u} + u \hat{J}_c(1) \big] \right].$$

The equation on the left yields $\hat{J}_c(1) = c$, and for $c \geq 1$, the minimization in the equation on the right takes the form

$$\min_{u \in (0,1]} \big[1 - \sqrt{u} + uc \big].$$

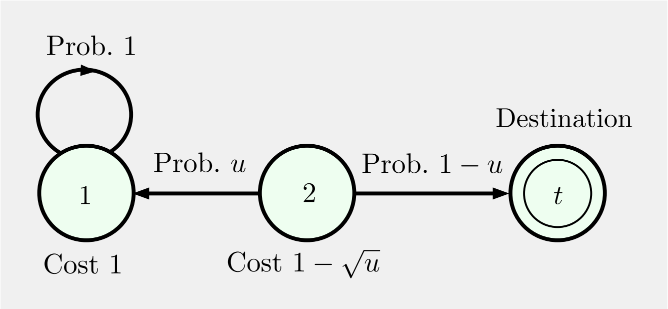


Figure 4.1.2. The SSP problem of Example 4.1.4. There are states 1 and 2, in addition to the termination state t. At state 2, upon selecting control u, we incur cost $1-\sqrt{u}$, and move to state 1 with probability u and to t with probability $1-u$. At state 1 we incur cost 1 and stay at 1 with probability 1.

By setting to 0 the derivative with respect to u, we see that the minimum is attained at $u = 1/(2c)^2$, yielding

$$\hat{J}_c(2) = 1 - \frac{1}{4c}, \qquad \text{for } c \geq 1.$$

Thus we have $\lim_{c\to\infty} \hat{J}_c(2) = 1$, while $J^*(2) = \infty$. This shows that without compactness of the control constraint sets we cannot have $\lim_{c\to\infty} \hat{J}_c = J^*$.

(b) $U(2) = [0,1]$, which is infinite and compact. Then we have $J^*(1) = \infty$, $J^*(2) = 1$. Similar to case (a), we calculate $\hat{J}_c(1)$ and $\hat{J}_c(2)$ from the Bellman equation. An essentially identical calculation to the one of case (a) yields the same results for $c \geq 1$:

$$\hat{J}_c(1) = c, \qquad \hat{J}_c(2) = 1 - \frac{1}{4c}.$$

Thus we have $\lim_{c\to\infty} \hat{J}_c(1) = J^*(1) = \infty$, and $\lim_{c\to\infty} \hat{J}_c(2) = J^*(2) = 1$. However, $\hat{J}_c(2) < J^*(2)$ for all c. This shows that finiteness of the control space is essential in order to have $\hat{J}_c = J^*$ for sufficiently large c. On the other hand, the property $\lim_{c\to\infty} \hat{J}_c = J^*$ can be shown in generality in finite-state problems under Assumption P when $U(x)$ is compact, and $p_{xy}(\cdot)$ and $g(x,\cdot)$ are continuous over $U(x)$; see [BeY15].

Proposition 4.1.13 suggests a procedure to solve a problem for which J^* is not real-valued, but the one-stage cost is nonnegative and the control space is finite:

(1) Compute the sets X_0 and X_+ using the algorithm of this section.

(2) Introduce for all $x \in X_+$ a stopping action with a cost $c > 0$.

(3) Solve the equivalent SSP problem and obtain a candidate optimal policy for the original problem using Prop. 4.1.11(b). This step can be done with the PI and VI algorithms of Chapter 3.

(4) Check that c is high enough, by testing to see if the candidate optimal policy changes as c is increased, or satisfies the optimality condition of Prop. 4.1.5. If it does, the current policy is optimal; if it does not, increase c by some fixed factor and repeat from Step (3).

By Prop. 4.1.13(b), this procedure terminates with an optimal policy in a finite number of steps.

4.2 CONTINUOUS-STATE DETERMINISTIC POSITIVE COST MODELS

In this section, we consider a positive cost problem, where the objective is to steer a deterministic system towards a cost-free and absorbing set of states. The system equation is

$$x_{k+1} = f(x_k, u_k), \qquad k = 0, 1, \ldots, \tag{4.27}$$

where x_k and u_k are the state and control at stage k, lying in 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 $U(x_k)$. The cost per stage, denoted $g(x, u)$, is assumed nonnegative:

$$0 \le g(x, u), \qquad x \in X,\ u \in U(x). \tag{4.28}$$

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.

Because the system is deterministic, given an initial state x_0, a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ when applied to the system (4.27), generates a unique sequence of state control pairs $\big(x_k, \mu_k(x_k)\big)$, $k = 0, 1, \ldots$. The corresponding cost function is

$$J_\pi(x_0) = \lim_{N\to\infty} \sum_{k=0}^{N-1} g\big(x_k, \mu_k(x_k)\big), \qquad x_0 \in X. \tag{4.29}$$

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

$$g(x, u) = 0, \qquad x = f(x, u), \qquad \forall\ x \in X_0,\ u \in U(x). \tag{4.30}$$

Clearly, $J^*(x) = 0$ for all $x \in X_0$, so the set X_0 may be viewed as a desirable set of termination states that we are trying to reach or approach with minimum total cost.

A major class of practical problems of this type are *regulation* problems in control applications, where the objective is to bring and maintain the state within a small region around a desired point. A popular formulation involves a deterministic linear system and a quadratic cost, such as the one discussed in Section 4.1 of Vol. I. Variations of this problem may involve a nonquadratic cost function, and state and control constraints. We will pause to discuss this type of problem in this section, and we will also discuss it in Section 4.3, in the context of *adaptive control*, where some of the system parameters are unknown.

Another major class of relevant practical problems is control of a dynamic system where the objective is to reach a goal state. Problems of this type are often called *planning* problems, and arise frequently in robotics, and in production scheduling and inventory control, among others.

Typically, the regulation and planing problems just described involve uncertainty, but they are often formulated as deterministic problems, and in practice they are combined with on-line replanning, to correct for disturbances and changes in the problem data.

The problem of this section is covered by the positive cost theory of Section 4.1 [cf. Eq. (4.28)]. Thus, from Props. 4.1.1 and 4.1.5, the optimal cost function J^* satisfies Bellman's equation:

$$J^*(x) = \min_{u \in U(x)} \big\{ g(x,u) + J^*\big(f(x,u)\big) \big\}, \qquad \forall\ x \in X,$$

and an optimal stationary policy may be obtained through the minimization in the right side of this equation. In this section we will focus on deriving conditions under which J^* is the *unique solution of this equation within a certain restricted class of functions*, whose value within the set X_0 is fixed at zero. We will also discuss how to compute J^* with the VI and PI algorithms.

The VI algorithm starts from some nonnegative function $J : X \mapsto [0,\infty]$, and generates a sequence of functions $\{J_k\}$ according to

$$J_{k+1} = \min_{u \in U(x)} \big\{ g(x,u) + J_k\big(f(x,u)\big) \big\}.$$

This sequence is also written as $J_k = T^k J$, where consistent with the notation of the preceding section, T is the mapping given by

$$(TJ)(x) = \min_{u \in U(x)} \big\{ g(x,u) + J\big(f(x,u)\big) \big\}, \qquad x \in X.$$

We will derive conditions under which J_k converges to J^* pointwise.

The PI algorithm 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

$$J_{\mu^k}(x) = g\big(x, \mu^k(x)\big) + J_{\mu^k}\Big(f\big(x, \mu^k(x)\big)\Big), \qquad x \in X, \tag{4.31}$$

interleaved with policy improvements to obtain μ^{k+1} from J_{μ^k} according to

$$\mu^{k+1}(x) = \arg\min_{u\in U(x)} \big\{ g(x,u) + J_{\mu^k}\big(f(x,u)\big)\big\}, \qquad x \in X. \tag{4.32}$$

Note that J_{μ^k} satisfies Eq. (4.31) by Prop. 4.1.2. Also for the PI algorithm to be well-defined, the minimum in Eq. (4.32) should be attained for each $x \in X$, which is true under some conditions that guarantee compactness of the level sets

$$\big\{ u \in U(x) \mid g(x,u) + J_{\mu^k}\big(f(x,u)\big) \leq \lambda \big\}, \qquad \lambda \in \Re;$$

cf. Prop. 4.1.8. We will derive conditions under which J_{μ^k} converges to J^*.

In our analysis, we will assume that $J^*(x) > 0$ for $x \notin X_0$, so that

$$X_0 = \big\{ x \in X \mid J^*(x) = 0 \big\}. \tag{4.33}$$

In the applications of primary interest, g is usually taken to be strictly positive outside of X_0 to encourage asymptotic convergence of the generated state sequence to X_0, so this assumption is natural and often easily verifiable. Besides X_0, another interesting subset of X is

$$X_f = \big\{ x \in X \mid J^*(x) < \infty \big\}.$$

Ordinarily, in practical applications, the states in X_f are those from which one can reach the stopping set X_0, at least asymptotically.

For an initial state x, we say that a policy π *terminates starting from* x if the state sequence $\{x_k\}$ generated starting from x and using π reaches X_0 in finite time, i.e., satisfies $x_{\bar{k}} \in X_0$ for some index $\bar{k}$. A key assumption in this section is that the optimal cost $J^*(x)$ (if it is finite) can be approximated arbitrarily closely by using policies that terminate from x. In particular, we assume the following throughput this section.

Assumption 4.2.1: (Asymptotic Termination) The cost per stage is nonnegative [cf. Eq. (4.28)], and for all states x outside the stopping set X_0 we have $J^*(x) > 0$ [cf. Eq. (4.33)]. Moreover, for every pair (x,ϵ) with $x \in X_f$ and $\epsilon > 0$, there exists a policy π that terminates starting from x and satisfies $J_\pi(x) \leq J^*(x) + \epsilon$.

Specific and easily verifiable conditions that imply this assumption will be given later. A prominent case is when X and U are finite, so the problem becomes a deterministic shortest path problem with nonnegative arc lengths. If all cycles of the state transition graph have positive length, all policies π that do not terminate from a state $x \in X_f$ must satisfy

$J_\pi(x) = \infty$, implying that there exists an optimal policy that terminates from all $x \in X_f$. Thus, in this case Assumption 4.2.1 is naturally satisfied.

When X is the n-dimensional Euclidean space $\Re^n$, a primary case of interest, it may easily happen that the optimal policies are not terminating from some $x \in X_f$, but instead the optimal state trajectories may approach X_0 asymptotically. This is true for example in the classical linear-quadratic optimal control problem, 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 cost is positive semidefinite quadratic. There the optimal policy is linear of the form $\mu^*(x) = Lx$, where L is some matrix obtained through the steady-state solution of the Riccati equation (see Section 4.1 of Vol. I, and also Section 4.3 in this chapter). Since the optimal closed-loop system is stable and 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, we will show later that the Assumption 4.2.1 is satisfied under some natural and easily verifiable conditions.

We denote by $E^+(X)$ the set of all functions $J : X \mapsto [0, \infty]$, and by $\mathcal{J}$ the set of functions

$$\mathcal{J} = \big\{ J \in E^+(X) \mid J(x) = 0,\ \forall\ x \in X_0 \big\}. \tag{4.34}$$

Since X_0 consists of cost-free and absorbing states [cf. Eq. (4.30)], the set $\mathcal{J}$ contains the cost function J_π of all policies π, as well as J^*. In our terminology, all equations, inequalities, and convergence limits involving functions are meant to be pointwise. Let us also denote for all $x \in X$,

$$\Pi_{R,x} = \big\{ \pi \in \Pi \mid \pi \text{ terminates from } x \big\}, \tag{4.35}$$

and note the following key implication of asymptotic termination Assumption 4.2.1:

$$J^*(x) = \min_{\pi \in \Pi_{R,x}} J_\pi(x), \qquad \forall\ x \in X_f. \tag{4.36}$$

In the subsequent proof arguments, the significance of policies that terminate starting from some initial state x_0 is that the corresponding generated sequences $\{x_k\}$ satisfy $J(x_k) = 0$ for all $J \in \mathcal{J}$ and k sufficiently large.

Our main results are given in the following three propositions.

Proposition 4.2.1: (Uniqueness of Solution of Bellman's Equation) Let Assumption 4.2.1 hold. The optimal cost function J^* is the unique solution of Bellman's equation within the set of functions $\mathcal{J}$.

Proof: Let $\hat J \in \mathcal{J}$ be a solution of Bellman's equation, so that

$$\hat J(x) \le g(x,u) + \hat J\big(f(x,u)\big), \qquad \forall\ x \in X,\ u \in U(x), \tag{4.37}$$

while by Prop. 4.1.3(a), $J^* \leq \hat{J}$. For any $x_0 \in X_f$ and policy $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi_{R,x_0}$, we have by using repeatedly Eq. (4.37),

$$J^*(x_0) \leq \hat{J}(x_0) \leq \hat{J}(x_k) + \sum_{t=0}^{k-1} g\big(x_t, \mu_t(x_t)\big), \quad k = 1, 2, \ldots,$$

where $\{x_k\}$ is the state sequence generated starting from x_0 and using π. Also, since $\pi \in \Pi_{R,x_0}$ and hence $x_k \in X_0$ and $\hat{J}(x_k) = 0$ for all sufficiently large k, we have

$$\limsup_{k\to\infty} \left\{ \hat{J}(x_k) + \sum_{t=0}^{k-1} g\big(x_t, \mu_t(x_t)\big) \right\} = \lim_{k\to\infty} \left\{ \sum_{t=0}^{k-1} g\big(x_t, \mu_t(x_t)\big) \right\} = J_\pi(x_0).$$

By combining the last two relations, we obtain

$$J^*(x_0) \leq \hat{J}(x_0) \leq J_\pi(x_0), \qquad \forall\ x_0 \in X_f,\ \pi \in \Pi_{R,x_0}.$$

Taking the minimum over $\pi \in \Pi_{R,x_0}$ and using Eq. (4.36), it follows that $J^*(x_0) = \hat{J}(x_0)$ for all $x_0 \in X_f$. Also for $x_0 \notin X_f$, we have $J^*(x_0) = \hat{J}(x_0) = \infty$ [since $J^* \leq \hat{J}$ by Prop. 4.1.3(a)], so we obtain $J^* = \hat{J}$. **Q.E.D.**

We give two examples where the asymptotic termination Assumption 4.2.1 is violated because there are states $x \notin X_0$ such that $J^*(x) = 0$, so the condition (4.33) does not hold. In both examples, in addition to J^*, there are other solutions of Bellman's equation within $\mathcal{J}$.

Example 4.2.1 (Shortest Path Example)

Consider the positive cost case of the deterministic Example 4.1.3. Here $X = \{1, t\}$, where t is a cost-free and absorbing state. We let $X_0 = \{t\}$, so that

$$\mathcal{J} = \big\{J \mid J(1) \geq 0,\ J(t) = 0\big\};$$

cf. Eq. (4.34). At state 1 there are two choices: u which leads to t at cost $b > 0$, and u' that self-transitions to 1 at cost 0. We have $J^*(1) = J^*(t) = 0$, so

$$X_0 \neq \big\{x \in X \mid J^*(x) = 0\big\} = X,$$

and the condition (4.33) is violated. Here Bellman's equation takes the form

$$J(1) = \min\big\{J(1),\, b + J(t)\big\}, \qquad J(t) = J(t),$$

and it can be seen that its set of solutions within $\mathcal{J}$ is the infinite set

$$\big\{J \mid 0 \leq J(1) \leq b,\ J(t) = 0\big\};$$

cf. Fig. 4.1.1.

Example 4.2.2 (Linear-Quadratic Example)

Consider the scalar system $x_{k+1} = \gamma x_k + u_k$ with $X = U(x) = \Re$, and the quadratic cost $g(x,u) = u^2$. We take $X_0 = \{0\}$, so that

$$\mathcal{J} = \big\{ J \geq 0 \mid J(0) = 0 \big\}.$$

We have $J^*(x) \equiv 0$, so

$$X_0 \neq \big\{ x \in X \mid J^*(x) = 0 \big\} = X,$$

and condition (4.33) is violated. 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. Let us assume that $\gamma > 1$ so the system is unstable (the instability of the system is important for the purpose of this example). Then it can be verified that the quadratic function

$$J(x) = (\gamma^2 - 1)x^2,$$

which belongs to $\mathcal{J}$, also solves Bellman's equation. This is the cost function of the suboptimal policy $\mu(x) = \frac{(1-\gamma^2)x}{\gamma}$, which yields the stable closed-loop system $x_{k+1} = \frac{1}{\gamma} x_k$. This policy can be verified to be optimal among policies that are linear and yield a stable closed-loop system, but it is not optimal among all policies [the optimal policy applies $\mu^*(x) = 0$ for all x].

In this case the algebraic Riccati equation associated with the problem has two nonnegative solutions because there is no cost on the state. This is a consequence of the violation of the standard observability condition for uniqueness of solution of the Riccati equation (cf. Section 4.1, Vol. I). If the cost per stage were $g(x,u) = qx^2 + u^2$, with $q > 0$, instead of $g(x,u) = u^2$, Assumption 4.2.1 would be satisfied, and by Prop. 4.2.1, Bellman's equation would have a unique solution within $\mathcal{J}$. Consistently with this fact, the Riccati equation would have a unique positive solution (although it would also have a negative solution, cf. Section 4.1, Vol. I).

Proposition 4.2.2: (Convergence of VI) Let Assumption 4.2.1 hold.

(a) The sequence $\{T^k J\}$ generated by VI starting from a function $J \in \mathcal{J}$ with $J \geq J^*$ converges to J^*.

(b) Assume further that the compactness condition of Prop. 4.1.8 holds, i.e., that the sets $U_k(x, \lambda)$ given by

$$U_k(x,\lambda) = \big\{ u \in U(x) \mid g(x,u) + (T^k J_0)\big(f(x,u)\big) \leq \lambda \big\},$$

are compact subsets of a metric space for all $x \in X$, $\lambda \in \Re$, and for all k greater than some integer $\overline{k}$, where J_0 is the identically zero function. Then the sequence $\{T^k J\}$ generated by VI starting from any function $J \in \mathcal{J}$ converges to J^*.

Proof: (a) Let J be a function in $\mathcal{J}$ with $J \geq J^*$, and let us apply the VI operation to both sides of the inequality $J \geq J^*$. Since J^* is a solution of Bellman's equation and VI has a monotonicity property that maintains the direction of functional inequalities, we see that $TJ \geq J^*$. Continuing similarly, we obtain $T^kJ \geq J^*$ for all k. Moreover, we clearly have $(T^kJ)(x) = 0$ for all $x \in X_0$, so $T^kJ \in \mathcal{J}$ for all k. We now argue that since T^kJ is produced by k steps of VI starting from J, it is the optimal cost function of the k-stage version of the problem with terminal cost function J. Therefore, we have for every $x_0 \in X$ and policy $\pi = \{\mu_0, \mu_1, \ldots\}$,

$$J^*(x_0) \leq (T^kJ)(x_0) \leq J(x_k) + \sum_{t=0}^{k-1} g\big(x_t, \mu_t(x_t)\big), \quad k = 1, 2, \ldots,$$

where $\{x_t\}$ is the state sequence generated starting from x_0 and using π. If $x_0 \in X_f$ and π is a policy in the set Π_{R,x_0} defined by Eq. (4.35), we have $x_k \in X_0$ and $J(x_k) = 0$ for all sufficiently large k, so that

$$\limsup_{k\to\infty} \left\{ J_0(x_k) + \sum_{t=0}^{k-1} g\big(x_t, \mu_t(x_t)\big) \right\} = \lim_{k\to\infty} \left\{ \sum_{t=0}^{k-1} g\big(x_t, \mu_t(x_t)\big) \right\} = J_\pi(x_0).$$

By combining the last two relations, we obtain

$$J^*(x_0) \leq \liminf_{k\to\infty} (T^kJ)(x_0) \leq \limsup_{k\to\infty} (T^kJ)(x_0) \leq J_\pi(x_0),$$

for all $x_0 \in X_f$ and $\pi \in \Pi_{R,x_0}$. Taking the minimum over $\pi \in \Pi_{R,x_0}$ and using Eq.(4.36), it follows that $\lim_{k\to\infty}(T^kJ)(x_0) = J^*(x_0)$ for all $x_0 \in X_f$. Since for $x_0 \notin X_f$, we have $J^*(x_0) = (T^kJ)(x_0) = \infty$, we obtain $T^kJ \to J^*$.

(b) Let J be any function in $\mathcal{J}$. By the monotonicity of the VI operation, $\{T^kJ\}$ lies between the sequence of VI iterates starting from the zero function [which converges to J^* from below by Prop. 4.1.8], and the sequence of VI iterates starting from $\hat{J} = \max\{J, J^*\}$ [which converges to J^* from above by part (a)]. **Q.E.D.**

The following example shows why the compactness assumption of part (b) is necessary to guarantee convergence of VI to J^*, starting from initial conditions $J \leq J^*$.

Example 4.2.3 (Counterexample for Convergence of VI)

Let $X = [0, \infty) \cup \{t\}$, with t being a cost-free and absorbing state, and let $U = (0, \infty) \cup \{\bar{u}\}$, where $\bar{u}$ is a special stopping control, which moves the system from states $x \geq 0$ to state t at unit cost. When u_k is not the stopping control $\bar{u}$, the system evolves according to

$$x_{k+1} = x_k + u_k, \qquad \text{if } x_k \geq 0 \text{ and } u_k \neq \bar{u},$$

The cost per stage has the form

$$g(x_k, u_k) = x_k, \qquad \text{if } x_k \geq 0 \text{ and } u_k \neq \bar{u},$$

and $g(x_k, \bar{u}) = 1$ when u_k is the stopping control $\bar{u}$. Let also $X_0 = \{t\}$. Then it can be verified that

$$J^*(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x = t, \end{cases}$$

and that an optimal policy is to use the stopping control $\bar{u}$ at every state (since using any other control at states $x \geq 0$, leads to unbounded accumulation of positive cost). Thus it can be seen that Assumption 4.2.1 is satisfied. On the other hand, the VI algorithm is

$$J_{k+1}(x) = \min \left\{ 1 + J_k(t), \min_{u \geq 0} \big\{ x + J_k(x+u) \big\} \right\}$$

for $x \geq 0$, and $J_{k+1}(t) = J_k(t)$, and it can be verified by induction that starting from $J_0 \equiv 0$, the sequence $\{J_k\}$ is given for all k by

$$J_k(x) = \begin{cases} \min\{1,\, kx\} & \text{if } x \geq 0, \\ 0 & \text{if } x = t. \end{cases}$$

Thus $J_k(0) = 0$ for all k, while $J^*(0) = 1$, so the VI algorithm fails to converge for the state $x = 0$. The difficulty here is that the compactness assumption of Prop. 4.2.2(b) is violated.

We next consider the convergence of the PI algorithm. We implicitly assume that the algorithm is well-defined in the sense that the minimization in the policy improvement operation (4.32) can be carried out for every $x \in X$. Easily verifiable conditions that guarantee this also guarantee the compactness condition of Prop. 4.2.2(b), and will be noted later. Moreover, we will prove later a similar convergence result for a variant of the PI algorithm where the policy evaluation is carried out approximately through a finite number of VIs.

Proposition 4.2.3: (Convergence of PI) Let Assumption 4.2.1 hold. A sequence $\{J_{\mu^k}\}$ generated by the PI algorithm (4.31), (4.32), satisfies $J_{\mu^k}(x) \downarrow J^*(x)$ for all $x \in X$.

Proof: If μ is a stationary policy and $\bar{\mu}$ satisfies the policy improvement equation

$$\bar{\mu}(x) = \arg \min_{u \in U(x)} \big\{ g(x,u) + J_\mu\big(f(x,u)\big) \big\}, \qquad x \in X,$$

[cf. Eq. (4.32)], we have for all $x \in X$,

$$\begin{aligned} J_\mu(x) &= g\big(x,\mu(x)\big) + J_\mu\Big(f\big(x,\mu(x)\big)\Big) \\ &\geq \min_{u\in U(x)} \Big\{g(x,u) + J_\mu\big(f(x,u)\big)\Big\} \\ &= g\big(x,\bar\mu(x)\big) + J_\mu\Big(f\big(x,\bar\mu(x)\big)\Big), \end{aligned} \tag{4.38}$$

where the first equality follows from Prop. 4.1.2 and the second equality follows from the definition of $\bar\mu$. Let us fix x and let $\{x_k\}$ be the sequence generated starting from x and using μ. By repeatedly applying Eq. (4.38), we see that the sequence $\big\{\tilde J_k(x)\big\}$ defined by

$$\tilde J_0(x) = J_\mu(x),$$

$$\tilde J_1(x) = J_\mu(x_1) + g\big(x,\bar\mu(x)\big),$$

and more generally,

$$\tilde J_k(x) = J_\mu(x_k) + \sum_{t=0}^{k-1} g\big(x_t,\bar\mu(x_t)\big), \qquad k = 1,2,\ldots,$$

is monotonically nonincreasing. Thus, using also Eq. (4.38), we have

$$J_\mu(x) \geq \min_{u\in U(x)} \Big\{g(x,u) + J_\mu\big(f(x,u)\big)\Big\} = \tilde J_1(x) \geq \tilde J_k(x),$$

for all $x \in X$ and $k \geq 1$. This implies that

$$\begin{aligned} J_\mu(x) &\geq \min_{u\in U(x)} \Big\{g(x,u) + J_\mu\big(f(x,u)\big)\Big\} \\ &\geq \lim_{k\to\infty} \tilde J_k(x) \\ &= \lim_{k\to\infty} \left\{ J_\mu(x_k) + \sum_{t=0}^{k-1} g\big(x_t,\bar\mu(x_t)\big) \right\} \\ &\geq \lim_{k\to\infty} \sum_{t=0}^{k-1} g\big(x_t,\bar\mu(x_t)\big) \\ &= J_{\bar\mu}(x), \end{aligned}$$

where the last inequality follows since $J_\mu \geq 0$. In conclusion, the cost function of $\bar\mu$ is no worse than the cost function of μ, and we have

$$J_\mu(x) \geq \min_{u\in U(x)} \Big\{g(x,u) + J_\mu\big(f(x,u)\big)\Big\} \geq J_{\bar\mu}(x), \quad x \in X.$$

Using μ^k and μ^{k+1} in place of μ and $\bar\mu$ in the preceding relation, we obtain for all $x \in X$,

$$J_{\mu^k}(x) \geq \min_{u\in U(x)} \Big\{g(x,u) + J_{\mu^k}\big(f(x,u)\big)\Big\} \geq J_{\mu^{k+1}}(x). \tag{4.39}$$

Thus 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, by taking the limit as $k \to \infty$ in Eq. (4.39), we have the two relations

$$J_\infty(x) \geq \min_{u \in U(x)} \bigl\{g(x,u) + J_\infty\bigl(f(x,u)\bigr)\bigr\}, \qquad x \in X,$$

and

$$g(x,u) + J_{\mu^k}\bigl(f(x,u)\bigr) \geq J_\infty(x), \qquad x \in X,\ u \in U(x).$$

We now take the limit in the second relation as $k \to \infty$, then the minimum over $u \in U(x)$, and then combine with the first relation, to obtain

$$J_\infty(x) = \min_{u \in U(x)} \bigl\{g(x,u) + J_\infty\bigl(f(x,u)\bigr)\bigr\}, \qquad x \in X.$$

Thus J_∞ is a solution of Bellman's equation, satisfying $J_\infty \in \mathcal{J}$ (since $J_{\mu^k} \in \mathcal{J}$ and $J_{\mu^k} \downarrow J_\infty$), so by the uniqueness result of Prop. 4.2.1, we have $J_\infty = J^*$. **Q.E.D.**

Example 4.2.4 (Counterexample for Convergence of PI)

Consider the deterministic two-state problem of Examples 4.1.3 and 4.2.1. Let μ be the suboptimal policy that moves from state 1 to state t. Then $J_\mu(1) = 1$, $J_\mu(t) = 0$, and it can be seen that μ satisfies the policy improvement equation

$$\mu(1) = \arg\min \bigl\{1 + J_\mu(t),\ J_\mu(1)\bigr\}.$$

Thus PI may stop with the suboptimal policy μ.

Conditions that Imply the Asymptotic Termination Assumption

We will now discuss readily verifiable conditions guaranteeing that Assumption 4.2.1 holds. As noted earlier, it holds when X and U are finite, a terminating policy exists from every x, and all cycles of the state transition graph have positive length. For the case where X is infinite, let us assume that X is a normed space with norm denoted $\|\cdot\|$, and 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

$$\lim_{k \to \infty} \operatorname{dist}(x_k, X_0) = 0,$$

where $\operatorname{dist}(x, X_0)$ denotes the minimum distance from x to X_0,

$$\operatorname{dist}(x, X_0) = \min_{y \in X_0} \|x - y\|, \qquad x \in X.$$

We have the following proposition.

Proposition 4.2.4: Assume that the cost per stage is nonnegative [cf. Eq. (4.28)], and for all states x outside the stopping set X_0 we have $J^*(x) > 0$ [cf. Eq. (4.33)]. Assume further the following:

(1) For every $x \in X_f$ 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 $x \in X_f$ 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 Assumption 4.2.1 holds.

Proof: Fix $x \in X_f$ 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 $x_{\bar{k}}$. **Q.E.D.**

Condition (1) of the preceding proposition requires that for states $x \in X_f$, the optimal cost $J^*(x)$ can be achieved arbitrarily closely with policies that asymptotically terminate from x. Problems for which condition (1) holds are those involving a cost per stage that is strictly positive outside of X_0. More precisely, condition (1) holds if for each $\delta > 0$ there exists $\epsilon > 0$ such that

$$\min_{u \in U(x)} g(x, u) \geq \epsilon, \qquad \forall\, x \in X \text{ such that } \text{dist}(x, X_0) \geq \delta. \tag{4.40}$$

Then for any x and policy π that does not asymptotically terminate from x, we will have $J_\pi(x) = \infty$, so that if $x \in X_f$, all policies π with $J_\pi(x) < \infty$

must be asymptotically terminating from x. In applications, condition (1) is natural and consistent with the aim of steering the state towards the terminal set X_0 with finite cost. Condition (2) is a "controllability" condition implying that the state can be steered into X_0 with arbitrarily small cost from a starting state that is sufficiently close to X_0.

Example 4.2.5 (Linear System Case)

Consider a linear system

$$x_{k+1} = Ax_k + Bu_k,$$

where A and B are given matrices, with the terminal set being the origin, i.e., $X_0 = \{0\}$. We assume the following:

(a) $X = \Re^n$, $U = \Re^m$, and there is an open sphere R centered at the origin such that $U(x)$ contains R for all $x \in X$.

(b) The system is controllable, i.e., one may drive the system from any state to the origin within at most n steps using suitable controls, or equivalently that the matrix $[B\ AB\ \cdots A^{n-1}B]$ has rank n (cf. Section 4.1 of Vol. I).

(c) g satisfies

$$0 \leq g(x,u) \leq \beta\big(\|x\|^p + \|u\|^p\big), \qquad \forall\ (x,u) \in V,$$

where V is some open sphere centered at the origin, β, p are some positive scalars, and $\|\cdot\|$ is the standard Euclidean norm.

Then condition (2) of Prop. 4.2.4 is satisfied, while $x = 0$ is cost-free and absorbing [cf. Eq. (4.30)]. Still, however, in the absence of additional assumptions, there may be multiple solutions to Bellman's equation within $\mathcal{J}$, as shown by Example 4.2.2. Assume now that in addition to (a)-(c), we have for some positive scalars γ, p, $\min_{u\in U(x)} g(x,u) \geq \gamma\|x\|^p$ for all $x \in \Re^n$. Then $J^*(x) > 0$ for all $x \neq 0$ [cf. Eq. (4.33)], while condition (1) of Prop. 4.2.4 is satisfied as well [cf. Eq. (4.40)]. Thus by Prop. 4.2.4, Assumption 4.2.1 holds, and Bellman's equation has a unique solution within $\mathcal{J}$.

Note that there are straightforward extensions of the conditions of the preceding example to a nonlinear system. Note also that even for a controllable system, it is possible that there exist states from which the terminal set cannot be reached, because $U(x)$ may imply constraints on the magnitude of the control vector. Still the preceding analysis allows for this case.

An Optimistic Form of PI

Let us consider a variant of PI where policies are evaluated inexactly, with a finite number of VIs (cf. Section 2.3.3). In particular, this algorithm starts with some $J_0 \in E(X)$, and generates a sequence of cost function and policy pairs $\{J_k, \mu^k\}$ as follows: Given J_k, we generate μ^k according to

$$\mu^k(x) = \arg \min_{u\in U(x)} \big\{g(x,u) + J_k\big(f(x,u)\big)\big\}, \qquad x \in X, \tag{4.41}$$

and then we obtain J_{k+1} with $m_k \geq 1$ VIs using μ^k:

$$J_{k+1}(x_0) = J_k(x_{m_k}) + \sum_{t=0}^{m_k-1} g\big(x_t, \mu^k(x_t)\big), \qquad x_0 \in X, \tag{4.42}$$

where $\{x_t\}$ is the sequence generated using μ^k and starting from x_0, and m_k are arbitrary positive integers. Here J_0 is a function in $\mathcal{J}$ that is required to satisfy

$$J_0(x) \geq \min_{u\in U(x)} \big\{g(x,u) + J_0\big(f(x,u)\big)\big\}, \ \ \forall\, x \in X,\ u \in U(x). \tag{4.43}$$

For example J_0 may be equal to the cost function of some stationary policy, or be the function that takes the value 0 for $x \in X_0$ and ∞ at $x \notin X_0$. Note that when $m_k \equiv 1$ the method is equivalent to VI, while the case $m_k = \infty$ corresponds to the standard PI considered earlier.

