

4

Noncontractive Models

Contents

4.1. Noncontractive Models - Problem Formulation	p. 217
4.2. Finite Horizon Problems	p. 219
4.3. Infinite Horizon Problems	p. 225
4.3.1. Fixed Point Properties and Optimality Conditions	p. 228
4.3.2. Value Iteration	p. 240
4.3.3. Exact and Optimistic Policy Iteration -	
λ -Policy Iteration	p. 244
4.4. Regularity and Nonstationary Policies	p. 249
4.4.1. Regularity and Monotone Increasing Models	p. 255
4.4.2. Nonnegative Cost Stochastic Optimal Control	p. 257
4.4.3. Discounted Stochastic Optimal Control	p. 260
4.4.4. Convergent Models	p. 262
4.5. Stable Policies for Deterministic Optimal Control	p. 266
4.5.1. Forcing Functions and p -Stable Policies	p. 270
4.5.2. Restricted Optimization over Stable Policies	p. 273
4.5.3. Policy Iteration Methods	p. 285
4.6. Infinite-Spaces Stochastic Shortest Path Problems	p. 291
4.6.1. The Multiplicity of Solutions of Bellman's Equation	p. 299
4.6.2. The Case of Bounded Cost per Stage	p. 301
4.7. Notes, Sources, and Exercises	p. 304

In this chapter, we consider abstract DP models that are similar to the ones of the earlier chapters, but we do not assume any contraction-like property. We discuss both finite and infinite horizon models, and introduce just enough assumptions (including monotonicity) to obtain some minimal results, which we will strengthen as we go along.

In Section 4.2, we consider a general type of finite horizon problem. Under some reasonable assumptions, we show the standard results that one may expect in an abstract setting.

In Section 4.3, we discuss an infinite horizon problem that is motivated by the well-known *positive* and *negative* DP models (see [Ber12a], Chapter 4). These are the special cases of the infinite horizon stochastic optimal control problem of Example 1.2.1, where the cost per stage g is uniformly nonpositive or uniformly nonnegative. For these models there is interesting theory (the validity of Bellman's equation and the availability of optimality conditions in a DP context), which we discuss in Section 4.3.1. There are also interesting computational methods, patterned after the VI and PI algorithms, which we discuss in Sections 4.3.2 and 4.3.3. However, the performance guarantees for these methods are not as powerful as in the contractive case, and their validity hinges upon certain additional assumptions.

In Section 4.4, we extend the notion of regularity of Section 3.2 so that it applies more broadly, including situations where nonstationary policies need to be considered. The mathematical reason for considering nonstationary policies is that for some of the noncontractive models of Section 4.3, stationary policies are insufficient in the sense that there may not exist ϵ -optimal policies that are stationary. In this section, we also discuss some applications, including some general types of optimal control problems with nonnegative cost per stage. Principal results here are that J^* is the unique solution of Bellman's equation within a certain class of functions, and related results regarding the convergence of the VI algorithm.

In Section 4.5, we discuss a nonnegative cost deterministic optimal control problem, which combines elements of the noncontractive models of Section 4.3 and the semicontractive models of Chapter 3 and Section 4.4. Within this setting we explore the structure and the multiplicity of solutions of Bellman's equation. We draw inspiration from the analysis of Section 4.4, but we also use a perturbation-based line of analysis, similar to the one of Section 3.4. In particular, our starting point is a perturbed version of the mapping T_μ that defines the “stable” policies, in place of a subset S that defines the S -regular policies. Still with a proper definition of S , the “stable” policies are S -regular.

Finally, in Section 4.6, we extend the ideas of Section 4.5 to stochastic optimal control problems, by generalizing the notion of a proper policy to the case of infinite state and control spaces.

4.1 NONCONTRACTIVE MODELS - PROBLEM FORMULATION

Throughout this chapter we will continue to use the model of Section 3.2, which involves the set of extended real numbers $\mathfrak{R}^* = \mathfrak{R} \cup \{\infty, -\infty\}$. To repeat some of the basic definitions, we denote by $\mathcal{E}(X)$ the set of all extended real-valued functions $J : X \mapsto \mathfrak{R}^*$, by $\mathcal{R}(X)$ the set of real-valued functions $J : X \mapsto \mathfrak{R}$, and by $\mathcal{B}(X)$ the set of real-valued functions $J : X \mapsto \mathfrak{R}$ that are bounded with respect to a given weighted sup-norm.

We have a set X of states and a set U of controls, and for each $x \in X$, the nonempty control constraint set $U(x) \subset U$. We denote by \mathcal{M} the set of all functions $\mu : X \mapsto U$ with $\mu(x) \in U(x)$, for all $x \in X$, and by Π the set of “nonstationary policies” $\pi = \{\mu_0, \mu_1, \dots\}$, with $\mu_k \in \mathcal{M}$ for all k . We refer to a stationary policy $\{\mu, \mu, \dots\}$ simply as μ .

We introduce a mapping $H : X \times U \times \mathcal{E}(X) \mapsto \mathfrak{R}^*$, and we define the mapping $T : \mathcal{E}(X) \mapsto \mathcal{E}(X)$ by

$$(TJ)(x) = \inf_{u \in U(x)} H(x, u, J), \quad \forall x \in X, J \in \mathcal{E}(X),$$

and for each $\mu \in \mathcal{M}$ the mapping $T_\mu : \mathcal{E}(X) \mapsto \mathcal{E}(X)$ by

$$(T_\mu J)(x) = H(x, \mu(x), J), \quad \forall x \in X, J \in \mathcal{E}(X).$$

We continue to use the following assumption throughout this chapter, without mentioning it explicitly in various propositions.

Assumption 4.1.1: (Monotonicity) If $J, J' \in \mathcal{E}(X)$ and $J \leq J'$, then

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

A fact that we will be using frequently is that for each $J \in \mathcal{E}(X)$ and scalar $\epsilon > 0$, there exists a $\mu_\epsilon \in \mathcal{M}$ such that for all $x \in X$,

$$(T_{\mu_\epsilon} J)(x) \leq \begin{cases} (TJ)(x) + \epsilon & \text{if } (TJ)(x) > -\infty, \\ -(1/\epsilon) & \text{if } (TJ)(x) = -\infty. \end{cases}$$

In particular, if J is such that

$$(TJ)(x) > -\infty, \quad \forall x \in X,$$

then for each $\epsilon > 0$, there exists a $\mu_\epsilon \in \mathcal{M}$ such that

$$(T_{\mu_\epsilon} J)(x) \leq (TJ)(x) + \epsilon, \quad \forall x \in X.$$

We will often use in our analysis the unit function e , defined by $e(x) \equiv 1$, so for example, we will write the above relation in shorthand as

$$T_{\mu\epsilon}J \leq TJ + \epsilon e.$$

We define cost functions for policies consistently with Chapters 2 and 3. In particular, we are given a function $\bar{J} \in \mathcal{E}(X)$, and we consider for every policy $\pi = \{\mu_0, \mu_1, \dots\} \in \Pi$ and positive integer N the function $J_{N,\pi} \in \mathcal{E}(X)$ defined by

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

and the function $J_\pi \in \mathcal{E}(X)$ defined by

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

We refer to $J_{N,\pi}$ as the N -stage cost function of π and to J_π as the infinite horizon cost function of π (or just “cost function” if the length of the horizon is clearly implied by the context). For a stationary policy $\pi = \{\mu, \mu, \dots\}$ we also write J_π as J_μ .

In Section 4.2, we consider the N -stage optimization problem

$$\begin{aligned} & \text{minimize} && J_{N,\pi}(x) \\ & \text{subject to} && \pi \in \Pi, \end{aligned} \tag{4.1}$$

while in Sections 4.3 and 4.4 we discuss its infinite horizon version

$$\begin{aligned} & \text{minimize} && J_\pi(x) \\ & \text{subject to} && \pi \in \Pi. \end{aligned} \tag{4.2}$$

For a fixed $x \in X$, we denote by $J_N^*(x)$ and $J^*(x)$ the optimal costs for these problems, i.e.,

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

We say that a policy $\pi^* \in \Pi$ is N -stage optimal if

$$J_{N,\pi^*}(x) = J_N^*(x), \quad \forall x \in X,$$

and (infinite horizon) optimal if

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

For a given $\epsilon > 0$, we say that π_ϵ is N -stage ϵ -optimal if

$$J_{\pi_\epsilon}(x) \leq \begin{cases} J_N^*(x) + \epsilon & \text{if } J_N^*(x) > -\infty, \\ -(1/\epsilon) & \text{if } J_N^*(x) = -\infty, \end{cases}$$

and we say that π_ϵ is ϵ -optimal if

$$J_{\pi_\epsilon}(x) \leq \begin{cases} J^*(x) + \epsilon & \text{if } J^*(x) > -\infty, \\ -(1/\epsilon) & \text{if } J^*(x) = -\infty. \end{cases}$$

4.2 FINITE HORIZON PROBLEMS

Consider the N -stage problem (4.1), where the cost function $J_{N,\pi}$ is defined by

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

Based on the theory of finite horizon DP, we expect that (at least under some conditions) the optimal cost function J_N^* is obtained by N successive applications of the DP mapping T on the initial function \bar{J} , i.e.,

$$J_N^* = \inf_{\pi \in \Pi} J_{N,\pi} = T^N \bar{J}.$$

This is the analog of Bellman's equation for the finite horizon problem in a DP context.

The Case Where Uniformly N -Stage Optimal Policies Exist

A favorable case where the analysis is simplified and we can easily show that $J_N^* = T^N \bar{J}$ is when the finite horizon DP algorithm yields an optimal policy during its execution. By this we mean that the algorithm that starts with \bar{J} , and sequentially computes $T\bar{J}, T^2\bar{J}, \dots, T^N\bar{J}$, also yields corresponding $\mu_{N-1}^*, \mu_{N-2}^*, \dots, \mu_0^* \in \mathcal{M}$ such that

$$T_{\mu_k^*} T^{N-k-1} \bar{J} = T^{N-k} \bar{J}, \quad k = 0, \dots, N-1. \quad (4.3)$$

While $\mu_{N-1}^*, \dots, \mu_0^* \in \mathcal{M}$ satisfying this relation need not exist (because the corresponding infimum in the definition of T is not attained), if they do exist, they both form an optimal policy and also guarantee that

$$J_N^* = T^N \bar{J}.$$

The proof is simple: we have for every $\pi = \{\mu_0, \mu_1, \dots\} \in \Pi$

$$J_{N,\pi} = T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J} \geq T^N \bar{J} = T_{\mu_0^*} \cdots T_{\mu_{N-1}^*} \bar{J}, \quad (4.4)$$

where the inequality follows from the monotonicity assumption and the definition of T , and the last equality follows from Eq. (4.3). Thus $\{\mu_0^*, \mu_1^*, \dots\}$ has no worse N -stage cost function than every other policy, so it is N -stage optimal and $J_N^* = T_{\mu_0^*} \cdots T_{\mu_{N-1}^*} \bar{J}$. By taking the infimum of the left-hand side over $\pi \in \Pi$ in Eq. (4.4), we obtain $J_N^* = T^N \bar{J}$.

The preceding argument can also be used to show that $\{\mu_k^*, \mu_{k+1}^*, \dots\}$ is $(N-k)$ -stage optimal for all $k = 0, \dots, N-1$. Such a policy is called *uniformly N -stage optimal*. The fact that the finite horizon DP algorithm provides an optimal solution of *all* the k -stage problems for $k = 1, \dots, N$, rather than just the last one, is a manifestation of the classical principle

of optimality, expounded by Bellman in the early days of DP (the tail portion of an optimal policy obtained by DP minimizes the corresponding tail portion of the finite horizon cost). Note, however, that there may exist an N -stage optimal policy that is not k -stage optimal for some $k < N$.

We state the result just derived as a proposition.

Proposition 4.2.1: Suppose that a policy $\{\mu_0^*, \mu_1^*, \dots\}$ satisfies the condition (4.3). Then this policy is uniformly N -stage optimal, and we have $J_N^* = T^N \bar{J}$.

While the preceding result is theoretically limited, it is very useful in practice, because the existence of a policy satisfying the condition (4.3) can often be established with a simple analysis. For example, this condition is trivially satisfied if the control space is finite. The following proposition provides a generalization.

Proposition 4.2.2: Let the control space U be a metric space, and assume that for each $x \in X$, $\lambda \in \Re$, and $k = 0, 1, \dots, N - 1$, the set

$$U_k(x, \lambda) = \{u \in U(x) \mid H(x, u, T^k \bar{J}) \leq \lambda\}$$

is compact. Then there exists a uniformly N -stage optimal policy.

Proof: We will show that the infimum in the relation

$$(T^{k+1} \bar{J})(x) = \inf_{u \in U(x)} H(x, u, T^k \bar{J})$$

is attained for all $x \in X$ and k . Indeed if $H(x, u, T^k \bar{J}) = \infty$ for all $u \in U(x)$, then every $u \in U(x)$ attains the infimum. If for a given $x \in X$,

$$\inf_{u \in U(x)} H(x, u, T^k \bar{J}) < \infty,$$

the corresponding part of the proof of Lemma 3.3.1 applies and shows that the above infimum is attained. The result now follows from Prop. 4.2.1.

Q.E.D.

The General Case

We now consider the case where there may not exist a uniformly N -stage optimal policy. By using the definitions of J_N^* and $T^N \bar{J}$, the equation

$J_N^* = T^N \bar{J}$ can be equivalently written as

$$\inf_{\mu_0, \dots, \mu_{N-1} \in \mathcal{M}} T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J} = \inf_{\mu_0 \in \mathcal{M}} T_{\mu_0} \left(\inf_{\mu_1 \in \mathcal{M}} T_{\mu_1} \left(\cdots \inf_{\mu_{N-1} \in \mathcal{M}} T_{\mu_{N-1}} \bar{J} \right) \right).$$

Thus we have $J_N^* = T^N \bar{J}$ if the operations \inf and T_μ can be interchanged in the preceding equation. We will introduce two alternative assumptions, which guarantee that this interchange is valid. Our first assumption is a form of continuity from above of H with respect to J .

Assumption 4.2.1: For each sequence $\{J_m\} \subset \mathcal{E}(X)$ with $J_m \downarrow J$ and $H(x, u, J_0) < \infty$ for all $x \in X$ and $u \in U(x)$, we have

$$\lim_{m \rightarrow \infty} H(x, u, J_m) = H(x, u, J), \quad \forall x \in X, u \in U(x). \quad (4.5)$$

Note that if $\{J_m\}$ is monotonically nonincreasing, the same is true for $\{T_\mu J_m\}$. It follows that

$$\inf_m J_m = \lim_{m \rightarrow \infty} J_m, \quad \inf_m (T_\mu J_m) = \lim_{m \rightarrow \infty} (T_\mu J_m),$$

so for all $\mu \in \mathcal{M}$, Eq. (4.5) implies that

$$\inf_m (T_\mu J_m) = \lim_{m \rightarrow \infty} (T_\mu J_m) = T_\mu \left(\lim_{m \rightarrow \infty} J_m \right) = T_\mu \left(\inf_m J_m \right).$$

This equality can be extended for any $\mu_1, \dots, \mu_k \in \mathcal{M}$ as follows:

$$\begin{aligned} \inf_m (T_{\mu_1} \cdots T_{\mu_k} J_m) &= T_{\mu_1} \left(\inf_m (T_{\mu_1} \cdots T_{\mu_k} J_m) \right) \\ &= \cdots \\ &= T_{\mu_1} T_{\mu_1} \cdots T_{\mu_{k-1}} \left(\inf_m (T_{\mu_k} J_m) \right) \\ &= T_{\mu_1} \cdots T_{\mu_k} \left(\inf_m J_m \right). \end{aligned} \quad (4.6)$$

We use this relation to prove the following proposition.

Proposition 4.2.3: Let Assumption 4.2.1 hold, and assume further that $J_{k,\pi}(x) < \infty$, for all $x \in X$, $\pi \in \Pi$, and $k \geq 1$. Then $J_N^* = T^N \bar{J}$.

Proof: We select for each $k = 0, \dots, N-1$, a sequence $\{\mu_k^m\} \subset \mathcal{M}$ such that

$$\lim_{m \rightarrow \infty} T_{\mu_k^m} (T^{N-k-1} \bar{J}) \downarrow T^{N-k} \bar{J}.$$

Since $J_N^* \leq T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J}$ for all $\mu_0, \dots, \mu_{N-1} \in \mathcal{M}$, we have using also Eq. (4.6) and the assumption $J_{k,\pi}(x) < \infty$, for all k, π , and x ,

$$\begin{aligned}
J_N^* &\leq \inf_{m_0} \cdots \inf_{m_{N-1}} T_{\mu_0}^{m_0} \cdots T_{\mu_{N-1}}^{m_{N-1}} \bar{J} \\
&= \inf_{m_0} \cdots \inf_{m_{N-2}} T_{\mu_0}^{m_0} \cdots T_{\mu_{N-2}}^{m_{N-2}} \left(\inf_{m_{N-1}} T_{\mu_{N-1}}^{m_{N-1}} \bar{J} \right) \\
&= \inf_{m_0} \cdots \inf_{m_{N-2}} T_{\mu_0}^{m_0} \cdots T_{\mu_{N-2}}^{m_{N-2}} T \bar{J} \\
&\quad \vdots \\
&= \inf_{m_0} T_{\mu_0}^{m_0} (T^{N-1} \bar{J}) \\
&= T^N \bar{J}.
\end{aligned}$$

On the other hand, it is clear from the definitions that $T^N \bar{J} \leq J_{N,\pi}$ for all N and $\pi \in \Pi$, so that $T^N \bar{J} \leq J_N^*$. Thus, $J_N^* = T^N \bar{J}$. **Q.E.D.**

We now introduce an alternative assumption, which in addition to $J_N^* = T^N \bar{J}$, guarantees the existence of an ϵ -optimal policy.

Assumption 4.2.2: We have

$$J_k^*(x) > -\infty, \quad \forall x \in X, k = 1, \dots, N.$$

Moreover, there exists a scalar $\alpha \in (0, \infty)$ such that for all scalars $r \in (0, \infty)$ and functions $J \in \mathcal{E}(X)$, we have

$$H(x, u, J + r e) \leq H(x, u, J) + \alpha r, \quad \forall x \in X, u \in U(x). \quad (4.7)$$

Proposition 4.2.4: Let Assumption 4.2.2 hold. Then $J_N^* = T^N \bar{J}$, and for every $\epsilon > 0$, there exists an ϵ -optimal policy.