Proposition 4.2.5: (Convergence of Optimistic PI) Let Assumption 4.2.1 hold. For the PI algorithm (4.41)-(4.42), where J_0 belongs to $\mathcal{J}$ and satisfies the condition (4.43), we have $J_k \downarrow J^*$.

Proof: We have for all $x \in X$,

$$\begin{aligned} J_0(x) &\geq \min_{u\in U(x)} \big\{g(x,u) + J_0\big(f(x,u)\big)\big\} \\ &= g\big(x,\mu^0(x)\big) + J_0\big(f(x,\mu^0(x))\big) \\ &\geq J_1(x) \\ &\geq g\big(x,\mu^0(x)\big) + J_1\big(f(x,\mu^0(x))\big) \\ &\geq \min_{u\in U(x)} \big\{g(x,u) + J_1\big(f(x,u)\big)\big\} \\ &= g\big(x,\mu^1(x)\big) + J_1\big(f(x,\mu^1(x))\big) \\ &\geq J_2(x), \end{aligned}$$

where the first inequality is the condition (4.43), the second and third inequalities follow because of the monotonicity of the m_0 VIs (4.42) for μ^0, and the fourth inequality follows from the policy improvement equation (4.41). Continuing similarly, we have

$$J_k(x) \geq \min_{u\in U(x)} \big\{g(x,u) + J_k\big(f(x,u)\big)\big\} \geq J_{k+1}(x),$$

for all $x \in X$ and k. Moreover, since $J_0 \in \mathcal{J}$, we have $J_k \in \mathcal{J}$ for all k. Thus $J_k \downarrow J_\infty$ for some $J_\infty \in \mathcal{J}$, and similar to the proof of Prop. 4.2.3, it follows that J_∞ is a solution of Bellman's equation. Hence, by the uniqueness result of Prop. 4.2.1, we have $J_\infty = J^*$. **Q.E.D.**

Minimax Control to a Terminal Set of States

Our analysis of this section can be readily extended to minimax problems with a terminal set of states. Here the system is

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

where w_k is the control of an antagonistic opponent that aims to maximize the cost function. We assume that w_k is chosen from a given set W to maximize the sum of costs per stage, which are assumed nonnegative:

$$0 \leq g(x, u, w), \qquad x \in X,\ U \in U(x),\ w \in W.$$

We wish to choose a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ to minimize the cost function

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

where $\big\{x_k, \mu_k(x_k)\big\}$ is a state-control sequence corresponding to π and the sequence $\{w_0, w_1, \ldots\}$. This problem admits a similar analysis to the one for stochastic problems under Assumption P in Section 4.1. In particular, the results of that section on the validity of Bellman's equation, the optimality conditions, and the convergence of VI, all have minimax analogs. The reason is that these results hold in the context of a more general abstract DP model that contains as special cases both stochastic and minimax models: the monotone increasing models of [Ber13], Section 4.3 (see also Appendix B).

To formulate a minimax problem similar to the one of the deterministic problem of the present section, we introduce a termination set X_0. The states in this set are cost-free and absorbing, i.e.,

$$g(x, u, w) = 0, \qquad x = f(x, u, w),$$

for all $x \in X_0$, $u \in U(x)$, $w \in W$, and we assume that all states outside X_0 have strictly positive optimal cost, so that

$$X_0 = \big\{x \in X \mid J^*(x) = 0\big\}.$$

We next adapt the definition of termination. In particular, given a state x, in the minimax context we say that a policy π terminates from x

if there exists an index $\bar{k}$ [which depends on (π, x)] such that the sequence $\{x_k\}$, which is generated starting from x and using π, satisfies $x_{\bar{k}} \in X_0$ for all sequences $\{w_0, \ldots, w_{\bar{k}-1}\}$ with $w_t \in W$ for all $t = 0, \ldots, \bar{k}-1$. Then the asymptotic termination Assumption 4.2.1 is modified to reflect this new definition of termination, and our results can be readily extended, with Props. 4.2.1, 4.2.2, 4.2.3, and 4.2.5, and their proofs, holding essentially as stated.

The main adjustment needed is to replace expressions of the forms

$$g(x,u) + J\big(f(x,u)\big)$$

and

$$J(x_k) + \sum_{t=0}^{k-1} g(x_t, u_t)$$

in these proofs with

$$\sup_{w \in W} \big\{g(x,u,w) + J\big(f(x,u,w)\big)\big\}$$

and

$$\sup_{\substack{w_t \in W\\ t=0,\ldots,k-1}} \left\{ J(x_k) + \sum_{t=0}^{k-1} g(x_t, u_t, w_t) \right\},$$

respectively.

If the state and control spaces are finite, then the assumption that all states outside X_0 have strictly positive optimal cost can be circumvented, by reformulating the minimax problem into another minimax problem where the assumption is satisfied. The technique for doing this is the same as the one of Section 4.1.4, and essentially lumps all the states x such that $J^*(x) = 0$ into a single termination state for the reformulated problem.

4.3 LINEAR-QUADRATIC PROBLEMS AND ADAPTIVE DP

In this section we consider the case of the linear system

$$x_{k+1} = Ax_k + Bu_k + w_k, \qquad k = 0, 1, \ldots,$$

where $x_k \in \Re^n$, $u_k \in \Re^m$ for all k, and the matrices A, B are known. As in Sections 4.1 and 5.2 of Vol. I, we assume that the random disturbances w_k are independent with zero mean and finite second moments. The cost function is quadratic and has the form

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

where $\alpha \in (0, 1]$ is the discount factor, Q is a positive semidefinite symmetric $n \times n$ matrix, and R is a positive definite symmetric $m \times m$ matrix. Clearly, the positive cost Assumption P of Section 4.1 holds, and the results of that section apply. We will consider two cases:

(a) The undiscounted case where $\alpha = 1$. Then we will assume the deterministic case where $w_k = 0$, so that the optimal cost function J^* is guaranteed to be real-valued.

(b) The discounted case where $0 < \alpha < 1$. Then the optimal cost function J^* is real-valued, even when w_k has nonzero second moment.

The Undiscounted Case

Consider first the case where $\alpha = 1$, and the problem is deterministic, so that $w_k = 0$ for all k. We will apply the theory of Sections 4.1 and 4.2 as follows:

(1) We use Prop. 4.1.8 to show that the sequence $\{T^k J_0\}$, generated by VI starting from the identically zero function J_0, converges to J^*, and that an optimal stationary policy exists. Indeed, based on the analysis of Section 4.1 of Vol. I, we have

$$(TJ_0)(x) = \min_u [x'Qx + u'Ru] = x'Qx,$$

$$\begin{aligned}(T^2 J_0)(x) &= \min_{u \in \Re^n} \big[x'Qx + u'Ru + (TJ_0)(Ax + Bu)\big] \\ &= \min_{u \in \Re^n} \big[x'Qx + u'Ru + (Ax + Bu)'Q(Ax + Bu)\big],\end{aligned}$$

and more generally,

$$(T^k J_0)(x) = x'P_k x, \quad k = 1, 2, \ldots,$$

where

$$P_1 = Q$$

and P_{k+1} is generated from P_k using the Riccati equation

$$P_{k+1} = A'\big(P_k - P_k B(B'P_k B + R)^{-1} B'P_k\big)A + Q, \qquad k = 1, 2, \ldots,$$

Clearly P_k is positive semidefinite, which in view of the positive definiteness of R, implies that the compactness condition of Prop. 4.1.8 is satisfied, so that $T^k J_0 \uparrow J^*$, and an optimal stationary policy exists.

(2) Assuming controllability of the pair (A, B), it also follows that for $k \geq n$, $J^*(x)$ is bounded from above by the cost corresponding to a control sequence that forces x to the origin in n steps and applies zero

control after that. Thus P_k converges to some positive semidefinite symmetric matrix P^*, and we have

$$J^*(x) = x'P^*x, \qquad x \in \Re^n. \tag{4.44}$$

Moreover, by taking the limit as $k \to \infty$ in the Riccati equation, we see that P^* is a solution of the steady-state Riccati equation

$$P = A'\big(P - PB(B'PB + R)^{-1}B'P\big)A + Q. \tag{4.45}$$

If in addition we assume observability of the pair (A, C), where $Q = CC'$, then Prop. 4.1.1 of Section 4.1 in Vol. I shows that P_k is positive definite for sufficiently large k, P^* is also positive definite, and it is the unique solution of the Riccati equation (4.45) within the class of positive semidefinite matrices (Example 4.2.2 shows that in the absence of the observability assumption, this may not be true and that the Riccati equation may have two positive semidefinite solutions).

(3) From Props. 4.1.5 and 4.1.8, it follows that there exists an optimal stationary policy μ^*, which is obtained by minimizing in the right-hand side of Bellman's equation,

$$x'P^*x = \min_{u \in \Re^m} \big[x'Qx + u'Ru + (Ax + Bu)'P^*(Ax + bu)\big],$$

assuming that the pair (A, B) is controllable so that J^* is real-valued. By setting the gradient of the minimized expression to 0, we obtain

$$\mu^*(x) = L^*x, \qquad x \in \Re^n, \tag{4.46}$$

where L^* is the matrix

$$L^* = -(B'P^*B + R)^{-1}B'P^*A. \tag{4.47}$$

This policy is attractive for practical implementation since it is linear and stationary. Also, from Prop. 4.1.1 of Section 4.1, Vol. I, we have that the corresponding optimal closed-loop system is stable if in addition the pair (A, C) is observable.

(4) We next view the problem within the framework of Section 4.2, with the terminal set of states being just the origin:

$$X_0 = \{0\},$$

assuming both the controllability and observability assumptions noted earlier. Clearly, when at state $x = 0$, it is optimal to stay at that state by applying $u = 0$. Thus we may view X_0 as a set of cost-free and absorbing states, as required by the framework of Section 4.2. To show that Assumption 4.2.1 is satisfied, we first note that the observability assumption guarantees that $J^*(x) > 0$ for all $x \notin X_0$ [cf. Eq. (4.44)]. Next, we note that by the stability of the optimal closed-loop system and the controllability assumption, for every pair (x, ϵ) with $x \in X_f$ and $\epsilon > 0$, there exists a policy π that terminates starting from x and satisfies $J_\pi(x) \leq J^*(x) + \epsilon$; this is the policy that follows the optimal policy up to getting within a sufficiently small distance from the origin, and then following a minimum cost policy that drives the system to the origin (such a policy exists by the controllability assumption).

(5) Having shown that Assumption 4.2.1 is satisfied, the theory of Section 4.2 applies. In particular:

(i) From Prop. 4.2.1, it follows that J^* is the unique solution of Bellman's equation within the class of functions

$$\mathcal{J} = \big\{ J \in E^+(X) \mid J(x) = 0, \, \forall \; x \in X_0 \big\}. \tag{4.48}$$

Note that for this the controllability and observability assumptions made earlier are necessary, cf. Example 4.2.2.

(ii) From Prop. 4.2.2, it follows that VI converges to J^* starting from any function in $\mathcal{J}$ (not just starting from the identically zero function, or an arbitrary positive semidefinite quadratic function, as implied by the convergence properties of the Riccati equation, cf. Prop. 4.1.1 of Section 4.1, Vol. I).

(iii) From Props. 4.2.3 and 4.2.5, PI and its optimistic variant converge to J^*.

Policy Iteration for Undiscounted Problems

Let us consider the deterministic undiscounted linear-quadratic problem, under the controllability and observability assumptions for the pairs (A, B) and (A, C), made earlier. Since the optimal policy μ^* of Eqs. (4.46)-(4.47) belongs to the class of linear policies with a stable corresponding closed-loop system, it makes sense to try to confine the PI algorithm within that class of policies. Indeed this is possible as we will now show.

Consider first the cost evaluation of a linear policy μ of the form

$$\mu(x) = L_\mu x.$$

The corresponding cost function J_μ is the unique solution of the Bellman equation

$$\begin{aligned} J_\mu(x) &= x'Qx + (L_\mu x)'RL_\mu x + J_\mu(Ax + BL_\mu x) \\ &= x'(Q + L_\mu{}'RL_\mu)x + J_\mu\big((A + BL_\mu)x\big), \end{aligned}$$

within the class of functions $\mathcal{J}$ of Eq. (4.48). Moreover, J_μ is the limit of the VI sequence $\{T_\mu^k J_0\}$, where J_0 is the identically zero function. Using the relation

$$(T_\mu^{k+1} J_0)(x) = x'(Q + L_\mu{}' R L_\mu)x + (T_\mu^k J_0)\big((A + BL_\mu)x\big), \quad k = 0, 1, \ldots,$$

we can verify by induction that each of the functions $T_\mu^k J_0$ is quadratic of the form

$$(T_\mu^k J_0)(x) = x' P_{\mu,k} x,$$

where $P_{\mu,0} = 0$ and

$$P_{\mu,k+1} = (A + BL_\mu)' P_{\mu,k} (A + BL_\mu) + Q + L_\mu' R L_\mu, \qquad k = 0, 1, \ldots.$$

If the closed-loop system corresponding to μ is stable, the eigenvalues of the matrix $A + BL_\mu$ lie strictly within the unit circle, and the preceding iteration involves a contraction with respect to some norm, so it converges to P_μ, the unique solution of the linear equation

$$P_\mu = (A + BL_\mu)' P_\mu (A + BL_\mu) + Q + L_\mu' R L_\mu. \tag{4.49}$$

Moreover, we have

$$J_\mu(x) = x' P_\mu x, \qquad x \in \Re^n.$$

Consider next the policy improvement phase of PI, given the current policy μ. It generates an improved policy $\overline{\mu}$ by finding for each x, the control $\overline{\mu}(x)$ that minimizes over $u \in \Re^m$ the Q-factor

$$Q_\mu(x, u) = x'Qx + u'Ru + (Ax + Bu)' P_\mu (Ax + Bu). \tag{4.50}$$

By setting to zero the gradient of this expression with respect to u, we see that $\overline{\mu}$ is linear of the form

$$\overline{\mu}(x) = L_{\overline{\mu}} x, \tag{4.51}$$

where $L_{\overline{\mu}}$ is given by

$$L_{\overline{\mu}} = -(B' P_\mu B + R)^{-1} B' P_\mu A; \tag{4.52}$$

cf. Eq. (4.47). Moreover it can be seen that the closed-loop system corresponding to $\overline{\mu}$ is stable (intuitively, this is so since the closed-loop system corresponding to μ is stable, and $\overline{\mu}$ is an improved policy over μ).

In conclusion, under our assumptions, the PI algorithm, starting with a linear policy μ^0 under which the closed-loop system is stable, it generates a sequence of linear policies $\{\mu^k\}$ and corresponding gain matrices L_{μ^k} that converge to the optimal, in the sense that $J_{\mu^k}(x) \downarrow J^*(x)$ for all x (cf. Prop. 4.2.3), while $L_{\mu^k} \to L^*$.

The Discounted Case

Consider next the case where $0 < \alpha < 1$ while the problem need not be deterministic. Then we can use the theory of Section 4.1, but not the theory of Section 4.2 (which applies only to deterministic problems). However, it turns out that the results for the deterministic case derived earlier still apply in suitably modified form.

In particular, we use the VI algorithm, starting from the identically zero function J_0, to obtain the sequence $\{T^k J_0\}$. We have

$$(TJ_0)(x) = \min_{u\in\Re^m} [x'Qx + u'Ru] = x'Qx,$$

$$\begin{aligned}(T^2J_0)(x) &= \min_{u\in\Re^m} E\big\{x'Qx + u'Ru + \alpha(TJ_0)(Ax + Bu + w)\big\}\\ &= \min_{u\in\Re^m} E\big\{x'Qx + u'Ru + \alpha(Ax + Bu + w)'Q(Ax + Bu + w)\big\},\\ &= \min_{u\in\Re^m} \big[x'Qx + u'Ru + \alpha(Ax + Bu)'Q(Ax + Bu)\big] + \alpha E\{w'Qw\}.\end{aligned}$$

Proceeding similarly for every k, we obtain more generally,

$$(T^kJ_0)(x) = x'P_kx + \sum_{t=0}^{k-1} \alpha^{k-t} E\{w'P_tw\}, \qquad k = 1, 2, \ldots, \tag{4.53}$$

where

$$P_1 = Q$$

and P_{k+1} is generated from P_k using the Riccati equation

$$P_{k+1} = A'\big(\alpha P_k - \alpha^2 P_kB(\alpha B'P_kB + R)^{-1}B'P_k\big)A + Q, \qquad k = 0, 1, \ldots$$

By defining $\tilde{R} = R/\alpha$ and $\tilde{A} = \sqrt{\alpha}A$, this equation may be written as

$$P_{k+1} = \tilde{A}'\big(P_k - P_kB(B'P_kB + \tilde{R})^{-1}B'P_k\big)\tilde{A} + Q,$$

and has the Riccati equation form considered in the undiscounted case. Thus the generated matrix sequence $\{P_k\}$ converges to a positive definite symmetric matrix P^*, provided the pairs $(\tilde{A}, B)$ and $(\tilde{A}, C)$, where $Q = C'C$, are controllable and observable, respectively. Since $\tilde{A} = \sqrt{\alpha}A$, controllability and observability of (A, B) or (A, C) are clearly equivalent to controllability and observability of $(\tilde{A}, B)$ or $(\tilde{A}, C)$, respectively.

Since the compactness condition of Prop. 4.1.8 is satisfied, it follows that $\{T^kJ_0\}$, given by Eq. (4.53), converges to J^*, so that

$$J^*(x) = x'P^*x + \lim_{k\to\infty} \sum_{t=0}^{k-1} \alpha^{k-t} E\{w'P_tw\}, \qquad x \in \Re^n,$$

where the matrix P^* is the unique solution of the steady-state Riccati equation

$$P = A'\big(\alpha P - \alpha^2 PB(\alpha B'PB + R)^{-1}B'P\big)A + Q.$$

Because $P_k \to P^*$, it can also be seen that the limit

$$c = \lim_{k\to\infty} \sum_{t=0}^{k-1} \alpha^{k-t} E\{w'P_t w\}$$

is well defined, and in fact

$$c = \frac{\alpha}{1-\alpha} E\{w'P^*w\}.$$

Finally, the optimal stationary policy μ^*, obtained by minimization in Bellman's equation, has the form

$$\mu^*(x) = -\alpha(\alpha B'KB + R)^{-1}B'KAx, \qquad x \in \Re^n.$$

Moreover, similar to the case where $\alpha = 1$, the problem can be solved using the PI algorithm (see Exercise 4.16).

4.3.1 Simulation-Based Adaptive Control

We will now consider the PI algorithm of Eqs. (4.49)-(4.52), for the case where the matrices A and B of the deterministic system

$$x_{k+1} = Ax_k + Bu_k$$

are *unknown*. Instead, we will assume that we have access to a computer simulator of the real system (or perhaps the real system itself), which for any given state-control pair (x, u), can generate the successor state $Ax+Bu$. This is a problem that can be addressed with the methodology of adaptive control that we have discussed in Section 6.1 of Vol. I. In this section, however, we will follow a different approach that is based on PI.

It is of course possible to consider a two-phase procedure: first to calculate exactly the matrices A and B by various estimation methods, and then apply the PI algorithm, but this may not be the preferred option in a given practical situation. For example it may be cumbersome to separate the optimization process into system identification and control calculation phases. Moreover in practice, the matrices A and B may change as the system is controlled, so the two-phase procedure may need to be repeated.

An alternative is to execute the PI algorithm by using the simulator to evaluate the Q-factors

$$\begin{aligned} Q_\mu(x,u) &= x'Qx + u'Ru + J_\mu(Ax + Bu) \\ &= x'Qx + u'Ru + (Ax + Bu)'P_\mu(Ax + Bu), \end{aligned} \tag{4.54}$$

of the current policy μ [cf. Eq. (4.50)], and use them to calculate the improved policy $\overline{\mu}$ by the minimization

$$\overline{\mu}(x) = \arg \min_{u \in \Re^m} Q_\mu(x, u). \tag{4.55}$$

For an on-line context, with A and B changing over time, this approach is more convenient than the two-phase procedure noted above, and seems to have better stability properties in practice.

We note that Q-factors were introduced in Section 2.2.3 and will play a major role in the simulation-based approximate DP methodology to be discussed in Chapters 6 and 7. An important fact for our purposes in this section is that the Q-factors of μ determine the cost function J_μ via the equation

$$J_\mu(x) = Q_\mu\big(x, \mu(x)\big),$$

which is just Bellman's equation for J_μ. Thus the equation (4.54) can be written as

$$Q_\mu(x, u) = x'Qx + u'Ru + Q_\mu\big((Ax + Bu), \mu(x)\big). \tag{4.56}$$

We now note from Eq. (4.54) that the Q-factor of any linear policy μ is quadratic,

$$Q_\mu(x, u) = (\, x' \quad u' \,)\, K_\mu \begin{pmatrix} x \\ u \end{pmatrix},$$

where K_μ is the $(n + m) \times (n + m)$ symmetric matrix

$$K_\mu = \begin{pmatrix} Q + A'P_\mu A & A'P_\mu B \\ B'P_\mu A & R + B'P_\mu B \end{pmatrix}.$$

Thus the Q-factors $Q_\mu(x, u)$ are determined from the entries of K_μ, which as we will show next, can be estimated by using the system simulator.

In particular, let us introduce the column vector r_μ, which consists of the $(n+m)^2$ entries of the matrix K_μ. Let also $\phi(x, u)$ be the column vector whose components are all the possible products of the scalar components of x and u, i.e.,

$$x_i x_j, \quad x_i u_\ell, \quad u_\ell x_i, \quad u_\ell u_t, \qquad i, j = 1, \ldots, m, \quad \ell, t = 1, \ldots, m.$$

For example in the scalar case where $n = m = 1$,

$$\phi(x, u) = \begin{pmatrix} x^2 \\ xu \\ ux \\ u^2 \end{pmatrix}.$$

Then the Q-factors can be expressed as

$$Q_\mu(x,u) = \phi(x,u)' r_\mu.$$

By using this inner product form for Q_μ into the Bellman equation (4.56), we have for all (x,u),

$$\phi(x,u)' r_\mu = x'Qx + u'Ru + \phi\big((Ax+Bu),\mu(x)\big)' r_\mu. \tag{4.57}$$

Given μ, this is an overdetermined system of an infinite number of equations for r_μ [one equation for each pair (x,u)]. To solve this system (exactly) we can generate by simulation a sufficiently large set of triples $(x,u,Ax+Bu)$, to form a corresponding nonsingular system of finite number of equations for r_μ. Having computed r_μ, the improved policy $\overline{\mu}$ can be generated according to the minimization (4.55), and the process can be repeated to compute $r_{\overline{\mu}}$ and so on. Note that this process generates identical results with the exact PI method of Eqs. (4.49)-(4.52), which could be applied if A and B were exactly known.

In practice, the simulation-based PI method of this section can be applied in other more general settings. For example, when the results of the simulation contain errors, we may prefer to solve the system of equations (4.57) approximately by generating a large number of (noisy) samples $(x,u,Ax+Bu)$, and obtain an approximate solution using a linear least squares method. In other cases, in order to deal with the potentially large dimension of r_μ, we may replace the vector $\phi(x,u)$ by a simpler vector. We refer to the literature for examples of application of this methodology.

4.4 STOCHASTIC SHORTEST PATHS UNDER WEAK CONDITIONS

In this section we consider undiscounted total cost problems where the cost per stage can take both positive and negative values, so the methodology of the earlier sections in this chapter does not apply. We will focus on the finite-state SSP problems of Chapter 3, but under weaker conditions that do not guarantee the powerful results of that chapter. In particular, we will replace the infinite cost assumption of improper policies with the condition that J^* is real-valued. Under our assumptions, J^* need not be a solution of Bellman's equation, and even when it is, it may not be obtained by VI starting from any initial condition other than J^* itself, while the standard form of PI may not be convergent.

We will show instead that $\hat{J}$, which is the optimal cost function over proper policies only, is the unique solution of Bellman's equation within the class of functions $J \geq \hat{J}$. Moreover VI converges to $\hat{J}$ from any initial condition $J \geq \hat{J}$, while a form of PI yields a sequence of proper policies that asymptotically attains the optimal value $\hat{J}$. Results of this type are exceptional in the context of our analysis in this book, where J^* is typically shown to satisfy Bellman's equation. Our line of analysis is also unusual, and relies on a perturbation argument, which induces a more effective discrimination between proper and improper policies in terms of finiteness of their cost functions. This argument depends critically on the assumption that J^* is real-valued.

We use the notation and terminology of Chapter 3. In particular, we consider an SSP problem with states $1, \ldots, n$, plus the termination state t. The control space is denoted by U, and the set of feasible controls at state i is denoted by $U(i)$. We assume that U is a finite set, although in a more advanced treatment of the problem, we may allow U to be infinite as long as $U(i)$ satisfies the compactness conditions discussed in Section 3.2 (see also [BeY15] for this analysis). From state i under control $u \in U(i)$, a transition to state j occurs with probability $p_{ij}(u)$ and incurs an expected one-stage cost $g(i,u)$. At state t we have $p_{tt}(u) = 1$, $g(t,u) = 0$, for all $u \in U(t)$, i.e., t is absorbing and cost-free.

As in Chapter 3, we define the total cost of $\pi = \{\mu_0, \mu_1, \ldots\}$ for initial state i to be

$$J_\pi(i) = \limsup_{N\to\infty} E\left[\sum_{k=0}^{N-1} g(x_k,u_k) \;\Big|\; x_0 = i\right], \tag{4.58}$$

where the expectation is with respect to the probability law induced by π. The use of lim sup in the definition of J_π is necessary because the limit as $N \to \infty$ of the N-stage costs may not exist. However, the statements of our results, our analysis, and our algorithms are also valid if lim sup is replaced by lim inf. The optimal cost at state i, denoted $J^*(i)$, is the minimum of $J_\pi(i)$ over π. Note that in general there may exist states i such that $J_\pi(i) = \infty$ or $J_\pi(i) = -\infty$ for some policies π, as well as $J^*(i) = \infty$ or $J^*(i) = -\infty$.

Let us review the abstract notation relating to the mappings that arise in optimality conditions and algorithms. We consider the mapping $H : \{1, \ldots, n\} \times U \times \mathcal{E}^n \mapsto \mathcal{E}$ defined by

$$H(i,u,J) = g(i,u) + \sum_{j=1}^{n} p_{ij}(u)J(j), \quad i = 1, \ldots, n,\ u \in U(i),\ J \in \mathcal{E}^n, \tag{4.59}$$

where we denote by $\mathcal{E}$ the set of extended real numbers, $\mathcal{E} = \Re \cup \{\infty, -\infty\}$, and by $\mathcal{E}^n$ the set of n-dimensional vectors with extended real-valued components. When working with vectors $J \in \mathcal{E}^n$, we make sure that the sum $\infty - \infty$ never appears in our analysis. We also consider the mappings T_μ, $\mu \in \mathcal{M}$, and T defined for all $i = 1, \ldots, n$, and $J \in \mathcal{E}^n$ by

$$(T_\mu J)(i) = H\big(i, \mu(i), J\big), \qquad (TJ)(i) = \min_{u \in U(i)} H(i,u,J).$$

We will frequently use the monotonicity property of T_μ and T, i.e.,

$$J \le J' \qquad \Rightarrow \qquad T_\mu J \le T_\mu J', \;\; TJ \le TJ'.$$

The fixed point equations $J^* = TJ^*$ and $J_\mu = T_\mu J_\mu$ are the Bellman equations for the optimal cost function and for the cost function of μ, respectively.

We will now summarize the SSP analysis of Chapter 3, and preview its differences from the analysis of this section. We recall that a policy μ is said to be *proper* if when using μ, there is positive probability that the termination state t will be reached after at most n stages, regardless of the initial state; i.e., if

$$\rho_\mu = \max_{i=1,\ldots,n} P\{x_n \ne t \mid x_0 = i, \mu\} < 1.$$

Otherwise, μ is said to be *improper*. An important property is that for a proper policy μ, T_μ is a weighted sup-norm contraction. 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.

The analysis of Chapter 3 was based on the following assumption, which we will aim to weaken in this section.

Infinite Cost Conditions:

(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.

An important consequence of these conditions is that a improper policy cannot be optimal [since the costs $J_\mu(i)$ of a proper policy μ are real-valued], so by focusing the analysis on the "well-behaved" proper policies, we can exploit their contraction properties, and obtain favorable algorithmic results that allow the computation of J^* and an optimal proper policy. In the absence of the infinite cost conditions we employ a different approach in order to focus the analysis on the proper policies. We introduce

$$\hat J(i) = \min_{\mu:\,\text{proper}} J_\mu(i), \qquad i = 1, \ldots, n, \tag{4.60}$$

the optimal cost function over proper policies, and show that it is the one that can be readily computed by VI and PI (and not J^*). In particular, it turns out that *when $\hat J \ne J^*$, our VI and PI algorithms will only be able to obtain $\hat J$, and not J^*.*

The following deterministic shortest path problem, also used in Example 4.1.3 to demonstrate exceptional behavior, illustrates some of the consequences of violation of the infinite cost conditions.

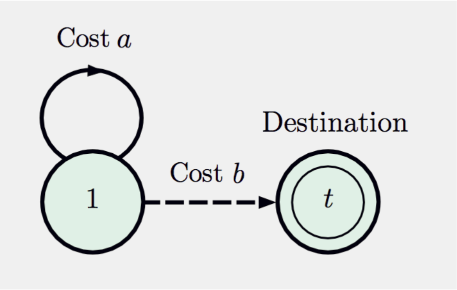


Figure 4.4.1. A deterministic shortest path problem with a single node 1 and a termination node t. At 1 there are two choices; a self-transition, which costs a, and a transition to t, which costs b.

Example 4.4.1 (Deterministic Shortest Path Problem)

Here there is a single state 1 in addition to the termination state t (cf. Fig. 4.4.1). At state 1 there are two choices: a self-transition which costs a and a transition to t, which costs b. The mapping H has the form

$$H(1,u,J) = \begin{cases} a+J & \text{if } u\text{: self transition,} \\ b & \text{if } u\text{: transition to } t, \end{cases} \qquad J \in \Re,$$

and the mapping T has the form

$$TJ = \min\{a+J,\, b\}, \qquad J \in \Re.$$

There are two policies: the policy that self-transitions at state 1, which is improper, and the policy that transitions from 1 to t, which is proper. When $a < 0$ the improper policy is optimal and we have $J^*(1) = -\infty$. The optimal cost is finite if $a > 0$ or $a = 0$, in which case the cycle has positive or zero length, respectively. Note the following:

(a) If $a > 0$, the infinite cost conditions are satisfied, and the optimal cost, $J^*(1) = b$, is the unique fixed point of T.

(b) If $a = 0$ and $b \geq 0$, the set of fixed points of T (which has the form $TJ = \min\{J,\, b\}$), is the interval $(-\infty, b]$. Here the improper policy is optimal for all $b \geq 0$, and the proper policy is also 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, $J^*(1) = 0$, lies in the interior of the set of fixed points of T, which is $(-\infty, b]$. Thus the VI method that generates $\{T^kJ\}$ 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 $b \neq J^*(1)$. 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 $(T_\mu J_\mu)(1) = b = \min\big\{J_\mu(1),\, b\big\} = (TJ_\mu)(1)$

[the other/optimal policy μ^* also satisfies $(T_{\mu^*}J_\mu)(1) = (TJ_\mu)(1)$, 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 $b < 0$, only the proper policy is optimal, and we have $J^*(1) = b$. Here it can be seen that the VI sequence $\{T^kJ\}$ converges to $J^*(1)$ 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 μ^* or the improper policy μ, the standard form of PI may oscillate, since $(T_{\mu^*}J_\mu)(1) = (TJ_\mu)(1)$ and $(T_\mu J_{\mu^*})(1) = (TJ_{\mu^*})(1)$, as can be easily verified [the optimal policy μ^* also satisfies $(T_{\mu^*}J_{\mu^*})(1) = (TJ_{\mu^*})(1)$ but it is not clear how to break the tie; compare also with case (c) above].

As we have seen in case (c) of the preceding example, VI may fail starting from $J \neq J^*$. Actually in cases (b)-(d) the one-stage costs are either all nonnegative or nonpositive, so they belong to the classes of positive and negative cost models of Section 4.1, respectively. From the results for these models (cf. Prop. 4.1.7), there is an initial condition, namely $J = 0$, starting from which VI converges to J^*. However, this is not necessarily the best initial condition; for example in deterministic shortest path problems, initial conditions $J \geq J^*$ are generally preferred and result in polynomial complexity computation assuming that all cycles have positive length. By contrast VI has only pseudopolynomial complexity when started from $J = 0$. We will also see later in this section that if there are both positive and negative one-stage costs, it may happen that J^* is not a fixed point of T, so it cannot be obtained by VI or PI.

In the next two subsections we weaken part (b) of the infinite cost conditions, by assuming that J^* is real-valued instead of requiring that each improper policy has infinite cost from some initial states. Under this assumption, we will discuss the properties of Bellman's equation, and the behavior of VI and PI. In particular, in Section 4.4.1, we show that $\hat{J}$ is the unique fixed point of T within the set $\{J \mid J \geq \hat{J}\}$, and can be computed by VI starting from any J within that set. We provide an example showing that J^* may not be a fixed point of T if g can take both positive and negative values, and the SSP problem is nondeterministic.

The idea of the analysis is to introduce an additive perturbation $\delta > 0$ to the cost of each transition. Since J^* is real-valued, the cost function of each improper policy becomes infinite for some states, thereby bringing to bear the infinite cost conditions for the perturbed problem, while the cost function of each proper policy changes by only an $O(\delta)$ amount. By considering the limit as $\delta \to 0$, this line of analysis yields results relating to the solution of Bellman's equation and the convergence of VI. The perturbation idea also applies in the context of PI, and in Section 4.4.2, we will propose a perturbed version of PI that converges to $\hat{J}$, as a replacement of the standard form of PI, which may fail as we have seen in case (d) of

Example 4.4.1.

4.4.1 A Perturbation Approach

In this section we allow the one-stage costs $g(i,u)$ to be both positive and negative, but assume that J^* is real-valued and that there exists at least one proper policy. As a result, by adding a positive perturbation δ to g, we are guaranteed to drive to ∞ the cost $J_\mu(i)$ of each improper policy μ, for at least one state i, thereby differentiating proper and improper policies.

We thus consider for each scalar $\delta > 0$ an SSP problem, referred to as the *δ-perturbed problem*, which is identical to the original problem, except that the cost per stage is

$$g_\delta(i,u) = \begin{cases} g(i,u) + \delta & \text{if } i = 1,\dots,n, \\ 0 & \text{if } i = t, \end{cases}$$

and the corresponding mappings $T_{\mu,\delta}$ and T_δ are given by

$$T_{\mu,\delta}J = T_\mu J + \delta e, \qquad T_\delta J = TJ + \delta e, \qquad \forall\ J \in \Re^n,$$

where e is the unit vector $[e(i) \equiv 1]$. This problem has the same proper and improper policies as the original. The corresponding cost function of a policy $\pi = \{\mu_0, \mu_1, \dots\} \in \Pi$ is given by

$$J_{\pi,\delta}(i) = \limsup_{N\to\infty} J_{\pi,\delta,N}(i), \qquad i = 1,\dots,n,$$

with

$$J_{\pi,\delta,N}(i) = E\left[\sum_{k=0}^{N-1} g_\delta(x_k,u_k) \;\middle|\; x_0 = i\right].$$

We denote by $J_{\mu,\delta}$ the cost function of a stationary policy μ for the δ-perturbed problem, and by J_δ^* the corresponding optimal cost function,

$$J_\delta^* = \min_{\pi\in\Pi} J_{\pi,\delta}.$$

Note that for every proper policy μ, the function $J_{\mu,\delta}$ is real-valued, and that

$$\lim_{\delta\downarrow 0} J_{\mu,\delta} = J_\mu.$$

This is because for a proper policy, the extra δ cost per stage will be incurred only a finite expected number of times prior to termination, starting from any state. This is not so for improper policies, and in fact the idea behind perturbations is that the addition of δ to the cost per stage in conjunction with the assumption that J^* is real-valued imply that if μ is improper, then

$$J_{\mu,\delta}(i) = \lim_{k\to\infty} (T_{\mu,\delta}^k J)(i) = \infty, \tag{4.61}$$

for all $i \neq t$ that are recurrent under μ and all $J \in \Re^n$. Thus part (b) of the infinite cost conditions holds for the δ-perturbed problem, and the associated strong results noted in Section 1 come into play. In particular, we have the following proposition.

Proposition 4.4.1: Assume that J^* is real-valued. Then for each $\delta > 0$:

(a) J_δ^* is the unique solution of the equation

$$J(i) = (TJ)(i) + \delta, \qquad i = 1, \ldots, n.$$

(b) A policy μ is optimal for the δ-perturbed problem ($J_{\mu,\delta} = J_\delta^*$) if and only if $T_\mu J_\delta^* = TJ_\delta^*$. Moreover, for the δ-perturbed problem, all optimal policies are proper and there exists at least one proper policy that is optimal.

(c) The optimal cost function over proper policies $\hat{J}$ [cf. Eq. (4.60)] satisfies

$$\hat{J}(i) = \lim_{\delta \downarrow 0} J_\delta^*(i), \qquad i = 1, \ldots, n.$$

(d) There exists a proper policy $\hat{\mu}$ that attains the minimum over all proper policies, i.e., $J_{\hat{\mu}} = \hat{J}$.

Proof: (a), (b) The proof of these parts follows from the discussion preceding the proposition, and the results of Chapter 3, which hold under the infinite cost conditions [the equation of part (a) is Bellman's equation for the δ-perturbed problem].

(c) For an optimal proper policy μ_δ^* of the δ-perturbed problem [cf. part (b)], we have

$$\hat{J} = \min_{\mu:\,\text{proper}} J_\mu \leq J_{\mu_\delta^*} \leq J_{\mu_\delta^*,\delta} = J_\delta^* \leq J_{\mu',\delta}, \qquad \forall\ \mu' : \text{proper}.$$

Since for every proper policy μ', we have $\lim_{\delta \downarrow 0} J_{\mu',\delta} = J_{\mu'}$, it follows that

$$\hat{J} \leq \lim_{\delta \downarrow 0} J_{\mu_\delta^*} \leq J_{\mu'}, \qquad \forall\ \mu' : \text{proper}.$$

By taking the minimum over all μ' that are proper, 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 proper policies for the δ_k-perturbed problems. Since the set of proper policies is finite, 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

$$\hat{J} \le J_{\hat{\mu}} \le J^*_{\delta_k},$$

for all k sufficiently large. Since by part (c), $J^*_{\delta_k} \downarrow \hat{J}$, it follows that $J_{\hat{\mu}} = \hat{J}$. **Q.E.D.**

Convergence of Value Iteration

The preceding perturbation-based analysis, can be used to investigate properties of $\hat{J}$ by using properties of J^*_δ and taking limit as $\delta \downarrow 0$. In particular, we use the preceding proposition to show that $\hat{J}$ is a fixed point of T, and can be obtained by VI starting from any $J \ge \hat{J}$.

Proposition 4.4.2: Assume that J^* is real-valued. Then:

(a) The optimal cost function over proper policies $\hat{J}$ is the unique fixed point of T within the set $\{J \in \Re^n \mid J \ge \hat{J}\}$.

(b) We have $T^k J \to \hat{J}$ for every $J \in \Re^n$ with $J \ge \hat{J}$.

(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 $T_\mu \hat{J} = T\hat{J}$.

Proof: (a), (b) For all proper μ, we have $J_\mu = T_\mu J_\mu \ge T_\mu \hat{J} \ge T\hat{J}$. Taking minimum over proper μ, we obtain $\hat{J} \ge T\hat{J}$. Conversely, for all $\delta > 0$ and $\mu \in \mathcal{M}$, we have

$$J^*_\delta = TJ^*_\delta + \delta e \le T_\mu J^*_\delta + \delta e.$$

Taking limit as $\delta \downarrow 0$, and using Prop. 4.4.1(c), we obtain $\hat{J} \le T_\mu \hat{J}$ for all $\mu \in \mathcal{M}$. Taking minimum over $\mu \in \mathcal{M}$, it follows that $\hat{J} \le T\hat{J}$. Thus $\hat{J}$ is a fixed point of T.

For all $J \in \Re^n$ with $J \ge \hat{J}$ and proper policies μ, we have by using the relation $\hat{J} = T\hat{J}$ just shown,

$$\hat{J} = \lim_{k\to\infty} T^k \hat{J} \le \lim_{k\to\infty} T^k J \le \lim_{k\to\infty} T^k_\mu J = J_\mu.$$

Taking the minimum over all proper μ, we obtain

$$\hat{J} \le \lim_{k\to\infty} T^k J \le \hat{J}, \qquad \forall\ J \ge \hat{J}.$$

This proves part (b) and also the claimed uniqueness property of $\hat{J}$ in part (a).

(c) If μ is a proper 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 $\lim_{k\to\infty} T_\mu^k \hat{J} = \hat{J}$. Since μ is proper, we have $J_\mu = \lim_{k\to\infty} T_\mu^k \hat{J}$, so $J_\mu = \hat{J}$. **Q.E.D.**

Note that there may exist an improper policy μ that is strictly suboptimal and yet satisfies the optimality condition $T_\mu J^* = TJ^*$ [cf. case (c) of Example 4.4.1], so properness of μ is an essential assumption in Prop. 4.4.2(c). The following proposition shows that starting from any $J \geq \hat{J}$, the convergence rate of VI to $\hat{J}$ is linear. The proposition also provides a corresponding error bound.

Proposition 4.4.3: (Convergence Rate of VI) Assume that J^* is real-valued and let $\hat{\mu}$ be a proper policy that is optimal within the class of proper policies, i.e., $J_{\hat{\mu}} = \hat{J}$ [cf. Prop. 4.4.1(d)]. Then

$$\|TJ - \hat{J}\|_v \leq \beta \|J - \hat{J}\|_v, \qquad \forall\, J \geq \hat{J}, \tag{4.62}$$

where $\|\cdot\|_v$ is a weighted sup-norm with respect to which $T_{\hat{\mu}}$ is a contraction and β is the corresponding modulus of contraction. Moreover we have

$$\|J - \hat{J}\|_v \leq \frac{1}{1-\beta} \max_{i=1,\ldots,n} \frac{J(i) - (TJ)(i)}{v(i)}, \qquad \forall\, J \geq \hat{J}. \tag{4.63}$$

Proof: By using the optimality of $\hat{\mu}$ and Prop. 4.4.2, we have $T_{\hat{\mu}}\hat{J} = T\hat{J} = \hat{J}$, so for all states i and $J \geq \hat{J}$,

$$\frac{(TJ)(i) - \hat{J}(i)}{v(i)} \leq \frac{(T_{\hat{\mu}}J)(i) - (T_{\hat{\mu}}\hat{J})(i)}{v(i)} \leq \beta \max_{i=1,\ldots,n} \frac{J(i) - \hat{J}(i)}{v(i)}.$$

By taking the maximum of the left-hand side over i, and by using the fact that the inequality $J \geq \hat{J}$ implies that $TJ \geq T\hat{J} = \hat{J}$, we obtain Eq. (4.62).

By using again the relation $T_{\hat{\mu}}\hat{J} = T\hat{J} = \hat{J}$, we have for all states i and all $J \geq \hat{J}$,

$$\begin{aligned} \frac{J(i) - \hat{J}(i)}{v(i)} &= \frac{J(i) - (TJ)(i)}{v(i)} + \frac{(TJ)(i) - \hat{J}(i)}{v(i)} \\ &\leq \frac{J(i) - (TJ)(i)}{v(i)} + \frac{(T_{\hat{\mu}}J)(i) - (T_{\hat{\mu}}\hat{J})(i)}{v(i)} \\ &\leq \frac{J(i) - (TJ)(i)}{v(i)} + \beta \|J - \hat{J}\|_v. \end{aligned}$$

By taking the maximum of both sides over i, and by using the inequality $J \geq \hat{J}$, we obtain Eq. (4.63). **Q.E.D.**

Note that if there exists a proper policy but J^* is not real-valued, the mapping T cannot have any real-valued fixed point. To see this, assume to arrive at a contradiction that $\tilde{J}$ is a real-valued fixed point, and let ϵ be a scalar such that $\tilde{J} \leq J_0 + \epsilon e$, where J_0 is the zero vector. Since $T(J_0+\epsilon e) \leq TJ_0+\epsilon e$, it follows that $\tilde{J} = T^N\tilde{J} \leq T^N(J_0+\epsilon e) \leq T^N J_0+\epsilon e \leq J_{\pi,N}+\epsilon e$ for any N and policy π. Taking lim sup with respect to N and then minimum over π, it follows that $\tilde{J} \leq J^* + \epsilon e$. Hence $J^*(i)$ cannot take the value $-\infty$ for any state i. Since $J^*(i)$ also cannot take the value ∞ (by the existence of a proper policy), this shows that J^* must be real-valued - a contradiction.

Another important special case where favorable results hold is when $g(i,u) \leq 0$ for all (i,u). Then from the analysis of Section 4.1, we know that J^* is the unique fixed point of T within the set $\{J \mid J^* \leq J \leq 0\}$, and the VI sequence $\{T^kJ\}$ converges to J^* starting from any J within that set. In the following proposition, we will use Prop. 4.4.2 to obtain related results for SSP problems where g may take both positive and negative values. An example is an optimal stopping problem, where at each state i we have cost $g(i,u) \geq 0$ for all u except one that leads to the termination state t with nonpositive cost. Classical problems of this type include searching among several sites for a valuable object, with nonnegative search costs and nonpositive stopping costs (stopping the search at every site is a proper policy guaranteeing that $\hat{J} \leq 0$).

Proposition 4.4.4: Assume that $\hat{J} \leq 0$, and that J^* is real-valued. Then J^* is equal to $\hat{J}$ and it is the unique fixed point of T within the set $\{J \in \Re^n \mid J \geq J^*\}$. Moreover, we have $T^kJ \to J^*$ for every $J \in \Re^n$ with $J \geq J^*$.

Proof: We first observe that the hypothesis $\hat{J} \leq 0$ implies that there exists at least one proper policy, so Prop. 4.4.2 applies, and shows that $\hat{J}$ is the unique fixed point of T within the set $\{J \in \Re^n \mid J \geq \hat{J}\}$ and that $T^kJ \to \hat{J}$ for all $J \in \Re^n$ with $J \geq \hat{J}$. We will prove the result by showing that $\hat{J} = J^*$. Since generically we have $\hat{J} \geq J^*$, it will suffice to show the reverse inequality.

Let J_0 denote the zero function. Since $\hat{J}$ is a fixed point of T and $\hat{J} \leq J_0$, we have

$$\hat{J} = \lim_{k\to\infty} T^k\hat{J} \leq \limsup_{k\to\infty} T^kJ_0. \tag{4.64}$$

Also, for each policy $\pi = \{\mu_0, \mu_1, \ldots\}$, we have

$$J_\pi = \limsup_{k \to \infty} T_{\mu_0} \cdots T_{\mu_{k-1}} J_0.$$

Since

$$T^k J_0 \le T_{\mu_0} \cdots T_{\mu_{k-1}} J_0, \qquad \forall\ k \ge 0,$$

it follows that $\limsup_{k\to\infty} T^k J_0 \le J_\pi$, so by taking the minimum over π, we have

$$\limsup_{k \to \infty} T^k J_0 \le J^*. \tag{4.65}$$

Combining Eqs. (4.64) and (4.65), it follows that $\hat{J} \le J^*$, so that $\hat{J} = J^*$. **Q.E.D.**

Finally, let us address the question of finding a function $J \ge \hat{J}$ with which to start VI. One possibility that may work is to use the cost function of a proper policy or an upper bound thereof. For example in a stopping problem we may use the cost function of the policy that stops at every state. More generally we may try to introduce an artificial high stopping cost, which was our approach in Section 4.1.4. If it can be guaranteed that $\hat{J} = J^*$ this approach will also yield J^*.

On the Solutions of Bellman's Equation

While $\hat{J}$ is a fixed point of T under our assumptions, as shown in Prop. 4.4.2(a), an interesting question is whether and under what conditions J^* is also a fixed point of T. It turns out that this is so in the special case of a deterministic shortest path problem, i.e., an SSP problem where for each i and $u \in U(i)$, there is a unique successor state denoted $f(i, u)$. For such a problem, the mapping T_μ takes the form

$$(T_\mu J)(i) = g\big(i, \mu(i)\big) + J\big(f(i, \mu(i))\big).$$

Moreover, by using the definition (4.58) of J_μ in terms of lim sup, we have for all μ (proper or improper),

$$J_\mu(i) = g\big(i, \mu(i)\big) + J_\mu\big(f(i, \mu(i))\big) = (T_\mu J_\mu)(i), \qquad i = 1, \ldots, n.$$

Indeed, for any policy $\pi = \{\mu_0, \mu_1, \ldots\}$, using the definition (4.58) of J_π in terms of lim sup, we have for all i,

$$J_\pi(i) = g\big(i, \mu_0(i)\big) + J_{\pi_1}\big(f(i, \mu_0(i))\big), \tag{4.66}$$

where $\pi_1 = \{\mu_1, \mu_2, \ldots\}$. By taking the minimum of the left-hand side over π, and the minimum of the right-hand side over π_1 and then over μ_0, we obtain $J^* = TJ^*$. Note that this argument does not require any assumptions other than the deterministic character of the problem, and holds even if the state space is infinite.

However, the following example shows that if $g(i,u)$ can take both positive and negative values, and the problem is stochastic, J^* may not be a fixed point of T. Moreover, J_μ need not be a fixed point of T_μ for an improper policy μ.

Example 4.4.2 (A Problem Where J^* is not a Fixed Point of T [BeY15])

Consider the SSP problem of Fig. 4.4.2, which involves an improper policy μ, whose transitions are marked by 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 3 or 5 with equal probability 1/2. Under the definition (4.58) of J_μ in terms of lim sup, it can be verified that

$$J_\mu(1) = 0, \quad J_\mu(2) = J_\mu(5) = 1, \quad J_\mu(3) = J_\mu(7) = 0, \quad J_\mu(4) = J_\mu(6) = 2,$$

so that the Bellman equation at state 1,

$$J_\mu(1) = \frac{1}{2}\big(J_\mu(2) + J_\mu(5)\big),$$

is not satisfied. Thus J_μ is not a fixed point of T_μ. If for $i = 1,4,7$, we introduce another control that leads from i to t with a cost $c > 2$, we create a proper policy that is strictly suboptimal, while not affecting J^*, which again is not a fixed point of T.

The mathematical reason why Bellman's equation $J_\mu = T_\mu J_\mu$ may not hold for stochastic problems and improper μ (cf. Example 4.4.2) is that lim sup may not commute with the expected value that is inherent in T_μ, and the proof argument given for deterministic problems breaks down.

The improper policy of Example 4.4.2 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 in Example 4.4.2 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 of Example 4.4.2, the optimal cost function ceases to be a solution of Bellman's equation and simultaneously the optimal cost at state 1 is improved!

4.4.2 A Policy Iteration Algorithm with Perturbations

We will now use our perturbation framework to deal with the oscillatory behavior of PI, which is illustrated in case (d) of Example 4.4.1. We will

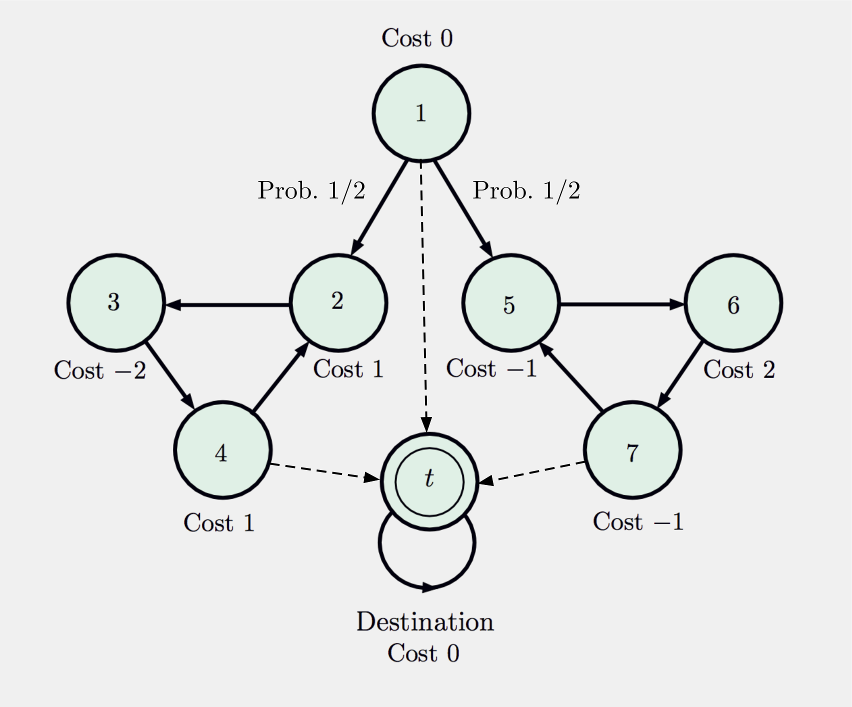


Figure 4.4.2. An example of an improper policy μ, where J_μ is not a fixed point of T_μ. All transitions under μ are shown by solid lines. These transitions are deterministic, except at state 1 where the next state is 2 or 5 with equal probability 1/2. There are additional high cost transitions from nodes 1, 4, and 7 to the destination, which create a suboptimal proper policy. We have $J^* = J_\mu$ and J^* is not a fixed point of T.

develop a perturbed version of the PI algorithm that generates a sequence of proper policies $\{\mu^k\}$ such that $J_{\mu^k} \to \hat{J}$, under the assumptions of Prop. 4.4.2, which include the existence of a proper policy and that J^* is real-valued. 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

$$T_{\mu^{k+1}} J_{\mu^k,\delta_k} = T J_{\mu^k,\delta_k}. \tag{4.67}$$

Note that since μ^k is proper, J_{μ^k,δ_k} is 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. (4.67) exists by the finiteness of the control space. We claim that μ^{k+1} is proper. To see this, note that

$$T_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k} = T J_{\mu^k,\delta_k} + \delta_k\, e \le T_{\mu^k} J_{\mu^k,\delta_k} + \delta_k\, e = J_{\mu^k,\delta_k},$$

so that

$$T^m_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k} \le T_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k} = TJ_{\mu^k,\delta_k} + \delta_k\, e \le J_{\mu^k,\delta_k}, \qquad \forall\ m \ge 1. \tag{4.68}$$

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 $(T^m_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k})(i) \to \infty$ for some i; cf. Eq. (4.61)]. Thus the sequence $\{\mu^k\}$ generated by the perturbed PI algorithm (4.67) is well-defined and consists of proper policies. We have the following proposition.

> **Proposition 4.4.5:** Assume that there exists at least one proper policy, and that J^* is real-valued. Then the sequence $\{J_{\mu^k}\}$ generated by the perturbed PI algorithm (4.67) satisfies $J_{\mu^k} \to \hat J$.

Proof: Using Eq. (4.68), we have

$$J_{\mu^{k+1},\delta_{k+1}} \le J_{\mu^{k+1},\delta_k} = \lim_{m\to\infty} T^m_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k} \le TJ_{\mu^k,\delta_k} + \delta_k\, e \le J_{\mu^k,\delta_k},$$

where the equality holds because μ^{k+1} is proper, as shown earlier. Taking the limit as $k \to \infty$, and noting that $J_{\mu^{k+1},\delta_{k+1}} \ge \hat J$, we see that $J_{\mu^k,\delta_k} \downarrow J^+$ for some $J^+ \ge \hat J$, and we obtain

$$\hat J \le J^+ = \lim_{k\to\infty} TJ_{\mu^k,\delta_k}. \tag{4.69}$$

We also have

$$\begin{aligned} \min_{u\in U(i)} H(i,u,J^+) &\le \lim_{k\to\infty} \min_{u\in U(i)} H(i,u,J_{\mu^k,\delta_k}) \\ &\le \min_{u\in U(i)} \lim_{k\to\infty} H(i,u,J_{\mu^k,\delta_k}) \\ &= \min_{u\in U(i)} H\big(i,u,\lim_{k\to\infty} J_{\mu^k,\delta_k}\big) \\ &= \min_{u\in U(i)} H(i,u,J^+), \end{aligned}$$

where the first inequality follows from the fact $J^+ \le J_{\mu^k,\delta_k}$, which implies that $H(i,u,J^+) \le H(i,u,J_{\mu^k,\delta_k})$, and the first equality follows from the continuity of $H(i,u,\cdot)$. Thus equality holds throughout above, so that

$$\lim_{k\to\infty} TJ_{\mu^k,\delta_k} = TJ^+. \tag{4.70}$$

Combining Eqs. (4.69) and (4.70), we obtain $\hat J \le J^+ = TJ^+$. Since by Prop. 4.4.2, $\hat J$ is the unique fixed point of T within $\{J \in \Re^n \mid J \ge \hat J\}$, it follows that $J^+ = \hat{J}$. Thus $J_{\mu^k,\delta_k} \downarrow \hat{J}$, and since $J_{\mu^k,\delta_k} \geq J_{\mu^k} \geq \hat{J}$, we have $J_{\mu^k} \to \hat{J}$. **Q.E.D.**

Proposition 4.4.5 guarantees the monotonic convergence $J_{\mu^k,\delta_k} \downarrow \hat{J}$ (see the preceding proof) and the (possibly nonmonotonic) convergence $J_{\mu^k} \to \hat{J}$. Moreover, Prop. 4.4.5 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 i such that $J_\mu(i) \geq J^*(i) + \epsilon$. This convergence behavior should be contrasted with the behavior of PI without perturbations, which may lead to difficulties, as noted earlier [cf. case (d) of Example 4.4.1].

4.5 AFFINE MONOTONIC AND RISK-SENSITIVE MODELS

In this section we discuss a broad class of DP models, called *affine monotonic*, which are in some ways similar to the abstract models of Sections 1.6 and 2.5. The main difference is that instead of involving a contractive mapping, they involve an affine mapping that is monotone but may not be contractive. Affine monotonic models are also similar to the SSP models of Chapter 3 and Section 4.4. The main similarity is that in both models, there is a special class of policies, which is well-behaved with respect to VI and plays a critical role: in SSP, it is the class of proper policies, while in affine monotonic, it is the class of stable policies, which are the policies whose affine mapping is a contraction; see the subsequent definition.

At an abstract mathematical level, SSP and affine monotonic models are very similar. We assume a finite state space $X = \{1, \ldots, n\}$, and a control constraint set $U(i) \subset U$ for each state i. We replace the SSP mapping H of Eq. (4.59) with

$$H(i,u,J) = b(i,u) + \sum_{j=1}^{n} A_{ij}(u)J(j),$$

where $b(i,u)$ and $A_{ij}(u)$ are given scalars, with

$$b(i,u) \geq 0, \qquad A_{ij}(u) \geq 0, \qquad \forall\, i,j = 1,\ldots,n,\ u \in U(i). \tag{4.71}$$

Thus $b(i,u)$ replaces the cost per stage $g(i,u)$ and $A_{ij}(u)$ replaces the transition probability $p_{ij}(u)$. The Bellman equation in affine monotonic models has the form

$$J(i) = \min_{u\in U(i)} H(i,u,J) = \min_{u\in U(i)} \left[b(i,u) + \sum_{j=1}^{n} A_{ij}(u)J(j) \right], \quad j = 1,\ldots,n. \tag{4.72}$$

We are interested in solutions of Bellman's equation (4.72) within the set of vectors with nonnegative extended real-valued components, which we denote by

$$\mathcal{E}_+^n = \big\{J \mid 0 \leq J(i) \leq \infty,\ i = 1,\ldots,n\big\}.$$

We will also focus on solutions of Bellman's equation within $\Re_+^n$, the set of vectors with nonnegative real-valued components,

$$\Re_+^n = \big\{J \mid 0 \leq J(i) < \infty,\ i = 1,\ldots,n\big\}.$$

As earlier, we denote by $\mathcal{M}$ the set of all functions $\mu : X \mapsto U$ with $\mu(i) \in U(i)$, for all $i \in X$, and we consider policies, which are sequences $\pi = \{\mu_0, \mu_1, \ldots\}$, with $\mu_k \in \mathcal{M}$ for all k. We denote by Π the set of all policies. Moreover, we also refer to any $\mu \in \mathcal{M}$ as a "policy" and use it in place of the stationary policy $\{\mu, \mu, \ldots\}$, when confusion cannot arise.

To formulate the DP model that corresponds to the Bellman equation (4.72), we introduce for each $\mu \in \mathcal{M}$ the mapping $T_\mu : \mathcal{E}_+^n \mapsto \mathcal{E}_+^n$ given by

$$T_\mu J = b_\mu + A_\mu J, \tag{4.73}$$

where b_μ is the vector of $\Re^n$ with components $b\big(i, \mu(i)\big)$, $i = 1,\ldots,n$, and A_μ is the $n \times n$ matrix with components $A_{ij}\big(\mu(i)\big)$, $i, j = 1,\ldots,n$.

We define the mapping $T : \mathcal{E}_+^n \mapsto \mathcal{E}_+^n$, where for each $J \in \mathcal{E}_+^n$, TJ is the vector of $\mathcal{E}_+^n$ with components

$$(TJ)(i) = \min_{\mu \in \mathcal{M}} (T_\mu J)(i), \qquad i = 1,\ldots,n,$$

where $\mathcal{M}$ is the set of stationary policies, or equivalently

$$(TJ)(i) = \min_{u \in U(i)} \left[b(i,u) + \sum_{j=1}^n A_{ij}(u) J(j) \right], \qquad i = 1,\ldots,n. \tag{4.74}$$

We now define a DP-like optimization problem that involves the mappings T_μ. We introduce a special vector $\bar{J} \in \Re_+^n$, and we define the cost function of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ by

$$J_\pi(i) = \limsup_{N \to \infty} (T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(i), \qquad i = 1,\ldots,n.$$

(We use lim sup because we are not assured that the limit exists; the analysis remains unchanged if lim sup is replaced by lim inf.) This definition bears similarity with the one of Section 1.6.1 for contractive abstract DP models, except that in the latter definition the choice of $\bar{J}$ is largely immaterial thanks to the contraction property. The optimal cost function J^* is defined by

$$J^*(i) = \min_{\pi \in \Pi} J_\pi(i), \qquad i = 1,\ldots,n,$$

where Π denotes the set of all policies. We wish to find J^* and a policy $\pi^* \in \Pi$ that is optimal, i.e., $J_{\pi^*} = J^*$. Note that by using the definition

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

we can verify by induction that for every $\pi = \{\mu_0, \mu_1, \ldots\}$, we have

$$T_{\mu_0} \cdots T_{\mu_{N-1}} \bar J = A_{\mu_0} \cdots A_{\mu_{N-1}} \bar J + b_{\mu_0} + \sum_{k=1}^{N-1} A_{\mu_0} \cdots A_{\mu_{k-1}} b_{\mu_k}, \quad (4.75)$$

so that

$$J_\pi = \limsup_{N\to\infty} \left(A_{\mu_0} \cdots A_{\mu_{N-1}} \bar J + b_{\mu_0} + \sum_{k=1}^{N-1} A_{\mu_0} \cdots A_{\mu_{k-1}} b_{\mu_k} \right).$$

The preceding affine monotonic optimization problem was introduced in the monograph [Ber13], Section 4.5, as a special class of abstract DP models with a broad variety of applications. The development of [Ber13] involves arbitrary state and control spaces, but similar to SSP problems, the strongest results are the ones obtained for the finite state space case of this section. Even stronger results can be obtained in the special cases where the additional assumption that $T_\mu \bar J \geq \bar J$ for all policies $\mu \in \mathcal{M}$ or the additional assumption that $T_\mu \bar J \leq \bar J$ for all policies $\mu \in \mathcal{M}$. These are the so called *monotone increasing* and *monotone decreasing* cases (see the exercises and [Ber13]).

Clearly, finite-state sequential stochastic control problems under Assumption P, and SSP problems with nonnegative cost per stage (cf. Chapter 3, and Sections 4.1, 4.4) are special cases of affine monotonic models where $\bar J$ is the identically zero function $[\bar J(i) \equiv 0]$. Also, variants of the stochastic control problem of Section 4.1 under Assumption P, which involve state and control-dependent discount factors (for example semi-Markov problems, cf. Section 7.5 of Vol. I), are special cases of the affine monotonic model, with the discount factors being absorbed within the scalars $A_{ij}(u)$. In all of these cases, A_μ is a substochastic matrix. There are also other special cases, where A_μ is not substochastic. They correspond to interesting classes of practical problems, including SSP-type problems involving a multiplicative or an exponential (rather than additive) cost function, which we proceed to discuss.