Proof: Note that since by assumption, $J_N^*(x) > -\infty$ for all $x \in X$, an N -stage ϵ -optimal policy $\pi_\epsilon \in \Pi$ is one for which

$$J_N^* \leq J_{N,\pi_\epsilon} \leq J_N^* + \epsilon e.$$

We use induction. The result clearly holds for $N = 1$. Assume that it holds for $N = k$, i.e., $J_k^* = T^k \bar{J}$ and for a given $\epsilon > 0$, there is a $\pi_\epsilon \in \Pi$

with $J_{k,\pi_\epsilon} \leq J_k^* + \epsilon e$. Using Eq. (4.7), we have for all $\mu \in \mathcal{M}$,

$$J_{k+1}^* \leq T_\mu J_{k,\pi_\epsilon} \leq T_\mu J_k^* + \alpha \epsilon e.$$

Taking the infimum over μ and then the limit as $\epsilon \rightarrow 0$, we obtain $J_{k+1}^* \leq T J_k^*$. By using the induction hypothesis $J_k^* = T^k \bar{J}$, it follows that $J_{k+1}^* \leq T^{k+1} \bar{J}$. On the other hand, we have clearly $T^{k+1} \bar{J} \leq J_{k+1,\pi}$ for all $\pi \in \Pi$, so that $T^{k+1} \bar{J} \leq J_{k+1}^*$, and hence $T^{k+1} \bar{J} = J_{k+1}^*$.

We now turn to the existence of an ϵ -optimal policy part of the induction argument. Using the assumption $J_k^*(x) > -\infty$ for all $x \in X$, for any $\bar{\epsilon} > 0$, we can choose $\bar{\pi} = \{\bar{\mu}_0, \bar{\mu}_1, \dots\}$ such that

$$J_{k,\bar{\pi}} \leq J_k^* + \frac{\bar{\epsilon}}{2\alpha} e, \quad (4.8)$$

and $\bar{\mu} \in \mathcal{M}$ such that

$$T_{\bar{\mu}} J_k^* \leq T J_k^* + \frac{\bar{\epsilon}}{2} e.$$

Let $\bar{\pi}_{\bar{\epsilon}} = \{\bar{\mu}, \bar{\mu}_0, \bar{\mu}_1, \dots\}$. Then

$$J_{k+1,\bar{\pi}_{\bar{\epsilon}}} = T_{\bar{\mu}} J_{k,\bar{\pi}} \leq T_{\bar{\mu}} J_k^* + \frac{\bar{\epsilon}}{2} e \leq T J_k^* + \bar{\epsilon} e = J_{k+1}^* + \bar{\epsilon} e,$$

where the first inequality is obtained by applying $T_{\bar{\mu}}$ to Eq. (4.8) and using Eq. (4.7). The induction is complete. **Q.E.D.**

We now provide some counterexamples showing that the conditions of the preceding propositions are necessary, and that for exceptional (but otherwise very simple) problems, the Bellman equation $J_N^* = T^N \bar{J}$ may not hold and/or there may not exist an ϵ -optimal policy.

Example 4.2.1 (Counterexample to Bellman's Equation I)

Let

$$X = \{0\}, \quad U(0) = (-1, 0], \quad \bar{J}(0) = 0,$$

$$H(0, u, J) = \begin{cases} u & \text{if } -1 < J(0), \\ J(0) + u & \text{if } J(0) \leq -1. \end{cases}$$

Then

$$(T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(0) = \mu_0(0),$$

and $J_N^*(0) = -1$, while $(T^N \bar{J})(0) = -N$ for every N . Here Assumption 4.2.1, and the condition (4.7) (cf. Assumption 4.2.2) are violated, even though the condition $J_k^*(x) > -\infty$ for all $x \in X$ (cf. Assumption 4.2.2) is satisfied.

Example 4.2.2 (Counterexample to Bellman's Equation II)

Let

$$X = \{0, 1\}, \quad U(0) = U(1) = (-\infty, 0], \quad \bar{J}(0) = \bar{J}(1) = 0,$$

$$H(0, u, J) = \begin{cases} u & \text{if } J(1) = -\infty, \\ 0 & \text{if } J(1) > -\infty, \end{cases} \quad H(1, u, J) = u.$$

Then

$$(T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(0) = 0, \quad (T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(1) = \mu_0(1), \quad \forall N \geq 1.$$

It can be seen that for $N \geq 2$, we have $J_N^*(0) = 0$ and $J_N^*(1) = -\infty$, but $(T^N \bar{J})(0) = (T^N \bar{J})(1) = -\infty$. Here Assumption 4.2.1, and the condition $J_k^*(x) > -\infty$ for all $x \in X$ (cf. Assumption 4.2.2) are violated, even though the condition (4.7) of Assumption 4.2.2 is satisfied.

In the preceding two examples, the anomalies are due to discontinuity of the mapping H with respect to J . In classical finite horizon DP, the mapping H is generally continuous when it takes finite values, but counterexamples arise in unusual problems where infinite values occur. The next example is a simple stochastic optimal control problem, which involves some infinite expected values of random variables and we have $J_2^* \neq T^2 \bar{J}$.

Example 4.2.3 (Counterexample to Bellman's Equation III)

Let

$$X = \{0, 1\}, \quad U(0) = U(1) = \mathfrak{R}, \quad \bar{J}(0) = \bar{J}(1) = 0,$$

let w be a real-valued random variable with $E\{w\} = \infty$, and let

$$H(x, u, J) = \begin{cases} E\{w + J(1)\} & \text{if } x = 0, \\ u + J(1) & \text{if } x = 1, \end{cases} \quad \forall x \in X, u \in U(x).$$

Then if J_m is real-valued for all m , and $J_m(1) \downarrow J(1) = -\infty$, we have

$$\lim_{m \rightarrow \infty} H(0, u, J_m) = \lim_{m \rightarrow \infty} E\{w + J_m(1)\} = \infty,$$

while

$$H\left(0, u, \lim_{m \rightarrow \infty} J_m\right) = E\{w + J(1)\} = -\infty,$$

so Assumption 4.2.1 is violated. Indeed, the reader may verify with a straightforward calculation that $J_2^*(0) = \infty$, $J_2^*(1) = -\infty$, while $(T^2 \bar{J})(0) = -\infty$, $(T^2 \bar{J})(1) = -\infty$, so $J_2^* \neq T^2 \bar{J}$. Note that Assumption 4.2.2 is also violated because $J_2^*(1) = -\infty$.

In the next counterexample, Bellman's equation holds, but there is no ϵ -optimal policy. This is an undiscounted deterministic optimal control problem of the type discussed in Section 1.1, where $J_k^*(x) = -\infty$ for some x and k , so Assumption 4.2.2 is violated. We use the notation introduced there.

Example 4.2.4 (Counterexample to Existence of an ϵ -Optimal Policy)

Let $\alpha = 1$ and

$$N = 2, \quad X = \{0, 1, \dots\}, \quad U(x) = (0, \infty), \quad \bar{J}(x) = 0, \quad \forall x \in X,$$

$$f(x, u) = 0, \quad \forall x \in X, u \in U(x),$$

$$g(x, u) = \begin{cases} -u & \text{if } x = 0, \\ x & \text{if } x \neq 0, \end{cases} \quad \forall u \in U(x),$$

so that

$$H(x, u, J) = g(x, u) + J(0).$$

Then for $\pi \in \Pi$ and $x \neq 0$, we have $J_{2,\pi}(x) = x - \mu_1(0)$, so that $J_2^*(x) = -\infty$ for all $x \neq 0$. Clearly, we also have $J_2^*(0) = -\infty$. Here Assumption 4.2.1, as well as Eq. (4.7) (cf. Assumption 4.2.2) are satisfied, and indeed we have $J_2^*(x) = (T^2 \bar{J})(x) = -\infty$ for all $x \in X$. However, the condition $J_k^*(x) > -\infty$ for all x and k (cf. Assumption 4.2.2) is violated, and it is seen that there does not exist a two-stage ϵ -optimal policy for any $\epsilon > 0$. The reason is that an ϵ -optimal policy $\pi = \{\mu_0, \mu_1\}$ must satisfy

$$J_{2,\pi}(x) = x - \mu_1(0) \leq -\frac{1}{\epsilon}, \quad \forall x \in X,$$

[in view of $J_2^*(x) = -\infty$ for all $x \in X$], which is impossible since the left-hand side above can become positive for x sufficiently large.

4.3 INFINITE HORIZON PROBLEMS

We now turn to the infinite horizon problem (4.2), where the cost function of a policy $\pi = \{\mu_0, \mu_1, \dots\}$ is

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

In this section one of the following two assumptions will be in effect.

Assumption I: (Monotone Increase)

(a) We have

$$-\infty < \bar{J}(x) \leq H(x, u, \bar{J}), \quad \forall x \in X, u \in U(x).$$

(b) For each sequence $\{J_m\} \subset \mathcal{E}(X)$ with $J_m \uparrow J$ and $\bar{J} \leq J_m$ for all $m \geq 0$, we have

$$\lim_{m \rightarrow \infty} H(x, u, J_m) = H(x, u, J), \quad \forall x \in X, u \in U(x).$$

- (c) There exists a scalar $\alpha \in (0, \infty)$ such that for all scalars $r \in (0, \infty)$ and functions $J \in \mathcal{E}(X)$ with $\bar{J} \leq J$, we have

$$H(x, u, J + r e) \leq H(x, u, J) + \alpha r, \quad \forall x \in X, u \in U(x).$$

Assumption D: (Monotone Decrease)

- (a) We have

$$\bar{J}(x) \geq H(x, u, \bar{J}), \quad \forall x \in X, u \in U(x).$$

- (b) For each sequence $\{J_m\} \subset \mathcal{E}(X)$ with $J_m \downarrow J$ and $J_m \leq \bar{J}$ for all $m \geq 0$, we have

$$\lim_{m \rightarrow \infty} H(x, u, J_m) = H(x, u, J), \quad \forall x \in X, u \in U(x).$$

Assumptions I and D apply to the positive and negative cost DP models, respectively (see [Ber12a], Chapter 4). These are the special cases of the infinite horizon stochastic optimal control problem of Example 1.2.1, where $\bar{J}(x) \equiv 0$ and the cost per stage g is uniformly nonnegative or uniformly nonpositive, respectively. The latter occurs often when we want to maximize positive rewards.

It is important to note that Assumptions I and D allow J_π to be defined as a limit rather than as a limsup. In particular, part (a) of the assumptions and the monotonicity of H imply that

$$\bar{J} \leq T_{\mu_0} \bar{J} \leq T_{\mu_0} T_{\mu_1} \bar{J} \leq \dots \leq T_{\mu_0} \dots T_{\mu_k} \bar{J} \leq \dots$$

under Assumption I, and

$$\bar{J} \geq T_{\mu_0} \bar{J} \geq T_{\mu_0} T_{\mu_1} \bar{J} \geq \dots \geq T_{\mu_0} \dots T_{\mu_k} \bar{J} \geq \dots$$

under Assumption D. Thus we have

$$J_\pi(x) = \lim_{k \rightarrow \infty} (T_{\mu_0} \dots T_{\mu_k} \bar{J})(x), \quad \forall x \in X,$$

with the limit being a real number or ∞ or $-\infty$.

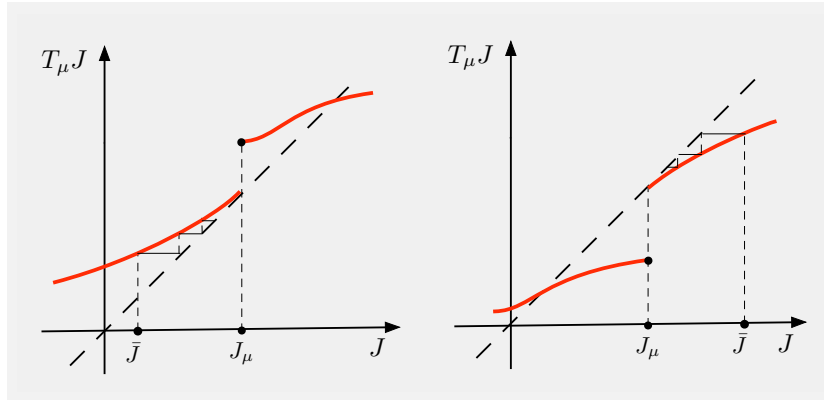


Figure 4.3.1. Illustration of the consequences of lack of continuity of T_μ from below or from above [cf. part (b) of Assumption I or D, respectively]. In the figure on the left, we have $\bar{J} \leq T_\mu \bar{J}$ but T_μ is discontinuous from below at J_μ , so Assumption I does not hold, and J_μ is not a fixed point of T_μ . In the figure on the right, we have $\bar{J} \geq T_\mu \bar{J}$ but T_μ is discontinuous from above at J_μ , so Assumption D does not hold, and J_μ is not a fixed point of T_μ .

The conditions of part (b) of Assumptions I and D are continuity assumptions designed to preclude some of the pathologies of the type encountered also in Chapter 3, and addressed with the use of S -regular policies. In particular, these conditions are essential for making a connection with fixed point theory: they ensure that J_μ is a fixed point of T_μ , as shown in the following proposition.

Proposition 4.3.1: Let Assumption I or Assumption D hold. Then for every policy $\mu \in \mathcal{M}$, we have

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

Proof: Let Assumption I hold. Then for all $k \geq 0$,

$$(T_\mu^{k+1} \bar{J})(x) = H(x, \mu(x), T_\mu^k \bar{J}), \quad x \in X,$$

and by taking the limit as $k \rightarrow \infty$, and using part (b) of Assumption I, and the fact $T_\mu^k \bar{J} \uparrow J_\mu$, we have for all $x \in X$,

$$J_\mu(x) = \lim_{k \rightarrow \infty} H(x, \mu(x), T_\mu^k \bar{J}) = H(x, \mu(x), \lim_{k \rightarrow \infty} T_\mu^k \bar{J}) = H(x, \mu(x), J_\mu),$$

or equivalently $J_\mu = T_\mu J_\mu$. The proof for the case of Assumption D is similar. **Q.E.D.**

Figure 4.3.1 illustrates how J_μ may fail to be a fixed point of T_μ if part (b) of Assumption I or D is violated. Note also that continuity of T_μ does not imply continuity of T , and for example, under Assumption I, T may be discontinuous from below. We will see later that as a result, the value iteration sequence $\{T^k \bar{J}\}$ may fail to converge to J^* in the absence of additional conditions (see Section 4.3.2). Part (c) of Assumption I is a technical condition that facilitates the analysis, and assures the existence of ϵ -optimal policies.

Despite the similarities between Assumptions I and D, the corresponding results that one may obtain involve some substantial differences. An important fact, which breaks the symmetry between the two cases, is that J^* is approached by $T^k \bar{J}$ from below in the case of Assumption I and from above in the case of Assumption D. Another important fact is that since the condition $\bar{J}(x) > -\infty$ for all $x \in X$ is part of Assumption I, all the functions J encountered in the analysis under this assumption (such as $T^k \bar{J}$, J_π , and J^*) also satisfy $J(x) > -\infty$, for all $x \in X$. In particular, if $J \geq \bar{J}$, we have

$$(TJ)(x) \geq (T\bar{J})(x) > -\infty, \quad \forall x \in X,$$

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

$$T_{\mu_\epsilon} J \leq TJ + \epsilon e.$$

This property is critical for the existence of an ϵ -optimal policy under Assumption I (see the next proposition) and is not available under Assumption D. It accounts in part for the different character of the results that can be obtained under the two assumptions.

4.3.1 Fixed Point Properties and Optimality Conditions

We first consider the question whether the optimal cost function J^* is a fixed point of T . This is indeed true, but the lines of proof are different under the Assumptions I and D. We begin with the proof under Assumption I, and as a preliminary step we show the existence of an ϵ -optimal policy, something that is of independent theoretical interest.

Proposition 4.3.2: Let Assumption I hold. Then given any $\epsilon > 0$, there exists a policy $\pi_\epsilon \in \Pi$ such that

$$J^* \leq J_{\pi_\epsilon} \leq J^* + \epsilon e.$$

Furthermore, if the scalar α in part (c) of Assumption I satisfies $\alpha < 1$, the policy π_ϵ can be taken to be stationary.

Proof: Let $\{\epsilon_k\}$ be a sequence such that $\epsilon_k > 0$ for all k and

$$\sum_{k=0}^{\infty} \alpha^k \epsilon_k = \epsilon. \quad (4.9)$$

For each $x \in X$, consider a sequence of policies $\{\pi_k[x]\} \subset \Pi$ of the form

$$\pi_k[x] = \{\mu_0^k[x], \mu_1^k[x], \dots\}, \quad (4.10)$$

such that for $k = 0, 1, \dots$,

$$J_{\pi_k[x]}(x) \leq J^*(x) + \epsilon_k. \quad (4.11)$$

Such a sequence exists, since we have assumed that $\bar{J}(x) > -\infty$, and therefore $J^*(x) > -\infty$, for all $x \in X$.

The preceding notation should be interpreted as follows. The policy $\pi_k[x]$ of Eq. (4.10) is associated with x . Thus $\mu_i^k[x]$ denotes for each x and k , a function in \mathcal{M} , while $\mu_i^k[x](z)$ denotes the value of $\mu_i^k[x]$ at an element $z \in X$. In particular, μ_i^kx denotes the value of $\mu_i^k[x]$ at $x \in X$.