Multiplicative and Exponential Cost SSP Problems

To describe a type of SSP problem, which is different than the one considered in Chapter 3 and Section 4.4, let us introduce in addition to the states $i = 1, \ldots, n$, a cost-free and absorbing state t. As in SSP, there are probabilistic state transitions among the states $i = 1, \ldots, n$, up to the first time a transition to state t occurs, in which case the state transitions terminate. We denote by $p_{it}(u)$ and $p_{ij}(u)$ the probabilities of transition from i to t and to j under u, respectively, so that

$$p_{it}(u)+\sum_{j=1}^{n}p_{ij}(u)=1, \qquad i=1,\ldots,n,\ u\in U(i).$$

Then we define the scalars $A_{ij}(u)$ and $b(i,u)$ by

$$A_{ij}(u)=p_{ij}(u)h(i,u,j), \qquad i,j=1,\ldots,n,\ u\in U(i), \tag{4.76}$$

and

$$b(i,u)=p_{it}(u)h(i,u,t), \quad i=1,\ldots,n,\ u\in U(i), \tag{4.77}$$

where the scalars $h(i,u,t)$ and $h(i,u,j)$ are nonnegative,

$$h(i,u,t)\ge 0, \qquad h(i,u,j)\ge 0, \qquad \forall\ i,j=1,\ldots,n,\ u\in U(i).$$

By letting

$$\bar J(i)=1, \qquad i=1,\ldots,n,$$

we obtain an affine monotonic problem. The cost function of this problem has a multiplicative character, as we show next.

Indeed, with the preceding definitions of $A_{ij}(u)$, $b(i,u)$, and $\bar J$, we will prove that the expression for the cost function of a policy $\pi=\{\mu_0,\mu_1,\ldots\}$,

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

can be written in the multiplicative form

$$J_\pi(x_0)=\limsup_{N\to\infty}\,E\left\{\prod_{k=0}^{N-1}h\big(x_k,\mu_k(x_k),x_{k+1}\big)\right\}, \qquad x_0=1,\ldots,n, \tag{4.78}$$

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 (4.78) for J_π, we write the N-stage cost vector

$$T_{\mu_0}\cdots T_{\mu_{N-1}}\bar{J}$$

as the sum

$$A_{\mu_0}\cdots A_{\mu_{N-1}}\bar{J}+b_{\mu_0}+\sum_{k=1}^{N-1}A_{\mu_0}\cdots A_{\mu_{k-1}}b_{\mu_k}, \tag{4.79}$$

[cf. Eq. (4.75)], and we interpret the ith component of each term in the sum as a conditional expected value of the expression

$$\prod_{k=0}^{N-1}h\big(x_k,\mu_k(x_k),x_{k+1}\big) \tag{4.80}$$

multiplied with the appropriate conditional probability. In particular:

(a) The ith component of the term $A_{\mu_0}\cdots A_{\mu_{N-1}}\bar{J}$ in Eq. (4.79) is the conditional expected value of the expression (4.80), 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 term b_{μ_0} in Eq. (4.79) is the conditional expected value of the expression (4.80), 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 term $A_{\mu_0}\cdots A_{\mu_{k-1}}b_{\mu_k}$ in Eq. (4.79) is the conditional expected value of the expression (4.80), 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 (4.78).

A special case of multiplicative SSP is the *risk-sensitive SSP problem with exponential cost function*, where for all $i=1,\ldots,n$, and $u\in U(i)$,

$$h(i,u,t)=\exp\big(g(i,u,t)\big),\qquad h(i,u,j)=\exp\big(g(i,u,j)\big), \tag{4.81}$$

and the function g can take both positive and negative values. The Bellman equation for this problem is

$$J(i) = \min_{u\in U(i)} \left[p_{it}(u)\mathrm{exp}\big(g(i,u,t)\big) + \sum_{j=1}^{n} p_{ij}(u)\mathrm{exp}\big(g(i,u,j)\big)J(j) \right], \tag{4.82}$$

for all $i = 1,\dots,n$. The exponential here provides the risk sensitivity, by assigning a far larger (nonlinear) penalty for large rather than small cost of a trajectory up to termination.

The deterministic version of this problem where for each $u \in U(i)$, only one of the transition probabilities $p_{it}(u), p_{i1}(u),\dots,p_{in}(u)$ 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). Some of the deterministic shortest path examples given earlier in this chapter to illustrate various pathological situations, such as multiplicity of fixed points of T, and failures of the VI and PI algorithms, can be translated to examples within the context of the risk-sensitive SSP problem with exponential cost (see the subsequent Example 4.5.1).

Stable Policies

The subsequent analysis of this section has much in common with the analysis of SSP problems under the assumptions of Chapter 3 and under the weaker assumptions of Section 4.4. The key is a generalization of the fundamental notion of a proper policy within the affine monotonic model, which we introduce next.

We say that a given stationary policy μ is *stable if $A_\mu^N \to 0$ as $N \to \infty$*. Equivalently, μ is stable if all the eigenvalues of A_μ lie strictly within the unit circle. Otherwise, μ is called *unstable*. As noted in Section 1.5, a policy μ is stable if and only if T_μ is a contraction with respect to some norm. Because $A_\mu \geq 0$, an equivalent statement is that μ is stable if and only if A_μ is a contraction with respect to some weighted sup-norm (see the discussion in Section 1.5.1, and [BeT89], Ch. 2, Cor. 6.2). We next derive an expression for the cost function of stable and unstable policies.

By repeatedly applying the equation $T_\mu J = b_\mu + A_\mu J$, we have

$$T_\mu^N J = A_\mu^N J + \sum_{k=0}^{N-1} A_\mu^k b_\mu, \qquad \forall\ J \in \Re^n,\ N = 1,2,\dots,$$

and hence

$$J_\mu = \limsup_{N\to\infty} T_\mu^N \bar J = \limsup_{N\to\infty} A_\mu^N \bar J + \sum_{k=0}^{\infty} A_\mu^k b_\mu. \tag{4.83}$$

From these expressions, it follows that if μ is stable, the initial function $\bar{J}$ in the definition of J_μ does not matter, and we have

$$J_\mu = \limsup_{N\to\infty} T_\mu^N J = \limsup_{N\to\infty} \sum_{k=0}^{N-1} A_\mu^k b_\mu, \qquad \forall\ \mu\text{: stable},\ J \in \Re^n.$$

Moreover, since for a stable μ, T_μ is a contraction with respect to some norm, the lim sup above can be replaced by lim, so that

$$J_\mu = \sum_{k=0}^{\infty} A_\mu^k b_\mu = (I - A_\mu)^{-1} b_\mu, \qquad \forall\ \mu\text{: stable.} \tag{4.84}$$

Thus if μ is stable, J_μ is real-valued as well as nonnegative. If μ is unstable, we have $J_\mu \in \mathcal{E}_+^n$ and it is possible that for some states i, $J_\mu(i) = \infty$. Note the analogy of proper and improper policies in SSP with stable and unstable policies in the affine monotonic context. In our analysis we will assume the following, which parallels Assumption 3.1.1 for SSP.

Assumption 4.5.1: There exists at least one stable policy.

Control Space Compactness

In our analysis of SSP in Chapter 3 and Section 4.4, we assumed that the control space is finite, although we noted in Section 3.2 that our basic results extend to the case of a compact control constraint set (assuming continuity of the cost per stage and the transition probabilities with respect to u). However, this extension requires a complicated proof for which we referred to [Ber91a]. It turns out that in the case of an affine monotonic model, the analysis is facilitated by the nonnegativity of the cost per stage, and the attendant lower boundedness of the optimal cost function. We will thus be able to give relatively simple proofs of our basic results while allowing an infinite control space, under the following assumption.

Assumption 4.5.2: (Compactness) The control space U is a metric space, and $p_{ij}(\cdot)$ and $g(i,\cdot)$ are continuous functions of u over $U(i)$, for all i and j. Moreover, for each state i, the sets

$$\left\{ u \in U(i) \;\middle|\; b(i,u) + \sum_{j=1}^{n} A_{ij}(u) J(j) \le \lambda \right\}$$

are compact subsets of U for all $\lambda \in \Re$ and $J \in \Re^n_+$.

Note that the preceding assumption is satisfied if the control space U is finite. One way to see this is to simply identify each $u \in U$ with a distinct integer from the real line. Another interesting case where the assumption is satisfied is when for all i, $U(i)$ is a compact subset of the metric space U, and the functions $b(i,\cdot)$ and $A_{ij}(\cdot)$ are continuous functions of u over $U(i)$ (cf. Assumption 3.2.1 in Section 3.2).

An advantage of allowing $U(i)$ to be infinite and compact is that it allows randomized policies for problems where there is a *finite* set of feasible actions at each state i, call it $C(i)$. We may then specify $U(i)$ to be the set of all probability distributions over $C(i)$, which is a compact subset of a Euclidean space. In this way, our results apply to finite-state and finite-action problems where randomization is allowed, and J^* is the optimal cost function over all randomized nonstationary policies. Note, however, that the optimal cost function may change when randomized policies are introduced in this way, as Example 4.4.2 shows. Basically, for our purposes, optimization over nonrandomized and over randomized policies over finite action sets $C(i)$ are two different problems, both of which are interesting and can be addressed with the methodology of this section. However, when the sets $C(i)$ are infinite, a different and mathematically more sophisticated framework is required in order to allow randomized nonstationary Markov policies. The reason is that randomized policies over the infinite action sets $C(i)$ must obey measurability restrictions, such as universal measurability, and a related mathematical formulation and analysis (see Appendix A).

Similar to earlier sections, the preceding compactness assumption guarantees some important properties of the mapping T. These are summarized in the following proposition, which bears a close relation to the results on convergence of VI under Assumption P (Props. 4.1.7 and 4.1.8).

Proposition 4.5.1: Let Assumptions 4.5.1 and 4.5.2 hold.

(a) The minimum of the expression

$$\min_{u\in U(i)} \left[b(i,u) + \sum_{j=1}^n A_{ij}(u)J(j) \right], \tag{4.85}$$

is attained for all $J \in \mathcal{E}^n_+$ and $i = 1,\ldots,n$.

(b) Let J_0 be the zero vector in $\Re^n$ [$J_0(i) \equiv 0$]. The VI sequence $\{T^k J_0\}$ is monotonically nondecreasing and converges to a limit $J_\infty \in \Re^n_+$ that satisfies $J_\infty \le J^*$ and $J_\infty = TJ_\infty$.

Proof: (a) If the expression (4.85) is equal to ∞, the minimum is attained for all $u \in U(i)$, so assume otherwise. The set of minima in Eq. (4.85) is the intersection $\cap_{m=1}^\infty U_m$ of the nested sequence of sets

$$U_m = \left\{ u \in U(i) \;\middle|\; b(i,u) + \sum_{j=1}^n A_{ij}(u)J(j) \le \lambda_m \right\}, \qquad m = 1, 2, \ldots,$$

where $\{\lambda_m\}$ is a monotonically decreasing scalar sequence such that

$$\lambda_m \downarrow \min_{u\in U(i)} \left[b(i,u) + \sum_{j=1}^n A_{ij}(u)J(j) \right].$$

Each set U_m is nonempty, and by Assumption 4.5.2, it is compact, so the intersection is nonempty (cf. the discussion preceding Prop. 4.1.8).

(b) By the nonnegativity of $b(i,u)$ and $A_{ij}(u)$, we have $J_0 \le TJ_0$, which by the monotonicity of T implies that $\{T^k J_0\}$ is monotonically nondecreasing to a limit $J_\infty \in \mathcal{E}^n_+$, and we have

$$J_0 \le TJ_0 \le \cdots \le T^k J_0 \le \cdots \le J_\infty. \tag{4.86}$$

For all policies $\pi = \{\mu_0, \mu_1, \ldots\}$, we have $T^k J_0 \le T^k \bar{J} \le T_{\mu_0} \cdots T_{\mu_{k-1}} \bar{J}$, so by taking the limit as $k \to \infty$, we obtain $J_\infty \le J_\pi$, and by taking the minimum over π, it follows that $J_\infty \le J^*$. By Assumption 4.5.1, there exists at least one stable policy μ, for which J_μ is real-valued [cf. Eq. (4.84)], so $J^* \in \Re^n_+$. It follows that the VI sequence $\{T^k J_0\}$ consists of vectors in $\Re^n_+$.

By applying T to both sides of Eq. (4.86), we obtain

$$(T^{k+1}J_0)(i) = \min_{u\in U(i)} \left[b(i,u) + \sum_{j=1}^n A_{ij}(u)(T^k J_0)(j) \right] \le (TJ_\infty)(i),$$

and by taking the limit as $k \to \infty$, it follows that $J_\infty \le TJ_\infty$. Assume to arrive at a contradiction that there exists a state $\tilde{i}$ such that

$$J_\infty(\tilde{i}) < (TJ_\infty)(\tilde{i}). \tag{4.87}$$

Consider the sets

$$U_k(\tilde{i}) = \left\{ u \in U(\tilde{i}) \;\middle|\; b(\tilde{i}, u) + \sum_{j=1}^{n} A_{\tilde{i}j}(u)(T^k J_0)(j) \le J_\infty(\tilde{i}) \right\},$$

for $k \ge 0$. Since by Eq. (4.87) we have $J_\infty(\bar{i}) < \infty$, it follows by Assumption 4.5.2 and Eq. (4.86) that $\big\{U_k(\tilde{i})\big\}$ is a nested sequence of compact sets. Let also u_k be a control attaining the minimum in

$$\min_{u \in U(\tilde{i})} \left[b(\tilde{i}, u) + \sum_{j=1}^{n} A_{\tilde{i}j}(u)(T^k J_0)(j) \right];$$

[such a point exists by part (a)]. From Eq. (4.86), it follows that for all $m \ge k$,

$$b(\tilde{i}, u_m) + \sum_{j=1}^{n} A_{\tilde{i}j}(u_m)(T^k J_0)(j) \le b(\tilde{i}, u_m) + \sum_{j=1}^{n} A_{\tilde{i}j}(u_m)(T^m J_0)(j) \le J_\infty(\tilde{i}).$$

Therefore $\{u_m\}_{m=k}^{\infty} \subset U_k(\tilde{i})$, and since $U_k(\tilde{i})$ is compact, all the limit points of $\{u_m\}_{m=k}^{\infty}$ belong to $U_k(\tilde{i})$ and at least one such limit point exists. Hence the same is true of the limit points of the entire sequence $\{u_m\}_{m=0}^{\infty}$. It follows that if $\tilde{u}$ is a limit point of $\{u_m\}_{m=0}^{\infty}$ then

$$\tilde{u} \in \cap_{k=0}^{\infty} U_k(\tilde{i}).$$

This implies that for all $k \ge 0$

$$(T^{k+1} J_0)(\tilde{i}) \le b(\tilde{i}, \tilde{u}) + \sum_{j=1}^{n} A_{\bar{i}j}(\tilde{u})(T^k J_0)(j) \le J_\infty(\tilde{i}).$$

By taking the limit in this relation as $k \to \infty$, we obtain

$$J_\infty(\tilde{i}) = b(\tilde{i}, \tilde{u}) + \sum_{j=1}^{n} A_{\bar{i}j}(\tilde{u}) J_\infty(j).$$

Since the right-hand side is greater than or equal to $(T J_\infty)(\tilde{i})$, Eq. (4.87) is contradicted, implying that $J_\infty = T J_\infty$. **Q.E.D.**

Analysis When Unstable Policies Have Infinite Cost

We now turn to questions relating to Bellman's equation, VI convergence, and optimality conditions, similar to the ones we have addressed for SSP problems. The mappings T_μ are similar in the SSP and affine monotonic cases, but there are two mathematical differences that affect the analysis, and play a significant role in the subsequent proofs.

(a) In SSP problems, the vector b_μ can have negative components, so for some initial states the optimal cost may be $-\infty$.

(b) In affine monotonic problems, A_μ need not be nonexpansive, so for some $J \in \Re^n_+$, $A_\mu^N J$ may become unbounded as $N \to \infty$.

We will consider the following assumption, which parallels Assumptions 3.1.1 and 3.1.2 in Section 3.1 for SSP.

Assumption 4.5.3: (Infinite Cost Condition) For every unstable policy μ, there is at least one state such that the corresponding components of the vector $\sum_{k=0}^\infty A_\mu^k b_\mu$ is equal to ∞.

Note that the preceding assumption guarantees that for every unstable policy μ, we have $J_\mu(i) = \infty$ for at least one state i [cf. Eq. (4.83)]. The reverse is not true, however: $J_\mu(i) = \infty$ does not imply that the ith component of $\sum_{k=0}^\infty A_\mu^k b_\mu$ is equal to ∞, since there is the possibility that $A_\mu^N \bar{J}$ may become unbounded as $N \to \infty$ [cf. Eq. (4.83)]. This is a difference from Chapter 3 for SSP, where Assumption 3.1.2 requires that for every improper policy μ, we have $J_\mu(i) = \infty$ for at least one state i. However, in the SSP context, $\sum_{k=0}^\infty A_\mu^k b_\mu$ has an infinite component if and only if J_μ does, so the Assumption 4.5.3, when specialized to the SSP problem of Chapter 3, becomes identical to the Assumption 3.1.2 in Section 3.1. More generally, if A_μ is a nonexpansive mapping with respect to some norm, $\sum_{k=0}^\infty A_\mu^k b_\mu$ has an infinite component if and only if J_μ does, and Assumption 4.5.3 can be accordingly restated.

Under Assumptions 4.5.1-4.5.3, we will derive results that closely parallel the ones for SSP in Chapter 3. We have the following characterization of stable policies, which parallels Prop. 3.2.1 for proper policies.

Proposition 4.5.2: (Properties of Stable and Unstable Policies) Let Assumption 4.5.3 hold.

(a) For a stable policy μ, the associated cost vector J_μ satisfies

$$\lim_{k\to\infty} (T_\mu^k J)(i) = J_\mu(i), \qquad i = 1,\ldots,n,$$

for every vector $J \in \Re^n$. Furthermore,

$$J_\mu = T_\mu J_\mu,$$

and J_μ is the unique solution of this equation within $\Re^n$.

(b) A stationary policy μ is stable if and only if it satisfies

$$J(i) \geq (T_\mu J)(i), \qquad i = 1,\ldots,n,$$

for some vector $J \in \Re^n_+$.

Proof: (a) Follows from Eqs. (4.75) and (4.84).

(b) If μ is stable, by part (a) we have $J \geq T_\mu J$ for $J = J_\mu$. Conversely, let J be a vector in $\Re^n_+$ with $J \geq T_\mu J$, and assume to arrive at a contradiction that μ is unstable. Then the monotonicity of T_μ and Eq. (4.75) imply that

$$J \geq T_\mu^N J = A_\mu^N J + \sum_{k=0}^{N-1} A_\mu^k b_\mu, \qquad N = 1,2,\ldots.$$

Since μ is unstable, by Assumption 4.5.3, some component of $\sum_{k=0}^{N-1} A_\mu^k b_\mu$ diverges to ∞ as $N \to \infty$, while $A_\mu^N J \geq 0$, which contradicts the preceding relation. **Q.E.D.**

The following proposition parallels Prop. 3.2.2, the main result of Chapter 3 for SSP. Under Assumption 4.5.3, it shows existence and uniqueness of the solution of Bellman's equation within $\Re^n_+$, as well as the convergence of VI. The proof of the proposition bears similarity to the one of Prop. 3.2.2 (with stable policies replacing proper policies), and relies on Prop. 4.5.2 in the same way that the proof of Prop. 3.2.2 relies on Prop. 3.2.1.

The infinite cost Assumption 4.5.3 is essential for the proposition to hold; for example if $\min_\mu b_\mu = 0$, Bellman's equation always has the zero vector as a solution, and there may be other solutions as well (see the subsequent Example 4.5.1). However, we will provide later a perturbation-based analysis and related results, similar to the one of Section 4.4, which will require just the existence of at least one stable policy.

Proposition 4.5.3: (Bellman's Equation and Optimality Conditions) Let Assumptions 4.5.1, 4.5.2, and 4.5.3 hold.

(a) The optimal cost vector J^* satisfies Bellman's equation

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

Furthermore, J^* is the unique solution of this equation within $\Re_+^n$.

(b) We have

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

for every vector $J \in \Re_+^n$.

(c) A stationary policy μ is optimal if and only if

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

Moreover there exists an optimal stationary policy, and all optimal stationary policies are stable.

(d) If a vector $J \in \Re_+^n$ is such that $J \le TJ$, we have $J \le J^*$, and if $J \ge TJ$, we have $J \ge J^*$.

Proof: (a) We first show that T has at most one fixed point within $\Re_+^n$. Indeed, if J and J' are two fixed points, then we select μ and μ' such that $J = TJ = T_\mu J$ and $J' = TJ' = T_{\mu'} J'$; this is possible because of Prop. 4.5.1(a). By Prop. 4.5.2(b), we have that μ and μ' are stable, and Prop. 4.5.2(a) implies that $J = J_\mu$ and $J' = J_{\mu'}$. We also have $J = T^k J \le T_{\mu'}^k J$ for all $k \ge 1$, and by Prop. 4.5.2(a), we obtain $J \le \lim_{k\to\infty} T_{\mu'}^k J = J_{\mu'} = J'$. Similarly, $J' \le J$, showing that $J = J'$ and that T has at most one fixed point within $\Re_+^n$.

We next show that T has at least one fixed point within $\Re_+^n$. Let μ be a stable policy (there exists one by Assumption 4.5.1). Choose μ' such that

$$T_{\mu'} J_\mu = TJ_\mu.$$

Then we have $J_\mu = T_\mu J_\mu \ge T_{\mu'} J_\mu$. By Prop. 4.5.2(b), μ' is stable, and using the monotonicity of $T_{\mu'}$ and Prop. 4.5.2(a), we obtain

$$J_\mu \ge \lim_{k\to\infty} T_{\mu'}^k J_\mu = J_{\mu'}. \tag{4.88}$$

Continuing in the same manner, we construct a sequence $\{\mu^k\}$ such that each μ^k is stable and

$$J_{\mu^k} \ge T_{\mu^{k+1}} J_{\mu^k} = TJ_{\mu^k} \ge J_{\mu^{k+1}}, \qquad k = 0, 1, \ldots \tag{4.89}$$

Since the sequence $\{J_{\mu^k}\}$ is monotonically nonincreasing and bounded below by the zero vector, it converges to some vector $J_\infty \in \Re^n_+$.

We now claim that the sequence of vectors $\big(\mu^k(1),\ldots,\mu^k(n)\big)$ has a limit point $\big(\overline{\mu}(1),\ldots,\overline{\mu}(n)\big)$, with $\overline{\mu}$ being a feasible policy. Indeed, using Eq. (4.89) and the fact $J_\infty \le J_{\mu^{k-1}}$, we have for all $k=1,2,\ldots,$

$$T_{\mu^k}J_\infty \le T_{\mu^k}J_{\mu^{k-1}} = TJ_{\mu^{k-1}} \le T_{\mu^{k-1}}J_{\mu^{k-1}} = J_{\mu^{k-1}} \le J_{\mu^0},$$

so $\mu^k(i)$ belongs to the set

$$\hat{U}(i) = \left\{ u \in U(i) \;\middle|\; b(i,u) + \sum_{j=1}^n A_{ij}(u)J_\infty(j) \le J_{\mu^0}(i) \right\},$$

which is compact by Assumption 4.5.2. Hence the sequence $\{\mu^k\}$ belongs to the compact set $\hat{U}(1)\times\cdots\times\hat{U}(n)$, and has a limit point $\overline{\mu}$, which is a feasible policy. In what follows, without loss of generality, we assume that the entire sequence $\{\mu^k\}$ converges to $\overline{\mu}$. We will show that $\overline{\mu}$ is stable, and that $J_{\overline{\mu}} = J_\infty = TJ_\infty$, so that $J_{\overline{\mu}}$ is the unique fixed point of T within $\Re^n_+$.

Indeed, by taking limit as $k\to\infty$ in Eq. (4.89), and using the continuity part of Assumption 4.5.2, we obtain $J_\infty = T_{\overline{\mu}}J_\infty$. It follows from Prop. 4.5.2(b) that $\overline{\mu}$ is stable, and that $J_{\overline{\mu}}$ is equal to J_∞. To show that $J_{\overline{\mu}}$ is a fixed point of T, we note that from the right side of Eq. (4.89), we have for all policies μ, $T_\mu J_{\mu^k} \ge J_{\mu^{k+1}}$, which by taking limit as $k\to\infty$ yields $T_\mu J_{\overline{\mu}} \ge J_{\overline{\mu}}$. By taking minimum over μ, we obtain $TJ_{\overline{\mu}} \ge J_{\overline{\mu}}$. Combining this with the relation $J_{\overline{\mu}} = T_{\overline{\mu}}J_{\overline{\mu}} \ge TJ_{\overline{\mu}}$, it follows that $J_{\overline{\mu}} = TJ_{\overline{\mu}}$. Since T can have at most one fixed point in $\Re^n_+$, as shown earlier, $J_{\overline{\mu}}$ is the unique fixed point of T within $\Re^n_+$.

We will now show that $J_{\overline{\mu}}$ is equal to the optimal cost vector J^* (which also implies the optimality of the policy $\overline{\mu}$, obtained from the preceding PI process starting from a stable policy). Let J_0 be the identically zero vector [$J_0(i)\equiv 0$]. By Prop. 4.5.1(b), the sequence T^kJ_0 converges monotonically to some $J_\infty \in \Re^n_+$ that is a fixed point of T. By the uniqueness property shown earlier, J_∞ must be equal to $J_{\overline{\mu}}$. Also, for every policy $\pi = \{\mu_0,\mu_1,\ldots\}$, we have

$$T^kJ_0 \le T^k\bar{J} \le T_{\mu_0}\cdots T_{\mu_{k-1}}\bar{J}, \qquad k=0,1,\ldots,$$

and by taking the limit as $k\to\infty$, we obtain $J_{\overline{\mu}} = J_\infty = \lim_{k\to\infty} T^kJ_0 \le J_\pi$ for all π, showing that $J_{\overline{\mu}} = J^*$. Thus J^* is the unique fixed point of T within $\Re^n_+$.

(b) In the proof of part (a), we showed that $T^kJ_0 \to J^*$, which implies that

$$\lim_{k\to\infty} T^kJ = J^*, \qquad \forall\ J \in \Re^n_+ \text{ with } J \le J^*. \tag{4.90}$$

Also, for any $J \in \Re_+^n$ with $J \geq J^*$, we have

$$T_{\overline{\mu}}^k J \geq T^k J \geq T^k J^* = J^* = J_{\overline{\mu}},$$

where $\overline{\mu}$ is the stable optimal policy obtained by PI in the proof of part (a). By taking the limit as $k \to \infty$ and using the fact $T_{\overline{\mu}}^k J \to J_{\overline{\mu}}$ (which follows from the stability of $\overline{\mu}$), we obtain

$$\lim_{k \to \infty} T^k J = J^*, \qquad \forall\ J \in \Re_+^n \text{ with } J \geq J^*. \tag{4.91}$$

Finally, given any $J \in \Re_+^n$, we have from Eqs. (4.90) and (4.91),

$$\lim_{k \to \infty} T^k\big(\min\{J, J^*\}\big) = J^*, \qquad \lim_{k \to \infty} T^k\big(\max\{J, J^*\}\big) = J^*,$$

and since J lies between $\min\{J, J^*\}$ and $\max\{J, J^*\}$, it follows that $T^k J \to J^*$.

(c) If μ is optimal, then $J_\mu = J^*$ and since J^* is real-valued, by Prop. 4.5.3(c), μ is stable. Therefore, by Prop. 4.5.2(a),

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

Conversely, if $J^* = TJ^* = T_\mu J^*$, it follows from Prop. 4.5.2(b) that μ is stable, and by using Prop. 4.5.2(a), we obtain $J^* = J_\mu$. Therefore μ is optimal. The existence of an optimal policy follows from Prop. 4.5.1(a), implying that the minimum in Bellman's equation is attained for all i.

(d) If $J \in \Re_+^n$ 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 $J \geq J^*$ if $J \geq TJ$ is similar. **Q.E.D.**

Note that the preceding proof also established the validity of PI for affine monotonic models starting from a stable policy. This is similar to the proof of Prop. 3.2.2, which established the validity of PI for SSP starting from a proper policy.

We finally note that VI admits an asynchronous implementation, similar to the one given in Section 3.4.2 for SSP. The convergence can be proved similar to the SSP case, by appealing to the asynchronous convergence theorem (Prop. 2.6.1), and by using the monotonicity of T. Moreover, the PI algorithm of Section 3.5.4, which involves asynchronous iteration of costs and Q-factors, admits a straightforward extension to affine monotonic problems.

Perturbation Analysis Assuming Existence of at Least One Stable Policy

We will now eliminate Assumption 4.5.3, thus allowing unstable policies with real-valued cost functions. We will prove results that closely parallel the ones obtained for SSP under the weak assumptions of Section 4.4. An important notion in this regard is the optimal cost that can be achieved with stable policies only, i.e., the vector $\hat{J}$ with components given by

$$\hat{J}(i) = \min_{\mu:\,\text{stable}} J_\mu(i), \qquad i = 1, \ldots, n. \tag{4.92}$$

This is similar to Section 4.4, but with stable policies instead of proper policies; cf. Eq. (4.60).

We will show that $\hat{J}$ is a solution of Bellman's equation (Example 4.4.2 shows that J^* is not necessarily a solution). To this end we use the perturbation line of analysis of Section 4.4, by adding a constant $\delta > 0$ to all components of b_μ, thus obtaining what we call the *δ-perturbed affine monotonic model* (in analogy with the δ-perturbed SSP of Section 4.4). An important property of unstable policies in this regard is given by the following proposition.

Proposition 4.5.4: If μ is an unstable policy and all the components of the vector b_μ are strictly positive, then there is at least one state i such that the corresponding component of the vector $\sum_{k=0}^{\infty} A_\mu^k b_\mu$ is ∞.

Proof: According to the Perron-Frobenius Theorem, the nonnegative matrix A_μ has a real eigenvalue λ, which is equal to its spectral radius, and a corresponding nonnegative eigenvector $\xi \neq 0$ (see e.g., [BeT89], Chapter 2, Prop. 6.6). 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 unstable), 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.**

We denote by $J_{\mu,\delta}$ and by J_δ^* the cost function of μ and the optimal cost function of this δ-perturbed model. Then, in view of Prop. 4.5.4, all unstable policies in this model have infinite cost for all $\delta > 0$, and we can prove the following analog of Prop. 4.4.1, with an essentially identical proof.

Proposition 4.5.5: Let Assumptions 4.5.1 and 4.5.2 hold. Then for each $\delta > 0$:

(a) J_δ^* is the unique solution within $\Re_+^n$ of the equation

$$J(i) = (TJ)(i) + \delta, \qquad i = 1, \ldots, n.$$

(b) A policy μ is optimal for the δ-perturbed problem (i.e., $J_{\mu,\delta} = J_\delta^*$) if and only if $T_\mu J_\delta^* = TJ_\delta^*$. Moreover, for the δ-perturbed problem, all optimal policies are stable and there exists at least one stable policy that is optimal.

(c) The optimal cost function over stable policies $\hat{J}$ [cf. Eq. (4.92)] satisfies

$$\hat{J}(i) = \lim_{\delta \downarrow 0} J_\delta^*(i), \qquad i = 1, \ldots, n.$$

(d) If the control constraint set $U(i)$ is finite for all states $i = 1, \ldots, n$, there exists a stable policy $\hat{\mu}$ that attains the minimum over all stable policies, i.e., $J_{\hat{\mu}} = \hat{J}$.

We note that if $U(i)$ is infinite it is possible that $\hat{J} = J^*$, but the only optimal policy is unstable, even if the compactness Assumption 4.5.2 holds. This is shown in the following example, which is adapted from the paper [BeY15] (Example 2.1).

Example 4.5.1

Consider an exponentiated cost SSP problem with two states 1 and 2, in addition to the termination state t; see Fig. 4.5.1. From state 1 we transition to t with cost -1. At state 2 we must choose $u \in [0, 1]$, with cost equal to u. Then, we transition to state 1 with probability e^{-u}, and we self-transition to state 2 with probability $1 - e^{-u}$. It can be seen that a policy μ is stable if and only if it applies $\mu(2) = u > 0$ at state 2. Thus, from Eqs. (4.72) and (4.81), the Bellman equation for the corresponding exponentiated cost problem is

$$J(1) = e^{-1}, \qquad J(2) = \min_{u \in [0,1]} \big[e^{-u} e^u J(1) + (1 - e^{-u}) e^u J(2) \big],$$

from which

$$J(2) = \min_{u \in [0,1]} \big[e^{-1} + (e^u - 1) J(2) \big].$$

A stationary policy μ that applies $\mu(2) = u \in [0, 1]$ has cost equal to

$$J_\mu(2) = \frac{e^{-1}}{2 - e^u}.$$

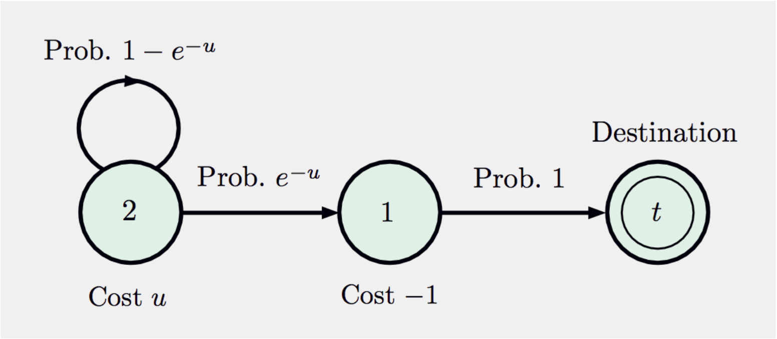


Figure 4.5.1. An exponentiated cost SSP problem with two states 1, 2, and a termination state t. Here we have $\hat{J} = J^*$ and the compact control constraint set $U(2) = [0, 1]$, but there is no stable policy that attains this cost.

We have

$$J^*(2) = \hat{J}(2) = e^{-1},$$

but the only optimal policy is the unstable policy that applies $u = 0$, while there is no stable policy that attains this cost. Note that in this example, the compactness Assumption 4.5.2 is satisfied.

Finally, in analogy with Prop. 4.4.2, we can prove the following proposition, again with an essentially identical proof.

Proposition 4.5.6: Let Assumptions 4.5.1 and 4.5.2 hold. Then:

(a) The optimal cost function over stable policies $\hat{J}$ is the unique fixed point of T within the set $\{J \in \Re^n_+ \mid J \geq \hat{J}\}$.

(b) We have $T^k J \to \hat{J}$ for every $J \in \Re^n_+$ with $J \geq \hat{J}$.

(c) Let μ be a stable policy. Then μ is optimal within the class of stable policies (i.e., $J_\mu = \hat{J}$) if and only if $T_\mu \hat{J} = T\hat{J}$.

Algorithms

The discussion of Section 4.4 regarding algorithms applies to affine monotonic problems with no essential changes. In particular, the convergence rate result of Prop. 4.4.3 is extended as follows, assuming that there exists a stable policy $\hat{\mu}$ that is optimal within the class of stable policies, i.e., $J_{\hat{\mu}} = \hat{J}$. This is true if the sets $U(i)$ are finite [cf. Prop. 4.5.5(d)], but not necessarily under the compactness Assumption 4.5.2; cf. Example 4.5.1, where the only optimal policy is unstable.

Proposition 4.5.7: (Convergence Rate of VI) Let Assumptions 4.5.1 and 4.5.2 hold, and assume that there exists a stable policy $\hat{\mu}$ that is optimal within the class of stable policies, i.e., $J_{\hat{\mu}} = \hat{J}$. Then

$$\|TJ - \hat{J}\|_v \le \beta \|J - \hat{J}\|_v, \qquad \forall\ J \ge \hat{J},$$

where $\|\cdot\|_v$ is a weighted sup-norm with respect to which T_{μ^*} is a contraction and β is the corresponding modulus of contraction. Moreover we have

$$\|J - \hat{J}\|_v \le \frac{1}{1-\beta} \max_{i=1,\ldots,n} \frac{J(i) - (TJ)(i)}{v(i)}, \qquad \forall\ J \ge \hat{J}.$$

We also note that the PI algorithm with perturbations for SSP developed in Section 4.4 (cf. Prop. 4.4.5) can be readily adapted to affine monotonic problems. In particular, we have the following proposition.

Proposition 4.5.8: Let Assumptions 4.5.1 and 4.5.2 hold, and assume that the control constraint sets $U(i)$, $i = 1, \ldots, n$, are finite. Let $\{\delta_k\}$ be a positive sequence with $\delta_k \downarrow 0$, and let μ^0 be any stable policy.Then the sequence $\{J_{\mu^k}\}$ generated by the perturbed PI algorithm

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

[cf. Eq. (4.67)] satisfies $J_{\mu^k} \to \hat{J}$.

Both the preceding proposition and its SSP counterpart (Prop. 4.4.5) can also be proved assuming the compactness Assumption 4.5.2 [instead of finiteness of $U(i)$]. The proof that $J_{\mu^k} \to \hat{J}$ is essentially the same in both cases. However, by using Example 4.5.1, it can be shown that the sequence of stable policies μ^k generated by the perturbed PI algorithm need not converge to a stable policy. We refer to the paper [BeY15] (Prop. 2.6) for a proof and for further discussion.

The following example involves the same two-node deterministic shortest path problem as Example 4.4.1, but with exponentiated cost of a path. It demonstrates the preceding analysis, both in the case where the infinite cost Assumption 4.5.3 is satisfied and for the case where it is not. The example also illustrates that the framework of this section allows the treatment of deterministic shortest path problems with cycles that have either positive and negative cycles (and also zero length cycles if a perturbation approach is used)!

Example 4.5.2 (Shortest Paths with Risk-Sensitive Cost)

Consider the shortest path problem of Fig. 4.4.1, where there is a single state 1 in addition to the termination state t, and at state 1 there are two choices: a self-transition, which costs a, and a transition to t, which costs b. However, the cost of a policy is equal to the exponentiated length of the path of the policy, thereby obtaining an affine monotonic model with one state, $i = 1$, and $\bar{J} = 1$. There are two policies denoted μ and $\overline{\mu}$: the 1st policy is $1 \to t$, while the 2nd policy is the self-transition $1 \to 1$. The corresponding mappings T_μ and $T_{\overline{\mu}}$ are given by

$$T_\mu J = e^b, \qquad T_{\overline{\mu}} J = e^a J,$$

corresponding to

$$A_\mu = 0, \quad b_\mu = e^b, \qquad A_{\overline{\mu}} = e^a, \quad b_{\overline{\mu}} = 0;$$

cf. Eq. (4.82). Clearly μ is stable, while $\overline{\mu}$ is unstable when $a \geq 0$, and stable when $a < 0$. Note that when $\overline{\mu}$ is unstable, Assumption 4.5.3 is violated because we have $b_{\overline{\mu}} = 0$. The mapping T is given by

$$TJ = \min\big\{e^b,\, e^a J\big\}.$$

The cost functions of μ and $\overline{\mu}$ are

$$J_\mu = e^b, \qquad J_{\overline{\mu}} = \begin{cases} \infty & \text{if } a > 0, \\ 1 & \text{if } a = 0, \\ 0 & \text{if } a < 0. \end{cases}$$

Consider three cases:

(a) When $a < 0$, the self-transition cycle has negative cost, but Prop. 4.5.3 applies because both policies are stable. Indeed, consistent with Prop. 4.5.3(b), we have $J^* = \hat{J} = 0$, and $J^* = 0$ is the unique fixed point of T within $\Re_+$. Moreover, VI converges to J^* starting from any $J \in \Re_+$.

(b) When $a > 0$, or when $a = 0$ and $b \leq 0$, the stable policy μ is optimal, we have $J^* = \hat{J} = e^b$ and J^* is a fixed point of T. However, 0 is an additional fixed point of T [the fixed points of T are the scalars 0 and e^b when $a > 0$, and the interval $\big[0, e^b\big]$ when $a = 0$ and $b \leq 0$]. When $a > 0$, VI converges to J^* when started at $J > 0$, but stays at the extraneous fixed point 0 when started at 0! When $a = 0$ and $b \leq 0$, VI converges to J^* when started at $J \geq J^*$, but stays at J when started at $J \in [0, J^*]$. The difficulty here is that the Assumption 4.5.3 is violated because for the unstable policy $\overline{\mu}$ we have $b_{\overline{\mu}} = 0$. However, Prop. 4.5.6 applies, and we can find the stable optimal policy by VI starting from any $J \geq J^*$, or by the perturbed version of PI.

(c) When $a = 0$ and $b > 0$, we have $1 = J^* < \hat{J} = e^b$. Then all the points between J^* and $\hat{J}$ are fixed points of T, and in fact the set of fixed points within $\Re_+$ is the interval $[0, \hat{J}]$. Thus in this exceptional case there is an interval of fixed points of T containing points both above and below J^*. This means that J^* cannot be obtained by VI, starting either from above or from below! Here, like the preceding case, the Assumption 4.5.3 is violated because for the unstable policy $\overline{\mu}$ we have $b_{\overline{\mu}} = 0$. However, Prop. 4.5.6 applies. In this case we can find $\hat{J}$ and the optimal policy within the set of stable policies by VI starting from any $J \geq \hat{J}$, or by the perturbed version of PI. Nonetheless, these algorithms will not necessarily find J^* and the optimal unstable policy $\overline{\mu}$.

4.6 EXTENSIONS AND APPLICATIONS

In this section we will elaborate on some of the methodology of the preceding sections, and we will discuss a number of applications.

4.6.1 Optimal Stopping

Consider an infinite horizon version of the stopping problems of Section 4.4 of Vol. I. At each state x, we must choose between two actions: pay a cost $s(x)$ and *stop* with no further cost incurred, or pay a cost $c(x)$ and *continue* the process according to the system equation

$$x_{k+1} = f_c(x_k, w_k), \qquad k = 0, 1, \ldots \tag{4.93}$$

The objective is to find the optimal stopping policy that minimizes the total expected cost over an infinite number of stages. It is assumed that the input disturbances w_k have the same probability distribution for all k, which depends only on the current state x_k.

This problem may be viewed as a special case of the SSP problem of Section 3.1, but here we will not assume that the state space is finite and the other assumptions of Section 3.1 regarding proper and improper policies. Instead we will rely on the general theory of unbounded cost problems developed in Section 4.1.

To put the problem within the framework of the total cost infinite horizon problem, we introduce an additional state t (the termination state) and we complete the system equation (4.93) as in Section 4.4 of Vol. I by letting

$$x_{k+1} = t, \qquad \text{if } u_k = \text{stop} \;\text{ or }\; x_k = t.$$

Once the system reaches the termination state, it remains there at no cost.

We first assume that

$$s(x) \geq 0, \qquad c(x) \geq 0, \qquad \text{for all } x \in X, \tag{4.94}$$

thus coming under the framework of Assumption P of Section 4.1. [The case corresponding to Assumption N, where $s(x) \leq 0$ and $c(x) \leq 0$ for all $x \in X$ will be considered later.] Actually, whenever there exists an $\epsilon > 0$ such that $c(x) \geq \epsilon$ for all $x \in X$, the results to be obtained under the assumption (4.94) apply also to the case where $s(x)$ is bounded below by some scalar rather than bounded by zero. The reason is that, if $c(x)$ is assumed to be greater than $\epsilon > 0$ for all $x \in X$, any policy that will not stop within a finite expected number of stages results in infinite cost and can be excluded from consideration. As a result, if we reformulate the problem and add a constant r to $s(x)$ so that $s(x) + r \geq 0$ for all $x \in X$, the optimal cost $J^*(x)$ will merely be increased by r, while optimal policies will remain unaffected.

The mapping T that defines the DP algorithm takes the form

$$(TJ)(x) = \begin{cases} \min\Big[s(x),\, c(x) + E\big\{J\big(f_c(x,w)\big)\big\}\Big] & \text{if } x \neq t, \\ 0 & \text{if } x = t, \end{cases} \tag{4.95}$$

where $s(x)$ is the cost of the stopping action, and $c(x) + E\big\{J\big(f_c(x,w)\big)\big\}$ is the cost of the continuation action. Since the control space has only two elements, by Prop. 4.1.7(a), we have

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

where J_0 is the zero function [$J_0(x) = 0$, for all $x \in X$]. By Prop. 4.1.5, there exists a stationary optimal policy given by

$$\begin{aligned} &\text{stop} && \text{if } s(x) < c(x) + E\Big\{J^*\big(f_c(x,w)\big)\Big\}, \\ &\text{continue} && \text{if } s(x) \geq c(x) + E\Big\{J^*\big(f_c(x,w)\big)\Big\}. \end{aligned}$$

Let us denote by S^* the optimal stopping set (which may be empty)

$$S^* = \Big\{x \in X \mid s(x) < c(x) + E\Big\{J^*\big(f_c(x,w)\big)\Big\}\Big\}.$$

Consider also the sets

$$S_k = \Big\{x \in X \mid s(x) < c(x) + E\Big\{(T^k J_0)\big(f_c(x,w)\big)\Big\}\Big\}$$

that determine the optimal policy for finite horizon versions of the stopping problem. Since we have

$$J_0 \leq TJ_0 \leq \cdots \leq T^k J_0 \leq \cdots \leq J^*,$$

it follows that

$$S_1 \subset S_2 \subset \cdots \subset S_k \subset \cdots \subset S^*$$

and therefore $\cup_{k=1}^{\infty} S_k \subset S^*$. Also, if $\tilde{x} \notin \cup_{k=1}^{\infty} S_k$, then we have

$$s(\tilde{x}) \geq c(\tilde{x}) + E\Big\{(T^k J_0)\big(f_c(\tilde{x}, w)\big)\Big\}, \qquad k = 0, 1, \ldots$$

By taking the limit as $k \to \infty$, and by using the monotone convergence theorem and the fact $T^k J_0 \to J^*$, we obtain

$$s(\tilde{x}) \geq c(\tilde{x}) + E\Big\{J^*\big(f_c(\tilde{x}, w)\big)\Big\},$$

from which $\tilde{x} \notin S^*$. Hence

$$S^* = \cup_{k=1}^{\infty} S_k.$$

In other words, the *optimal stopping set S^* for the infinite horizon problem is equal to the union of all the finite horizon stopping sets S_k.*

Consider now, as in Section 4.4 of Vol. I, the one-step-to-go stopping set

$$\tilde{S}_1 = \Big\{x \in S \mid s(x) \leq c(x) + E\Big\{s\big(f_c(x, w)\big)\Big\}\Big\} \tag{4.96}$$

and assume that $\tilde{S}_1$ is *absorbing* in the sense

$$f_c(x, w) \in \tilde{S}_1, \qquad \text{for all } x \in \tilde{S}_1, \quad w \in W, \tag{4.97}$$

and that the (monotonically nonincreasing) sequence $(T^k s)(x)$ converges to $J^*(x)$ for all $x \in X$. Then, as in Section 4.4 of Vol. I, it follows that the one-step lookahead policy

$$\text{stop if and only if } x \in \tilde{S}_1$$

is optimal. We now provide some examples.

Example 4.6.1 (Asset Selling)

Consider the version of the asset selling example of Sections 4.4 and 7.3 of Vol. I, where the rate of interest r is zero and there is instead a maintenance cost $c > 0$ per period for which the house remains unsold. Furthermore, past offers can be accepted at any future time. We have the following optimality equation:

$$J^*(x) = \max\Big[x, \ -c + E\big\{J^*\big(\max(x, w)\big)\big\}\Big].$$

In this case we consider maximization of total expected reward, the continuation cost is strictly negative, and the stopping reward x is positive. Hence the assumption (4.94) is not satisfied. If, however, we assume that x takes values in a bounded interval $[0, M]$, where M is an upper bound on the possible values of offers, our analysis is still applicable [cf. the discussion following Eq. (4.94)]. Consider the one-step-to-go stopping set given by

$$\tilde{S}_1 = \Big\{x \mid x \geq -c + E\big\{\max(x, w)\big\}\Big\}.$$

After a calculation similar to the one given in Section 4.4 of Vol. I, we see that

$$\tilde{S}_1 = \{x \mid x \geq \overline{a}\},$$

where $\overline{\alpha}$ is the scalar satisfying

$$\overline{\alpha} = P(\overline{\alpha})\overline{\alpha} + \int_{\overline{\alpha}}^{\infty} w \; dP(w) - c.$$

Clearly, $\tilde{S}_1$ is absorbing in the sense of Eq. (4.97), and therefore the one-step lookahead policy, which accepts the first offer that is greater or equal to $\overline{\alpha}$ is optimal.

Example 4.6.2 (Sequential Hypothesis Testing)

Consider the hypothesis testing problem of Example 5.4.4 of Vol. I for the case where the number of possible observations is unlimited. Here the states are x^0 and x^1 (true distribution of the observations is f_0 and f_1, respectively). The set X is the interval $[0,1]$ and corresponds to the sufficient statistic

$$p_k = P(x_k = x^0 \mid z_0, z_1, \ldots, z_k).$$

To each $p \in [0,1]$ we may assign the stopping cost

$$s(p) = \min\big[(1-p)L_0, \, pL_1\big],$$

i.e., the cost associated with optimal choice between the distributions f_0 and f_1. The mapping T of Eq. (4.95) takes the form

$$(TJ)(p) = \min\left[(1-p)L_0, \, pL_1, \, c + \mathop{E}_{z}\left\{J\left(\frac{pf_0(z)}{pf_0(z) + (1-p)f_1(z)}\right)\right\}\right]$$

for all $p \in [0,1]$, where the expectation over z is taken with respect to the probability distribution

$$P(z) = pf_0(z) + (1-p)f_1(z), \qquad z \in Z.$$

The optimal cost function J^* satisfies Bellman's equation

$$J^*(p) = \min\left[(1-p)L_0, pL_1, c + \mathop{E}_{z}\left\{J^*\left(\frac{pf_0(z)}{pf_0(z) + (1-p)f_1(z)}\right)\right\}\right]$$

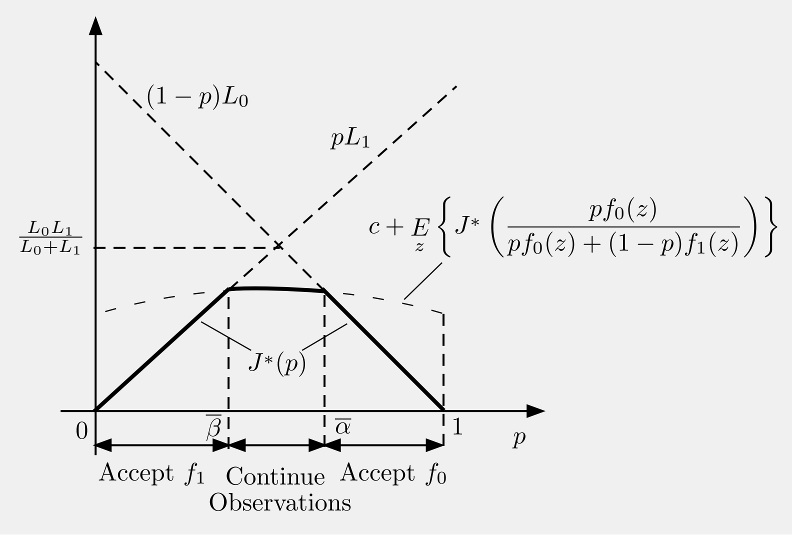


Figure 4.6.1 Derivation of the sequential probability ratio test.

and is obtained in the limit through the equation

$$J^*(p) = \lim_{k\to\infty} (T^k J_0)(p), \qquad p \in [0,1],$$

where J_0 is the zero function on $[0,1]$.

Now consider the functions $T^k J_0$, $k = 0, 1, \ldots$ It is clear that

$$J_0 \leq TJ_0 \leq \cdots \leq T^k J_0 \leq \cdots \leq \min\big[(1-p)L_0, pL_1\big].$$

Furthermore, in view of the analysis of Section 5.5 of Vol. I, we have that the function $T^k J_0$ is concave on $[0,1]$ for all k. Hence the pointwise limit function J^* is also concave on $[0,1]$. In addition, Bellman's equation implies that

$$J^*(0) = J^*(1) = 0,$$

$$J^*(p) \leq \min\big[(1-p)L_0, pL_1\big].$$

Using the reasoning illustrated in Fig. 4.6.1 it follows that [provided $c < L_0L_1/(L_0 + L_1)$] there exist two scalars $\overline{\alpha}$, $\overline{\beta}$ with $0 < \overline{\beta} \leq \overline{\alpha} < 1$, that determine an optimal stationary policy of the form

$$\begin{array}{rl} \text{accept } f_0 & \text{if } p \leq \overline{\alpha}, \\ \text{accept } f_1 & \text{if } p \leq \overline{\beta}, \\ \text{continue the observations} & \text{if } \overline{\beta} < p < \overline{\alpha}. \end{array}$$

In view of the optimality of the preceding stationary policy, the sequential probability ratio test described in Section 5.5 of Vol. I is justified when the number of possible observations is infinite.

The Case of Negative Transition Costs

We now consider the stopping problem under Assumption N, i.e.,

$$s(x) \leq 0, \qquad c(x) \leq 0, \qquad \text{for all } x \in X.$$

Under these circumstances there is no penalty for continuing operation of the system (although by not stopping at a given state, a favorable opportunity may be missed). The mapping T is given by

$$(TJ)(x) = \min \Big[s(x), c(x) + E\big\{ J\big(f_c(x, w)\big) \big\} \Big].$$

The optimal cost function J^* satisfies

$$J^*(x) \leq s(x), \qquad x \in X,$$

and by using Props. 4.1.1 and 4.1.7(b), we have

$$J^* = TJ^*, \qquad J^* = \lim_{k \to \infty} T^k J_0 = \lim_{k \to \infty} T^k s,$$

where J_0 is the zero function. It can also be seen that if the one-step-to-go stopping set $\tilde{S}_1$ is *absorbing* [cf. Eq. (4.97)], a one-step lookahead policy is optimal.

Example 4.6.3 (The Rational Burglar)

This example was considered at the end of Section 4.4 of Vol. I where it was shown that a one-step lookahead policy is optimal for any finite horizon length. The optimality equation is

$$J^*(x) = \max \Big[x,\, (1-p) E\big\{ J^*(x + w) \big\} \Big].$$

The problem is equivalent to a minimization problem where

$$s(x) = -x, \qquad c(x) = 0,$$

so Assumption N holds. From the preceding analysis, we have that $T^k s \to J^*$ and that a one-step lookahead policy is optimal if the one-step stopping set is absorbing [cf. Eqs. (4.96) and (4.97)]. It can be shown (see the analysis of Section 4.4 of Vol. I) that this condition holds, so the finite horizon optimal policy whereby the burglar retires when his accumulated earnings reach or exceed $(1-p)\overline{w}/p$ is optimal for an infinite horizon as well.

Example 4.6.4 (A Problem with no Optimal Policy)

This is an example of a deterministic stopping problem where Assumption N holds, and an optimal policy does not exist, even though only two controls are available at each state (stop and continue). The states are the positive integers, and continuation from state i leads to state $i+1$ with certainty and no cost, i.e., $X = \{1, 2, \ldots\}$, $c(i) = 0$, and $f_c(i, w) = i + 1$ for all $i \in X$ and $w \in W$. The stopping cost is $s(i) = -1 + (1/i)$ for all $i \in X$, so that there is an incentive to delay stopping at every state. We have $J^*(i) = -1$ for all i, and the optimal cost -1 can be approached arbitrarily closely by postponing the stopping action for a sufficiently long time. However, there is no policy that attains the optimal cost.

4.6.2 Inventory Control

Let us consider a discounted, infinite horizon version of the inventory control problem of Section 4.2 in Vol. I. Inventory stock evolves according to the equation

$$x_{k+1} = x_k + u_k - w_k, \qquad k = 0, 1, \ldots$$

We assume that the successive demands w_k are independent and bounded, and have identical probability distributions. We also assume for simplicity that there is no fixed cost. The case of a nonzero fixed cost can be treated similarly. The cost function is

$$J_\pi(x_0) = \lim_{N\to\infty} \underset{k=0,1,\ldots,N-1}{\underset{w_k}{E}} \left\{ \sum_{k=0}^{N-1} \alpha^k \big(c\mu_k(x_k) + H\big(x_k + \mu(x_k) - w_k\big)\big) \right\},$$

where

$$H(y) = p \max(0, -y) + h \max(0, y).$$

The DP algorithm is given by

$$J_0(x) = 0,$$

$$(T^{k+1}J_0)(x) = \min_{0\le u} E\big\{cu + H(x+u-w) + \alpha(T^k J_0)(x+u-w)\big\}.$$

We first show that the optimal cost is finite for all initial states:

$$J^*(x_0) = \min_{\pi} J_\pi(x_0) < \infty, \qquad \text{for all } x_0 \in X.$$

Indeed, consider the policy $\tilde\pi = \{\tilde\mu, \tilde\mu, \ldots\}$, where $\tilde\mu$ is defined by

$$\tilde\mu(x) = \begin{cases} 0 & \text{if } x \ge 0, \\ -x & \text{if } x < 0. \end{cases}$$

Since w_k is nonnegative and bounded, it follows that the inventory stock x_k when the policy $\tilde{\pi}$ is used satisfies

$$-w_{k-1} \leq x_k \leq \max(0, x_0), \qquad k = 1, 2, \ldots,$$

and is bounded. Hence $\tilde{\mu}(x_k)$ is also bounded. It follows that the cost per stage incurred when $\tilde{\pi}$ is used is bounded, and in view of the presence of the discount factor we have

$$J_{\tilde{\pi}}(x_0) < \infty, \qquad x_0 \in X.$$

Since $J^* \leq J_{\tilde{\pi}}$, the finiteness of the optimal cost follows.

Next we observe that, under the assumption $c < p$, the functions $T^k J_0$ are real-valued and convex. Indeed, we have

$$J_0 \leq TJ_0 \leq \cdots \leq T^k J_0 \leq \cdots \leq J^*,$$

which implies that $T^k J_0$ is real-valued. Convexity follows by induction as shown in Section 4.2 of Vol. I.

Consider now the sets

$$U_k(x, \lambda) = \Big\{ u \geq 0 \mid E\big\{cu + H(x+u-w) + \alpha(T^k J_0)(x_u - w)\big\} \leq \lambda \Big\}. \quad (4.98)$$

These sets are bounded since the expected value within the braces above tends to ∞ as $u \to \infty$. Also, the sets $U_k(x, \lambda)$ are closed since the expected value in Eq. (4.98) is a continuous function of u [recall that $T^k J_0$ is a real-valued convex and hence continuous function]. Thus we may invoke Prop. 4.1.8 and assert that

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

It follows from the convexity of the functions $T^k J_0$ that the limit function J^* is a real-valued convex function. Furthermore, an optimal stationary policy μ^* can be obtained by minimizing in the right-hand side of Bellman's equation

$$J^*(x) = \min_{u \geq 0} E\big\{cu + H(x + u - w) + \alpha J^*(x + u - w)\big\}.$$

We have

$$\mu^*(x) = \begin{cases} S^* - x & \text{if } x \leq S^*, \\ 0 & \text{otherwise,} \end{cases}$$

where S^* is a minimizing point of

$$G^*(y) = cy + L(y) + \alpha E\{J^*(y - w)\},$$

with

$$L(y) = E\big\{H(y-w)\big\}.$$

It can be seen that if $p > c$, we have $\lim_{|y|\to\infty} G^*(y) = \infty$, so that such a minimizing point exists. Furthermore, by using the observation made near the end of Section 4.1, it follows that a minimizing point S^* of $G^*(y)$ may be obtained as a limit point of a sequence $\{S_k\}$, where for each k the scalar S_k minimizes

$$G_k(y) = cy + L(y) + \alpha E\big\{(T^k J_0)(y-w)\big\}$$

and is obtained by means of the VI method.

It turns out that the critical level S^* has a simple characterization. It can be shown that S^* minimizes over y the expression $(1-\alpha)cy + L(y)$, and it can be essentially obtained in closed form (see Exercise 4.18, and the book [HeS84], Ch. 2).

In the case where there is a positive fixed cost ($K > 0$), the same line of argument may be used. Similarly, we prove that J^* is a real-valued K-convex function. A separate argument is necessary to prove that J^* is also continuous (this is intuitively clear and is left for the reader). Once K-convexity and continuity of J^* are established, the optimality of a stationary (s^*, S^*) policy follows from the equation

$$J^*(x) = \min_{u\geq 0} E\big\{C(u) + H(x+u-w) + \alpha J^*(x+u-w)\big\},$$

where $C(u) = K + cu$ if $u > 0$ and $C(0) = 0$.

4.6.3 Optimal Gambling Strategies

A gambler enters a certain game played as follows. The gambler may stake at any time k any amount $u_k \geq 0$ that does not exceed his current fortune x_k (defined to be his initial capital plus his gain or minus his loss thus far). He wins his stake back and as much more with probability p and he loses his stake with probability $(1-p)$. Thus the gambler's fortune evolves according to the equation

$$x_{k+1} = x_k + w_k u_k, \qquad k = 0, 1, \ldots, \tag{4.99}$$

where $w_k = 1$ with probability p and $w_k = -1$ with probability $(1-p)$. Several games, such as playing red and black in roulette, fit this description.

The gambler enters the game with an initial capital x_0, and his goal is to increase his fortune up to a level X. He continues gambling until he either reaches his goal or loses his entire initial capital, at which point he leaves the game. The problem is to determine the optimal gambling strategy for maximizing the probability of reaching his goal. By a gambling strategy, we mean a rule that specifies what the stake should be at time k when the gambler's fortune is x_k, for every x_k with $0 < x_k < X$.

The problem may be cast within the total cost, infinite horizon framework, where we consider maximization in place of minimization. Let us assume for convenience that fortunes are normalized so that $X = 1$. The state space is the set $[0,1] \cup \{t\}$, where t is a termination state to which the system moves with certainty from both states 0 and 1 with corresponding rewards 0 and 1. When $x_k \neq 0$, $x_k \neq 1$, the system evolves according to Eq. (4.99). The control constraint set is specified by

$$0 \leq u_k \leq x_k, \qquad 0 \leq u_k \leq 1 - x_k.$$

The reward per stage when $x_k \neq 0$ and $x_k \neq 1$ is zero. Under these circumstances the probability of reaching the goal is equal to the total expected reward. Assumption N holds since our problem is equivalent to a problem of minimizing expected total cost with nonpositive costs per stage.

The mapping T defining the DP algorithm takes the form

$$(TJ)(x) = \begin{cases} \max_{\substack{0 \leq u \leq x \\ 0 \leq u \leq 1-x}} \big[pJ(x+u) + (1-p)J(x-u)\big] & \text{if } x \in (0,1), \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x = 1, \end{cases}$$

for any function $J : [0,1] \mapsto [0,\infty]$.

Consider now the case where

$$0 < p < \frac{1}{2},$$

i.e., the game is unfair to the gambler. A discretized version of the case where $1/2 \leq p < 1$ is considered in Exercise 4.21. When $0 < p < 1/2$, it is intuitively clear that if the gambler follows a very conservative strategy and stakes a very small amount at each time, he is all but certain to lose his capital. For example, if the gambler adopts a strategy of betting $1/n$ at each time, then it may be shown (see Exercise 4.21 or Ash [Ash70], p. 182) that his probability of attaining the target fortune of 1 starting with an initial capital i/n, $0 < i < n$, is given by

$$\left(\left(\frac{1-p}{p}\right)^i - 1\right)\left(\left(\frac{1-p}{p}\right)^n - 1\right)^{-1}.$$

If $0 < p < 1/2$, n tends to infinity, and i/n tends to a constant, the above probability tends to zero, thus indicating that placing consistently small bets is a bad strategy.

We are thus led to a policy that places large bets and, in particular, the *bold strategy* whereby the gambler stakes at each time k his entire fortune x_k or just enough to reach his goal, whichever is least. In other words, the bold strategy is the stationary policy μ^* given by

$$\mu^*(x) = \begin{cases} x & \text{if } 0 < x \le 1/2, \\ 1 - x & \text{if } 1/2 \le x < 1. \end{cases}$$

We will prove that the bold strategy is indeed an optimal policy. To this end it is sufficient to show that for every initial fortune $x \in [0,1]$ the value of the reward function $J_{\mu^*}(x)$ corresponding to the bold strategy μ^* satisfies the sufficiency condition (cf. Prop. 4.1.6)

$$TJ_{\mu^*} = J_{\mu^*},$$

or equivalently

$$J_{\mu^*}(0) = 0, \qquad J_{\mu^*}(1) = 1,$$

$$J_{\mu^*}(x) \ge pJ_{\mu^*}(x+u) + (1-p)J_{\mu^*}(x-u),$$

for all $x \in (0,1)$ and $u \in [0,x] \cap [0, 1-x]$.

By using the definition of the bold strategy, Bellman's equation

$$J_{\mu^*} = T_{\mu^*}J_{\mu^*},$$

is written as

$$J_{\mu^*}(0) = 0, \qquad J_{\mu^*}(1) = 1, \tag{4.100}$$

$$J_{\mu^*}(x) = \begin{cases} pJ_{\mu^*}(2x) & \text{if } 0 < x \le 1/2, \\ p + (1-p)J_{\mu^*}(2x-1) & \text{if } 1/2 \le x < 1. \end{cases} \tag{4.101}$$

The following lemma shows that J_{μ^*} is uniquely defined from these relations.

Lemma 4.6.1: For every p, with $0 < p \le 1/2$, there is only one bounded function on $[0,1]$ satisfying Eqs. (4.100) and (4.101), the function J_{μ^*}. Furthermore, J_{μ^*} is continuous and strictly increasing on $[0,1]$.