Consider the functions $\bar{\mu}_k$ defined by

$$\bar{\mu}_k(x) = \mu_0^kx, \quad \forall x \in X, \quad (4.12)$$

and the functions \bar{J}_k defined by

$$\bar{J}_k(x) = H\left(x, \bar{\mu}_k(x), \lim_{m \rightarrow \infty} T_{\mu_1^k[x]} \cdots T_{\mu_m^k[x]} \bar{J}\right), \quad \forall x \in X, \quad k = 0, 1, \dots \quad (4.13)$$

By using Eqs. (4.11), (4.12), and part (b) of Assumption I, we obtain for all $x \in X$ and $k = 0, 1, \dots$

$$\begin{aligned} \bar{J}_k(x) &= \lim_{m \rightarrow \infty} (T_{\mu_0^k[x]} \cdots T_{\mu_m^k[x]} \bar{J})(x) \\ &= J_{\pi_k[x]}(x) \\ &\leq J^*(x) + \epsilon_k. \end{aligned} \quad (4.14)$$

From Eqs. (4.13), (4.14), and part (c) of Assumption I, we have for all $x \in X$ and $k = 1, 2, \dots$,

$$\begin{aligned} (T_{\bar{\mu}_{k-1}} \bar{J}_k)(x) &= H(x, \bar{\mu}_{k-1}(x), \bar{J}_k) \\ &\leq H(x, \bar{\mu}_{k-1}(x), J^* + \epsilon_k e) \\ &\leq H(x, \bar{\mu}_{k-1}(x), J^*) + \alpha \epsilon_k \\ &\leq H\left(x, \bar{\mu}_{k-1}(x), \lim_{m \rightarrow \infty} T_{\mu_1^{k-1}[x]} \cdots T_{\mu_m^{k-1}[x]} \bar{J}\right) + \alpha \epsilon_k \\ &= \bar{J}_{k-1}(x) + \alpha \epsilon_k, \end{aligned}$$

and finally

$$T_{\bar{\mu}_{k-1}} \bar{J}_k \leq \bar{J}_{k-1} + \alpha \epsilon_k e, \quad k = 1, 2, \dots$$

Using this inequality and part (c) of Assumption I, we obtain

$$\begin{aligned} T_{\bar{\mu}_{k-2}} T_{\bar{\mu}_{k-1}} \bar{J}_k &\leq T_{\bar{\mu}_{k-2}} (\bar{J}_{k-1} + \alpha \epsilon_k e) \\ &\leq T_{\bar{\mu}_{k-2}} \bar{J}_{k-1} + \alpha^2 \epsilon_k e \\ &\leq \bar{J}_{k-2} + (\alpha \epsilon_{k-1} + \alpha^2 \epsilon_k) e. \end{aligned}$$

Continuing in the same manner, we have for $k = 1, 2, \dots$,

$$T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{k-1}} \bar{J}_k \leq \bar{J}_0 + (\alpha \epsilon_1 + \cdots + \alpha^k \epsilon_k) e \leq J^* + \left(\sum_{i=0}^k \alpha^i \epsilon_i \right) e.$$

Since $\bar{J} \leq \bar{J}_k$, it follows that

$$T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{k-1}} \bar{J} \leq J^* + \left(\sum_{i=0}^k \alpha^i \epsilon_i \right) e.$$

Denote $\pi_\epsilon = \{\bar{\mu}_0, \bar{\mu}_1, \dots\}$. Then by taking the limit in the preceding inequality and using Eq. (4.9), we obtain

$$J_{\pi_\epsilon} \leq J^* + \epsilon e.$$

If $\alpha < 1$, we take $\epsilon_k = \epsilon(1-\alpha)$ for all k , and $\pi_k[x] = \{\mu_0[x], \mu_1[x], \dots\}$ in Eq. (4.11). The stationary policy $\pi_\epsilon = \{\bar{\mu}, \bar{\mu}, \dots\}$, where $\bar{\mu}(x) = \mu_0x$ for all $x \in X$, satisfies $J_{\pi_\epsilon} \leq J^* + \epsilon e$. **Q.E.D.**

Note that the assumption $\alpha < 1$ is essential in order to be able to take π_ϵ stationary in the preceding proposition. As an example, let $X = \{0\}$, $U(0) = (0, \infty)$, $\bar{J}(0) = 0$, $H(0, u, J) = u + J(0)$. Then $J^*(0) = 0$, but for any $\mu \in \mathcal{M}$, we have $J_\mu(0) = \infty$.

By using Prop. 4.3.2 we can prove the following.

Proposition 4.3.3: Let Assumption I hold. Then

$$J^* = T J^*.$$

Furthermore, if $J' \in \mathcal{E}(X)$ is such that $J' \geq \bar{J}$ and $J' \geq T J'$, then $J' \geq J^*$.

Proof: For every $\pi = \{\mu_0, \mu_1, \dots\} \in \Pi$ and $x \in X$, we have using part (b) of Assumption I,

$$\begin{aligned} J_\pi(x) &= \lim_{k \rightarrow \infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} \bar{J})(x) \\ &= T_{\mu_0} \left(\lim_{k \rightarrow \infty} T_{\mu_1} \cdots T_{\mu_k} \bar{J} \right)(x) \\ &\geq (T_{\mu_0} J^*)(x) \\ &\geq (T J^*)(x). \end{aligned}$$

By taking the infimum of the left-hand side over $\pi \in \Pi$, we obtain

$$J^* \geq T J^*.$$

To prove the reverse inequality, let ϵ_1 and ϵ_2 be any positive scalars, and let $\bar{\pi} = \{\bar{\mu}_0, \bar{\mu}_1, \dots\}$ be such that

$$T_{\bar{\mu}_0} J^* \leq T J^* + \epsilon_1 e, \quad J_{\bar{\pi}_1} \leq J^* + \epsilon_2 e,$$

where $\bar{\pi}_1 = \{\bar{\mu}_1, \bar{\mu}_2, \dots\}$ (such a policy exists by Prop. 4.3.2). The sequence $\{T_{\bar{\mu}_1} \cdots T_{\bar{\mu}_k} \bar{J}\}$ is monotonically nondecreasing, so by using the preceding relations and part (c) of Assumption I, we have

$$\begin{aligned} T_{\bar{\mu}_0} T_{\bar{\mu}_1} \cdots T_{\bar{\mu}_k} \bar{J} &\leq T_{\bar{\mu}_0} \left(\lim_{k \rightarrow \infty} T_{\bar{\mu}_1} \cdots T_{\bar{\mu}_k} \bar{J} \right) \\ &= T_{\bar{\mu}_0} J_{\bar{\pi}_1} \\ &\leq T_{\bar{\mu}_0} J^* + \alpha \epsilon_2 e \\ &\leq T J^* + (\epsilon_1 + \alpha \epsilon_2) e. \end{aligned}$$

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

$$J^* \leq J_{\bar{\pi}} = \lim_{k \rightarrow \infty} T_{\bar{\mu}_0} T_{\bar{\mu}_1} \cdots T_{\bar{\mu}_k} \bar{J} \leq T J^* + (\epsilon_1 + \alpha \epsilon_2) e.$$

Since ϵ_1 and ϵ_2 can be taken arbitrarily small, it follows that

$$J^* \leq T J^*.$$

Hence $J^* = T J^*$.

Assume that $J' \in \mathcal{E}(X)$ satisfies $J' \geq \bar{J}$ and $J' \geq T J'$. Let $\{\epsilon_k\}$ be any sequence with $\epsilon_k > 0$ for all k , and consider a policy $\bar{\pi} = \{\bar{\mu}_0, \bar{\mu}_1, \dots\} \in \Pi$ such that

$$T_{\bar{\mu}_k} J' \leq T J' + \epsilon_k e, \quad k = 0, 1, \dots$$

We have from part (c) of Assumption I

$$\begin{aligned}
J^* &= \inf_{\pi \in \Pi} \lim_{k \rightarrow \infty} T_{\mu_0} \cdots T_{\mu_k} \bar{J} \\
&\leq \inf_{\pi \in \Pi} \liminf_{k \rightarrow \infty} T_{\mu_0} \cdots T_{\mu_k} J' \\
&\leq \liminf_{k \rightarrow \infty} T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_k} J' \\
&\leq \liminf_{k \rightarrow \infty} T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{k-1}} (TJ' + \epsilon_k e) \\
&\leq \liminf_{k \rightarrow \infty} T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{k-1}} (J' + \epsilon_k e) \\
&\leq \liminf_{k \rightarrow \infty} (T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{k-1}} J' + \alpha^k \epsilon_k e) \\
&\vdots \\
&\leq \lim_{k \rightarrow \infty} \left(TJ' + \left(\sum_{i=0}^k \alpha^i \epsilon_i \right) e \right) \\
&\leq J' + \left(\sum_{i=0}^k \alpha^i \epsilon_i \right) e.
\end{aligned}$$

Since we may choose $\sum_{i=0}^k \alpha^i \epsilon_i$ as small as desired, it follows that $J^* \leq J'$.

Q.E.D.

The following counterexamples show that parts (b) and (c) of Assumption I are essential for the preceding proposition to hold.

Example 4.3.1 (Counterexample to Bellman's Equation I)

Let

$$\begin{aligned}
X &= \{0, 1\}, & U(0) &= U(1) = (-1, 0], & \bar{J}(0) &= \bar{J}(1) = -1, \\
H(0, u, J) &= \begin{cases} u & \text{if } J(1) \leq -1, \\ 0 & \text{if } J(1) > -1, \end{cases} & H(1, u, J) &= u.
\end{aligned}$$

Then for $N \geq 1$,

$$(T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(0) = 0, \quad (T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(1) = \mu_0(1).$$

Thus

$$J^*(0) = 0, \quad J^*(1) = -1, \quad (TJ^*)(0) = -1, \quad (TJ^*)(1) = -1,$$

and hence $J^* \neq TJ^*$. Notice also that \bar{J} is a fixed point of T , while $\bar{J} \leq J^*$ and $\bar{J} \neq J^*$, so the second part of Prop. 4.3.3 fails when $\bar{J} = J'$. Here parts (a) and (b) of Assumption I are satisfied, but part (c) is violated, since $H(0, u, \cdot)$ is discontinuous at $J = -1$ when $u < 0$.

Example 4.3.2 (Counterexample to Bellman’s Equation II)

Let

$$X = \{0, 1\}, \quad U(0) = U(1) = \{0\}, \quad \bar{J}(0) = \bar{J}(1) = 0,$$

$$H(0, 0, J) = \begin{cases} 0 & \text{if } J(1) < \infty, \\ \infty & \text{if } J(1) = \infty, \end{cases} \quad H(1, 0, J) = J(1) + 1.$$

Here there is only one policy, which we denote by μ . For all $N \geq 1$, we have

$$(T_\mu^N \bar{J})(0) = 0, \quad (T_\mu^N \bar{J})(1) = N,$$

so $J^*(0) = 0, J^*(1) = \infty$. On the other hand, we have $(TJ^*)(0) = (TJ^*)(1) = \infty$ and $J^* \neq TJ^*$. Here parts (a) and (c) of Assumption I are satisfied, but part (b) is violated.

As a corollary to Prop. 4.3.3 we obtain the following.

Proposition 4.3.4: Let Assumption I hold. Then for every $\mu \in \mathcal{M}$, we have

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

Furthermore, if $J' \in \mathcal{E}(X)$ is such that $J' \geq \bar{J}$ and $J' \geq T_\mu J'$, then $J' \geq J_\mu$.

Proof: Consider the variant of the infinite horizon problem where the control constraint set is $U_\mu(x) = \{\mu(x)\}$ rather than $U(x)$ for all $x \in X$. Application of Prop. 4.3.3 yields the result. **Q.E.D.**

We now provide the counterpart of Prop. 4.3.3 under Assumption D. We first prove a preliminary result regarding the convergence of the value iteration method, which is of independent interest (we will see later that this result need not hold under Assumption I).

Proposition 4.3.5: Let Assumption D hold. Then $T^N \bar{J} = J_N^*$, where J_N^* is the optimal cost function for the N -stage problem. Moreover

$$J^* = \lim_{N \rightarrow \infty} J_N^*.$$

Proof: By repeating the proof of Prop. 4.2.3, we have $T^N \bar{J} = J_N^*$ [part (b) of Assumption D is essentially identical to the assumption of that proposition]. Clearly we have $J^* \leq J_N^*$ for all N , and hence $J^* \leq \lim_{N \rightarrow \infty} J_N^*$.

Also for all $\pi = \{\mu_0, \mu_1, \dots\} \in \Pi$, we have

$$T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J} \geq J_N^*,$$

so by taking the limit of both sides as $N \rightarrow \infty$, we obtain $J_\pi \geq \lim_{N \rightarrow \infty} J_N^*$, and by taking infimum over π , $J^* \geq \lim_{N \rightarrow \infty} J_N^*$. Thus $J^* = \lim_{N \rightarrow \infty} J_N^*$.

Q.E.D.

Proposition 4.3.6: Let Assumption D hold. Then

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

Furthermore, if $J' \in \mathcal{E}(X)$ is such that $J' \leq \bar{J}$ and $J' \leq TJ'$, then $J' \leq J^*$.

Proof: For any $\pi = \{\mu_0, \mu_1, \dots\} \in \Pi$, we have

$$J_\pi = \lim_{k \rightarrow \infty} T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} \bar{J} \geq \lim_{k \rightarrow \infty} T_{\mu_0} T^k \bar{J} \geq T_{\mu_0} J^*,$$

where the last inequality follows from the fact $T^k \bar{J} \downarrow J^*$ (cf. Prop. 4.3.5). Taking the infimum of both sides over $\pi \in \Pi$, we obtain $J^* \geq TJ^*$.

To prove the reverse inequality, we select any $\mu \in \mathcal{M}$, and we apply T_μ to both sides of the equation $J^* = \lim_{N \rightarrow \infty} T^N \bar{J}$ (cf. Prop. 4.3.5). By using part (b) of assumption D, we obtain

$$T_\mu J^* = T_\mu \left(\lim_{N \rightarrow \infty} T^N \bar{J} \right) = \lim_{N \rightarrow \infty} T_\mu T^N \bar{J} \geq \lim_{N \rightarrow \infty} T^{N+1} \bar{J} = J^*.$$

Taking the infimum of the left-hand side over $\mu \in \mathcal{M}$, we obtain $TJ^* \geq J^*$, showing that $TJ^* = J^*$.

To complete the proof, let $J' \in \mathcal{E}(X)$ be such that $J' \leq \bar{J}$ and $J' \leq TJ'$. Then we have

$$\begin{aligned} J^* &= \inf_{\pi \in \Pi} \lim_{N \rightarrow \infty} T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J} \\ &\geq \lim_{N \rightarrow \infty} \inf_{\pi \in \Pi} T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J} \\ &\geq \lim_{N \rightarrow \infty} \inf_{\pi \in \Pi} T_{\mu_0} \cdots T_{\mu_{N-1}} J' \\ &\geq \lim_{N \rightarrow \infty} T^N J' \\ &\geq J', \end{aligned}$$

where the last inequality follows from the hypothesis $J' \leq TJ'$. Thus $J^* \geq J'$. **Q.E.D.**

Counterexamples to Bellman's equation can be readily constructed if part (b) of Assumption D (continuity from above) is violated. In particular, in Examples 4.2.1 and 4.2.2, part (a) of Assumption D is satisfied but part (b) is not. In both cases we have $J^* \neq TJ^*$, as the reader can verify with a straightforward calculation.

Similar to Prop. 4.3.4, we obtain the following.

Proposition 4.3.7: Let Assumption D hold. Then for every $\mu \in \mathcal{M}$, we have

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

Furthermore, if $J' \in \mathcal{E}(X)$ is such that $J' \leq \bar{J}$ and $J' \leq T_\mu J'$, then $J' \leq J_\mu$.

Proof: Consider the variation of our problem where the control constraint set is $U_\mu(x) = \{\mu(x)\}$ rather than $U(x)$ for all $x \in X$. Application of Prop. 4.3.6 yields the result. **Q.E.D.**

An examination of the proof of Prop. 4.3.6 shows that the only point where we need part (b) of Assumption D was in establishing the relations

$$\lim_{N \rightarrow \infty} TJ_N^* = T \left(\lim_{N \rightarrow \infty} J_N^* \right)$$

and

$$J_N^* = T^N \bar{J}.$$

If these relations can be established independently, then the result of Prop. 4.3.6 follows. In this manner we obtain the following proposition.

Proposition 4.3.8: Let part (a) of Assumption D hold, assume that X is a finite set, and that $J^*(x) > -\infty$ for all $x \in X$. Assume further that there exists a scalar $\alpha \in (0, \infty)$ such that for all scalars $r \in (0, \infty)$ and functions $J \in \mathcal{E}(X)$ with $J \leq \bar{J}$, we have

$$H(x, u, J) - \alpha r \leq H(x, u, J - r e), \quad \forall x \in X, u \in U(x). \quad (4.15)$$

Then

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

Furthermore, if $J' \in \mathcal{E}(X)$ is such that $J' \leq \bar{J}$ and $J' \leq TJ'$, then $J' \leq J^*$.

Proof: A nearly verbatim repetition of Prop. 4.2.4 shows that under our assumptions we have $J_N^* = T^N \bar{J}$ for all N . We will show that

$$\lim_{N \rightarrow \infty} H(x, u, J_N^*) \leq H\left(x, u, \lim_{N \rightarrow \infty} J_N^*\right), \quad \forall x \in X, u \in U(x).$$

Then the result follows as in the proof of Prop. 4.3.6.

Assume the contrary, i.e., that for some $\tilde{x} \in X$, $\tilde{u} \in U(\tilde{x})$, and $\epsilon > 0$, there holds

$$H(\tilde{x}, \tilde{u}, J_k^*) - \epsilon > H\left(\tilde{x}, \tilde{u}, \lim_{N \rightarrow \infty} J_N^*\right), \quad k = 1, 2, \dots$$

From the finiteness of X and the fact

$$J^*(x) = \lim_{N \rightarrow \infty} J_N^*(x) > -\infty, \quad \forall x \in X,$$

it follows that for some integer $\bar{k} > 0$

$$J_k^* - (\epsilon/\alpha)e \leq \lim_{N \rightarrow \infty} J_N^*, \quad \forall k \geq \bar{k}.$$

By using the condition (4.15), we obtain for all $k \geq \bar{k}$

$$H(\tilde{x}, \tilde{u}, J_k^*) - \epsilon \leq H(\tilde{x}, \tilde{u}, J_k^* - (\epsilon/\alpha)e) \leq H\left(\tilde{x}, \tilde{u}, \lim_{N \rightarrow \infty} J_N^*\right),$$

which contradicts the earlier inequality. **Q.E.D.**

Characterization of Optimal Policies

We now provide necessary and sufficient conditions for optimality of a stationary policy. These conditions are markedly different under Assumptions I and D.

Proposition 4.3.9: Let Assumption I hold. Then a stationary policy μ is optimal if and only if

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

Proof: If μ is optimal, then $J_\mu = J^*$ so that the equation $J^* = T J^*$ (cf. Prop. 4.3.3) implies that $J_\mu = T J_\mu$. Since $J_\mu = T_\mu J_\mu$ (cf. Prop. 4.3.4), it follows that $T_\mu J^* = T J^*$.

Conversely, if $T_\mu J^* = TJ^*$, then since $J^* = TJ^*$, it follows that $T_\mu J^* = J^*$. By Prop. 4.3.4, it follows that $J_\mu \leq J^*$, so μ is optimal. **Q.E.D.**

Proposition 4.3.10: Let Assumption D hold. Then a stationary policy μ is optimal if and only if

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

Proof: If μ is optimal, then $J_\mu = J^*$, so that the equation $J^* = TJ^*$ (cf. Prop. 4.3.6) can be written as $J_\mu = TJ_\mu$. Since $J_\mu = T_\mu J_\mu$ (cf. Prop. 4.3.4), it follows that $T_\mu J_\mu = TJ_\mu$.