Proof: Suppose that there existed two bounded functions $J_1 : [0,1] \mapsto \Re$ and $J_2 : [0,1] \mapsto \Re$ such that $J_i(0) = 0$, $J_i(1) = 1$, $i = 1,2$, and

$$J_i(x) = \begin{cases} pJ_i(2x) & \text{if } 0 < x \le 1/2, \\ p + (1-p)J_i(2x-1) & \text{if } 1/2 \le x < 1, \end{cases} \qquad i = 1,2.$$

Then we have

$$J_1(2x) - J_2(2x) = \frac{J_1(x) - J_2(x)}{p}, \qquad \text{if } 0 \le x \le 1/2, \tag{4.102}$$

$$J_1(2x-1) - J_2(2x-1) = \frac{J_1(x) - J_2(x)}{1-p}, \qquad \text{if } 1/2 \le x \le 1. \tag{4.103}$$

Let z be any real number with $0 \le z \le 1$. Define

$$z_1 = \begin{cases} 2z & \text{if } 0 \le z \le 1/2, \\ 2z - 1 & \text{if } 1/2 < z \le 1, \end{cases}$$

$$\vdots$$

$$z_k = \begin{cases} 2z_{k-1} & \text{if } 0 \le z_{k-1} \le 1/2, \\ 2z_{k-1} - 1 & \text{if } 1/2 < z_{k-1} \le 1, \end{cases}$$

for $k = 1, 2, \ldots$ Then from Eqs. (4.102) and (4.103) it follows (using $p \le 1/2$) that

$$\big|J_1(z_k) - J_1(z_k)\big| \ge \frac{\big|J_1(z) - J_2(z)\big|}{(1-p)^k}, \qquad k = 1, 2, \ldots$$

Since $J_1(z_k) - J_2(z_k)$ is bounded, it follows that $J_1(z) - J_2(z) = 0$, for otherwise the right side of the inequality would tend to ∞. Since $z \in [0,1]$ is arbitrary, we obtain $J_1 = J_2$. Hence J_{μ^*} is the unique bounded function on $[0,1]$ satisfying Eqs. (4.100) and (4.101).

To show that J_{μ^*} is strictly increasing and continuous, we consider the mapping T_{μ^*}, which operates on functions $J : [0,1] \mapsto [0,1]$ and is defined by

$$(T_{\mu^*}J)(x) = \begin{cases} pJ(2x) + (1-p)J(0) & \text{if } 0 < x \le 1/2, \\ pJ(1) + (1-p)J(2x-1) & \text{if } 1/2 \le x < 1, \end{cases}$$

$$(T_{\mu^*}J)(0) = 0, \qquad (T_{\mu^*}J)(1) = 1. \tag{4.104}$$

Consider the functions $J_0, T_\mu^* J_0, \cdots, T_{\mu^*}^k J_0, \ldots$, where J_0 is the zero function $[J_0(x) = 0$ for all $x \in [0,1]\,]$. We have

$$J_{\mu^*}(x) = \lim_{k \to \infty} (T_{\mu^*}^k J_0)(x), \qquad x \in [0,1]. \tag{4.105}$$

Furthermore, the functions $T_{\mu^*}^k J_0$ can be shown to be monotonically nondecreasing in the interval $[0,1]$. Hence, by Eq. (4.105), J_{μ^*} is also monotonically nondecreasing.

Consider now for $n = 0, 1, \ldots$ the sets

$$X_n = \big\{x \in [0,1] \mid x = k2^{-n}, k = \text{nonnegative integer}\big\}.$$

It is straightforward to verify that

$$(T_{\mu^*}^m J_0)(x) = (T_{\mu^*}^n J_0)(x), \qquad x \in X_{n-1}, \quad m \ge n \ge 1.$$

As a result of this equality and Eq. (4.105),

$$J_{\mu^*}(x) = (T^n_{\mu^*} J_0)(x), \qquad x \in X_{n-1}, \quad n \geq 1. \tag{4.106}$$

A further fact that may be verified by using induction and Eqs. (4.104) and (4.106) is that for any nonnegative integers k, n for which $0 \leq k2^{-n} < (k+1)2^{-n} \leq 1$, we have

$$p^n \leq J_{\mu^*}\big((k+1)2^{-n}\big) - J_{\mu^*}\big(k2^{-n}\big) \leq (1-p)^n. \tag{4.107}$$

Since any number in $[0,1]$ can be approximated arbitrarily closely from above and below by numbers of the form $k2^{-n}$, and since J_{μ^*} has been shown to be monotonically nondecreasing, it follows from Eq. (4.107) that J_{μ^*} is continuous and strictly increasing. **Q.E.D.**

We are now in a position to prove the following proposition.

Proposition 4.6.1: The bold strategy is an optimal stationary gambling policy.

Proof: We will prove the sufficiency condition

$$J_{\mu^*}(x) \geq pJ_{\mu^*}(x+u)+(1-p)J_{\mu^*}(x-u), \qquad x \in [0,1], \quad u \in [0,1]\cap[0,1-x]. \tag{4.108}$$

In view of the continuity of J_{μ^*} established in the previous lemma, it it sufficient to establish Eq. (4.108) for all $x \in [0,1]$ and $u \in [0,x] \cap [0,1-x]$ that belong to the union $\cup_{n=0}^{\infty} X_n$ of the sets X_n defined by

$$X_n = \big\{z \in [0,1] \mid z = k2^{-n},\ k = \text{nonnegative integer}\big\}.$$

We will use induction. By using the fact that $J_{\mu^*}(0) = 0$, $J_{\mu^*}(1/2) = p$, and $J_{\mu^*}(1) = 1$, we can show that Eq. (4.108) holds for all x and u in X_0 and X_1. Assume that Eq. (4.108) holds for all $x, u \in X_n$. We will show that it holds for all $x, u \in X_{n+1}$.

For any $x, u \in X_{n+1}$ with $u \in [0,x] \cap [0,1-x]$, there are four possibilities:

1. $x + u \leq 1/2$,
2. $x - u \geq 1/2$,
3. $x - u \leq x \leq 1/2 \leq x + u$,
4. $x - u \leq 1/2 \leq x \leq x + u$,

We will prove Eq. (4.108) for each of these cases.

Case 1. If $x, u \in X_{n+1}$, then $2x \in X_n$, and $2u \in X_n$, and by the induction hypothesis

$$J_{\mu^*}(2x) - pJ_{\mu^*}(2x+2u) - (1-p)J_{\mu^*}(2x-2u) \geq 0. \tag{4.109}$$

If $x + u \leq 1/2$, then by Eq. (4.101)

$$\begin{aligned} &J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u) \\ &\qquad\qquad = p\big(J_{\mu^*}(2x) - pJ_{\mu^*}(2x+2u) - (1-p)J_{\mu^*}(2x-2u)\big) \end{aligned}$$

and using Eq. (4.109), the desired relation Eq. (4.108) is proved for the case under consideration.

Case 2. If $x, u \in X_{n+1}$, then $(2x-1) \in X_n$ and $2u \in X_n$, and by the induction hypothesis

$$J_{\mu^*}(2x-1) - pJ_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u-1) \geq 0.$$

If $x - u \geq 1/2$, then by Eq. (4.101)

$$\begin{aligned} &J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u) \\ &\quad = p + (1-p)J_{\mu^*}(2x-1) - p\big(p + (1-p)J_{\mu^*}(2x+2u-1)\big) \\ &\qquad - (1-p)\big(p + (1-p)J_{\mu^*}(2x-2u-1)\big) \\ &\quad = (1-p)\big(J_{\mu^*}(2x-1) - pJ_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u-1)\big) \\ &\quad \geq 0, \end{aligned}$$

and Eq. (4.108) follows from the preceding relations.

Case 3. Using Eq. (4.101), we have

$$\begin{aligned} &J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u) \\ &\quad = pJ_{\mu^*}(2x) - p\big(p + (1-p)J_{\mu^*}(2x+2u-1)\big) - p(1-p)J_{\mu^*}(2x-2u) \\ &\quad = p\big(J_{\mu^*}(2x) - p - (1-p)J_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u)\big). \end{aligned}$$

Now we must have $x \geq \frac{1}{4}$, for otherwise $u < \frac{1}{4}$ and $x + u < 1/2$. Hence $2x \geq 1/2$ and the sequence of equalities can be continued as follows:

$$\begin{aligned} &J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u) \\ &\qquad = p\big(p + (1-p)J_{\mu^*}(4x-1) - p \\ &\qquad\quad - (1-p)J_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u)\big) \\ &\qquad = p(1-p)\big(J_{\mu^*}(4x-1) - J_{\mu^*}(2x+2u-1) - J_{\mu^*}(2x-2u)\big) \\ &\qquad = (1-p)\big(J_{\mu^*}(2x-1/2) - pJ_{\mu^*}(2x+2u-1) - pJ_{\mu^*}(2x-2u)\big). \end{aligned}$$

Since $p \leq (1-p)$, the last expression is greater than or equal to both

$$(1-p)\big(J_{\mu^*}(2x-1/2) - pJ_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u)\big)$$

and

$$(1-p)\big(J_{\mu^*}(2x-1/2)-(1-p)J_{\mu^*}(2x+2u-1)-pJ_{\mu^*}(2x-2u)\big).$$

Now for $x, u \in X_{n+1}$, and $n \geq 1$, we have $(2x-1/2) \in X_n$ and $(2u-1/2) \in X_n$ if $(2u-1/2) \in [0,1]$, and $(1/2-2u) \in X_n$ if $(1/2-2u) \in [0,1]$. By the induction hypothesis, the first or the second of the preceding expressions is nonnegative, depending on whether $2x+2u-1 \geq 2x-1/2$ or $2x-2u \geq 2x-1/2$ (i.e., $u \geq \frac{1}{4}$ or $u \leq \frac{1}{4}$). Hence Eq. (4.108) is proved for case 3.

Case 4. The proof resembles the one for case 3. Using Eq. (4.101), we have

$$\begin{aligned} J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u) &\\ = p + (1-p)J_{\mu^*}(2x-1) - p\big(p + (1-p)J_{\mu^*}(2x+2u-1)\big) &\\ - (1-p)pJ_{\mu^*}(2x-2u) &\\ = p(1-p) &\\ + (1-p)\big(J_{\mu^*}(2x-1) - pJ_{\mu^*}(2x+2u-1) - pJ_{\mu^*}(2x-2u)\big). \end{aligned}$$

We must have $x \leq \frac{3}{4}$ for otherwise $u < \frac{1}{4}$ and $x - u > \frac{1}{2}$. Hence $0 \leq 2x-1 \leq 1/2 \leq 2x-1/2 \leq 1$, and using Eq. (4.101) we have

$$(1-p)J_{\mu^*}(2x-1) = (1-p)pJ_{\mu^*}(4x-2) = p\big(J_{\mu^*}(2x-1/2) - p\big).$$

Using the preceding relations, we obtain

$$\begin{aligned} J_{\mu^*}(x) - pJ_{\mu^*}(x+u) - (1-p)J_{\mu^*}(x-u) &\\ = p(1-p) + p\big(J_{\mu^*}(2x-1/2) - p\big) - p(1-p)J_{\mu^*}(2x+2u-1) &\\ - p(1-p)J_{\mu^*}(2x-2u) &\\ = p\big((1-2p) + J_{\mu^*}(2x-1/2) - (1-p)J_{\mu^*}(2x+2u-1) &\\ - (1-p)J_{\mu^*}(2x-2u)\big). \end{aligned}$$

These relations are equal to both

$$\begin{aligned} p\big((1-2p)\big(1 - J_{\mu^*}(2x+2u-1)\big) &\\ + J_{\mu^*}(x-1/2) - pJ_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u)\big) \end{aligned}$$

and

$$\begin{aligned} p\big((1-2p)\big(1 - J_{\mu^*}(2x-2u)\big) &\\ + J_{\mu^*}(2x-1/2) - (1-p)J_{\mu^*}(2x+2u-1) - pJ_{\mu^*}(2x-2u)\big). \end{aligned}$$

Since $0 \leq J_{\mu^*}(2x+2u-1) \leq 1$ and $0 \leq J_{\mu^*}(2x-2u) \leq 1$, these expressions are greater than or equal to both

$$p\big(J_{\mu^*}(2x-1/2) - pJ_{\mu^*}(2x+2u-1) - (1-p)J_{\mu^*}(2x-2u)\big)$$

and

$$p\big(J_{\mu^*}(2x - 1/2) - (1-p)J_{\mu^*}(2x + 2u - 1) - pJ_{\mu^*}(2x - 2u)\big)$$

and the result follows as in case 3. **Q.E.D.**

We note that the bold strategy is not the unique optimal stationary gambling strategy. For a characterization of all optimal strategies, see the book [DuS65], p. 90. Several other gambling problems where strategies of the bold type are optimal are described in [DuS65], Chapters 5 and 6.

4.6.4 Continuous-Time Problems - Control of Queues

Problems of optimal control of queues often involve unbounded costs per stage. While by using the theory of Section 1.5.2, it is possible to address some of these problems similar to discounted problems with bounded cost per stage, this may involve restrictive assumptions. The alternative is to use the line of analysis of Section 1.4 based on uniformization. We recall from that section that if the time between transitions is exponentially distributed, the discounted continuous-time problem may be converted into an equivalent discrete-time problem (with bounded or unbounded cost per stage). In Section 1.4, we discussed some examples where the cost per stage is bounded. Here we consider some queueing applications where the cost per stage is unbounded.

Example 4.6.5 (M/M/1 Queue with Controlled Service Rate)

Consider a single-server queueing system where customers arrive according to a Poisson process with rate λ. The service time of a customer is exponentially distributed with parameter μ (called the service rate). Service times of customers are independent and are also independent of customer interarrival times. The service rate μ can be selected from a closed subset M of an interval $[0, \overline{\mu}]$ and can be changed at the times when a customer arrives or when a customer departs from the system. There is a cost $q(\mu)$ per unit time for using rate μ and a waiting cost $c(i)$ per unit time when there are i customers in the system (waiting in queue or undergoing service). The idea is that one should be able to cut down on the customer waiting costs by choosing a faster service rate, which presumably costs more. The problem, roughly, is to select the service rate so that the service cost is optimally traded off with the customer waiting cost.

We assume the following:

1. For some $\mu \in M$ we have $\mu > \lambda$. (In words, there is available a service rate that is fast enough to keep up with the arrival rate, thereby maintaining the queue length bounded.)

2. The waiting cost function c is nonnegative, monotonically nondecreasing, and "convex-like" in the sense

$$c(i+2) - c(i+1) \geq c(i+1) - c(i), \qquad i = 0, 1, \ldots$$

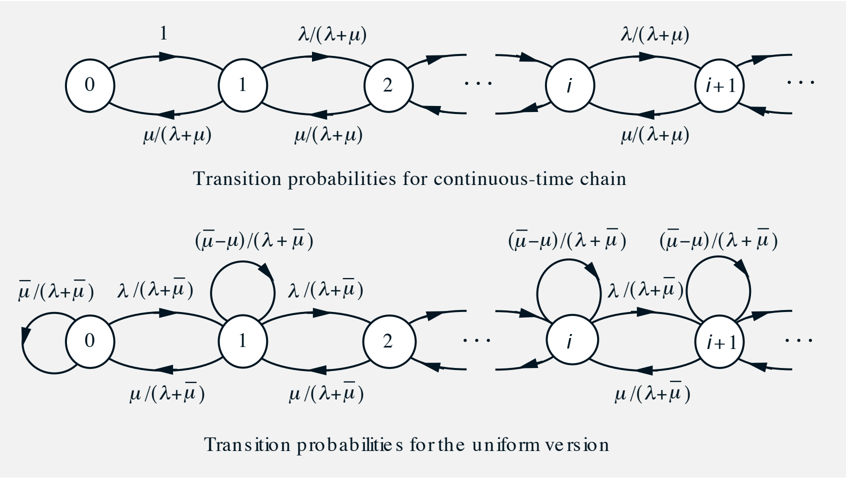


Figure 4.6.2 Continuous-time Markov chain and uniform version for Example 4.6.5 when the service rate is equal to μ. The transition rates of the original Markov chain are $\nu_i(\mu) = \lambda + \mu$ for states $i \geq 1$, and $\nu_0(\mu) = \lambda$ for state 0. The transition rate for the uniform version is $\nu = \lambda + \overline{\mu}$.

3. The service rate cost function q is nonnegative, and continuous on $[0, \overline{\mu}]$, with $q(0) = 0$.

Here the state is the number of customers in the system, and the control is the choice of service rate following a customer arrival or departure. The transition rate at state i is

$$\nu_i(\mu) = \begin{cases} \lambda & \text{if } i = 0, \\ \lambda + \mu & \text{if } i \geq 1. \end{cases}$$

The transition probabilities of the Markov chain and its uniform version for the choice

$$\nu = \lambda + \overline{\mu}$$

are shown in Fig. 4.6.2.

The effective discount factor is

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

and the cost per stage is

$$\frac{1}{\beta + \nu}\big(c(i) + q(\mu)\big).$$

Bellman's equation takes the form

$$J(0) = \frac{1}{\beta + \nu}\big(c(0) + (\nu - \lambda)J(0) + \lambda J(1)\big)$$

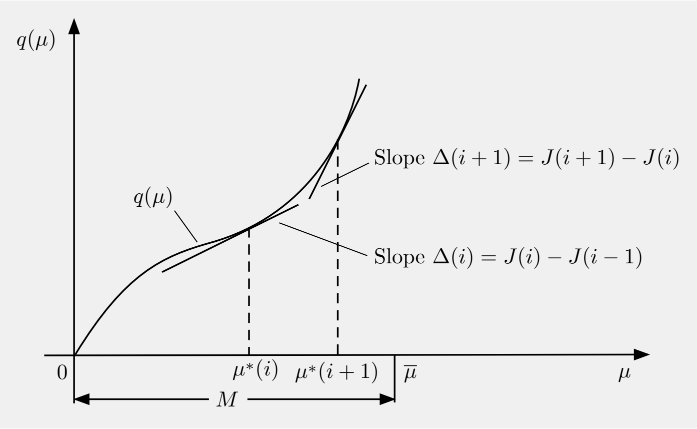

$q(\mu)$
$q(\mu)$
Slope $\Delta(i+1) = J(i+1) - J(i)$
Slope $\Delta(i) = J(i) - J(i-1)$
0
$\mu^*(i)$
$\mu^*(i+1)$
$\overline{\mu}$
μ
M

Figure 4.6.3 Determining the optimal service rate at states i and $(i+1)$ in Example 4.6.5. The optimal service rate $\mu^*(i)$ tends to increase as the system becomes more crowded (i increases).

and for $i = 1, 2, \ldots$,

$$J(i) = \frac{1}{\beta + \nu} \min_{\mu \in M} \big[c(i) + q(\mu) + \mu J(i-1) + (\nu - \lambda - \mu) J(i) + \lambda J(i+1) \big].$$

An optimal policy is to use at state i the service rate that minimizes the expression on the right. Thus it is optimal to use at state i the service rate

$$\mu^*(i) = \arg\min_{\mu \in M} \big\{ q(\mu) - \mu \Delta(i) \big\}, \tag{4.110}$$

where $\Delta(i)$ is the differential of the optimal cost

$$\Delta(i) = J(i) - J(i-1), \qquad i = 1, 2, \ldots$$

[When the minimum in Eq. (4.110) is attained by more than one service rate μ we choose by convention the smallest.] We will demonstrate shortly that $\Delta(i)$ *is monotonically nondecreasing*. It will then follow from Eq. (4.110) (see Fig. 4.6.3) that *the optimal service rate* $\mu^*(i)$ *is monotonically nondecreasing*. Thus, as the queue length increases, it is optimal to use a faster service rate.

To show that $\Delta(i)$ is monotonically nondecreasing, we use the DP recursion to generate a sequence of functions J_k from the starting function

$$J_0(i) = 0, \qquad i = 0, 1, \ldots$$

For $k = 0, 1, \ldots$, we have

$$J_{k+1}(0) = \frac{1}{\beta + \nu} \big(c(0) + (\nu - \lambda) J_k(0) + \lambda J_k(1) \big),$$

and for $i = 1, 2, \ldots$,

$$J_{k+1}(i) = \frac{1}{\beta+\nu} \min_{\mu\in M} \bigl[c(i) + q(\mu) + \mu J_k(i-1) + (\nu-\lambda-\mu)J_k(i) + \lambda J_k(i+1)\bigr]. \tag{4.111}$$

For $k = 0, 1, \ldots$ and $i = 1, 2, \ldots$, let

$$\Delta_k(i) = J_k(i) - J_k(i-1).$$

For completeness of notation, define also $\Delta_k(0) = 0$. From the theory of Section 4.1 (see Prop. 4.1.7), we have $J_k(i) \to J(i)$ as $k \to \infty$. It follows that

$$\lim_{k\to\infty} \Delta_k(i) = \Delta(i), \qquad i = 1, 2, \ldots$$

Therefore, it will suffice to show that $\Delta_k(i)$ is monotonically nondecreasing for every k. For this we use induction. The assertion is trivially true for $k = 0$. Assuming that $\Delta_k(i)$ is monotonically nondecreasing, we show that the same is true for $\Delta_{k+1}(i)$. Let

$$\mu^k(0) = 0,$$

$$\mu^k(i) = \arg\min_{\mu\in M} \bigl[q(\mu) - \mu\Delta_k(i)\bigr], \qquad i = 1, 2, \ldots$$

From Eq. (4.111) we have, for all $i = 0, 1, \ldots$,

$$\begin{aligned}
\Delta_{k+1}(i+1) &= J_{k+1}(i+1) - J_{k+1}(i) \\
&\geq \frac{1}{\beta+\nu}\Bigl(c(i+1) + q\bigl(\mu^k(i+1)\bigr) + \mu^k(i+1)J_k(i) \\
&\quad + \bigl(\nu-\lambda-\mu^k(i+1)\bigr)J_k(i+1) \\
&\quad + \lambda J_k(i+2) - c(i) - q\bigl(\mu^k(i+1)\bigr) - \mu^k(i+1)J_k(i-1) \\
&\quad - \bigl(\nu-\lambda-\mu^k(i+1)\bigr)J_k(i) - \lambda J_k(i+1)\Bigr) \\
&= \frac{1}{\beta+\nu}\Bigl(c(i+1) - c(i) + \lambda\Delta_k(i+2) + (\nu-\lambda)\Delta_k(i+1) \\
&\quad - \mu^k(i+1)\bigl(\Delta_k(i+1) - \Delta_k(i)\bigr)\Bigr).
\end{aligned} \tag{4.112}$$

Similarly, we obtain, for $i = 1, 2, \ldots$,

$$\begin{aligned}
\Delta_{k+1}(i) \leq \frac{1}{\beta+\nu}\Bigl(c(i) - c(i-1) + \lambda\Delta_k(i+1) + (\nu-\lambda)\Delta_k(i) \\
- \mu^k(i-1)\bigl(\Delta_k(i) - \Delta_k(i-1)\bigr)\Bigr).
\end{aligned}$$

Subtracting the last two inequalities, we obtain, for $i = 1, 2, \ldots$,

$$\begin{aligned}
(\beta+\nu)\bigl(\Delta_{k+1}(i+1) - \Delta_{k+1}(i)\bigr) \geq{}& \bigl(c(i+1) - c(i)\bigr) - \bigl(c(i) - c(i-1)\bigr) \\
&+ \lambda\bigl(\Delta_k(i+2) - \Delta_k(i+1)\bigr) \\
&+ \bigl(\nu-\lambda-\mu^k(i+1)\bigr)\bigl(\Delta_k(i+1) - \Delta_k(i)\bigr) \\
&+ \mu^k(i-1)\bigl(\Delta_k(i) - \Delta_k(i-1)\bigr).
\end{aligned}$$

Using our convexity assumption on $c(i)$, the fact $\nu - \lambda - \mu^k(i+1) = \overline{\mu} - \mu^k(i+1) \geq 0$, and the induction hypothesis, we see that every term on the right-hand side of the preceding inequality is nonnegative. Therefore, $\Delta_{k+1}(i+1) \geq \Delta_{k+1}(i)$ for $i = 1, 2, \ldots$ From Eq. (4.112) we can also show that $\Delta_{k+1}(1) \geq 0 = \Delta_{k+1}(0)$, and the induction proof is complete.

To summarize, the optimal service rate $\mu^*(i)$ is given by Eq. (4.110) and tends to become faster as the system becomes more crowded (i increases).

Example 4.6.6 (M/M/1 Queue with Controlled Arrival Rate)

Consider the same queueing system as in the previous example with the difference that the service rate μ is fixed, but the arrival rate λ can be controlled. We assume that λ is chosen from a closed subset Λ of an interval $[0, \overline{\lambda}]$, and there is a cost $q(\lambda)$ per unit time. All other assumptions of Example 4.6.5 are also in effect. What we have here is a problem of flow control, whereby we want to trade off optimally the cost for throttling the arrival process with the customer waiting cost.

This problem is very similar to the one of Example 4.6.5. We choose as uniform transition rate

$$\nu = \overline{\lambda} + \mu$$

and construct the uniform version of the Markov chain. Bellman's equation takes the form

$$J(0) = \frac{1}{\beta + \nu} \min_{\lambda \in \Lambda} \bigl[c(0) + q(\lambda) + (\nu - \lambda)J(0) + \lambda J(1)\bigr],$$

$$J(i) = \frac{1}{\beta + \nu} \min_{\lambda \in \Lambda} \bigl[c(i) + q(\lambda) + \mu J(i-1) + (\nu - \lambda - \mu)J(i) + \lambda J(i+1)\bigr].$$

An optimal policy is to use at state i the arrival rate

$$\lambda^*(i) = \arg\min_{\lambda \in \Lambda} \bigl[q(\lambda) + \lambda \Delta(i+1)\bigr], \tag{4.113}$$

where, as before, $\Delta(i)$ is the differential of the optimal cost

$$\Delta(i) = J(i) - J(i-1), \qquad i = 1, 2, \ldots$$

As in Example 4.6.5, we can show that $\Delta(i)$ is monotonically nondecreasing; so from Eq. (4.113) we see that *the optimal arrival rate tends to decrease as the system becomes more crowded* (i increases).

Example 4.6.7 (Optimal Routing for a Two-Station System)

Consider the system consisting of two queues shown in Fig. 4.6.4. Customers arrive according to a Poisson process with rate λ and are routed upon arrival to one of the two queues. Service times are independent and exponentially

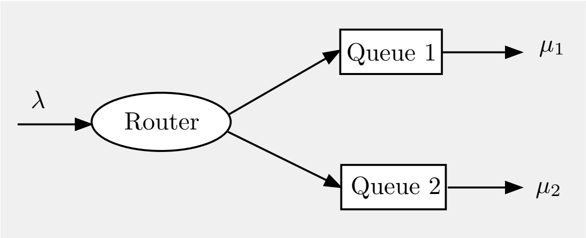


Figure 4.6.4 Queueing system of Example 4.6.7. The problem is to route each arriving customer to queue 1 or 2 so as to minimize the total average discounted waiting cost.

distributed with parameter μ_1 in the first queue and μ_2 in the second queue. The cost is

$$\lim_{N\to\infty} E\left\{\int_0^{t_N} e^{-\beta t}\big(c_1x_1(t)+c_2x_2(t)\big)dt\right\},$$

where β, c_1, and c_2 are given positive scalars, and $x_1(t)$ and $x_2(t)$ denote the number of customers at time t in queues 1 and 2, respectively.

As earlier, we construct the uniform version of this problem with uniform rate

$$\nu = \lambda + \mu_1 + \mu_2$$

and the transition probabilities shown in Fig. 4.6.5. We take as state space the set of pairs (i,j) of customers in queues 1 and 2. Bellman's equation takes the form

$$\begin{aligned} J(i,j) = {} & \frac{1}{\beta+\nu}\Big(c_1 i + c_2 j + \mu_1 J\big((i-1)^+,j\big) + \mu_2 J\big(i,(j-1)^+\big)\Big) \\ & + \frac{\lambda}{\beta+\nu}\min\big[J(i+1,j), J(i,j+1)\big], \end{aligned} \tag{4.114}$$

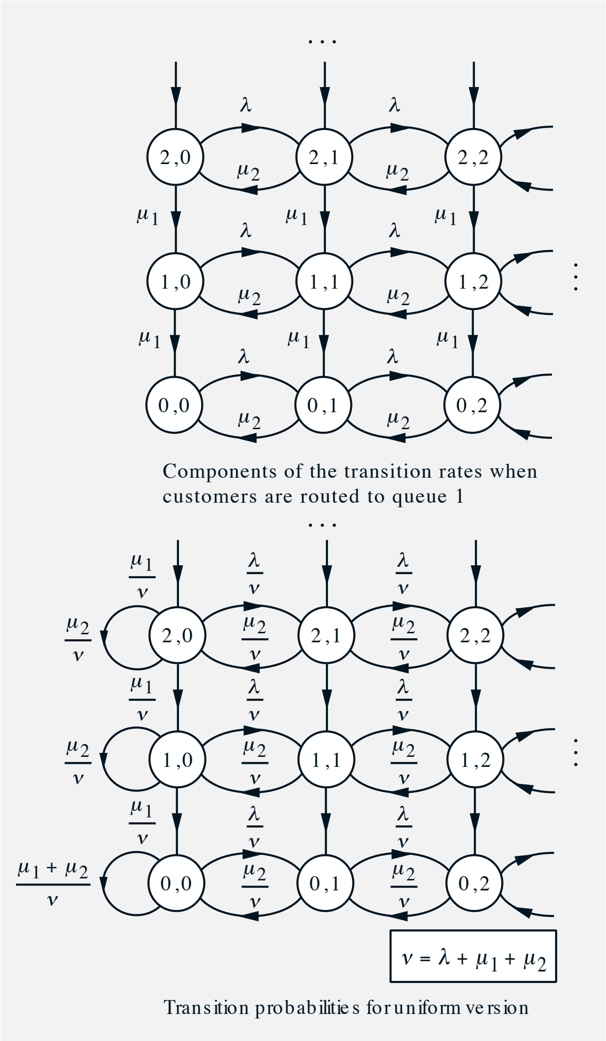
where for any x we denote

$$(x)^+ = \max(0,x).$$

From this equation we see that an optimal policy is to route an arriving customer to queue 1 if and only if the state (i,j) at the time of arrival belongs to the set

$$X_1 = \big\{(i,j) \mid J(i+1,j) \le J(i,j+1)\big\}. \tag{4.115}$$

This optimal policy can be characterized better by some further analysis. Intuitively, one expects that optimal routing can be achieved by sending a customer to the queue that is "less crowded" in some sense. It is therefore natural to conjecture that, if it is optimal to route to the first queue when the state is (i,j), it must be optimal to do the same when the first queue is even less crowded; i.e., the state is $(i-m,j)$ with $m \ge 1$. This is equivalent



Components of the transition rates when customers are routed to queue 1

Transition probabilities for uniform version

Figure 4.6.5 Continuous-time Markov chain and uniform version for Example 4.6.7 when customers are routed to the first queue. The states are the pairs of customer numbers in the two queues.

to saying that the set of states X_1 for which it is optimal to route to the first queue is characterized by a monotonically nondecreasing *threshold function* F by means of

$$X_1 = \{(i, u) \mid i = F(j)\} \tag{4.116}$$

(see Fig. 4.6.6). Accordingly, we call the corresponding optimal policy a *threshold policy.*

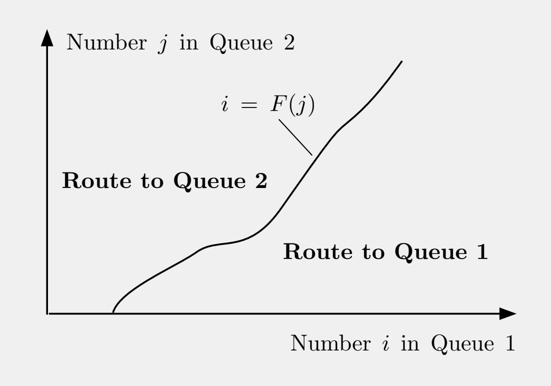


Figure 4.6.6 A threshold policy characterized by a threshold function F.

We will demonstrate the existence of a threshold optimal policy by showing that the functions

$$\Delta_1(i,j) = J(i+1,j) - J(i,j+1),$$

$$\Delta_2(i,j) = J(i,j+1) - J(i+1,j)$$

are monotonically nondecreasing in i for each fixed j, and in j for each fixed i, respectively. We will show this property for Δ_1; the proof for Δ_2 is analogous. It will be sufficient to show that for all $k = 0, 1, \ldots$, the functions

$$\Delta_1^k(i,j) = J_k(i+1,j) - J_k(i,j+1) \tag{4.117}$$

are monotonically nondecreasing in i for each fixed j, where J_k is generated by the DP recursion starting from the zero function; i.e., $J_{k+1}(i,j) = (TJ_k)(i,j)$, where T is the DP mapping defining Bellman's equation (4.114) and $J_0 = 0$. This is true because $J_k(i,j) \to J(i,j)$ for all i,j as $k \to \infty$ [Prop. 4.1.7(a)]. To prove that $\Delta_1^k(i,j)$ has the desired property, it is useful to first verify that $J_k(i,j)$ is monotonically nondecreasing in i (or j) for fixed j (or i). This is simple to show by induction or by arguing from first principles using the fact that $J_k(i,j)$ has a k-stage optimal cost interpretation. Next we use Eqs. (4.114) and (4.117) to write

$$\begin{aligned}(\beta+\nu)\Delta_1^{k+1}(i,j) = c_1 - c_2 & \\ &+ \mu_1\Big(J_k(i,j) - J_k\big((i-1)^+, j+1\big)\Big) \\ &+ \mu_2\Big(J_k\big(i+1,(j-1)^+\big) - J_k(i,j)\Big) \\ &+ \lambda\Big(\min\big[J_k(i+2,j), J_k(i+1,j+1)\big] \\ &- \min\big[J_k(i+1,j+1), J_k(i,j+2)\big]\Big).\end{aligned} \tag{4.118}$$

We now argue by induction. We have $\Delta_1^0(i,j) = 0$ for all (i,j). We assume that $\Delta_1^k(i,j)$ is monotonically nondecreasing in i for fixed j, and show that the same is true for $\Delta_1^{k+1}(i,j)$. This can be verified by showing that each of the terms in the right-hand side of Eq. (4.118) is monotonically nondecreasing in i for fixed j. Indeed, the first term is constant, and the second and third terms are seen to be monotonically nondecreasing in i using the induction hypothesis for the case where $i,j>0$ and the earlier shown fact that $J_k(i,j)$ is monotonically nondecreasing in i for the case where $i=0$ or $j=0$. The last term on the right-hand side of Eq. (4.118) can be written as

$$
\begin{aligned}
\lambda\Big(J_k(i+1,j+1) &+ \min\big[J_k(i+2,j) - J_k(i+1,j+1), 0\big] \\
&\quad - J_k(i+1,j+1) - \min\big[0, J_k(i,j+2) - J_k(i+1,j+1)\big]\Big) \\
&= \lambda\Big(\min\big[0, J_k(i+1,j) - J_k(i+1,j+1)\big] \\
&\quad + \max\big[0, J_k(i+1,j+1) - J_k(i,j+2)\big]\Big) \\
&= \lambda\Big(\min\big[0, \Delta_1^k(i+1,j)\big] + \max\big[0, \Delta_1^k(i,j+1)\big]\Big).
\end{aligned}
$$

Since $\Delta_1^k(i+1,j)$ and $\Delta_1^k(i,j+1)$ are monotonically nondecreasing in i by the induction hypothesis, the same is true for the preceding expression. Therefore, each of the terms on the right-hand side of Eq. (4.118) is monotonically nondecreasing in i, and the induction proof is complete. Thus the existence of an optimal threshold policy is established.

There are a number of generalizations of the routing problem of this example that admit a similar analysis and for which there exist optimal policies of the threshold type. For example, suppose that there are additional Poisson arrival processes with rates λ_1 and λ_2 at queues 1 and 2, respectively. The existence of an optimal threshold policy can be shown by a nearly verbatim repetition of our analysis. A more substantive extension is obtained when there is additional service capacity μ that can be switched at the times of transition due to an arrival or service completion to serve a customer in queue 1 or 2. Then we can similarly prove that it is optimal to route to queue 1 if and only if $(i,j) \in X_1$ and to switch the additional service capacity to queue 2 if and only if $(i+1,j+1) \in X_1$, where X_1 is given by Eq. (4.115) and is characterized by a threshold function as in Eq. (4.116). For a proof of this and further extensions, we refer to [Haj84], which generalizes and unifies several earlier results on the subject.

4.6.5 Nonstationary and Periodic Problems

Our standing assumption so far has been that the problem involves a stationary system and a stationary cost per stage (except for the presence of the discount factor). Problems with nonstationary system or cost per stage arise occasionally in practice or in theoretical studies and are thus of some interest. It turns out that such problems can be converted to stationary ones by a simple reformulation. We can then obtain results analogous to those obtained earlier for stationary problems.

Consider a nonstationary system of the form

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

and a cost function of the form

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

In these equations, for each k, x_k belongs to a space X_k, u_k belongs to a space U_k and satisfies $u_k \in U_k(x_k)$ for all $x_k \in X_k$, and w_k belongs to a countable space W_k. The sets X_k, U_k, $U_k(x_k)$, W_k may differ from one stage to the next. The random disturbances w_k are characterized by probabilities $P_k(\cdot \mid x_k, u_k)$, which depend on x_k and u_k as well as the time index k. The set of admissible policies Π is the set of all sequences $\pi = \{\mu_0, \mu_1, \ldots\}$ with $\mu_k : X_k \mapsto U_k$ and $\mu_k(x_k) \in U_k(x_k)$ for all $x_k \in X_k$ and $k = 0, 1, \ldots$ The functions $g_k : X_k \times U_k \times W_k \mapsto \Re$ are given and are assumed to satisfy one of the following three assumptions:

Assumption D′: We have $\alpha < 1$, and the functions g_k satisfy, for all $k = 0, 1, \ldots$,

$$\big|g_k(x_k, u_k, w_k)\big| \le M, \qquad \text{for all } (x_k, u_k, w_k) \in X_k \times U_k \times W_k,$$

where M is some scalar.

Assumption P′: The functions g_k satisfy, for all $k = 0, 1, \ldots$,

$$0 \le g_k(x_k, u_k, w_k), \qquad \text{for all } (x_k, u_k, w_k) \in X_k \times U_k \times W_k.$$

Assumption N′: The functions g_k satisfy, for all $k = 0, 1, \ldots$,

$$g_k(x_k, u_k, w_k) \le 0, \qquad \text{for all } (x_k, u_k, w_k) \in X_k \times U_k \times W_k.$$

We will refer to the problem formulated as the *nonstationary problem* (NSP for short). We can get an idea on how the NSP can be converted to a stationary problem by considering the special case where the state space is the same for each stage (i.e., $X_k = X$ for all k). We consider an augmented state

$$\tilde{x} = (x, k),$$

where $x \in X$, and k is the time index. The new state space is $\tilde{X} = X \times K$, where K denotes the set of nonnegative integers. The augmented system evolves according to

$$(x, k) \to \big(f_k(x, u_k, w_k), k+1\big), \qquad (x, k) \in \tilde{X}.$$

Similarly, we can define a cost per stage as

$$\tilde{g}\big((x, k), u_k, w_k\big) = g_k(x, u_k, w_k), \qquad (x, k) \in \tilde{X}.$$

It is evident that the problem corresponding to the augmented system is stationary. If we restrict attention to initial states $\tilde{x}_0 \in X \times \{0\}$, it can be seen that this stationary problem is equivalent to the NSP.

Let us now consider the more general case. To simplify notation, we will assume that the state spaces X_i, $i = 0, 1, \ldots$, the control spaces U_i, $i = 0, 1, \ldots$, and the disturbance spaces W_i, $i = 0, 1, \ldots$, are all mutually disjoint. This assumption does not involve a loss of generality since, if necessary, we may relabel the elements of X_i, U_i, and W_i without affecting the structure of the problem. Define now a new state space X, a new control space U, and a new (countable) disturbance space W by

$$X = \cup_{i=0}^{\infty} X_i, \qquad U = \cup_{i=0}^{\infty} U_i, \qquad W = \cup_{i=0}^{\infty} W_i.$$

Introduce a new (stationary) system

$$\tilde{x}_{k+1} = f(\tilde{x}_k, \tilde{u}_k, \tilde{w}_k), \qquad k = 0, 1, \ldots,$$

where $\tilde{x}_k \in X$, $\tilde{u}_k \in U$, $\tilde{w}_k \in W$, and the system function $f : X \times U \times W \mapsto X$ is defined by

$$f(\tilde{x}, \tilde{u}, \tilde{w}) = f_i(\tilde{w}, \tilde{u}, \tilde{w}), \qquad \text{if } \tilde{x} \in X_i, \quad \tilde{u} \in U_i, \quad \tilde{w} \in W_i, \quad i = 0, 1, \ldots$$

For triplets $(\tilde{x}, \tilde{u}, \tilde{w})$, where for some $i = 0, 1, \ldots$, we have $\tilde{x} \in X_i$, but $\tilde{u} \notin U_i$ or $\tilde{w} \notin W_i$, the definition of f is immaterial; any definition is adequate for our purposes in view of the control constraints to be introduced. The control constraint is taken to be $\tilde{u} \in U(\tilde{x})$ for all $\tilde{x} \in X$, where $U(\cdot)$ is defined by

$$U(\tilde{x}) = U_i(\tilde{x}), \qquad \text{if } \tilde{x} \in X_i, \quad i = 0, 1, \ldots$$

The disturbance $\tilde{w}$ is characterized by probabilities $P(\tilde{w} \mid \tilde{x}, \tilde{u})$ such that

$$P(\tilde{w} \in W_i \mid \tilde{x} \in X_i, \tilde{u} \in U_i) = 1, \qquad i = 0, 1, \ldots,$$

$$P(\tilde{w} \notin W_i \mid \tilde{x} \in X_i, \tilde{u} \in U_i) = 0, \qquad i = 0, 1, \ldots$$

Furthermore, for any $w_i \in W_i$, $x_i \in X_i$, $u_i \in U_i$, $i = 0, 1, \ldots$, we have

$$P(w_i \mid x_i, u_i) = P_i(w_i \mid x_i, u_i).$$

We also introduce a new cost function

$$\tilde{J}_{\tilde{\pi}}(\tilde{x}_0) = \lim_{N\to\infty} \mathop{E}_{\substack{w_k \\ k=0,1,\ldots,N-1}} \left\{ \sum_{k=0}^{N-1} \alpha^k g\big(\tilde{x}_k, \mu_k(\tilde{x}_k), \tilde{w}_k\big) \right\},$$

where the (stationary) cost per stage $g : X \times U \times W \mapsto \Re$ is defined for all $i = 0, 1, \ldots,$ by

$$g(\tilde{x}, \tilde{u}, \tilde{w}) = g_i(\tilde{x}, \tilde{u}, \tilde{w}), \quad \text{if } \tilde{x} \in X_i, \quad \tilde{u} \in U_i, \quad \tilde{w} \in W_i.$$

For triplets $(\tilde{x}, \tilde{u}, \tilde{w})$, where for some $i = 0, 1, \ldots,$ we have $\tilde{x} \in X_i$ but $\tilde{u} \notin U_i$ or $\tilde{w} \notin W_i$, any definition of g is adequate provided $\big|g(\tilde{x}, \tilde{u}, \tilde{w})\big| \leq M$ for all $(\tilde{x}, \tilde{u}, \tilde{w})$ when Assumption D′ holds, $0 \leq g(\tilde{x}, \tilde{u}, \tilde{w})$ when P′ holds, and $g(\tilde{x}, \tilde{u}, \tilde{w}) \leq 0$ when N′ holds. The set of admissible policies $\tilde{\Pi}$ for the new problem consists of all sequences $\tilde{\pi} = \{\tilde{\mu}_0, \tilde{\mu}_1, \ldots\}$, where $\tilde{\mu}_k : X \mapsto U$ and $\tilde{\mu}_k(\tilde{x}) \in U(\tilde{x})$ for all $\tilde{x} \in X$ and $k = 0, 1, \ldots$.

The construction given defines a problem that clearly fits the framework of the infinite horizon total cost problem. We will refer to this problem as the *stationary problem* (SP for short).

It is important to understand the nature of the intimate connection between the NSP and the SP formulated here. Let $\pi = \{\mu_0, \mu_1 \ldots\}$ be an admissible policy for the NSP. Also, let $\tilde{\pi} = \{\tilde{\mu}_0, \tilde{\mu}_1, \ldots\}$ be an admissible policy for the SP such that

$$\tilde{\mu}_i(\tilde{x}) = \mu_i(\tilde{x}), \qquad \text{if } \tilde{x} \in X_i, \quad i = 0, 1, \ldots \tag{4.119}$$

Let $x_0 \in X_0$ be the initial state for the NSP and consider the same initial state for the SP (i.e., $\tilde{x}_0 = x_0 \in X_0$). Then the sequence of states $\{\tilde{x}_i\}$ generated in the SP will satisfy $\tilde{x}_i \in X_i$, $i = 0, 1, \ldots,$ with probability 1 (i.e., the system will move from the set X_0 to the set X_1, then to X_2, etc., just as in the NSP). Furthermore, the probabilistic law of generation of states and costs is identical in the NSP and the SP. As a result, it is easy to see that for any admissible policies π and $\tilde{\pi}$ satisfying Eq. (4.119) and initial states x_0, $\tilde{x}_0$ satisfying $x_0 = \tilde{x}_0 \in X_0$, the sequence of generated states in the NSP and the SP is the same ($x_i = \tilde{x}_i$, for all i) provided the generated disturbances w_i and $\tilde{w}_i$ are also the same for all i ($w_i = \tilde{w}_i$, for all i). Furthermore, if π and $\tilde{\pi}$ satisfy Eq. (4.119), we have $J_\pi(x_0) = \tilde{J}_\pi(\tilde{x}_0)$ if $x_0 = \tilde{x}_0 \in X_0$. Let us also consider the optimal cost functions for the NSP and the SP:

$$J^*(x_0) = \min_{\pi\in\Pi} J_\pi(x_0), \qquad x_0 \in X_0,$$

$$\tilde{J}^*(\tilde{x}_0) = \min_{\tilde{\pi}\in\tilde{\Pi}} J_\pi(\tilde{x}_0), \qquad \tilde{x}_0 \in X_0.$$

Then it follows from the construction of the SP that

$$\tilde{J}^*(\tilde{x}_0) = \tilde{J}^*(\tilde{x}_0, i), \qquad \text{if } \tilde{x}_0 \in X_i, \qquad i = 0, 1, \ldots,$$

where, for all $i = 0, 1, \ldots$,

$$\tilde{J}^*(\tilde{x}_0, i) = \min_{\pi \in \Pi} \lim_{N \to \infty} \mathop{E}_{\substack{w_k \\ k=0,1,\ldots,N-1}} \left\{ \sum_{k=i}^{N-1} \alpha^{k-i} g_k\big(x_k, \mu_k(x_k), w_k\big) \right\}, \tag{4.120}$$

if $\tilde{x}_0 = x_i \in X_i$. Note that in this equation, the right-hand side is defined in terms of the data of the NSP. As a special case of this equation, we obtain

$$\tilde{J}^*(\tilde{x}_0) = \tilde{J}^*(\tilde{x}_0, 0) = J^*(x_0), \qquad \text{if } \tilde{x}_0 = x_0 \in X_0.$$

Thus *the optimal cost function J^* of the NSP can be obtained from the optimal cost function $\tilde{J}^*$ of the SP*. Furthermore, if $\tilde{\pi}^* = \{\tilde{\mu}_0^*, \tilde{\mu}_1^*, \ldots\}$ is an optimal policy for the SP, then the policy $\pi^* = \{\mu_0^*, \mu_1^*, \ldots\}$ defined by

$$\mu_i^*(x_i) = \tilde{\mu}_i^*(x_i), \qquad \text{for all } x_i \in X_i, \qquad i = 0, 1, \ldots, \tag{4.121}$$

is an optimal policy for the NSP. *Thus optimal policies for the SP yield optimal policies for the NSP via Eq. (4.121).* Another point to be noted is that *if Assumption D′ (P′, N′) is satisfied for the NSP, then Assumption D (P, N) introduced earlier in this chapter is satisfied for the SP.*

These observations show that one may analyze the NSP by means of the SP. Every result given in the preceding sections when applied to the SP yields a corresponding result for the NSP. We will just provide the form of the optimality equation for the NSP in the following proposition.

Proposition 4.6.2: Under Assumption D′ (P′, N′), there holds

$$J^*(x_0) = \tilde{J}^*(x_0, 0), \qquad x_0 \in X_0,$$

where for all $i = 0, 1, \ldots$, the functions $\tilde{J}^*(\cdot, i)$ map X_i into $\Re$ ($[0, \infty]$, $[-\infty, 0]$), are given by Eq. (4.120), and satisfy for all $x_i \in X_i$ and $i = 0, 1, \ldots$,

$$\tilde{J}^*(x_i, i) = \min_{u_i \in U_i(x_i)} \mathop{E}_{w_i} \big\{ g_i(x_i, u_i, w_i) + \alpha \tilde{J}^*\big(f_i(x_i, u_i, w_i), i+1\big)\big\}. \tag{4.122}$$

Under Assumption D′ the functions $\tilde{J}^*(\cdot, i)$, $i = 0, 1, \ldots$, are the unique bounded solutions of the set of equations Eq. (4.122). Furthermore, under Assumption D′ or P′, if $\mu_i^*(x_i) \in U_i(x_i)$ attains the minimum in Eq. (4.122) for all $x_i \in X_i$ and i, then the policy $\pi^* = \{\mu_0^*, \mu_1^*, \ldots\}$ is optimal for the NSP.

Periodic Problems

Assume within the framework of the NSP that there exists an integer $p \geq 2$ (called the *period*) such that for all integers i and j with

$$|i - j| = mp, \qquad m = 1, 2, \ldots,$$

we have

$$X_i = X_j, \qquad U_i = U_j, \qquad W_i = W_j, \qquad U_i(\cdot) = U_j(\cdot),$$

$$f_i = f_j, \qquad g_i = g_j, \qquad P_i(\cdot \mid x, j) = P_j(\cdot \mid x, u), \qquad (x, u) \in X_i \times U_i.$$

We assume that the spaces X_i, U_i, W_i, $i = 0, 1, \ldots, p-1$, are mutually disjoint. We define new state, control, and disturbance spaces by

$$X = \cup_{i=0}^{p-1} X_i, \qquad U = \cup_{i=0}^{p-1} U_i, \qquad W = \cup_{i=0}^{p-1} W_i.$$

The optimality equation for the equivalent stationary problem reduces to the system of p equations

$$\tilde{J}^*(x_0, 0) = \min_{u_0 \in U_0(x_0)} \mathop{E}_{w_0} \bigl\{ g_0(x_0, u_0, w_0) + \alpha \tilde{J}^*\bigl(f_0(x_0, u_0, w_0), 1\bigr) \bigr\},$$

$$\tilde{J}^*(x_1, 1) = \min_{u_1 \in U_1(x_1)} \mathop{E}_{w_1} \bigl\{ g(x_1, u_1, w_1) + \alpha \tilde{J}^*\bigl(f_1(x_1, u_1, w_1), 2\bigr) \bigr\},$$

$$\vdots$$

$$\begin{aligned} \tilde{J}^*\bigl(x_{p-1}, p-1\bigr) = \min_{u_{p-1} \in U_{p-1}(x_{p-1})} \mathop{E}_{w_{p-1}} & \bigl\{ g_{p-1}(x_{p-1}, u_{p-1}, w_{p-1}) \\ & + \alpha \tilde{J}^*\bigl(f_{p-1}(x_{p-1}, u_{p-1}, w_{p-1}), 0\bigr) \bigr\}. \end{aligned}$$

These equations may be used to obtain (under Assumption D′ or P′) a periodic policy of the form $\{\mu_0^*, \ldots, \mu_{p-1}^*, \mu_0^*, \ldots, \mu_{p-1}^*, \ldots\}$ whenever the minimum of the right-hand side is attained for all x_i, $i = 0, 1, \ldots, p-1$. When all spaces involved are finite and $\alpha < 1$, an optimal policy may be found by means of the algorithms of Chapter 2, appropriately adapted to the corresponding SP.

4.7 NOTES, SOURCES, AND EXERCISES

Undiscounted problems and discounted problems with unbounded cost per stage were first analyzed systematically by Blackwell [Bla65], Dubins and Savage [DuS65], and Strauch [Str66] (who was Blackwell's PhD student). The sufficient conditions for convergence of the VI method under Assumption P [cf. Props. 4.1.7(a) and 4.1.8], together with necessary conditions for convergence were derived by the author in [Ber75], [Ber77]. Further results were given by Schal [Sch75] and Whittle [Whi80]. The convergence of VI from above condition of Prop. 4.1.9 was obtained by Yu and Bertsekas [YuB13]. Problems involving convexity assumptions are analyzed in the author's [Ber73b]. The optimistic PI convergence analysis under Assumption N was given in the author's monograph [Ber13], Section 4.3.3, which also proposed a related method based on λ-policy iteration (see Exercise 6.13 in Chapter 3).

The paper by Blackwell [Bla65] also introduced the Borel space framework to deal with the issues that arise when the disturbance space is uncountably infinite, so that measurability restrictions must be placed on the policies (see Appendix A). These issues were also addressed in a number of subsequent works, including the monographs by Hinderer [Hin70], Stiebel [Str75], Bertsekas and Shreve [BeS78], Dynkin and Yushkevich [DuY79], and Hernandez-Lerma [Her89], and the papers by Strauch [Str66], and Blackwell, Freedman, and Orkin [BFO74]. The monograph [BeS78] provides an extensive treatment (summarized in Appendix A), based on the use of universally measurable policies.

A question left open within the Borel and universally measurable frameworks is the validity of PI [it is not guaranteed that the policy improvement operation can produce a universally measurable policy, even if the minimum is attained for all $x \in X$ when computing $(TJ_\mu)(x)$, because J_μ may be lower semianalytic; see Appendix A]. The recent paper by Yu and Bertsekas [YuB13] resolves this difficulty by using a combined VI and PI approach that bears some similarity with the approach of Section 2.6.3. This paper also derived a number of other results relating to VI and PI within a Borel and a universal measurability framework, including the result regarding convergence of VI from above under Assumption P, given in Prop. 4.1.9.

The analysis of Section 4.1.4, which converts a finite-state finite-control stochastic control problem under Assumption P to an equivalent SSP, is joint work of the author with H. Yu. See the paper [BeY15], which also considers the finite-state infinite-control case under Assumption P, where the control constraint set is assumed to satisfy a compactness condition like the one of Prop. 4.1.8. Example 4.1.4 illustrates what may happen under these circumstances.

We have bypassed a number of complex theoretical issues under Assumptions P and N, which historically have played an important role and relate to stationary policies. The main question is to what extent is it possible to restrict attention to such policies. Much theoretical work has been done on this question (Bertsekas and Shreve [BeS79], Blackwell [Bla65], Blackwell [Bla70], Dubins and Savage [DuS65], Feinberg [Fei78], [Fei92a], [Fei92b], Ornstein [Orn69]), and some aspects are still open. Suppose, for example, that we are given an $\epsilon > 0$. One issue is whether there exists an

ϵ-optimal stationary policy, i.e., a stationary policy μ such that

$$J_\mu(x) \le J^*(x) + \epsilon, \qquad \text{for all } x \in X \text{ with } J^*(x) > -\infty,$$

$$J_\mu(x) \le -\frac{1}{\epsilon}, \qquad \text{for all } x \in X \text{ with } J^*(x) = -\infty.$$

The answer is positive under any one of the following conditions:

1. Assumption P holds and $\alpha < 1$ (see Exercise 4.5).
2. Assumption N holds, X is a finite set, $\alpha = 1$, and $J^*(x) > -\infty$ for all $x \in X$ (see Exercise 4.6 or Blackwell [Bla65], [Bla70], and Ornstein [Orn69]).
3. Assumption N holds, X is a countable set, $\alpha = 1$, and the problem is deterministic (see Bertsekas and Shreve [BeS79]).

The answer can be negative under any one of the following conditions:

1. Assumption P holds and $\alpha = 1$ (see Exercise 4.5).
2. Assumption N holds and $\alpha < 1$ (see Exercise 4.6, or Bertsekas and Shreve [BeS79]).

The existence of an ϵ-optimal stationary policy for SSP problems with a finite state space, but under somewhat different assumptions than the ones of Section 3.1, is analyzed by Feinberg [Fei92b].

Another issue is whether one can confine the search for an optimal policy within the class of stationary policies, i.e., whether there exists an optimal stationary policy when there exists an optimal policy for each initial state. This is true under Assumption P (see Exercise 4.10). It is also true (but very hard to prove) under Assumption N if $J^*(x) > -\infty$ for all $x \in X$, $\alpha = 1$, and the disturbance space W is countable (Blackwell [Bla70], Dubins and Savage [DuS65], Ornstein [Orn69]). Simple two-state examples can be constructed showing that the result fails to hold if $\alpha = 1$ and $J^*(x) = -\infty$ for some state x (see Exercises 4.11 and 4.12). However, these examples rely on the presence of a stochastic element in the problem. If the problem is deterministic, stronger results are available; one can find an optimal stationary policy if there exists an optimal policy at each initial state and either $\alpha = 1$ or $\alpha < 1$ and $J^*(x) > -\infty$ for all $x \in X$. These results also require a difficult proof (Bertsekas and Shreve [BeS79]).

There has been much research on VI and PI algorithms for discrete-time deterministic optimal control (Section 4.2), and more recently in the context of adaptive DP (cf. Section 4.3). For a selective list of recent references, which themselves contain extensive lists of other references, see the book by Vrabie, Vamvoudakis, and Lewis [VVL13], the papers by Jiang and Jiang [JiJ13], [JiJ14], Heydari [Hey14a], [Hey14b], Liu and Wei [LiW13], Wei et al. [WWL14], the survey papers in the edited volumes by Si et al. [SBP04], and Lewis and Liu [LeL13], and the special issue edited by Lewis, Liu, and Lendaris [LLL08]. Some of these works relate to continuous-time problems as well, and in their treatment of algorithmic convergence, typically assume that X and U are Euclidean spaces, as well as continuity and other conditions on g, special structure of the system, etc. Important antecedents of these works, which deal with PI algorithms for deterministic continuous-time optimal control are the papers by Rekasius [Rek64], and Saridis and Lee [SaL79], and the thesis by Beard [Bea95] (supervised by Saridis).

The results and analysis of Section 4.2 on deterministic optimal control to a terminal set of states were given in the author's paper [Ber15b]. The line of analysis of this section is inspired by extensions of the abstract DP theory given in Sections 1.6 and 2.5 for contractive problems. This extended analysis is outlined in Appendix B, and may be applied more broadly, to noncontractive problems, including the SSP problems of Chapter 3 and Section 4.4. Other examples are positive cost and negative cost problems, such as the ones of Section 4.1, minimax problems of the shortest path type (see [Ber14] and Appendix B), and risk sensitive shortest path-type problems (see Section 4.5). Central in this analysis are notions of *regularity*, which extend the notion of a proper policy in SSP problems, and were developed in the author's abstract DP monograph [Ber13] and the paper [Ber15a].

Let us describe the regularity idea briefly, as given in [Ber15a], and its connection to the analysis of Section 4.2 (see also Appendix B). Given a set of functions $S \in E^+(X)$, we say that a collection $\mathcal{C}$ of policy-state pairs (π, x_0), with $\pi \in \Pi$ and $x_0 \in X$, is S*-regular* if for all $(\pi, x_0) \in \mathcal{C}$ and $J \in S$, we have

$$J_\pi(x_0) = \lim_{N\to\infty} \left\{ J(x_N) + \sum_{k=0}^{N-1} g\big(x_k, \mu_k(x_k)\big) \right\}.$$

In words, for all $(\pi, x_0) \in \mathcal{C}$, $J_\pi(x_0)$ can be obtained in the limit by VI starting from any $J \in S$. The favorable properties with respect to VI of an S-regular collection $\mathcal{C}$ can be translated into interesting properties relating to solutions of Bellman's equation and convergence of VI. In particular, the optimal cost function over the set of policies $\big\{\pi \mid (\pi, x) \in \mathcal{C}\big\}$,

$$J^*_{\mathcal{C}}(x) = \min_{\{\pi \,|\, (\pi,x)\in\mathcal{C}\}} J_\pi(x), \qquad x \in X,$$

under appropriate problem-dependent assumptions, is the unique solution of Bellman's equation within the set

$$\big\{J \in S \mid J \geq J^*_{\mathcal{C}}\big\},$$

and can be obtained by VI starting from any J within that set (cf. the SSP analysis of Section 4.4).

Within the deterministic optimal control context of Section 4.2, it works well to choose $\mathcal{C}$ to be the set of all (π, x) such that $x \in X_f$ and π is terminating starting from x, and to choose S to be $\mathcal{J}$, as defined by Eq. (4.34). Then, in view of Assumption 4.2.1, we have $J^*_{\mathcal{C}} = J^*$, and the favorable properties of $J^*_{\mathcal{C}}$ are shared by J^*. For other types of problems different choices of $\mathcal{C}$ may be appropriate, and corresponding results relating to the uniqueness of solutions of Bellman's equation and the validity of VI and PI may be obtained; see [Ber15a].

The proposal of PI for infinite horizon linear-quadratic problems (cf. Section 4.3) is due to Kleinman [Kle68]. Using simulation-based PI for adaptive control of linear-quadratic models, the starting point for the methodology of adaptive DP (cf. Section 4.3.1), was first proposed by Bradtke, Ydstie, and Barto [BYB94]. There has been a lot of followup work based on this idea, some of which was noted earlier in connection with Section 4.2.

The SSP material and the perturbation-based analysis of Section 4.4 is joint work of the author with H. Yu. Together with the connection of SSP with finite state stochastic control problems under Assumption P developed in Section 4.1.4, it was given in the paper [BeY15]. This paper develops the theory in a more general setting where the control space may be infinite, but the compactness conditions given in Section 3.2 are assumed.

A class of problems that includes the positive and negative cost problems of Section 4.1 and the SSP problem of Section 4.4, is obtained when we allow both positive and negative transition costs, but without explicitly assuming a termination state. The survey by Feinberg [Fei02] overviews the theory of these problems, and the paper by Yu [Yu14] addresses the associated intricacies of the convergence of VI; see also the theory of positive bounded MDP discussed in Section 7.2 of [Put94]. These works require certain cost function convergence assumptions, which may not be satisfied in some SSP contexts. In particular, when specialized to deterministic shortest path problems, these assumptions require that each zero length cycle consists of zero length transitions.

The affine monotonic model of Section 4.5 was introduced in more general form (infinite state space) in the author's abstract DP monograph [Ber13]. The treatment given here also shares ideas with the material of Section 4.5, which is joint work of the author with H. Yu [BeY15]. The exponentiated cost special case of the affine monotonic model has received considerable attention; see Denardo and Rothblum [DeR79], [DeR06], who use a different line of analysis based on linear programming, and Patek [Pat01], who considers the monotone increasing case where $T\bar{J} \geq \bar{J}$.

A general analysis of (possibly infinite-state) shortest path-type problems, which relies on notions of regularity that extend the notion of a proper policy, has been developed in Chapter 3 of the author's abstract DP monograph [Ber13]. Problems that admit such an analysis were called *semicontractive* in that monograph, in view of the fact that for some policies μ (such as the ones that are proper, stable, etc) the mapping T_μ is a contraction, while for other policies (such as improper, unstable, etc) it is not. Aside from SSP and affine monotonic, semicontractive models include shortest path-type minimax problems; see Bertsekas [Ber14] and Appendix B.

The material of Sections 4.6.1, 4.6.2, and 4.6.5 is classical, and dates to the early days of DP. The material on the gambling problem of Section 4.6.3 is taken from the important work of Dubins and Savage [DuS65]. A surprising property of the optimal reward function J^* for this problem has been shown by Billingsley [Bil83]: J^* is almost everywhere differentiable with derivative zero, yet it is strictly increasing, taking values that range from 0 to 1. Control of queueing systems and problems of priority assignment and routing (Section 4.6.4) have been researched extensively. We give some representative references: Ayoun and Rosberg [AyR91], Baras, Dorsey, and Makowski [BDM83], Bhattacharya and Ephremides [BhE91], Courcoubetis and Varaiya [CoV84], Cruz and Chuah [CrC91], Ephremides, Varaiya, and Walrand [EVW80], Ephremides and Verd'u [EpV89], Hajek [Haj84], Harrison [Har75a], [Har75b], Lin and Kumar [LiK84], Pattipati and Kleinman [PaK81], Stidham and Prabhu [StP74], Stidham and Weber [StW93], Suk and Cassandras [SuC91], Towsley, Sparaggis, and Cassandras [TSC92], Tsitsiklis [Tsi84], Viniotis and Ephremides [ViE88], and Walrand [Wal88].

EXERCISES

4.1 (VI Convergence Counterexample Under P)

Let $X = [0, \infty)$ and $U = U(x) = (0, \infty)$ be the state and control spaces, respectively, let the system equation be

$$x_{k+1} = \left(\frac{2}{\alpha}\right) x_k + u_k, \qquad k = 0, 1, \ldots,$$

where $\alpha \in (0, 2)$, and let

$$g(x_k, u_k) = x_k + u_k$$

be the cost per stage. Show that for this deterministic problem, Assumption P holds and that $J^*(x) = \infty$ for all $x \in X$, but $(T^k J_0)(0) = 0$ for all k [J_0 is the zero function, $J_0(x) = 0$, for all $x \in X$].