Conversely, if $T_\mu J_\mu = TJ_\mu$, then since $J_\mu = T_\mu J_\mu$, it follows that $J_\mu = TJ_\mu$. By Prop. 4.3.7, it follows that $J_\mu \leq J^*$, so μ is optimal. **Q.E.D.**

An example showing that under Assumption I, the condition $T_\mu J_\mu = TJ_\mu$ does not guarantee optimality of μ is given in Exercise 4.3. Under Assumption D, we note that by Prop. 4.3.1, we have $J_\mu = T_\mu J_\mu$ for all μ , so if μ is a stationary optimal policy, the fixed point equation

$$J^*(x) = \inf_{u \in U(x)} H(x, u, J^*), \quad \forall x \in X, \quad (4.16)$$

and the optimality condition of Prop. 4.3.10, yield

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

Thus under D, a stationary optimal policy attains the infimum in the fixed point Eq. (4.16) for all x . However, there may exist nonoptimal stationary policies also attaining the infimum for all x ; an example is the shortest path problem of Section 3.1.1 for the case where $a = 0$ and $b = 1$. Moreover, it is possible that this infimum is attained but no optimal policy exists, as shown by Fig. 4.3.2.

Proposition 4.3.9 shows that under Assumption I, there exists a stationary optimal policy if and only if the infimum in the optimality equation

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

is attained for every $x \in X$. When the infimum is not attained for some $x \in X$, this optimality equation can still be used to yield an ϵ -optimal policy, which can be taken to be stationary whenever the scalar α in Assumption I(c) is strictly less than 1. This is shown in the following proposition.

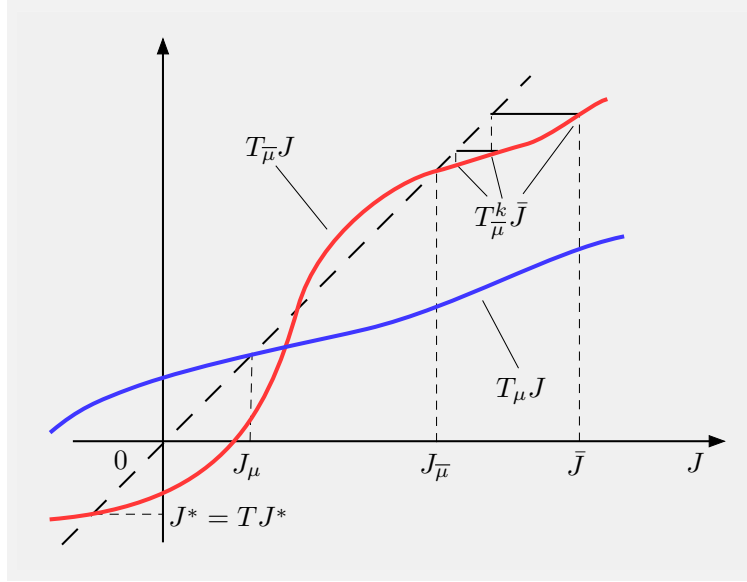


Figure 4.3.2. An example where nonstationary policies are dominant under Assumption D. Here there is only one state and $S = \mathfrak{R}$. There are two stationary policies μ and $\bar{\mu}$ with cost functions J_μ and $J_{\bar{\mu}}$ as shown. However, by considering a nonstationary policy of the form $\pi_k = \{\bar{\mu}, \dots, \bar{\mu}, \mu, \mu, \dots\}$, with a number k of policies $\bar{\mu}$, we can obtain a sequence $\{J_{\pi_k}\}$ that converges to the value J^* shown. Note that here there is no optimal policy, stationary or not.

Proposition 4.3.11: Let Assumption I hold. Then:

- (a) If $\epsilon > 0$, the sequence $\{\epsilon_k\}$ satisfies $\sum_{k=0}^{\infty} \alpha^k \epsilon_k = \epsilon$, and $\epsilon_k > 0$ for all k , and the policy $\pi^* = \{\mu_0^*, \mu_1^*, \dots\} \in \Pi$ is such that

$$T_{\mu_k^*} J^* \leq T J^* + \epsilon_k e, \quad \forall k = 0, 1, \dots,$$

then

$$J^* \leq J_{\pi^*} \leq J^* + \epsilon e.$$

- (b) If $\epsilon > 0$, the scalar α in part (c) of Assumption I is strictly less than 1, and $\mu^* \in \mathcal{M}$ is such that

$$T_{\mu^*} J^* \leq T J^* + \epsilon(1 - \alpha) e,$$

then

$$J^* \leq J_{\mu^*} \leq J^* + \epsilon e.$$

Proof: (a) Since $TJ^* = J^*$, we have

$$T_{\mu_k^*} J^* \leq J^* + \epsilon_k e,$$

and applying $T_{\mu_{k-1}^*}$ to both sides, we obtain

$$T_{\mu_{k-1}^*} T_{\mu_k^*} J^* \leq T_{\mu_{k-1}^*} J^* + \alpha \epsilon_k e \leq J^* + (\epsilon_{k-1} + \alpha \epsilon_k) e.$$

Applying $T_{\mu_{k-2}^*}$ throughout and repeating the process, we obtain for every $k = 1, 2, \dots$,

$$T_{\mu_0^*} \cdots T_{\mu_k^*} J^* \leq J^* + \left(\sum_{i=0}^k \alpha^i \epsilon_i \right) e, \quad k = 1, 2, \dots$$

Since $\bar{J} \leq J^*$, it follows that

$$T_{\mu_0^*} \cdots T_{\mu_k^*} \bar{J} \leq J^* + \left(\sum_{i=0}^k \alpha^i \epsilon_i \right) e, \quad k = 1, 2, \dots$$

By taking the limit as $k \rightarrow \infty$, we obtain $J_{\pi^*} \leq J^* + \epsilon e$.

(b) This part is proved by taking $\epsilon_k = \epsilon(1 - \alpha)$ and $\mu_k^* = \mu^*$ for all k in the preceding argument. **Q.E.D.**

Under Assumption D, the existence of an ϵ -optimal policy is harder to establish, and requires some restrictive conditions.

Proposition 4.3.12: Let Assumption D hold, and let the additional assumptions of Prop. 4.3.8 hold. Then for any $\epsilon > 0$, there exists an ϵ -optimal policy.

Proof: For each N , denote

$$\epsilon_N = \frac{\epsilon}{2(1 + \alpha + \cdots + \alpha^{N-1})},$$

and let

$$\pi_N = \{\mu_0^N, \mu_1^N, \dots, \mu_{N-1}^N, \mu, \mu \dots\}$$

be such that $\mu \in \mathcal{M}$, and for $k = 0, \dots, N-1$, $\mu_k^N \in \mathcal{M}$ and

$$T_{\mu_k^N} T^{N-k-1} \bar{J} = T^{N-k} \bar{J} + \epsilon_N e.$$

We have $T_{\mu_{N-1}}^N \bar{J} \leq T\bar{J} + \epsilon_N e$, and applying $T_{\mu_{N-2}}^N$ to both sides, we obtain

$$T_{\mu_{N-2}}^N T_{\mu_{N-1}}^N \bar{J} \leq T_{\mu_{N-2}}^N T\bar{J} + \alpha \epsilon_N e \leq T^2 \bar{J} + (1 + \alpha) \epsilon_N e.$$

Continuing in the same manner, we have

$$T_{\mu_0}^N \cdots T_{\mu_{N-1}}^N \bar{J} \leq T^N \bar{J} + (1 + \alpha + \cdots + \alpha^{N-1}) \epsilon_N e,$$

from which we obtain for $N = 0, 1, \dots$,

$$J_{\pi_N} \leq T^N \bar{J} + (\epsilon/2) e.$$

By Prop. 4.3.5, we have $J^* = \lim_{N \rightarrow \infty} T^N \bar{J}$, so let \bar{N} be such that

$$T^{\bar{N}} \bar{J} \leq J^* + (\epsilon/2) e$$

[such a \bar{N} exists using the assumptions of finiteness of X and $J^*(x) > -\infty$ for all $x \in X$]. Then we obtain $J_{\pi_{\bar{N}}} \leq J^* + \epsilon e$, and $\pi_{\bar{N}}$ is the desired policy. **Q.E.D.**

4.3.2 Value Iteration

We will now discuss algorithms for abstract DP under Assumptions I and D. We first consider the VI algorithm, which consists of successively generating $T\bar{J}, T^2\bar{J}, \dots$. Note that because T need not be a contraction, it may have multiple fixed points J all of which satisfy $J \geq J^*$ under Assumption I (cf. Prop. 4.3.3) or $J \leq J^*$ under Assumption D (cf. Prop. 4.3.6). Thus, in the absence of additional conditions (to be discussed in Sections 4.4 and 4.5), it is essential to start VI with \bar{J} or an initial J_0 such that $\bar{J} \leq J_0 \leq J^*$ under Assumption I or $\bar{J} \geq J_0 \geq J^*$ under Assumption D. In the next two propositions, we show that for such initial conditions, we have convergence of VI to J^* under Assumption D, and with an additional compactness condition, under Assumption I.

Proposition 4.3.13: Let Assumption D hold, and assume that $J_0 \in \mathcal{E}(X)$ is such that $\bar{J} \geq J_0 \geq J^*$. Then

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

Proof: The condition $\bar{J} \geq J_0 \geq J^*$ implies that $T^k \bar{J} \geq T^k J_0 \geq J^*$ for all k . By Prop. 4.3.5, $T^k \bar{J} \rightarrow J^*$, and the result follows. **Q.E.D.**

The convergence of VI under I requires an additional compactness condition, which is satisfied in particular if $U(x)$ is a finite set for all $x \in X$.

Proposition 4.3.14: Let Assumption I hold, let U be a metric space, and assume that the sets

$$U_k(x, \lambda) = \{u \in U(x) \mid H(x, u, T^k \bar{J}) \leq \lambda\} \quad (4.17)$$

are compact for every $x \in X$, $\lambda \in \mathfrak{R}$, and for all k greater than some integer \bar{k} . Assume that $J_0 \in \mathcal{E}(X)$ is such that $\bar{J} \leq J_0 \leq J^*$. Then

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

Furthermore, there exists a stationary optimal policy.

Proof: Similar to the proof of Prop. 4.3.13, it will suffice to show that $T^k \bar{J} \rightarrow J^*$. Since $\bar{J} \leq J^*$, we have $T^k \bar{J} \leq T^k J^* = J^*$, so that

$$\bar{J} \leq T \bar{J} \leq \dots \leq T^k \bar{J} \leq \dots \leq J^*.$$

Thus we have $T^k \bar{J} \uparrow J_\infty$ for some $J_\infty \in \mathcal{E}(X)$ satisfying $T^k \bar{J} \leq J_\infty \leq J^*$ for all k . Applying T to this relation, we obtain

$$(T^{k+1} \bar{J})(x) = \min_{u \in U(x)} H(x, u, T^k \bar{J}) \leq (T J_\infty)(x),$$

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

$$J_\infty \leq T J_\infty.$$

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

$$J_\infty(\tilde{x}) < (T J_\infty)(\tilde{x}). \quad (4.18)$$

Similar to Lemma 3.3.1, there exists a point u_k attaining the minimum in

$$(T^{k+1} \bar{J})(\tilde{x}) = \inf_{u \in U(\tilde{x})} H(\tilde{x}, u, T^k \bar{J});$$

i.e., u_k is such that

$$(T^{k+1} \bar{J})(\tilde{x}) = H(\tilde{x}, u_k, T^k \bar{J}).$$

Clearly, by Eq. (4.18), we must have $J_\infty(\tilde{x}) < \infty$. For every k , consider the set

$$U_k(\tilde{x}, J_\infty(\tilde{x})) = \left\{ u \in U(\tilde{x}) \mid H(\tilde{x}, u, T^k \bar{J}) \leq J_\infty(\tilde{x}) \right\},$$

and the sequence $\{u_i\}_{i=k}^\infty$. Since $T^k \bar{J} \uparrow J_\infty$, it follows that for all $i \geq k$,

$$H(\tilde{x}, u_i, T^k \bar{J}) \leq H(\tilde{x}, u_i, T^i \bar{J}) \leq J_\infty(\tilde{x}).$$

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

$$\tilde{u} \in \bigcap_{k=0}^\infty U_k(\tilde{x}, J_\infty(\tilde{x})).$$

By Eq. (4.17), this implies that for all $k \geq \bar{k}$

$$J_\infty(\tilde{x}) \geq H(\tilde{x}, \tilde{u}, T^k \bar{J}) \geq (T^{k+1} \bar{J})(\tilde{x}).$$

Taking the limit as $k \rightarrow \infty$, and using part (b) of Assumption I, we obtain

$$J_\infty(\tilde{x}) \geq H(\tilde{x}, \tilde{u}, J_\infty) \geq (T J_\infty)(\tilde{x}), \quad (4.19)$$

which contradicts Eq. (4.18). Hence $J_\infty = T J_\infty$, which implies that $J_\infty \geq J^*$ in view of Prop. 4.3.3. Combined with the inequality $J_\infty \leq J^*$, which was shown earlier, we have $J_\infty = J^*$.

To show that there exists an optimal stationary policy, observe that the relation $J^* = J_\infty = T J_\infty$ and Eq. (4.19) [whose proof is valid for all $\tilde{x} \in X$ such that $J^*(\tilde{x}) < \infty$] imply that \tilde{u} attains the infimum in

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

for all $\tilde{x} \in X$ with $J^*(\tilde{x}) < \infty$. For $\tilde{x} \in X$ such that $J^*(\tilde{x}) = \infty$, every $u \in U(\tilde{x})$ attains the preceding minimum. Hence by Prop. 4.3.9 an optimal stationary policy exists. **Q.E.D.**

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

$$(T^{k+1} \bar{J})(\tilde{x}) = \inf_{u \in U(\tilde{x})} H(\tilde{x}, u, T^k \bar{J}),$$

and $\mu^*(\tilde{x})$ is a limit point of $\{\mu_k(\tilde{x})\}$, for every $\tilde{x} \in X$, then the stationary policy μ^* is optimal. Furthermore, $\{\mu_k(\tilde{x})\}$ has at least one limit point for every $\tilde{x} \in X$ for which $J^*(\tilde{x}) < \infty$. Thus the VI algorithm under the assumption of Prop. 4.3.14 yields in the limit not only the optimal cost function J^* but also an optimal stationary policy.

On the other hand, under Assumption I but in the absence of the compactness condition (4.17), $T^k \bar{J}$ need not converge to J^* . What is happening here is that while the mappings T_μ are continuous from below as required by Assumption I(b), T may not be, and a phenomenon like the one illustrated in the left-hand side of Fig. 4.3.1 may occur, whereby

$$\lim_{k \rightarrow \infty} T^k \bar{J} \leq T \left(\lim_{k \rightarrow \infty} T^k \bar{J} \right),$$

with strict inequality for some $x \in X$. This can happen even in simple deterministic optimal control problems, as shown by the following example.

Example 4.3.3 (Counterexample to Convergence of VI)

Let

$$X = [0, \infty), \quad U(x) = (0, \infty), \quad \bar{J}(x) = 0, \quad \forall x \in X,$$

and

$$H(x, u, J) = \min \{1, x + J(2x + u)\}, \quad \forall x \in X, u \in U(x).$$

Then it can be verified that for all $x \in X$ and policies μ , we have $J_\mu(x) = 1$, as well as $J^*(x) = 1$, while it can be seen by induction that starting with \bar{J} , the VI algorithm yields

$$(T^k \bar{J})(x) = \min \{1, (1 + 2^{k-1})x\}, \quad \forall x \in X, k = 1, 2, \dots$$

Thus we have $0 = \lim_{k \rightarrow \infty} (T^k \bar{J})(0) \neq J^*(0) = 1$.

The range of convergence of VI may be expanded under additional assumptions. In particular, in Chapter 3, under various conditions involving the existence of optimal S -regular policies, we showed that VI converges to J^* assuming that the initial condition J_0 satisfies $J_0 \geq J^*$. Thus if the assumptions of Prop. 4.3.14 hold in addition, we are guaranteed convergence of VI starting from any J satisfying $J \geq \bar{J}$. Results of this type will be obtained in Sections 4.4 and 4.5, where semicontractive models satisfying Assumption I will be discussed.

Asynchronous Value Iteration

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

$$S(k) = \{J \mid T^k \bar{J} \leq J \leq J^*\}, \quad k = 0, 1, \dots$$

Assuming that $T^k \bar{J} \rightarrow J^*$ (cf. Prop. 4.3.14), it follows that the asynchronous form of VI converges pointwise to J^* starting from any function in $S(0)$. This result can also be shown for the case where J^* is not real-valued, by using a simple extension of Prop. 2.6.1, where the set of real-valued functions $\mathcal{R}(X)$ is replaced by the set of all $J \in \mathcal{E}(X)$ with $\bar{J} \leq J \leq J^*$.

Under Assumption D similar conclusions hold for the asynchronous version of VI that starts with a function J with $J^* \leq J \leq \bar{J}$. Asynchronous pointwise convergence to J^* can be shown, based on an extension of the asynchronous convergence theorem (Prop. 2.6.1), where $\mathcal{R}(X)$ is replaced by the set of all $J \in \mathcal{E}(X)$ with $J^* \leq J \leq \bar{J}$.

4.3.3 Exact and Optimistic Policy Iteration - λ -Policy Iteration

Unfortunately, in the absence of further conditions, the PI algorithm is not guaranteed to yield the optimal cost function and/or an optimal policy under either Assumption I or D. However, there are convergence results for nonoptimistic and optimistic variants of PI under some conditions. In what follows in this section we will provide an analysis of various types of PI, mainly under Assumption D. The analysis of PI under Assumption I will be given primarily in the next two sections, as it requires different assumptions and methods of proof, and will be coupled with regularity ideas relating to the semicontractive models of Chapter 3.

Optimistic Policy Iteration Under D

A surprising fact under Assumption D is that nonoptimistic/exact PI may generate a policy that is strictly inferior over the preceding one. Moreover there may be an oscillation between nonoptimal policies even when the state and control spaces are finite. An illustrative example is the shortest path example of Section 3.1.1, where it can be verified that exact PI may oscillate between the policy that moves to the destination from node 1 and the policy that does not. For a mathematical explanation, note that under Assumption D, we may have $T_\mu J^* = T J^*$ without μ being optimal, so starting from an optimal policy, we may obtain a nonoptimal policy by PI.

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

$$T_{\mu^k} J_k = T J_k, \quad J_{k+1} = T_{\mu^k}^{m_k} J_k, \quad (4.20)$$

where m_k is a positive integer. We assume that the algorithm starts with a function $J_0 \in \mathcal{E}(X)$ that satisfies $\bar{J} \geq J_0 \geq J^*$ and $J_0 \geq T J_0$. For example, we may choose $J_0 = \bar{J}$. We have the following proposition.

Proposition 4.3.15: Let Assumption D hold and let $\{J_k, \mu^k\}$ be a sequence generated by the optimistic PI algorithm (4.20), assuming that $\bar{J} \geq J_0 \geq J^*$ and $J_0 \geq T J_0$. Then $J_k \downarrow J^*$.

Proof: We have

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

† As with all PI algorithms in this book, we assume that the policy improvement operation is well-defined, in the sense that there exists μ^k such that $T_{\mu^k} J_k = T J_k$ for all k .

where the first, second, and third inequalities hold because the assumption $J_0 \geq TJ_0 = T_{\mu^0}J_0$ implies that

$$T_{\mu^0}^m J_0 \geq T_{\mu^0}^{m+1} J_0, \quad \forall m \geq 0.$$

Continuing similarly we obtain

$$J_k \geq TJ_k \geq J_{k+1}, \quad \forall k \geq 0.$$

Moreover, we can show by induction that $J_k \geq J^*$. Indeed this is true for $k = 0$ by assumption. If $J_k \geq J^*$, we have

$$J_{k+1} = T_{\mu^k}^{m_k} J_k \geq T_{\mu^k} J_k \geq T_{\mu^k} J^* = J^*, \quad (4.21)$$

where the last equality follows from the fact $TJ^* = J^*$ (cf. Prop. 4.3.6), thus completing the induction. By combining the preceding two relations, we have

$$J_k \geq TJ_k \geq J_{k+1} \geq J^*, \quad \forall k \geq 0. \quad (4.22)$$

We will now show by induction that

$$T^k J_0 \geq J_k \geq J^*, \quad \forall k \geq 0. \quad (4.23)$$

Indeed this relation holds by assumption for $k = 0$, and assuming that it holds for some $k \geq 0$, we have by applying T to it and by using Eq. (4.22),

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

thus completing the induction. By applying Prop. 4.3.13 to Eq. (4.23), we obtain $J_k \downarrow J^*$. **Q.E.D.**

λ -Policy Iteration Under \mathbf{D}

We now consider the λ -PI algorithm. It involves a scalar $\lambda \in (0, 1)$ and a corresponding multistep mapping, which bears a relation to temporal differences and the proximal algorithm (cf. Section 1.2.5). It is defined by

$$T_{\mu^k} J_k = TJ_k, \quad J_{k+1} = T_{\mu^k}^{(\lambda)} J_k, \quad (4.24)$$

where for any policy μ and scalar $\lambda \in (0, 1)$, $T_{\mu}^{(\lambda)}$ is the mapping defined by

$$(T_{\mu}^{(\lambda)} J)(x) = (1 - \lambda) \sum_{t=0}^{\infty} \lambda^t (T_{\mu}^{t+1} J)(x), \quad x \in X.$$

Here we assume that T_μ maps $\mathcal{R}(X)$ to $\mathcal{R}(X)$, and that for all $\mu \in \mathcal{M}$ and $J \in \mathcal{R}(X)$, the limit of the series above is well-defined as a function in $\mathcal{R}(X)$.

We discussed this algorithm in connection with semicontractive problems in Section 3.2.4, where we assumed that

$$T_\mu(T_\mu^{(\lambda)} J) = T_\mu^{(\lambda)}(T_\mu J), \quad \forall \mu \in \mathcal{M}, J \in \mathcal{R}(X). \quad (4.25)$$

We will show that for undiscounted finite-state MDP, the algorithm can be implemented by using matrix inversion, just like nonoptimistic PI for discounted finite-state MDP. It turns out that this can be an advantage in some settings, including approximate simulation-based implementations.

As noted earlier, λ -PI and optimistic PI are similar: they just use the mapping T_{μ^k} to apply VI in different ways. In view of this similarity, it is not surprising that it has the same type of convergence properties as the earlier optimistic PI method (4.20). Similar to Prop. 4.3.15, we have the following.

Proposition 4.3.16: Let Assumption D hold and let $\{J_k, \mu^k\}$ be a sequence generated by the λ -PI algorithm (4.24), assuming Eq. (4.25), and that $\bar{J} \geq J_0 \geq J^*$ and $J_0 \geq T J_0$. Then $J_k \downarrow J^*$.

Proof: As in the proof of Prop. 4.3.15, by using Assumption D, the monotonicity of T_μ , and the hypothesis $J_0 \geq T J_0$, we have

$$J_0 \geq T J_0 = T_{\mu^0} J_0 \geq T_{\mu^0}^{(\lambda)} J_0 = J_1 \geq T_{\mu^0} J_1 \geq T J_1 = T_{\mu^1} J_1 \geq T_{\mu^1}^{(\lambda)} J_1 = J_2,$$

where for the third inequality, we use the relation $J_0 \geq T_{\mu^0} J_0$, the definition of J_1 , and the assumption (4.25). Continuing in the same manner,

$$J_k \geq T J_k \geq J_{k+1}, \quad \forall k \geq 0.$$

Similar to the proof of Prop. 4.3.15, we show by induction that $J_k \geq J^*$, using the fact that if $J_k \geq J^*$, then

$$J_{k+1} = T_{\mu^k}^{(\lambda)} J_k \geq T_{\mu^k}^{(\lambda)} J^* = (1 - \lambda) \sum_{t=0}^{\infty} \lambda^t T^{t+1} J^* = J^*,$$

[cf. the induction step of Eq. (4.21)]. By combining the preceding two relations, we obtain Eq. (4.22), and the proof is completed by using the argument following that equation. **Q.E.D.**

The λ -PI algorithm has a useful property, which involves the mapping $W_k : \mathcal{R}(X) \mapsto \mathcal{R}(X)$ given by

$$W_k J = (1 - \lambda) T_{\mu^k} J_k + \lambda T_{\mu^k} J. \quad (4.26)$$

In particular J_{k+1} is a fixed point of W_k . Indeed, using the definition

$$J_{k+1} = T_{\mu^k}^{(\lambda)} J_k$$

[cf. Eq. (4.24)], and the linearity assumption (4.25), we have

$$\begin{aligned} W_k J_{k+1} &= (1 - \lambda)T_{\mu^k} J_k + \lambda T_{\mu^k} \left(T_{\mu^k}^{(\lambda)} J_k \right) \\ &= (1 - \lambda)T_{\mu^k} J_k + \lambda T_{\mu^k}^{(\lambda)} (T_{\mu^k} J_k) \\ &= T_{\mu^k}^{(\lambda)} J_k \\ &= J_{k+1}. \end{aligned}$$

Thus J_{k+1} can be calculated as a fixed point of W_k .

Consider now the case where T_{μ^k} is nonexpansive with respect to some norm. Then from Eq. (4.26), it is seen that W_k is a contraction of modulus λ with respect to that norm, so J_{k+1} is the unique fixed point of W_k . Moreover, if the norm is a weighted sup-norm, J_{k+1} can be found using the methods of Chapter 2 for contractive models. The following example applies this idea to finite-state SSP problems. The interesting aspect of this example is that it implements the policy evaluation portion of λ -PI through solution of a system of linear equations, similar to the exact policy evaluation method of classical PI.

Example 4.3.4 (Stochastic Shortest Path Problems with Nonpositive Costs)

Consider the SSP problem of Example 1.2.6 with states $1, \dots, n$, plus the termination state 0. For all $u \in U(x)$, the state following x is y with probability $p_{xy}(u)$ and the expected cost incurred is nonpositive. This problem arises when we wish to maximize nonnegative rewards up to termination. It includes a classical search problem where the aim, roughly speaking, is to move through the state space looking for states with favorable termination rewards.

We view the problem within our abstract framework with $\bar{J}(x) \equiv 0$ and

$$T_{\mu} J = g_{\mu} + P_{\mu} J, \quad (4.27)$$

with $g_{\mu} \in \Re^n$ being the corresponding nonpositive one-stage cost vector, and P_{μ} being an $n \times n$ substochastic matrix. The components of P_{μ} are the probabilities $p_{xy}(\mu(x))$, $x, y = 1, \dots, n$. Clearly Assumption D holds.

Consider the λ -PI method (4.24), with J_{k+1} computed by solving the fixed point equation $J = W_k J$, cf. Eq. (4.26). This is a nonsingular n -dimensional system of linear equations, and can be solved by matrix inversion, just like in exact PI for discounted n -state MDP. In particular, using Eqs. (4.26) and (4.27), we have

$$J_{k+1} = (I - \lambda P_{\mu^k})^{-1} (g_{\mu^k} + (1 - \lambda) P_{\mu^k} J_k). \quad (4.28)$$

For a small number of states n , this matrix inversion-based policy evaluation may be simpler than the optimistic PI policy evaluation equation

$$J_{k+1} = T_{\mu^k}^{m_k} J_k$$

[cf. Eq. (4.20)], which points to an advantage of λ -PI.

Note that based on the relation between the multistep mapping $T_{\mu}^{(\lambda)}$ and the proximal mapping, discussed in Section 1.2.5 and Exercise 1.2, the policy evaluation Eq. (4.28) may be viewed as an extrapolated proximal iteration. Note also that as $\lambda \rightarrow 1$, the policy evaluation Eq. (4.28) resembles the policy evaluation equation

$$J_{\mu^k} = (I - \lambda P_{\mu^k})^{-1} g_{\mu^k}$$

for λ -discounted n -state MDP. An important difference, however, is that for a discounted finite-state MDP, exact PI will find an optimal policy in a finite number of iterations, while this is not guaranteed for λ -PI. Indeed λ -PI does not require that there exists an optimal policy or even that $J^*(x)$ is finite for all x .

Policy Iteration Under I

Contrary to the case of Assumption D, the important cost improvement property of PI holds under Assumption I. Thus, if μ is a policy and $\bar{\mu}$ satisfies the policy improvement equation $T_{\bar{\mu}} J_{\mu} = T J_{\mu}$, we have

$$J_{\mu} = T_{\mu} J_{\mu} \geq T J_{\mu} = T_{\bar{\mu}} J_{\mu},$$

from which we obtain

$$J_{\mu} \geq \lim_{k \rightarrow \infty} T_{\bar{\mu}}^k J_{\mu}.$$

Since $J_{\mu} \geq \bar{J}$ and $J_{\bar{\mu}} = \lim_{k \rightarrow \infty} T_{\bar{\mu}}^k \bar{J}$, it follows that

$$J_{\mu} \geq T J_{\mu} \geq J_{\bar{\mu}}. \quad (4.29)$$

However, this cost improvement property is not by itself sufficient for the validity of PI under Assumption I (see the deterministic shortest path example of Section 3.1.1). Thus additional conditions are needed to guarantee convergence. To this end we may use the semicontractive framework of Chapter 3, and take advantage of the fact that under Assumption I, J^* is known to be a fixed point of T .

In particular, suppose that we have a set $S \subset \mathcal{E}(X)$ such that $J_S^* = J^*$. Then J_S^* is a fixed point of T and the theory of Section 3.2 comes into play. Thus, by Prop. 3.2.1 the following hold:

- (a) We have $T^k J \rightarrow J^*$ for every $J \in \mathcal{E}(X)$ such that $J^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$.
- (b) J^* is the only fixed point of T within the set of all $J \in \mathcal{E}(X)$ such that $J^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$.

Moreover, by Prop. 3.2.4, if S has the weak PI property and for each sequence $\{J_m\} \subset \mathcal{E}(X)$ with $J_m \downarrow J$ for some $J \in \mathcal{E}(X)$, we have

$$H(x, u, J) = \lim_{m \rightarrow \infty} H(x, u, J_m),$$

then every sequence of S -regular policies $\{\mu^k\}$ that can be generated by PI satisfies $J_{\mu^k} \downarrow J^*$. If in addition the set of S -regular policies is finite, there exists $\bar{k} \geq 0$ such that $\mu^{\bar{k}}$ is optimal.

For these properties to hold, it is of course critical that $J_S^* = J^*$. If this is not so, but J_S^* is still a fixed point of T , the VI and PI algorithms may converge to J_S^* rather than to J^* (cf. the linear quadratic problem of Section 3.5.4).

4.4 REGULARITY AND NONSTATIONARY POLICIES

In this section, we will extend the notion of regularity of Section 3.2 so that it applies more broadly. We will use this notion as our main tool for exploring the structure of the solution set of Bellman's equation. We will then discuss some applications involving mostly monotone increasing models in this section, as well as in Sections 4.5 and 4.6. We continue to focus on the infinite horizon case of the problem of Section 4.1, but we do not impose for the moment any additional assumptions, such as Assumption I or D.

We begin with the following extension of the definition of S -regularity, which we will use to prove a general result regarding the convergence properties of VI in the following Prop. 4.4.1. We will apply this result in the context of various applications in Sections 4.4.2-4.4.4, as well as in Sections 4.5 and 4.6.

Definition 4.4.1: For a nonempty set of functions $S \subset \mathcal{E}(X)$, we say that a nonempty collection \mathcal{C} of policy-state pairs (π, x) , with $\pi \in \Pi$ and $x \in X$, is *S -regular* if

$$J_\pi(x) = \limsup_{k \rightarrow \infty} (T_{\mu_0} \cdots T_{\mu_k} J)(x), \quad \forall (\pi, x) \in \mathcal{C}, J \in S.$$

The essence of the preceding definition of S -regularity is similar to the one of Chapter 3 for stationary policies: *for an S -regular collection of pairs (π, x) , the value of $J_\pi(x)$ is not affected if the starting function is changed from \bar{J} to any $J \in S$.* It is important to extend the definition of regularity to nonstationary policies because in noncontractive models, stationary policies are generally not sufficient, i.e., the optimal cost over

stationary policies may not be the same as the one over nonstationary policies (cf. Prop. 4.3.2, and the subsequent example). Generally, when referring to an S -regular collection \mathcal{C} , we implicitly assume that S and \mathcal{C} are nonempty, although on occasion we may state explicitly this fact for emphasis.

For a given set \mathcal{C} of policy-state pairs (π, x) , let us consider the function $J_{\mathcal{C}}^* \in \mathcal{E}(X)$, given by

$$J_{\mathcal{C}}^*(x) = \inf_{\{\pi \mid (\pi, x) \in \mathcal{C}\}} J_{\pi}(x), \quad x \in X.$$

Note that $J_{\mathcal{C}}^*(x) \geq J^*(x)$ for all $x \in X$ [for those $x \in X$ for which the set of policies $\{\pi \mid (\pi, x) \in \mathcal{C}\}$ is empty, we have by convention $J_{\mathcal{C}}^*(x) = \infty$].

For an important example, note that in the analysis of Chapter 3, the set of S -regular policies \mathcal{M}_S of Section 3.2 defines the S -regular collection

$$\mathcal{C} = \{(\mu, x) \mid \mu \in \mathcal{M}_S, x \in X\},$$

and the corresponding restricted optimal cost function J_S^* is equal to $J_{\mathcal{C}}^*$. In Sections 3.2-3.4 we saw that when J_S^* is a fixed point of T , then favorable results are obtained. Similarly, in this section we will see that for an S -regular collection \mathcal{C} , when $J_{\mathcal{C}}^*$ is a fixed point of T , interesting results are obtained.

The following two propositions play a central role in our analysis on this section and the next two, and may be compared with Prop. 3.2.1, which played a pivotal role in the analysis of Chapter 3.

Proposition 4.4.1: (Well-Behaved Region Theorem) Given a nonempty set $S \subset \mathcal{E}(X)$, let \mathcal{C} be a nonempty collection of policy-state pairs (π, x) that is S -regular. Then:

(a) For all $J \in \mathcal{E}(X)$ such that $J \leq \tilde{J}$ for some $\tilde{J} \in S$, we have

$$\limsup_{k \rightarrow \infty} T^k J \leq J_{\mathcal{C}}^*.$$

(b) For all $J' \in \mathcal{E}(X)$ with $J' \leq T J'$, and all $J \in \mathcal{E}(X)$ such that $J' \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$, we have

$$J' \leq \liminf_{k \rightarrow \infty} T^k J \leq \limsup_{k \rightarrow \infty} T^k J \leq J_{\mathcal{C}}^*.$$

Proof: (a) Using the generic relation $TJ \leq T_{\mu}J$, $\mu \in \mathcal{M}$, and the monotonicity of T and T_{μ} , we have for all k

$$(T^k \tilde{J})(x) \leq (T_{\mu_0} \cdots T_{\mu_{k-1}} \tilde{J})(x), \quad \forall (\pi, x) \in \mathcal{C}, \tilde{J} \in S.$$

By letting $k \rightarrow \infty$ and by using the definition of S -regularity, it follows that for all $(\pi, x) \in \mathcal{C}$, $J \in \mathcal{E}(X)$, and $\tilde{J} \in S$ with $J \leq \tilde{J}$,

$$\limsup_{k \rightarrow \infty} (T^k J)(x) \leq \limsup_{k \rightarrow \infty} (T^k \tilde{J})(x) \leq \limsup_{k \rightarrow \infty} (T_{\mu_0} \cdots T_{\mu_{k-1}} \tilde{J})(x) = J_\pi(x),$$

and by taking infimum of the right side over $\{\pi \mid (\pi, x) \in \mathcal{C}\}$, we obtain the result.

(b) Using the hypotheses $J' \leq T J'$, and $J' \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$, and the monotonicity of T , we have

$$J'(x) \leq (T J')(x) \leq \cdots \leq (T^k J')(x) \leq (T^k J)(x).$$

Letting $k \rightarrow \infty$ and using part (a), we obtain the result. **Q.E.D.**

Let us discuss some interesting implications of part (b) of the proposition. Suppose we are given a set $S \subset \mathcal{E}(X)$, and a collection \mathcal{C} that is S -regular. Then:

- (1) $J_{\mathcal{C}}^*$ is an upper bound to every fixed point J' of T that lies below some $\tilde{J} \in S$ (i.e., $J' \leq \tilde{J}$). Moreover, for such a fixed point J' , the VI algorithm, starting from any J with $J_{\mathcal{C}}^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$, ends up asymptotically within the region

$$\{J \in \mathcal{E}(X) \mid J' \leq J \leq J_{\mathcal{C}}^*\}.$$

Thus the convergence of VI is characterized by the *well-behaved region*

$$\mathcal{W}_{S,\mathcal{C}} = \{J \in \mathcal{E}(X) \mid J_{\mathcal{C}}^* \leq J \leq \tilde{J} \text{ for some } \tilde{J} \in S\}, \quad (4.30)$$

(cf. the corresponding definition in Section 3.2), and the *limit region*

$$\begin{aligned} &\{J \in \mathcal{E}(X) \mid J' \leq J \leq J_{\mathcal{C}}^* \text{ for all fixed points } J' \text{ of } T \\ &\text{with } J' \leq \tilde{J} \text{ for some } \tilde{J} \in S\}. \end{aligned}$$

The VI algorithm, starting from the former, ends up asymptotically within the latter; cf. Figs. 4.4.1 and 4.4.2.