4.2 (Existence of an Optimal Stationary Policy Under P)

Let Assumption P hold and consider the finite-state case $X = \{1, \ldots, n\}$, $x_{k+1} = w_k$. The mapping T is represented as

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

where $p_{ij}(u)$ denotes the transition probability that the next state will be j when the current state is i and control u is applied. Assume that the sets $U(i)$ are compact subsets of $\Re^n$ for all i, and that $p_{ij}(u)$ and $g(i,u)$ are continuous on $U(i)$ for all i and j. Use Prop. 4.1.8 to show that $\lim_{k \to \infty} (T^k J_0)(i) = J^*(i)$, where J_0 is the zero vector, and that there exists an optimal stationary policy.

4.3 (Existence of an Optimal Stationary Policy Under N)

This exercise explores further Example 4.6.4, which involves a deterministic stopping problem where Assumption N holds, and an optimal policy does not exist, even though only two controls are available at each state (stop and continue). The state space is $X = \{1, 2, \ldots\}$. Continuation from state i leads to state $i + 1$ with certainty and no cost, while the stopping cost is $-1 + (1/i)$, so that there is an incentive to delay stopping at every state.

(a) Verify that $J^*(i) = -1$ for all i, but there is no policy (stationary or not) that attains the optimal cost starting from i.

(b) Let μ be the policy that stops at every state. Show that the next policy $\overline{\mu}$ generated by PI is to continue at every state, and that we have

$$J_{\overline{\mu}}(i) = 0 > -1 + \frac{1}{i} = J_\mu(i), \qquad i = 1, 2, \ldots.$$

Moreover, the method oscillates between the policies μ and $\overline{\mu}$, none of which is optimal.

4.4 (Error Bound Under P)

Under Assumption P, let μ be such that for all $x \in X$, $\mu(x) \in U(x)$ and

$$(T_\mu J^*)(x) \le (TJ^*)(x) + \epsilon,$$

where ϵ is some positive scalar. Show that, if $\alpha < 1$,

$$J_\mu(x) \le J^*(x) + \frac{\epsilon}{1 - \alpha}, \qquad x \in X.$$

Hint: Show that $(T_\mu^k J^*)(x) \le J^*(x) + \sum_{i=0}^{k-1} \alpha^i \epsilon$.

4.5 (Existence of ϵ-Optimal Policies Under P)

Under Assumption P, show that, given $\epsilon > 0$, there exists a policy $\pi \in \Pi$ such that $J_\pi(x) \leq J^*(x) + \epsilon$ for all $x \in X$, and that for $\alpha < 1$, π can be taken stationary. Give an example where $\alpha = 1$ and for each stationary policy μ we have $J_\mu(x) = \infty$, while $J^*(x) = 0$ for all x. *Hint*: See the proof of Prop. 4.1.1.

4.6 (Existence of ϵ-Optimal Policies Under N)

Let Assumption N hold and assume that $J^*(x) > -\infty$ for all $x \in X$.

(a) Show that if X is a finite set and $\alpha \leq 1$, then given $\epsilon > 0$, there exists a policy $\pi \in \Pi$ such that $J_\pi(x) \leq J^*(x) + \epsilon$ for all $x \in X$, and that for $\alpha < 1$ π can be taken stationary (cf. the result of Exercise 4.5). *Hint*: Consider an integer N such that the N-stage optimal cost J_N satisfies

$$J_N(x) \leq J^*(x) + \epsilon, \qquad x \in X.$$

(b) Construct a counterexample to show that the result of part (a) can fail to hold if X is countable and $\alpha < 1$. *Hint*: Consider a stopping problem with $X = \{0, 1, \ldots\}$, and a stopping cost at state $i \geq 0$ equal to $1 - (1/\alpha)^i$. If we do not stop at state i, we move to state $i + 1$ at no cost. See [BeS79], p. 609.

4.7

Under Assumption P or N, show that if $\alpha < 1$ and $J' : X \mapsto \Re$ is a *bounded* function satisfying $J' = TJ'$, then $J' = J^*$. *Hint*: Under P, let r be a scalar such that $J^* + re \geq J'$. Argue that $J^* \geq J'$ and use Prop. 4.1.3(a).

4.8

We want to find a scalar sequence $\{u_0, u_1, \ldots\}$ that satisfies $\sum_{k=0}^\infty u_k \leq c$, $u_k \geq 0$, for all k, and maximizes $\sum_{k=0}^\infty g(u_k)$, where $c > 0$ and $g(u) \geq 0$ for all $u \geq 0$, $g(0) = 0$. Assume that g is monotonically nondecreasing on $[0, \infty)$. Show that the optimal value of the problem is $J^*(c)$, where J^* is a monotonically nondecreasing function on $[0, \infty)$ satisfying $J^*(0) = 0$ and

$$J^*(x) = \max_{0 \leq u \leq x} \big\{g(u) + J^*(x - u)\big\}, \qquad x \in [0, \infty).$$

4.9

Let Assumption P hold and assume that $\pi^* = \{\mu_0^*, \mu_1^*, \ldots\} \in \Pi$ satisfies $J^* = T_{\mu_k^*} J^*$ for all k. Show that π^* is optimal, i.e., $J_{\pi^*} = J^*$.

4.10

Under Assumption P, show that if there exists an optimal policy (a policy $\pi^* \in \Pi$ such that $J_{\pi^*} = J^*$), then there exists an optimal stationary policy.

4.11

Use the following counterexample to show that the result of Exercise 4.10 may fail to hold under Assumption N if $J^*(x) = -\infty$ for some $x \in X$. Let $X = D = \{0, 1\}$, $f(x, u, w) = w$, $g(x, u, w) = u$, $U(0) = (-\infty, 0]$, $U(1) = \{0\}$, $p(w = 0 \mid x = 0, u) = \frac{1}{2}$, and $p(w = 1 \mid x = 1, u) = 1$. Show that $J^*(0) = -\infty$, $J^*(1) = 0$ and that the admissible nonstationary policy $\{\mu_0^*, \mu_1^*, \ldots\}$ with $\mu_k^*(0) = -(2/\alpha)^k$ is optimal. Show that every stationary policy μ satisfies $J_\mu(0) = \big(2/(2-\alpha)\big)\mu(0)$, $J_\mu(1) = 0$ (see [Bla70], [DuS65], and [Orn69] for related analysis).

4.12 (The Blackmailer's Dilemma)

Consider Example 3.2.1. Here, there are two states, state 1 and a termination state t. At state 1, we can choose a control u with $0 < u \leq 1$; we then move to state t at no cost with probability $p(u)$, and stay in state 1 at a cost $-u$ with probability $1 - p(u)$.

(a) Let $p(u) = u^2$. For this case it was shown in Example 3.2.1 that the optimal costs are $J^*(1) = -\infty$ and $J^*(t) = 0$. Furthermore, it was shown that there is no optimal stationary policy, although there is an optimal nonstationary policy. Find the set of solutions to Bellman's equation and verify the conclusion of Prop. 4.1.3(b).

(b) Let $p(u) = u$. Find the set of solutions to Bellman's equation and use Prop. 4.1.3(b) to show that the optimal costs are $J^*(1) = -1$ and $J^*(t) = 0$. Show that there is no stationary optimal policy.

4.13 (Linear Systems and Discounted Positive Cost)

Consider a deterministic problem involving a linear system

$$x_{k+1} = Ax_k + Bu_k, \qquad k = 0, 1, \ldots,$$

where the pair (A, B) is controllable and $x_k \in \Re^n$, $u_k \in \Re^m$. Assume no constraints on the control and a cost per stage g satisfying

$$0 \leq g(x, u), \qquad (x, u) \in \Re^n \times \Re^m.$$

Assume furthermore that g is continuous in x and u, and that $g(x_n, u_n) \to \infty$ if $\{x_n\}$ is bounded and $\|u_n\| \to \infty$.

(a) Show that for a discount factor $\alpha < 1$, the optimal cost satisfies $0 \leq J^*(x) < \infty$, for all $x \in \Re^n$. Furthermore, there exists an optimal stationary policy and

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

(b) Show that the same is true, except perhaps for $J^*(x) < \infty$, when the system is of the form $x_{k+1} = f(x_k, u_k)$, with $f : \Re^n \times \Re^m \mapsto \Re^n$ being a continuous function.

(c) Prove the same results assuming that the control is constrained to lie in a compact set $U \in \Re^m$ [$U(x) = U$ for all x] in place of the assumption $g(x_n, u_n) \to \infty$ if $\{x_n\}$ is bounded and $\|u_n\| \to \infty$. *Hint*: Show that $T^k J_0$ is real valued and continuous for every k, and use Prop. 4.1.8.

4.14 (Periodic Linear-Quadratic Problems)

Consider the linear system

$$x_{k+1} = A_k x_k + B_k u_k + w_k, \qquad k = 0, 1, \ldots,$$

and the quadratic cost

$$J_\pi(x_0) = \lim_{N\to\infty} \underset{\substack{w_k \\ k=0,\ldots N-1}}{E} \left\{ \sum_{k=0}^{N-1} \alpha^k (x_k' Q_k x_k + u_k' R_k u_k) \right\},$$

where the matrices have appropriate dimensions, Q_k and R_k are positive semidefinite and positive definite symmetric, respectively, for all k, and $0 < \alpha < 1$. Assume that the system and cost are periodic with period p (cf. Section 4.7), that the controls are unconstrained, and that the disturbances are independent, and have zero mean and finite covariance. Assume further that the following (controllability) condition is in effect.

For any state $\overline{x}_0$, there exists a finite sequence of controls $\{\overline{u}_0, \overline{u}_1, \ldots, \overline{u}_r\}$ such that $\overline{x}_{r+1} = 0$, where $\overline{x}_{r+1}$ is generated by

$$\overline{x}_{k+1} = A_k \overline{x}_k + B_k \overline{u}_k, \qquad k = 0, 1, \ldots, r.$$

Show that there is an optimal periodic policy π^* of the form

$$\pi^* = \{\mu_0^*, \mu_1^*, \ldots, \mu_{p-1}^*, \mu_0^*, \mu_1^*, \ldots, \mu_{p-1}^*, \ldots\},$$

where $\mu_0^*, \mu_1^*, \ldots, \mu_{p-1}^*$ are given by

$$\mu_i^*(x) = -\alpha(\alpha B_i' K_{i+1} B_i + R_i)^{-1} B_i' K_{i+1} A_i x, \qquad i = 0, \ldots, p-2,$$

$$\mu_{p-1}^*(x) = -\alpha(\alpha B_{p-1}' K_0 B_{p-1} + R_{p-1})^{-1} B_{p-1}' K_0 A_{p-1} x,$$

and the matrices $K_0, K_1, \ldots, K_{p-1}$ satisfy the coupled set of p algebraic Riccati equations given for $i = 0, 1, \ldots, p-1$ by

$$K_i = A_i' \big(\alpha K_{i+1} - \alpha^2 K_{i+1} B_i (\alpha B_i' K_{i+1} B_i + R_i)^{-1} B_i' K_{i+1} A_i\big) + Q_i,$$

with

$$K_p = K_0.$$

4.15 (Linear-Quadratic Problems – Imperfect State Information)

Consider the linear-quadratic problem of Section 4.2 with the difference that the controller, instead of having perfect state information, has access to measurements of the form

$$z_k = Cx_k + v_k, \qquad k = 0, 1, \ldots$$

As in Section 5.2 of Vol. I, the disturbances v_k are independent and have identical statistics, zero mean, and finite covariance matrix. Assume that for every admissible policy π the matrices

$$E\Big\{\big(x_k - E\{x_k \mid I_k\}\big)\big(x_k - E\{x_k \mid I_k\}\big)' \mid \pi\Big\}$$

are uniformly bounded over k, where I_k is the information vector defined in Section 5.2 of Vol. I. Show that the stationary policy μ^* given by

$$\mu^*(I_k) = -\alpha(\alpha B'KB + R)^{-1}B'KAE\{x_k \mid I_k\}, \quad \text{for all } I_k,\ k = 0, 1, \ldots$$

is optimal. Show also that the same is true if w_k and v_k are nonstationary with zero mean and covariance matrices that are uniformly bounded over k. *Hint*: Combine the theory of Section 5.2 of Vol. I and Section 4.2.

4.16 (PI for Linear-Quadratic Problems [Kle68])

Consider the problem of Section 4.2 and let L_0 be an $m \times n$ matrix such that the matrix $(A + BL_0)$ has eigenvalues strictly within the unit circle.

(a) Show that the cost corresponding to the stationary policy μ_0, where $\mu_0(x) = L_0x$ is of the form

$$J_{\mu_0}(x) = x'K_0x + \text{constant},$$

where K_0 is a positive semidefinite symmetric matrix satisfying the (linear) equation

$$K_0 = \alpha(A + BL_0)'K_0(A + BL_0) + Q + L_0'RL_0.$$

(b) Let $\mu_1(x)$ attain the minimum for each x in the expression

$$\min_u \big\{u'Ru + \alpha(Ax + Bu)'K_0(Ax + bu)\big\}.$$

Show that for all x we have

$$J_{\mu_1}(x) = x'K_1x + \text{constant} \ \le J_{\mu_0}(x),$$

where K_1 is some positive semidefinite symmetric matrix.

(c) Show that the PI process described in parts (a) and (b) yields a sequence $\{K_k\}$ such that

$$K_k \to K,$$

where K is the optimal cost matrix of the problem.

4.17 (An Affine Monotonic Problem where J^* does not Satisfy Bellman's Equation)

Consider the shortest path problem of Example 4.4.2 and the policy μ illustrated in Fig. 4.4.2, but assume that the cost function is the limit, as $N \to \infty$, of the expected value of the exponential of the N-stage sum of costs [cf. Eqs. (4.81) and (4.82)].

(a) Show that

$$J_\mu(0) = \frac{1}{2}(e^1 + e^{-1}), \qquad J_\mu(2) = J_\mu(5) = e^1.$$

(b) Verify that the Bellman equation for J_μ at state 0 is

$$J_\mu(0) = \frac{1}{2}\big(J_\mu(2) + J_\mu(5)\big),$$

and that it is violated.

(c) Introduce a high cost terminating action that leads to t from every other state, and verify that in the resulting problem, the optimal cost $\hat{J}$ over stable policies satisfies $\hat{J} = T\hat{J}$.

4.18 (Periodic Inventory Control Problems)

In the inventory control problem of Section 4.3, consider the case where the statistics of the demands w_k, the prices c_k, and the holding and the shortage costs are periodic with period p. Show that there exists an optimal periodic policy of the form $\pi^* = \{\mu_0^*, \ldots, \mu_{p-1}^*, \mu_0^*, \ldots, \mu_{p-1}^*, \ldots\}$,

$$\mu_i^*(x) = \begin{cases} S_i^* - x & \text{if } x \le S_i^*, \\ 0 & \text{if otherwise,} \end{cases} \qquad i = 0, 1, \ldots, p-1,$$

where $S_0^*, \ldots, S_{p-1}^*$ are appropriate scalars.

4.19 [HeS84]

Show that the critical level S^* for the inventory problem with zero fixed cost of Section 4.3 minimizes $(1-\alpha)cy + L(y)$ over y. *Hint*: Show that the cost can be expressed as

$$J_\pi(x_0) = E\left\{ \sum_{k=0}^{\infty} \alpha^k \big((1-\alpha)cy_k + L(y_k)\big) + \frac{c\alpha}{1-\alpha} E\{w\} - cx_0 \right\},$$

where $y_k = x_k + \mu_k(x_k)$.

4.20

Consider a machine that may break down and can be repaired. When it operates over a time unit, it costs -1 (i.e., it produces a benefit of 1 unit), and it may break down with probability 0.1. When it is in the breakdown mode, it may be repaired with an effort u. The probability of making it operative over one time unit is then u, and the cost is Cu^2. Determine the optimal repair effort over an infinite time horizon with discount factor $\alpha < 1$.

4.21

Let $z_0, z_1, \ldots$ be a sequence of independent and identically distributed random variables taking values on a finite set Z. We know that the probability distribution of the z_k's is one out of n distributions $f_1, \ldots, f_n$, and we are trying to decide which distribution is the correct one. At each time k after observing $z_1, \ldots, z_k$, we may either stop the observations and accept one of the n distributions as correct, or take another observation at a cost $c > 0$. The cost for accepting f_i given that f_j is correct is L_{ij}, $i, j = 1, \ldots, n$. We assume $L_{ij} > 0$ for $i \neq j$, $L_{ii} = 0$, $i = 1, \ldots, n$. The a priori distribution of $f_1, \ldots, f_n$ is denoted

$$P_0 = \{p_0^1, p_0^2, \ldots, p_0^n\}, \qquad p_0^i \geq 0, \qquad \sum_{i=1}^n p_0^i = 1.$$

Show that the optimal cost $J^*(P_0)$ is a concave function of P_0. Characterize the optimal acceptance regions and show how they can be obtained in the limit by means of a VI method.

4.22 (Gambling Strategies for Favorable Games)

A gambler plays a game such as the one of Section 4.5, but where the probability of winning p satisfies $1/2 \leq p < 1$. His objective is to reach a final fortune n, where n is an integer with $n \geq 2$. His initial fortune is an integer i with $0 < i < n$, and his stake at time k can take only integer values u_k satisfying $0 \leq u_k \leq x_k$, $0 \leq u_k \leq n - x_k$, where x_k is his fortune at time k. Show that the strategy that always stakes one unit is optimal [i.e., $\mu^*(x) = 1$ for all integers x with $0 < x < n$ is optimal]. *Hint*: Show that if $p \in (1/2, 1)$,

$$J_{\mu^*}(i) = \left[\left(\frac{1-p}{p}\right)^i - 1\right]\left[\left(\frac{1-p}{p}\right)^n - 1\right]^{-1}, \qquad 0 \leq i \leq n,$$

and if $p = 1/2$,

$$J_{\mu^*}(i) = \frac{i}{n}, \qquad 0 \leq i \leq n,$$

(or see [Ash70], p. 182, for a proof). Then use the sufficiency condition of Prop. 4.1.6.

4.23 [Sch81]

Consider a network of n queues whereby a customer at queue i upon completion of service is routed to queue j with probability p_{ij}, and exits the network with probability $1 - \sum_j p_{ij}$. For each queue i denote:

r_i: the external customer arrival rate,

$\frac{1}{\mu_i}$: the average customer service time,

λ_i: the customer departure rate,

a_i: the total customer arrival rate (sum of external rate and departure rates from upstream queues weighted by the corresponding probabilities).

We have

$$a_i = r_i + \sum_{j=1}^{n} \lambda_j p_{ji}, \qquad \text{for all } i,$$

and we assume that any portion of the arrival rate a_i in excess of the service rate μ_i is lost; so the departure rate at queue i satisfies

$$\lambda_i = \min[\mu_i, a_i] = \min \left[\mu_i, r_i + \sum_{j=1}^{n} \lambda_j p_{ji} \right].$$

Assume that $r_i > 0$ for at least one i, and that for every queue i_1 with $r_{i_1} > 0$, there is a queue i with $1 - \sum_j p_{ij} > 0$, and a sequence $i_1, i_2, \ldots, i_k$, i such that $p_{i_1 i_2} > 0, \ldots, p_{i_k i} > 0$. Show that the departure rates λ_i satisfying the preceding equations are unique and can be found by VI or PI. *Hint*: This problem does not quite fit our framework because we may have $\sum_j p_{ji} > 1$ for some i. However, it is possible to carry out an analysis based on m-stage contraction mappings.

4.24 (Infinite Time Reachability [Ber71], [Ber72])

Consider the stationary system

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

where the disturbance space W is an arbitrary (not necessarily countable) set. The disturbances w_k can take values in a subset $W(x_k, u_k)$ of W that may depend on x_k and u_k. This problem deals with the following question: Given a nonempty subset X of the state space S, under what conditions does there exist an admissible policy that keeps the state of the (closed-loop) system

$$x_{k+1} = f\big(x_k, \mu_k(x_k), w_k\big) \tag{4.123}$$

in the set X for all k and all possible values $w_k \in W\big(x_k, \mu_k(x_k)\big)$, i.e.,

$$x_k \in X, \qquad \text{for all } w_k \in W\big(x_k, \mu_k(x_k)\big), \qquad k = 0, 1, \ldots \tag{4.124}$$

The set X is said to be *infinitely reachable* if there exists an admissible policy $\{\mu_0, \mu_1, \ldots\}$ and *some* initial state $x_0 \in X$ for which the above relations are satisfied. It is said to be *strongly reachable* if there exists an admissible policy $\{\mu_0, \mu_1, \ldots\}$ such that for *all* initial states $x_0 \in X$ the above relations are satisfied.

Consider the function R mapping any subset Z of the state space S into a subset $R(Z)$ of S defined by

$$R(Z) = \big\{x \mid \text{for some } u \in U(x),\ f(x,u,w) \in Z,\ \text{for all } w \in W(x,u)\big\} \cap Z.$$

(a) Show that the set X is strongly reachable if and only if $R(X) = X$.

(b) Given X, consider the set X^* defined as follows: $x_0 \in X^*$ if and only if $x_0 \in X$ and there exists an admissible policy $\{\mu_0, \mu_1, \ldots\}$ such that that Eqs. (4.123) and (4.124) are satisfied when x_0 is taken as the initial state of the system. Show that a set X is infinitely reachable if and only if it contains a nonempty strongly reachable set. Furthermore, the largest such set is X^* in the sense that X^* is strongly reachable whenever nonempty, and if $\tilde{X} \in X$ is another strongly reachable set, then $\tilde{X} \subset X^*$.

(c) Show that if X is infinitely reachable, there exists an admissible stationary policy μ such that if the initial state x_0 belongs to X^*, then all subsequent states of the closed-loop system $x_{k+1} = f\big(x_k, \mu(x_k), w_k\big)$ are guaranteed to belong to X^*.

(d) Given X, consider the sets $R^k(X)$, $k = 1, 2, \ldots$, where $R^k(X)$ denotes the set obtained after k applications of the mapping R on X. Show that

$$X^* \subset \cap_{k=1}^{\infty} R^k(X).$$

(e) Given X, consider for each $x \in X$ and $k = 1, 2, \ldots$ the set

$$U_k(x) = \big\{u \mid f(x,u,w) \in R^k(X) \text{ for all } w \in W(x,u)\big\}.$$

Show that, if there exists an index $\overline{k}$ such that for all $x \in X$ and $k \geq \overline{k}$ the set $U_k(x)$ is a compact subset of a Euclidean space, then $X^* = \cap_{k=1}^{\infty} R^k(X)$.

4.25 (Infinite Time Reachability for Linear Systems [Ber71])

Consider the linear stationary system

$$x_{k+1} = Ax_k + Bu_k + Gw_k,$$

where $x_k \in \Re^n$, $u_k \in \Re^m$, and $w_k \in \Re^r$, and the matrices A, B, and G are known and have appropriate dimensions. The matrix A is assumed invertible. The controls u_k and the disturbances w_k are restricted to take values in the ellipsoids $U = \{u \mid u'Ru \leq 1\}$ and $W = \{w \mid w'Qw \leq 1\}$, respectively, where R and Q are positive definite symmetric matrices of appropriate dimensions. Show that in order for the ellipsoid $X = \{x \mid x'Kx \leq 1\}$, where K is a positive definite symmetric matrix, to be strongly reachable (in the terminology of Exercise 4.24), it is sufficient that for some positive definite symmetric matrix M and for some scalar $\beta \in (0,1)$ we have

$$K = A' \left[(1-\beta)K^{-1} - \frac{1-\beta}{\beta} G Q^{-1} G' + B R^{-1} B' \right]^{-1} A + M,$$

$$K^{-1} - \frac{1}{\beta} G Q^{-1} G' : \text{positive definite.}$$

Show also that if the above relations are satisfied, the linear stationary policy μ^*, where $\mu^*(x) = Lx$ and

$$L = -(R + B'FB)^{-1} B'FA,$$

$$F = \left[(1-\beta)K^{-1} - \frac{1-\beta}{\beta} G Q^{-1} G' \right]^{-1},$$

achieves reachability of the ellipsoid $X = \{x \mid x'Kx \leq 1\}$. Furthermore, the matrix $(A + BL)$ has all its eigenvalues strictly within the unit circle. (For a proof together with a computational procedure for finding matrices K satisfying the above, see [Ber71] and [Ber72].)

4.26

Consider the $M/M/1$ queueing problem with variable service rate (Example 4.6.5). Assume that no arrivals are allowed ($\lambda = 0$), and one can either serve a customer at rate μ or refuse service ($M = \{0, \mu\}$). Let the cost rates for customer waiting and service be $c(i) = ci$ and $q(\mu)$, respectively, with $q(0) = 0$.

(a) Show that an optimal policy is to always serve an available customer if

$$\frac{q(\mu)}{\mu} \leq \frac{c}{\beta},$$

and to always refuse service otherwise.

(b) Analyze the problem when the cost rate for waiting is instead $c(i) = ci^2$.

4.27

An enterprising financier dreams of making it big in the currency market. He may trade between n currencies $c_1, ..., c_n$ and can convert a unit of c_i to r_{ij} units of c_j, for any currency pair (c_i, c_j) (we assume $r_{ij} > 0$ for all i and j). He is looking for a cycle of currencies

$$c_{i_1} \rightarrow c_{i_2} \rightarrow \ldots \rightarrow c_{i_k} \rightarrow c_{i_1}$$

which is such that

$$r_{i_1 i_2} \cdot r_{i_2 i_3} \cdots r_{i_{k?1} i_k} \cdot r_{i_k i_1} > 1.$$

(Such a cycle is also known as an arbitrage opportunity, i.e., an opportunity of sure profit).

(a) Formulate a multiplicative cost SSP problem, which has strictly positive optimal costs if and only if there is no arbitrage opportunity.

(b) Give an algorithm that detects the existence of an arbitrage opportunity.

4.28 (Monotone Increasing Affine Monotonic Models [Ber13])

Consider the affine monotonic problem of Section 4.5 under the compactness Assumption 4.5.2 and the condition $T_\mu \bar{J} \geq \bar{J}$ for all $\mu \in \mathcal{M}$. Derive the following generalizations of the results of Section 4.1 that were obtained under Assumption P, by following the lines of proof of that section.

(a) We have $J^* = TJ^*$ and $J_\mu = T_\mu J_\mu$ for every stationary policy μ.

(b) A stationary policy μ is optimal if and only if $TJ^* = T_\mu J^*$.

(c) We have $T^k J \to J^*$ for all $J \in \mathcal{E}_+^n$ satisfying $\bar{J} \leq J \leq J^*$.

(d) We have $T^k J \to J^*$ for all $J \in \Re_+^n$ if in addition there exists an optimal policy that is stable.

4.29 (Monotone Decreasing Affine Monotonic Models [Ber13])

Consider the affine monotonic problem of Section 4.5 under the condition $T_\mu \bar{J} \leq \bar{J}$ for all $\mu \in \mathcal{M}$. Derive the following generalizations of the results of Section 4.1 that were obtained under Assumption N, by following the lines of proof of that section.

(a) We have $J^* = TJ^*$ and $J_\mu = T_\mu J_\mu$ for every stationary policy μ.

(b) A stationary policy μ is optimal if and only if $TJ_\mu = T_\mu J_\mu$.

(c) We have $T^k J \to J^*$ for all $J \in \mathcal{E}_+^n$ satisfying $\bar{J} \geq J \geq J^*$.

Additional References

The following are new references for this chapter, which are not already included in the reference list of the printed book.

[BFO74] Blackwell, D., Freedman, D., and Orkin, M., 1974. "The Optimal Reward Operator in Dynamic Programming," Annals of Probability, Vol. 2, pp. 926-941.

[BYB94] Bradtke, S. J., Ydstie, B. E., and Barto, A. G., 1994. "Adaptive Linear Quadratic Control Using Policy Iteration," Proc. IEEE American Control Conference, Vol. 3, pp. 3475-3479.

[BeY15] Bertsekas, D. P., and Yu, H., 2015. "Stochastic Shortest Path Problems Under Weak Conditions," Lab. for Information and Decision Systems Report LIDS-2909; to appear in Math. of OR.

[Bea95] Beard, R. W., 1995. Improving the Closed-Loop Performance of Nonlinear Systems, Ph.D. Thesis, Rensselaer Polytechnic Institute, Troy, NY.

[Ber75] Bertsekas, D. P., 1975. "Monotone Mappings in Dynamic Programming," Proc. 1975 IEEE Conference on Decision and Control, Houston, TX, pp. 20-25.

[Ber13] Bertsekas, D. P., 2013. Abstract Dynamic Programming, Athena Scientific, Belmont, MA.

[Ber14] Bertsekas, D. P., 2014. "Robust Shortest Path Planning and Semicontractive Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-P-2915, MIT; revised Jan. 2015.

[Ber15a] Bertsekas, D. P., 2015. "Regular Policies in Abstract Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-P-3173, MIT, May 2015.

[Ber15b] Bertsekas, D. P., 2015. "Value and Policy Iteration in Deterministic Optimal Control and Adaptive Dynamic Programming", to appear in IEEE Transactions on Neural Networks and Learning Systems; arXiv preprint arXiv:1507.01026.

[DeR79] Denardo, E. V., and Rothblum, U. G., 1079. “Optimal Stopping, Exponential Utility, and Linear Programming,” Math. Programming, Vol. 16, pp. 228-244.

[DeR06] Denardo, E. V., and Rothblum, U. G., 2006. “A Turnpike Theorem for a Risk-Sensitive Markov Decision Process with Stopping,” SIAM J. on Control and Optimization, Vol. 45, pp. 414-431.

[Fei02] Feinberg, E. A., 2002. “Total Reward Criteria,” in E. A. Feinberg and A. Shwartz, (Eds.), Handbook of Markov Decision Processes, Springer, N. Y.

[Hey14a] Heydari, A., 2014. “Revisiting Approximate Dynamic Programming and its Convergence,” IEEE Transactions on Cybernetics, Vol. 44, pp. 2733-2743.

[Hey14b] Heydari, A., 2014. “Stabilizing Value Iteration With and Without Approximation Errors,” available at arXiv:1412.5675.

[JiJ13] Jiang, Y., and Jiang, Z. P., 2013. “Robust Adaptive Dynamic Programming for Linear and Nonlinear Systems: An Overview,” Eur. J. Control, Vol. 19, pp. 417-425.

[JiJ14] Jiang, Y., and Jiang, Z. P., 2014. “Robust Adaptive Dynamic Programming and Feedback Stabilization of Nonlinear Systems,” IEEE Trans. on Neural Networks and Learning Systems, Vol. 25, pp. 882-893.

[LeL13] Lewis, F. L., and Liu, D., (Eds), 2013. Reinforcement Learning and Approximate Dynamic Programming for Feedback Control, Wiley, Hoboken, N. J.

[LiW13] Liu, D., and Wei, Q., 2013. “Finite-Approximation-Error-Based Optimal Control Approach for Discrete-Time Nonlinear Systems, IEEE Transactions on Cybernetics, Vol. 43, pp. 779-789.

[Pat01] Patek, S. D., 2001. “On Terminating Markov Decision Processes with a Risk Averse Objective Function,” Automatica, Vol. 37, pp. 1379-1386.

[Rek64] Rekasius, Z. V., 1964. “Suboptimal Design of Intentionally Nonlinear Controllers,” IEEE Trans. on Automatic Control, Vol. 9, pp. 380-386.

[Rot84] Rothblum, U. G., 1984. “Multiplicative Markov Decision Chains,” Math. of OR, Vol. 9, pp. 6-24.

[SaL79] Saridis, G. N., and Lee, C.-S. G., 1979. “An Approximation Theory of Optimal Control for Trainable Manipulators,” IEEE Trans. Syst., Man, Cybernetics, Vol. 9, pp. 152-159.

[Str75] Striebel, C., 1975. Optimal Control of Discrete-Time Stochastic Systems, Springer-Verlag, New York.

[VVL13] Vrabie, D., 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.

[WWL14] Wei, Q., Wang, F. Y., Liu, D., and Yang, X., 2014. "Finite-Approximation-Error-Based Discrete-Time Iterative Adaptive Dynamic Programming," IEEE Transactions on Cybernetics, Vol. 44, pp. 2820-2833.

[Whi80] Whittle, P., 1980. "Stability and Characterization Conditions in Negative Programming," Journal of Applied Probability, Vol. 17, pp. 635-645.

[YuB13] Yu, H., and Bertsekas, D. P., 2013. "A Mixed Value and Policy Iteration Method for Stochastic Control with Universally Measurable Policies," Lab. for Information and Decision Systems Report LIDS-P-2905, MIT, July 2013; to appear in Math. of OR.

[Yu14] Yu, H., 2014. "On Convergence of Value Iteration for a Class of Total Cost Markov Decision Processes," Technical Report, University of Alberta; arXiv preprint arXiv:1411.1459; SIAM J. Control and Optimization, Vol. 53, pp. 1982-2016.