- (2) If $J_{\mathcal{C}}^*$ is a fixed point of T (a common case in our subsequent analysis), then the VI-generated sequence $\{T^k J\}$ converges to $J_{\mathcal{C}}^*$ starting from any J in the well-behaved region. If $J_{\mathcal{C}}^*$ is not a fixed point of T , we only have $\limsup_{k \rightarrow \infty} T^k J \leq J_{\mathcal{C}}^*$ for all J in the well-behaved region.
- (3) If the well-behaved region is unbounded above in the sense that $\mathcal{W}_{S,\mathcal{C}} = \{J \in \mathcal{E}(X) \mid J_{\mathcal{C}}^* \leq J\}$, which is true for example if $S = E(X)$, then $J' \leq J_{\mathcal{C}}^*$ for every fixed point J' of T . The reason is that for every fixed point J' of T we have $J' \leq J$ for some $J \in \mathcal{W}_{S,\mathcal{C}}$, and hence also $J' \leq \tilde{J}$ for some $\tilde{J} \in S$, so observation (1) above applies.

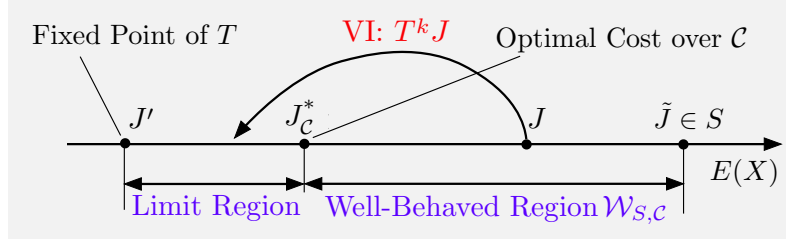


Figure 4.4.1. Schematic illustration of Prop. 4.4.1. Neither J_C^* nor J^* need to be fixed points of T , but if \mathcal{C} is S -regular, and there exists $\tilde{J} \in S$ with $J_C^* \leq \tilde{J}$, then J_C^* demarcates from above the range of fixed points of T that lie below \tilde{J} .

For future reference, we state these observations as a proposition, which should be compared to Prop. 3.2.1, the stationary special case where \mathcal{C} is defined by the set of S -regular stationary policies, i.e., $\mathcal{C} = \{(\mu, x) \mid \mu \in \mathcal{M}_S, x \in X\}$. Figures 4.4.2 and 4.4.3 illustrate some of the consequences of Prop. 4.4.1 for two cases, respectively: when $S = E(X)$ while J_C^* is not a fixed point of T , and when S is a strict subset of $E(X)$ while J_C^* is a fixed point of T .

Proposition 4.4.2: (Uniqueness of Fixed Point of T and Convergence of VI) Given a set $S \subset \mathcal{E}(X)$, let \mathcal{C} be a collection of policy-state pairs (π, x) that is S -regular. Then:

- (a) If J' is a fixed point of T with $J' \leq \tilde{J}$ for some $\tilde{J} \in S$, then $J' \leq J_C^*$. Moreover, J_C^* is the only possible fixed point of T within $\mathcal{W}_{S,\mathcal{C}}$.
- (b) We have $\limsup_{k \rightarrow \infty} T^k J \leq J_C^*$ for all $J \in \mathcal{W}_{S,\mathcal{C}}$, and if J_C^* is a fixed point of T , then $T^k J \rightarrow J_C^*$ for all $J \in \mathcal{W}_{S,\mathcal{C}}$.
- (c) If $\mathcal{W}_{S,\mathcal{C}}$ is unbounded from above in the sense that

$$\mathcal{W}_{S,\mathcal{C}} = \{J \in \mathcal{E}(X) \mid J_C^* \leq J\},$$

then $J' \leq J_C^*$ for every fixed point J' of T . In particular, if J_C^* is a fixed point of T , then J_C^* is the largest fixed point of T .

Proof: (a) The first statement follows from Prop. 4.4.1(b). For the second statement, let J' be a fixed point of T with $J' \in \mathcal{W}_{S,\mathcal{C}}$. Then from the definition of $\mathcal{W}_{S,\mathcal{C}}$, we have $J_C^* \leq J'$ as well as $J' \leq \tilde{J}$ for some $\tilde{J} \in S$, so from Prop. 4.4.1(b) it follows that $J' \leq J_C^*$. Hence $J' = J_C^*$.

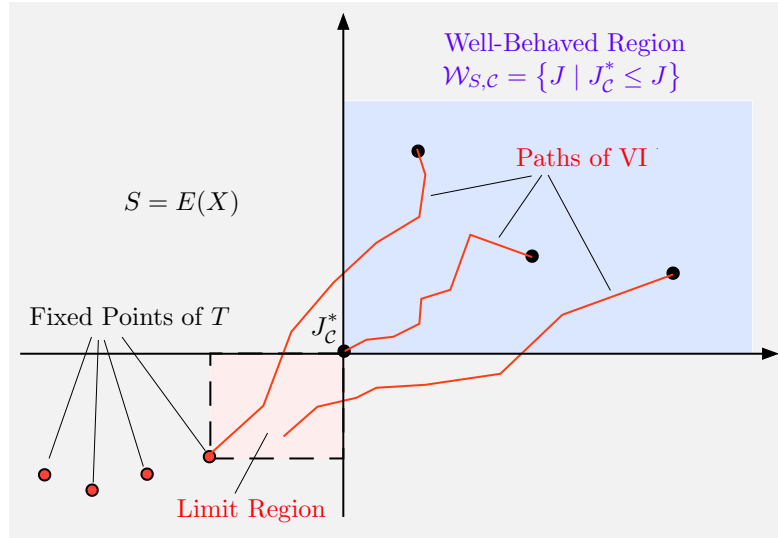


Figure 4.4.2. Schematic illustration of Prop. 4.4.2, for the case where $S = E(X)$ so that $\mathcal{W}_{S,C}$ is unbounded above, i.e., $\mathcal{W}_{S,C} = \{J \in E(X) \mid J_C^* \leq J\}$. In this figure J_C^* is not a fixed point of T . The VI algorithm, starting from the well-behaved region $\mathcal{W}_{S,C}$, ends up asymptotically within the limit region.

- (b) The result follows from Prop. 4.4.1(a), and in the case where J_C^* is a fixed point of T , from Prop. 4.4.1(b), with $J' = J_C^*$.
- (c) See observation (3) in the discussion preceding the proposition. **Q.E.D.**

Examples and counterexamples illustrating the preceding proposition are provided by the problems of Section 3.1 for the stationary case where

$$\mathcal{C} = \{(\mu, x) \mid \mu \in \mathcal{M}_S, x \in X\}.$$

Similar to the analysis of Chapter 3, the preceding proposition takes special significance when J^* is a fixed point of T and \mathcal{C} is rich enough so that $J_C^* = J^*$, as for example in the case where \mathcal{C} is the set $\Pi \times X$ of all (π, x) , or other choices to be discussed later. It then follows that every fixed point J' of T that belongs to S satisfies $J' \leq J^*$, and that VI converges to J^* starting from any $J \in \mathcal{E}(X)$ such that $J^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$. However, there will be interesting cases where $J_C^* \neq J^*$, as in shortest path-type problems (see Sections 3.5.1, 4.5, and 4.6).

Note that Prop. 4.4.2 does not say anything about fixed points of T that lie below J_C^* , and does not give conditions under which J_C^* is a fixed point. Moreover, it does not address the question whether J^* is a fixed point of T , or whether VI converges to J^* starting from \bar{J} or from below J^* . Generally, it can happen that both, only one, or none of the two

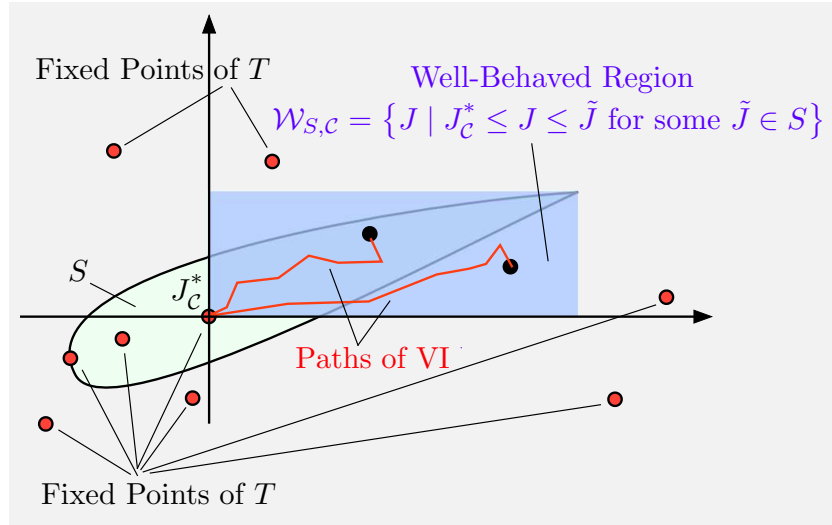


Figure 4.4.3. Schematic illustration of Prop. 4.4.2, and the set $\mathcal{W}_{S,\mathcal{C}}$ of Eq. (4.30), for a case where $J_{\mathcal{C}}^*$ is a fixed point of T and S is a strict subset of $E(X)$. Every fixed point of T that lies below some $\tilde{J} \in S$ should lie below $J_{\mathcal{C}}^*$. Also, the VI algorithm converges to $J_{\mathcal{C}}^*$ starting from within $\mathcal{W}_{S,\mathcal{C}}$. If S were unbounded from above, as in Fig. 4.4.2, $J_{\mathcal{C}}^*$ would be the largest fixed point of T .

functions $J_{\mathcal{C}}^*$ and J^* is a fixed point of T , as can be seen from the examples of Section 3.1.

The Case Where $J_{\mathcal{C}}^* \leq \bar{J}$

We have seen in Section 4.3 that the results for monotone increasing and monotone decreasing models are markedly different. In the context of S -regularity of a collection \mathcal{C} , it turns out that there are analogous significant differences between the cases $J_{\mathcal{C}}^* \geq \bar{J}$ and $J_{\mathcal{C}}^* \leq \bar{J}$. The following proposition establishes some favorable aspects of the condition $J_{\mathcal{C}}^* \leq \bar{J}$ in the context of VI. These can be attributed to the fact that \bar{J} can always be added to S without affecting the S -regularity of \mathcal{C} , so \bar{J} can serve as the element \tilde{J} of S in Props. 4.4.1 and 4.4.2 (see the subsequent proof). The following proposition may also be compared with the result on convergence of VI under Assumption D (cf. Prop. 4.3.13).

Proposition 4.4.3: Given a set $S \subset \mathcal{E}(X)$, let \mathcal{C} be a collection of policy-state pairs (π, x) that is S -regular, and assume that $J_{\mathcal{C}}^* \leq \bar{J}$. Then:

(a) For all $J' \in \mathcal{E}(X)$ with $J' \leq TJ'$, we have

$$J' \leq \liminf_{k \rightarrow \infty} T^k \bar{J} \leq \limsup_{k \rightarrow \infty} T^k \bar{J} \leq J_{\mathcal{C}}^*$$

(b) If $J_{\mathcal{C}}^*$ is a fixed point of T , then $J_{\mathcal{C}}^* = J^*$ and we have $T^k \bar{J} \rightarrow J^*$ as well as $T^k J \rightarrow J^*$ for every $J \in \mathcal{E}(X)$ such that $J^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$.

Proof: (a) If S does not contain \bar{J} , we can replace S with $\bar{S} = S \cup \{\bar{J}\}$, and \mathcal{C} will still be \bar{S} -regular. By applying Prop. 4.4.1(b) with S replaced by \bar{S} and $\tilde{J} = \bar{J}$, the result follows.

(b) Assume without loss of generality that $\bar{J} \in S$ [cf. the proof of part (a)]. By using Prop. 4.4.2(b) with $\tilde{J} = \bar{J}$, we have $J_{\mathcal{C}}^* = \lim_{k \rightarrow \infty} T^k \bar{J}$. Thus for every policy $\pi = \{\mu_0, \mu_1, \dots\} \in \Pi$,

$$J_{\mathcal{C}}^* = \lim_{k \rightarrow \infty} T^k \bar{J} \leq \limsup_{k \rightarrow \infty} T_{\mu_0} \cdots T_{\mu_{k-1}} \bar{J} = J_{\pi},$$

so by taking the infimum over $\pi \in \Pi$, we obtain $J_{\mathcal{C}}^* \leq J^*$. Since generically $J_{\mathcal{C}}^* \geq J^*$, it follows that $J_{\mathcal{C}}^* = J^*$. Finally, from Prop. 4.4.2(b), $T^k J \rightarrow J^*$ for all $J \in \mathcal{W}_{S, \mathcal{C}}$, implying the result. **Q.E.D.**

As a special case of the preceding proposition, we have that if $J^* \leq \bar{J}$ and J^* is a fixed point of T , then $J^* = \lim_{k \rightarrow \infty} T^k \bar{J}$, and for every other fixed point J' of T we have $J' \leq J^*$ (apply the proposition with $\mathcal{C} = \Pi \times X$ and $S = \{\bar{J}\}$, in which case $J_{\mathcal{C}}^* = J^* \leq \bar{J}$). This occurs, among others, in the monotone decreasing models, where $T_{\mu} \bar{J} \leq \bar{J}$ for all $\mu \in \mathcal{M}$. A special case is the convergence of VI under Assumption D (cf. Prop. 4.3.5).

The preceding proposition also applies to a classical type of search problem with both positive and negative costs per stage. This is the SSP problem, where at each $x \in X$ we have $\text{cost } E\{g(x, u, w)\} \geq 0$ for all u except one that leads to a termination state with probability 1 and non-positive cost; here $\bar{J}(x) = 0$ and $J_{\mathcal{C}}^*(x) \leq 0$ for all $x \in X$, but Assumption D need not hold.

4.4.1 Regularity and Monotone Increasing Models

We will now return to the monotone increasing model, cf. Assumption I. For this model, we know from Section 4.3 that J^* is the smallest fixed point of T within the class of functions $J \geq \bar{J}$, under certain relatively mild assumptions. However, VI may not converge to J^* starting from below J^* (e.g., starting from \bar{J}), and also starting from above J^* . In this section

we will address the question of convergence of VI from above J^* by using regularity ideas, and in Section 4.5 we will consider the characterization of the largest fixed point of T in the context of deterministic optimal control and infinite-space shortest path problems. We summarize the results of Section 4.3 that are relevant to our development in the following proposition (cf. Props. 4.3.2, 4.3.3, 4.3.9, and 4.3.14).

Proposition 4.4.4: Let Assumption I hold. Then:

- (a) $J^* = TJ^*$, and if $J' \in \mathcal{E}(X)$ is such that $J' \geq \bar{J}$ and $J' \geq TJ'$, then $J' \geq J^*$.
- (b) For all $\mu \in \mathcal{M}$ we have $J_\mu = T_\mu J_\mu$, and if $J' \in \mathcal{E}(X)$ is such that $J' \geq \bar{J}$ and $J' \geq T_\mu J'$, then $J' \geq J_\mu$.
- (c) $\mu^* \in \mathcal{M}$ is optimal if and only if $T_{\mu^*} J^* = TJ^*$.
- (d) If U is a metric space and the sets

$$U_k(x, \lambda) = \{u \in U(x) \mid H(x, u, T^k \bar{J}) \leq \lambda\}$$

are compact for all $x \in X$, $\lambda \in \mathfrak{R}$, and k , then there exists at least one optimal stationary policy, and we have $T^k J \rightarrow J^*$ for all $J \in \mathcal{E}(X)$ with $J \leq J^*$.

- (e) Given any $\epsilon > 0$, there exists a policy $\pi_\epsilon \in \Pi$ such that

$$J^* \leq J_{\pi_\epsilon} \leq J^* + \epsilon e.$$

Furthermore, if the scalar α in part (c) of Assumption I satisfies $\alpha < 1$, the policy π_ϵ can be taken to be stationary.

Since under Assumption I there may exist fixed points J' of T with $J^* \leq J'$, VI may not converge to J^* starting from above J^* . However, convergence of VI to J^* from above, if it occurs, is often much faster than convergence from below, so starting points $J \geq J^*$ may be desirable. One well-known such case is deterministic finite-state shortest path problems where major algorithms, such as the Bellman-Ford method or other label correcting methods have polynomial complexity, when started from J above J^* , but only pseudopolynomial complexity when started from J below J^* [see e.g., [BeT89] (Prop. 1.2 in Ch.4), [Ber98] (Exercise 2.7)].

In the next two subsections, we will consider discounted and undiscounted optimal control problems with nonnegative cost per stage, and we will establish conditions under which J^* is the unique nonnegative fixed point of T , and VI converges to J^* from above. Our analysis will proceed as follows:

- (a) Define a collection \mathcal{C} such that $J_{\mathcal{C}}^* = J^*$.
- (b) Define a set $S \subset \mathcal{E}^+(X)$ such that $J^* \in S$ and \mathcal{C} is S -regular.
- (c) Use Prop. 4.4.2 (which shows that $J_{\mathcal{C}}^*$ is the largest fixed point of T within S) in conjunction with Prop. 4.4.4(a) (which shows that J^* is the smallest fixed point of T within S) to show that J^* is the unique fixed point of T within S . Use also Prop. 4.4.2(b) to show that the VI algorithm converges to J^* starting from $J \in S$ such that $J \geq J^*$.
- (d) Use the compactness condition of Prop. 4.4.4(d), to enlarge the set of functions starting from which VI converges to J^* .

4.4.2 Nonnegative Cost Stochastic Optimal Control

Let us consider the undiscounted stochastic optimal control problem that involves the mapping

$$H(x, u, J) = E\{g(x, u, w) + J(f(x, u, w))\}, \quad (4.31)$$

where g is the one-stage cost function and f is the system function. The expected value is taken with respect to the distribution of the random variable w (which takes values in a countable set W). We assume that

$$0 \leq E\{g(x, u, w)\} < \infty, \quad \forall x \in X, u \in U(x), w \in W.$$

We consider the abstract DP model with H as above, and with $\bar{J}(x) \equiv 0$. Using the nonnegativity of g , we can write the cost function of a policy $\pi = \{\mu_0, \mu_1, \dots\}$ in terms of a limit,

$$J_{\pi}(x_0) = \lim_{k \rightarrow \infty} E_{x_0}^{\pi} \left\{ \sum_{m=0}^k g(x_m, \mu_m(x_m), w_m) \right\}, \quad x_0 \in X, \quad (4.32)$$

where $E_{x_0}^{\pi}\{\cdot\}$ denotes expected value with respect to the probability distribution induced by π under initial state x_0 .

We will apply the analysis of this section with

$$\mathcal{C} = \{(\pi, x) \mid J_{\pi}(x) < \infty\},$$

for which $J_{\mathcal{C}}^* = J^*$. We assume that \mathcal{C} is nonempty, which is true if and only if J^* is not identically ∞ , i.e., $J^*(x) < \infty$ for some $x \in X$. Consider the set

$$S = \left\{ J \in \mathcal{E}^+(X) \mid E_{x_0}^{\pi} \{J(x_k)\} \rightarrow 0, \forall (\pi, x_0) \in \mathcal{C} \right\}. \quad (4.33)$$

One interpretation is that the functions J that are in S have the character of Lyapounov functions for the policies π for which the set $\{x_0 \mid J_{\pi}(x_0) < \infty\}$ is nonempty.

Note that S is the largest set with respect to which \mathcal{C} is regular in the sense that \mathcal{C} is S -regular and if \mathcal{C} is S' -regular for some other set S' , then $S' \subset S$. To see this we write for all $J \in \mathcal{E}^+(X)$, $(\pi, x_0) \in \mathcal{C}$, and k ,

$$(T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x_0) = E_{x_0}^{\pi} \{J(x_k)\} + E_{x_0}^{\pi} \left\{ \sum_{m=0}^{k-1} g(x_m, \mu_m(x_m), w_m) \right\},$$

where μ_m , $m = 0, 1, \dots$, denote generically the components of π . The rightmost term above converges to $J_{\pi}(x_0)$ as $k \rightarrow \infty$ [cf. Eq. (4.32)], so by taking upper limit, we obtain

$$\limsup_{k \rightarrow \infty} (T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x_0) = \limsup_{k \rightarrow \infty} E_{x_0}^{\pi} \{J(x_k)\} + J_{\pi}(x_0). \quad (4.34)$$

In view of the definition (4.33) of S , this implies that for all $J \in S$, we have

$$\limsup_{k \rightarrow \infty} (T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x_0) = J_{\pi}(x_0), \quad \forall (\pi, x_0) \in \mathcal{C}, \quad (4.35)$$

so \mathcal{C} is S -regular. Moreover, if \mathcal{C} is S' -regular and $J \in S'$, Eq. (4.35) holds, so that [in view of Eq. (4.34) and $J \in \mathcal{E}^+(X)$] $\lim_{k \rightarrow \infty} E_{x_0}^{\pi} \{J(x_k)\} = 0$ for all $(\pi, x_0) \in \mathcal{C}$, implying that $J \in S$.

From Prop. 4.4.2, the fixed point property of J^* [cf. Prop. 4.4.4(a)], and the fact $J_C^* = J^*$, it follows that $T^k J \rightarrow J^*$ for all $J \in S$ that satisfy $J \geq J^*$. Moreover, if the sets $U_k(x, \lambda)$ of Eq. (4.17) are compact, the convergence of VI starting from below J^* will also be guaranteed. We thus have the following proposition, which in addition shows that J^* belongs to S and is the unique fixed point of T within S .

Proposition 4.4.5: (Uniqueness of Fixed Point of T and Convergence of VI) Consider the problem corresponding to the mapping (4.31) with $g \geq 0$, and assume that J^* is not identically ∞ . Then:

- (a) J^* belongs to S and is the unique fixed point of T within S . Moreover, we have $T^k J \rightarrow J^*$ for all $J \geq J^*$ with $J \in S$.
- (b) If U is a metric space, and the sets $U_k(x, \lambda)$ of Eq. (4.17) are compact for all $x \in X$, $\lambda \in \mathfrak{R}$, and k , we have $T^k J \rightarrow J^*$ for all $J \in S$, and an optimal stationary policy is guaranteed to exist.

Proof: (a) We first show that $J^* \in S$. Given a policy $\pi = \{\mu_0, \mu_1, \dots\}$, we denote by π_k the policy

$$\pi_k = \{\mu_k, \mu_{k+1}, \dots\}.$$

We have for all $(\pi, x_0) \in \mathcal{C}$

$$J_\pi(x_0) = E_{x_0}^\pi \{g(x_0, \mu_0(x_0), w_0)\} + E_{x_0}^\pi \{J_{\pi_1}(x_1)\}, \quad (4.36)$$

and for all $m = 1, 2, \dots$,

$$E_{x_0}^\pi \{J_{\pi_m}(x_m)\} = E_{x_0}^\pi \{g(x_m, \mu_m(x_m), w_m)\} + E_{x_0}^\pi \{J_{\pi_{m+1}}(x_{m+1})\}, \quad (4.37)$$

where $\{x_m\}$ is the sequence generated starting from x_0 and using π . By using repeatedly the expression (4.37) for $m = 1, \dots, k-1$, and combining it with Eq. (4.36), we obtain for all $k = 1, 2, \dots$,

$$J_\pi(x_0) = E_{x_0}^\pi \{J_{\pi_k}(x_k)\} + \sum_{m=0}^{k-1} E_{x_0}^\pi \{g(x_m, \mu_m(x_m), w_m)\}, \quad \forall (\pi, x_0) \in \mathcal{C}.$$

The rightmost term above tends to $J_\pi(x_0)$ as $k \rightarrow \infty$, so we obtain

$$E_{x_0}^\pi \{J_{\pi_k}(x_k)\} \rightarrow 0, \quad \forall (\pi, x_0) \in \mathcal{C}.$$

Since $0 \leq J^* \leq J_{\pi_k}$, it follows that

$$E_{x_0}^\pi \{J^*(x_k)\} \rightarrow 0, \quad \forall x_0 \text{ with } J^*(x_0) < \infty.$$

Thus $J^* \in S$ while J^* (which is equal to $J_{\mathcal{C}}^*$) is a fixed point of T . For every other fixed point J' of T , we have $J' \geq J^*$ [by Prop. 4.4.4(b)], so if J' belongs to S , by Prop. 4.4.2(a), $J' \leq J^*$ and thus $J' = J^*$. Hence, J^* is the unique fixed point of T within the set S . By Prop. 4.4.2(b), we also have $T^k J \rightarrow J^*$ for all $J \in S$ with $J \geq J^*$.

(b) This part follows from part (a) and Prop. 4.4.4(d). **Q.E.D.**

Note that under the assumptions of the preceding proposition, either T has a unique fixed point within $\mathcal{E}^+(X)$ (namely J^*), or else all the additional fixed points of T within $\mathcal{E}^+(X)$ lie outside S . To illustrate the limitations of this result, consider the shortest path problem of Section 3.1.1 for the case where the choice at state 1 is either to stay at 1 at cost 0, or move to the destination at cost $b > 0$. Then Bellman's equation at state 1 is $J(1) = \min\{b, J(1)\}$, and its set of nonnegative solutions is the interval $[0, b]$, while we have $J^* = 0$. The set S of Eq. (4.33) here consists of just J^* and Prop. 4.4.5 applies, but it is not very useful. Similarly, in the linear-quadratic example of Section 3.1.4, where T has the two fixed points $J^*(x) = 0$ and $\hat{J}(x) = (\gamma^2 - 1)x^2$, the set S of Eq. (4.33) consists of just J^* .

Thus the regularity framework of this section is useful primarily in the favorable case where J^* is the unique nonnegative fixed point of T . In particular, Prop. 4.4.5 cannot be used to differentiate between multiple

fixed points of T , and to explain the unusual behavior in the preceding two examples. In Sections 4.5 and 4.6, we address this issue within the more restricted contexts of deterministic and stochastic optimal control, respectively.

A consequence of Prop. 4.4.5 is the following condition for VI convergence from above, first proved in the paper by Yu and Bertsekas [YuB15] (Theorem 5.1) within a broader context that also addressed universal measurability issues.

Proposition 4.4.6: Under the conditions of Prop. 4.4.5, we have $T^k J \rightarrow J^*$ for all $J \in \mathcal{E}^+(X)$ satisfying

$$J^* \leq J \leq cJ^*, \quad (4.38)$$

for some scalar $c > 1$. Moreover, J^* is the unique fixed point of T within the set

$$\{J \in \mathcal{E}^+(X) \mid J \leq cJ^* \text{ for some } c > 0\}.$$

Proof: Since $J^* \in S$ as shown in Prop. 4.4.5, any J satisfying Eq. (4.38), also belongs to the set S of Eq. (4.33), and the result follows from Prop. 4.4.5. **Q.E.D.**

Note a limitation of the preceding proposition: in order to find functions J satisfying $J^* \leq J \leq cJ^*$ we must essentially know the sets of states x where $J^*(x) = 0$ and $J^*(x) = \infty$.

4.4.3 Discounted Stochastic Optimal Control

We will now consider a discounted version of the stochastic optimal control problem of the preceding section. For a policy $\pi = \{\mu_0, \mu_1, \dots\}$ we have

$$J_\pi(x_0) = \lim_{k \rightarrow \infty} E_{x_0}^\pi \left\{ \sum_{m=0}^{k-1} \alpha^m g(x_m, \mu_m(x_m), w_m) \right\},$$

where $\alpha \in (0, 1)$ is the discount factor, and as earlier $E_{x_0}^\pi \{\cdot\}$ denotes expected value with respect to the probability measure induced by $\pi \in \Pi$ under initial state x_0 . We assume that the one-stage expected cost is non-negative,

$$0 \leq E\{g(x, u, w)\} < \infty, \quad \forall x \in X, u \in U(x), w \in W.$$

By defining the mapping H as

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

and $\bar{J}(x) \equiv 0$, we can view this problem within the abstract DP framework of this chapter where Assumption I holds.

Note that because of the discount factor, the existence of a terminal set of states is not essential for the optimal costs to be finite. Moreover, the nonnegativity of g is not essential for our analysis. Any problem where g can take both positive and negative values, but is bounded below, can be converted to an equivalent problem where g is nonnegative, by adding a suitable constant c to g . Then the cost of all policies will simply change by the constant $\sum_{k=0}^{\infty} \alpha^k c = c/(1 - \alpha)$.

The line of analysis of this section makes a connection between the S -regularity notion of Definition 4.4.1 and a notion of stability, which is common in feedback control theory and will be explored further in Section 4.5. We assume that X is a normed space, so that boundedness within X is defined with respect to its norm. We introduce the set

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

which we assume to be nonempty. Given a state $x \in X^*$, we say that a policy π is *stable from x* if there exists a bounded subset of X^* [that depends on (π, x)] such that the (random) sequence $\{x_k\}$ generated starting from x and using π lies with probability 1 within that subset. We consider the set of policy-state pairs

$$\mathcal{C} = \{(\pi, x) \mid x \in X^*, \pi \text{ is stable from } x\},$$

and we assume that \mathcal{C} is nonempty.

Let us say that a function $J \in \mathcal{E}^+(X)$ is *bounded on bounded subsets of X^** if for every bounded subset $\tilde{X} \subset X^*$ there is a scalar b such that $J(x) \leq b$ for all $x \in \tilde{X}$. Let us also introduce the set

$$S = \{J \in \mathcal{E}^+(X) \mid J \text{ is bounded on bounded subsets of } X^*\}.$$

We assume that \mathcal{C} is nonempty, $J^* \in S$, and for every $x \in X^*$ and $\epsilon > 0$, there exists a policy π that is stable from x and satisfies $J_\pi(x) \leq J^*(x) + \epsilon$ (thus implying that $J_{\mathcal{C}}^* = J^*$). We have the following proposition.

Proposition 4.4.7: Under the preceding assumptions, J^* is the unique fixed point of T within S , and we have $T^k J \rightarrow J^*$ for all $J \in S$ with $J^* \leq J$. If in addition U is a metric space, and the sets $U_k(x, \lambda)$ of Eq. (4.17) are compact for all $x \in X$, $\lambda \in \mathfrak{R}$, and k , we have $T^k J \rightarrow J^*$ for all $J \in S$, and an optimal stationary policy is guaranteed to exist.

Proof: We have for all $J \in \mathcal{E}(X)$, $(\pi, x_0) \in \mathcal{C}$, and k ,

$$(T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x_0) = \alpha^k E_{x_0}^\pi \{J(x_k)\} + E_{x_0}^\pi \left\{ \sum_{m=0}^{k-1} \alpha^m g(x_m, \mu_m(x_m), w_m) \right\}.$$

Since $(\pi, x_0) \in \mathcal{C}$, there is a bounded subset of X^* such that $\{x_k\}$ belongs to that subset with probability 1, so if $J \in S$ it follows that $\alpha^k E_{x_0}^\pi \{J(x_k)\} \rightarrow 0$. Thus by taking limit as $k \rightarrow \infty$ in the preceding relation, we have for all $(\pi, x_0) \in \mathcal{C}$ and $J \in S$,

$$\begin{aligned} \lim_{k \rightarrow \infty} (T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x_0) &= \lim_{k \rightarrow \infty} E_{x_0}^\pi \left\{ \sum_{m=0}^{k-1} \alpha^m g(x_m, \mu_m(x_m), w_m) \right\} \\ &= J_\pi(x_0), \end{aligned}$$

so \mathcal{C} is S -regular. Since $J_{\mathcal{C}}^*$ is equal to J^* , which is a fixed point of T , the result follows similar to the proof of Prop. 4.4.5. **Q.E.D.**

4.4.4 Convergent Models

In this section we consider a case of an abstract DP model that generalizes both the monotone increasing and the monotone decreasing models. The model is patterned after the stochastic optimal control problem of Example 1.2.1, where the cost per stage function g can take negative as well as positive values. Our main assumptions are that the cost functions of all policies are defined as limits (rather than upper limits), and that $-\infty < \bar{J}(x) \leq J^*(x)$ for all $x \in X$.

These conditions are somewhat restrictive and make the model more similar to the monotone increasing than to the monotone decreasing model, but are essential for the results of this section (for a discussion of the pathological behaviors that can occur without the condition $\bar{J} \leq J^*$, see the paper by H. Yu [Yu15]). We will show that J^* is a fixed point of T , and that there exists an ϵ -optimal policy for every $\epsilon > 0$. This will bring to bear the regularity ideas and results of Prop. 4.4.2, and will provide a convergence result for the VI algorithm.

In particular, we denote

$$\mathcal{E}_b(X) = \{J \in \mathcal{E}(X) \mid J(x) > -\infty, \forall x \in X\},$$

and we will assume the following.

Assumption 4.4.1:

(a) For all $\pi = \{\mu_0, \mu_1, \dots\} \in \Pi$, J_π can be defined as a limit:

$$J_\pi(x) = \lim_{k \rightarrow \infty} (T_{\mu_0} \cdots T_{\mu_k} \bar{J})(x), \quad \forall x \in X. \quad (4.39)$$

Furthermore, we have $\bar{J} \in \mathcal{E}_b(X)$ and

$$\bar{J} \leq J^*.$$

(b) For each sequence $\{J_m\} \subset \mathcal{E}_b(X)$ with $J_m \rightarrow J \in \mathcal{E}_b(X)$, we have

$$\lim_{m \rightarrow \infty} H(x, u, J_m) = H(x, u, J), \quad \forall x \in X, u \in U(x).$$

(c) There exists $\alpha > 0$ such that for all $J \in \mathcal{E}_b(X)$ and $r \in \mathfrak{R}$,

$$H(x, u, J + re) \leq H(x, u, J) + \alpha r, \quad \forall x \in X, u \in U(x),$$

where e is the unit function, $e(x) \equiv 1$.

For an example of a type of problem where the convergence condition (4.39) is satisfied, consider the stochastic optimal control problem of Example 1.2.1, assuming that the state space consists of two regions: X_1 where the cost per stage is nonnegative under all controls, and X_2 where the cost per stage is nonpositive. Assuming that once the system enters X_1 it can never return to X_2 , the convergence condition (4.39) is satisfied for all π . The same is true for the reverse situation, where once the system enters X_2 it can never return to X_1 . Optimal stopping problems and SSP problems are often of this type.

We first prove the existence of ϵ -optimal policies and then use it to establish that J^* is a fixed point of T . The proofs are patterned after the ones under Assumption I (cf. Props. 4.3.2 and 4.3.3).

Proposition 4.4.8: Let Assumption 4.4.1 hold. Given any $\epsilon > 0$, there exists a policy $\pi_\epsilon \in \Pi$ such that

$$J^* \leq J_{\pi_\epsilon} \leq J^* + \epsilon e.$$

Proof: Let $\{\epsilon_k\}$ be a sequence such that $\epsilon_k > 0$ for all k and

$$\sum_{k=0}^{\infty} \alpha^k \epsilon_k = \epsilon, \tag{4.40}$$

where α is the scalar of Assumption 4.4.1(c). For each $x \in X$, consider a sequence of policies $\{\pi_k[x]\} \subset \Pi$, with components of $\pi_k[x]$ (to emphasize

their dependence on x) denoted by $\mu_m^k[x]$, $m = 0, 1, \dots$,

$$\pi_k[x] = \{\mu_0^k[x], \mu_1^k[x], \dots\},$$

such that for $k = 0, 1, \dots$,

$$J_{\pi_k[x]}(x) \leq J^*(x) + \epsilon_k. \quad (4.41)$$

Such a sequence exists since $J^* \in \mathcal{E}_b(X)$.

Consider the functions $\bar{\mu}_k$ defined by

$$\bar{\mu}_k(x) = \mu_0^kx, \quad \forall x \in X, \quad (4.42)$$

and the functions \bar{J}_k defined by

$$\bar{J}_k(x) = H\left(x, \bar{\mu}_k(x), \lim_{m \rightarrow \infty} T_{\mu_1^k[x]} \cdots T_{\mu_m^k[x]} \bar{J}\right), \quad \forall x \in X, \quad k = 0, 1, \dots \quad (4.43)$$

By using Eqs. (4.41)-(4.43), and the continuity property of Assumption 4.4.1(b), we obtain for all $x \in X$ and $k = 0, 1, \dots$,

$$\begin{aligned} \bar{J}_k(x) &= H\left(x, \mu_0^kx, \lim_{m \rightarrow \infty} T_{\mu_1^k[x]} \cdots T_{\mu_m^k[x]} \bar{J}\right) \\ &= \lim_{m \rightarrow \infty} H\left(x, \mu_0^kx, T_{\mu_1^k[x]} \cdots T_{\mu_m^k[x]} \bar{J}\right) \\ &= \lim_{m \rightarrow \infty} (T_{\mu_0^k[x]} \cdots T_{\mu_m^k[x]} \bar{J})(x) \\ &= J_{\pi_k[x]}(x) \\ &\leq J^*(x) + \epsilon_k. \end{aligned} \quad (4.44)$$

From Eqs. (4.43), (4.44), and Assumption 4.4.1(c), we have for all $x \in X$ and $k = 1, 2, \dots$,

$$\begin{aligned} (T_{\bar{\mu}_{k-1}} \bar{J}_k)(x) &= H(x, \bar{\mu}_{k-1}(x), \bar{J}_k) \\ &\leq H(x, \bar{\mu}_{k-1}(x), J^* + \epsilon_k e) \\ &\leq H(x, \bar{\mu}_{k-1}(x), J^*) + \alpha \epsilon_k \\ &\leq H\left(x, \bar{\mu}_{k-1}(x), \lim_{m \rightarrow \infty} T_{\mu_1^{k-1}[x]} \cdots T_{\mu_m^{k-1}[x]} \bar{J}\right) + \alpha \epsilon_k \\ &= \bar{J}_{k-1}(x) + \alpha \epsilon_k, \end{aligned}$$

and finally

$$T_{\bar{\mu}_{k-1}} \bar{J}_k \leq \bar{J}_{k-1} + \alpha \epsilon_k e, \quad k = 1, 2, \dots$$

Using this inequality and Assumption 4.4.1(c), we obtain

$$\begin{aligned} T_{\bar{\mu}_{k-2}} T_{\bar{\mu}_{k-1}} \bar{J}_k &\leq T_{\bar{\mu}_{k-2}} (\bar{J}_{k-1} + \alpha \epsilon_k e) \\ &\leq T_{\bar{\mu}_{k-2}} \bar{J}_{k-1} + \alpha^2 \epsilon_k e \\ &\leq \bar{J}_{k-2} + (\alpha \epsilon_{k-1} + \alpha^2 \epsilon_k) e. \end{aligned}$$

Continuing in the same manner, we have for $k = 1, 2, \dots$,

$$T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{k-1}} \bar{J}_k \leq \bar{J}_0 + (\alpha \epsilon_1 + \cdots + \alpha^k \epsilon_k) e \leq J^* + \left(\sum_{i=0}^k \alpha^i \epsilon_i \right) e.$$

Since by Assumption 4.4.1(c), we have $\bar{J} \leq J^* \leq \bar{J}_k$, it follows that

$$T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{k-1}} \bar{J} \leq J^* + \left(\sum_{i=0}^k \alpha^i \epsilon_i \right) e.$$

Denote $\pi_\epsilon = \{\bar{\mu}_0, \bar{\mu}_1, \dots\}$. Then by taking the limit in the preceding inequality and using Eq. (4.40), we obtain

$$J_{\pi_\epsilon} \leq J^* + \epsilon e.$$

Q.E.D.

By using Prop. 4.4.8 we can prove the following.

Proposition 4.4.9: Let Assumption 4.4.1 hold. Then J^* is a fixed point of T .

Proof: For every $\pi = \{\mu_0, \mu_1, \dots\} \in \Pi$ and $x \in X$, we have using the continuity property of Assumption 4.4.1(b) and the monotonicity of H ,

$$\begin{aligned} J_\pi(x) &= \lim_{k \rightarrow \infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} \bar{J})(x) \\ &= T_{\mu_0} \left(\lim_{k \rightarrow \infty} T_{\mu_1} \cdots T_{\mu_k} \bar{J} \right)(x) \\ &\geq (T_{\mu_0} J^*)(x) \\ &\geq (T J^*)(x). \end{aligned}$$

By taking the infimum of the left-hand side over $\pi \in \Pi$, we obtain

$$J^* \geq T J^*.$$

To prove the reverse inequality, let ϵ_1 and ϵ_2 be any positive scalars, and let $\bar{\pi} = \{\bar{\mu}_0, \bar{\mu}_1, \dots\}$ be such that

$$T_{\bar{\mu}_0} J^* \leq T J^* + \epsilon_1 e, \quad J_{\bar{\pi}_1} \leq J^* + \epsilon_2 e,$$

where $\pi_1 = \{\bar{\mu}_1, \bar{\mu}_2, \dots\}$ (such a policy exists by Prop. 4.4.8). By using the preceding relations and Assumption 4.4.1(c), we have

$$\begin{aligned}
J^* &\leq J_{\bar{\pi}} \\
&= \lim_{k \rightarrow \infty} T_{\bar{\mu}_0} T_{\bar{\mu}_1} \cdots T_{\bar{\mu}_k} \bar{J} \\
&= T_{\bar{\mu}_0} \left(\lim_{k \rightarrow \infty} T_{\bar{\mu}_1} \cdots T_{\bar{\mu}_k} \bar{J} \right) \\
&= T_{\bar{\mu}_0} J_{\pi_1} \\
&\leq T_{\bar{\mu}_0} (J^* + \epsilon_2 e) \\
&\leq T_{\bar{\mu}_0} J^* + \alpha \epsilon_2 e \\
&\leq T J^* + (\epsilon_1 + \alpha \epsilon_2) e.
\end{aligned}$$

Since ϵ_1 and ϵ_2 can be taken arbitrarily small, it follows that

$$J^* \leq T J^*.$$

Hence $J^* = T J^*$. **Q.E.D.**

It is known that J^* may not be a fixed point of T if the convergence condition (a) of Assumption 4.4.1 is violated (see the example of Section 3.1.2). Moreover, J^* may not be a fixed point of T if either part (b) or part (c) of Assumption 4.4.1 is violated, even when the monotone increase condition $\bar{J} \leq T \bar{J}$ [and hence also the convergence condition of part (a)] is satisfied (see Examples 4.3.1 and 4.3.2). By applying Prop. 4.4.2, we have the following proposition.

Proposition 4.4.10: Let Assumption 4.4.1 hold, let \mathcal{C} be a set of policy-state pairs such that $J_{\mathcal{C}}^* = J^*$, and let S be any subset of $\mathcal{E}(X)$ such that \mathcal{C} is S -regular. Then:

- (a) J^* is the only possible fixed point of T within the set $\{J \in S \mid J \geq J^*\}$.
- (b) We have $T^k J \rightarrow J^*$ for every $J \in \mathcal{E}(X)$ such that $J^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$.

Proof: By Prop. 4.4.9, J^* is a fixed point of T . The result follows from Prop. 4.4.2. **Q.E.D.**

4.5 STABLE POLICIES AND DETERMINISTIC OPTIMAL CONTROL

In this section, we will consider the use of the regularity ideas of the preceding section in conjunction with a particularly favorable class of monotone

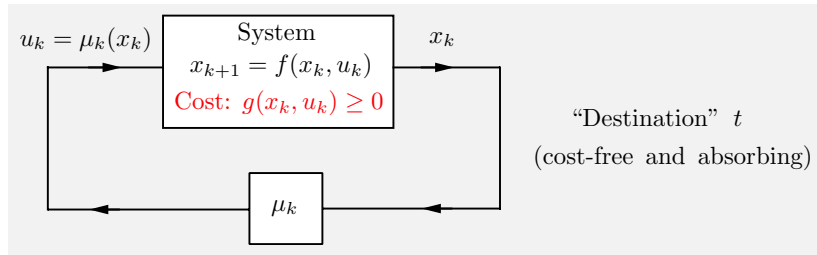


Figure 4.5.1 A deterministic optimal control problem with nonnegative cost per stage, and a cost-free and absorbing destination t .

increasing models. These are the discrete-time infinite horizon deterministic optimal control problems with nonnegative cost per stage, and a destination that is cost-free and absorbing. The classical linear-quadratic regulator problem of Section 3.5.4, as well as deterministic finite-state shortest path problems, are special cases. Except for the cost nonnegativity, our assumptions are very general, and allow the possibility that the optimal policy may not be stabilizing the system, e.g., may not reach the destination either asymptotically or in a finite number of steps. This situation is illustrated by the one-dimensional linear-quadratic example of Section 3.1.4, where we saw that the Riccati equation may have multiple nonnegative solutions, with the largest solution corresponding to the restricted optimal cost over just the stable policies.

Our approach is similar to the one of the preceding section. We use forcing functions and a perturbation line of analysis like the one of Section 3.4 to delineate collections \mathcal{C} of regular policy-state pairs such that the corresponding restricted optimal cost function $J_{\mathcal{C}}^*$ is a fixed point of T , as required by Prop. 4.4.2.

To this end, we introduce a new unifying notion of p -stability, which in addition to implying convergence of the generated states to the destination, quantifies the speed of convergence. Here is an outline of our analysis:

- (a) We consider the properties of several distinct cost functions: J^* , the overall optimal, and \hat{J}_p , the restricted optimal over just the p -stable policies. Different choices of p may yield different classes of p -stable policies, with different speeds of convergence.
- (b) We show that for any p and associated class of p -stable policies, \hat{J}_p is a solution of Bellman's equation, and we will characterize the smallest and the largest solutions: they are J^* , the optimal cost function, and \hat{J}^+ , the restricted optimal cost function over the class of (finitely) terminating policies.
- (c) We discuss modified versions of the VI and PI algorithms, as substitutes for the standard algorithms, which may not work in general.

Consider a deterministic discrete-time infinite horizon optimal control problem involving the system

$$x_{k+1} = f(x_k, u_k), \quad k = 0, 1, \dots, \quad (4.45)$$

where x_k and u_k are the state and control at stage k , which belong to sets X and U , referred to as the state and control spaces, respectively, and $f : X \times U \mapsto X$ is a given function. The control u_k must be chosen from a constraint set $U(x_k) \subset U$ that may depend on the current state x_k . The cost per stage $g(x, u)$ is assumed nonnegative and possibly extended real-valued:

$$0 \leq g(x, u) \leq \infty, \quad \forall x \in X, u \in U(x), k = 0, 1, \dots \quad (4.46)$$

We assume that X contains a special state, denoted t , which is referred to as the *destination*, and is cost-free and absorbing:

$$f(t, u) = t, \quad g(t, u) = 0, \quad \forall u \in U(t).$$

Except for the cost nonnegativity assumption (4.46), this problem is similar to the one of Section 3.5.5. It arises in many classical control applications involving regulation around a set point, and in finite-state and infinite-state versions of shortest path applications; see Fig. 4.5.1.

As earlier, we denote policies by π and stationary policies by μ . Given an initial state x_0 , a policy $\pi = \{\mu_0, \mu_1, \dots\}$ when applied to the system (4.45), generates a unique sequence of state-control pairs $(x_k, \mu_k(x_k))$, $k = 0, 1, \dots$. The cost of π starting from x_0 is

$$J_\pi(x_0) = \sum_{k=0}^{\infty} g(x_k, \mu_k(x_k)), \quad x_0 \in X,$$

[the series converges to some number in $[0, \infty]$ thanks to the nonnegativity assumption (4.46)]. The optimal cost function over the set of all policies Π is

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

We denote by $\mathcal{E}^+(X)$ the set of functions $J : X \mapsto [0, \infty]$. In our analysis, we will use the set of functions

$$\mathcal{J} = \{J \in \mathcal{E}^+(X) \mid J(t) = 0\}.$$

Since t is cost-free and absorbing, this set contains the cost function J_π of every $\pi \in \Pi$, as well as J^* .

Under the cost nonnegativity assumption (4.46), the problem can be cast as a special case of the monotone increasing model with

$$H(x, u, J) = g(x, u) + J(f(x, u)),$$

and the initial function \bar{J} being identically zero. Thus Prop. 4.4.4 applies and in particular J^* satisfies Bellman's equation:

$$J^*(x) = \inf_{u \in U(x)} \{g(x, u) + J^*(f(x, u))\}, \quad x \in X.$$

Moreover, an optimal stationary policy (if it exists) may be obtained through the minimization in the right side of this equation, cf. Prop. 4.4.4(c).

The VI method starts from some function $J_0 \in \mathcal{J}$, and generates a sequence of functions $\{J_k\} \subset \mathcal{J}$ according to

$$J_{k+1}(x) = \inf_{u \in U(x)} \{g(x, u) + J_k(f(x, u))\}, \quad x \in X, \quad k = 0, 1, \dots \quad (4.47)$$

From Prop. 4.4.6, we have that the VI sequence $\{J_k\}$ converges to J^* starting from any function $J_0 \in \mathcal{E}^+(X)$ that satisfies

$$J^* \leq J_0 \leq cJ^*,$$

for some scalar $c > 0$. We also have that VI converges to J^* starting from any J_0 with

$$0 \leq J_0 \leq J^*$$

under the compactness condition of Prop. 4.4.4(d). However, $\{J_k\}$ may not always converge to J^* because, among other reasons, Bellman's equation may have multiple solutions within \mathcal{J} .

The PI method starts from a stationary policy μ^0 , and generates a sequence of stationary policies $\{\mu^k\}$ via a sequence of policy evaluations to obtain J_{μ^k} from the equation

$$J_{\mu^k}(x) = g(x, \mu^k(x)) + J_{\mu^k}(f(x, \mu^k(x))), \quad x \in X, \quad (4.48)$$

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

$$\mu^{k+1}(x) \in \arg \min_{u \in U(x)} \{g(x, u) + J_{\mu^k}(f(x, u))\}, \quad x \in X. \quad (4.49)$$

Here, we implicitly assume that the minimum in Eq. (4.49) is attained for each $x \in X$, which is true under some compactness condition on either $U(x)$ or the level sets of the function $g(x, \cdot) + J_{\mu^k}(f(x, \cdot))$, or both. However, as noted in Section 4.3.3, PI may not produce a strict improvement of the cost function of a nonoptimal policy, a fact that was demonstrated with the simple deterministic shortest path example of Section 3.1.1.

The uniqueness of solution of Bellman's equation within \mathcal{J} , and the convergence of VI to J^* have been investigated as part of the analysis of Section 3.5.5. There we introduced conditions guaranteeing that J^* is the unique solution of Bellman's equation within a large set of functions

[the near-optimal termination Assumption 3.5.10, but not the cost nonnegativity assumption (4.46)]. Our approach here will make use of the cost nonnegativity but will address the problem under otherwise weaker conditions.

Our analytical approach will also be different than the approach of Section 3.5.5. Here, we will implicitly rely on the regularity ideas for nonstationary policies that we introduced in Section 4.4, and we will make a connection with traditional notions of feedback control system stability. Using nonstationary policies may be important in undiscounted optimal control problems with nonnegative cost per stage because it is not generally true that there exists a stationary ϵ -optimal policy [cf. the ϵ -optimality result of Prop. 4.4.4(e)].

4.5.1 Forcing Functions and p -Stable Policies

We will introduce a notion of stability that involves a function $p : X \mapsto [0, \infty)$ such that

$$p(t) = 0, \quad p(x) > 0, \quad \forall x \neq t.$$

As in Section 3.4, we refer to p as the *forcing function*, and we associate with it the p - δ -perturbed optimal control problem, where $\delta > 0$ is a given scalar. This is the same problem as the original, except that the cost per stage is changed to

$$g(x, u) + \delta p(x).$$

We denote by $J_{\pi, p, \delta}$ the cost function of a policy $\pi \in \Pi$ in the p - δ -perturbed problem:

$$J_{\pi, p, \delta}(x_0) = J_{\pi}(x_0) + \delta \sum_{k=0}^{\infty} p(x_k), \quad (4.50)$$

where $\{x_k\}$ is the sequence generated starting from x_0 and using π . We also denote by $\hat{J}_{p, \delta}$, the corresponding optimal cost function,

$$\hat{J}_{p, \delta}(x) = \inf_{\pi \in \Pi} J_{\pi, p, \delta}(x), \quad x \in X.$$

Definition 4.5.1: Let p be a given forcing function. For a state $x_0 \in X$, we say that a policy π is *p -stable from x_0* if for the sequence $\{x_k\}$ generated starting from x_0 and using π we have

$$J_{\pi}(x_0) < \infty \quad \text{and} \quad \sum_{k=0}^{\infty} p(x_k) < \infty, \quad (4.51)$$

or equivalently [using Eq. (4.50)]

$$J_{\pi,p,\delta}(x_0) < \infty, \quad \forall \delta > 0.$$

The set of all policies that are p -stable from x_0 is denoted by Π_{p,x_0} . We define the *restricted optimal cost function* \hat{J}_p by

$$\hat{J}_p(x) = \inf_{\pi \in \Pi_{p,x}} J_{\pi}(x), \quad x \in X, \quad (4.52)$$

(with the convention that the infimum over the empty set is ∞). We say that π is p -stable (without qualification) if $\pi \in \Pi_{p,x}$ for all $x \in X$ such that $\Pi_{p,x} \neq \emptyset$. The set of all p -stable policies is denoted by Π_p .

Note that since Eq. (4.51) does not depend on δ , we see that an equivalent definition of a policy π that is p -stable from x_0 is that $J_{\pi,p,\delta}(x_0) < \infty$ for *some* $\delta > 0$ (rather than all $\delta > 0$). Thus the set $\Pi_{p,x}$ of p -stable policies from x depends on p and x but not on δ . Let us make some observations:

- (a) *Rate of convergence to t using p -stable policies:* The relation (4.51) shows that the forcing function p quantifies the rate at which the destination is approached using the p -stable policies. As an example, let $X = \mathfrak{R}^n$ and

$$p(x) = \|x\|^\rho,$$

where $\rho > 0$ is a scalar. Then the policies $\pi \in \Pi_{p,x_0}$ are the ones that force x_k towards 0 at a rate faster than $O(1/k^\rho)$, so slower policies are excluded from Π_{p,x_0} .

- (b) *Approximation property of $J_{\pi,p,\delta}(x)$:* Consider a pair (π, x_0) with $\pi \in \Pi_{p,x_0}$. By taking the limit as $\delta \downarrow 0$ in the expression

$$J_{\pi,p,\delta}(x_0) = J_{\pi}(x_0) + \delta \sum_{k=0}^{\infty} p(x_k),$$

[cf. Eq. (4.50)] and by using Eq. (4.51), it follows that

$$\lim_{\delta \downarrow 0} J_{\pi,p,\delta}(x_0) = J_{\pi}(x_0), \quad \forall \text{ pairs } (\pi, x_0) \text{ with } \pi \in \Pi_{p,x_0}. \quad (4.53)$$

From this equation, we have that if $\pi \in \Pi_{p,x}$, then $J_{\pi,p,\delta}(x)$ is finite and differs from $J_{\pi}(x)$ by $O(\delta)$. By contrast, if $\pi \notin \Pi_{p,x}$, then $J_{\pi,p,\delta}(x) = \infty$ by the definition of p -stability, even though we may have $J_{\pi}(x) < \infty$.

- (c) *Limiting property of $\hat{J}_p(x_k)$:* Consider a pair (π, x_0) with $\pi \in \Pi_{p,x_0}$. By breaking down $J_{\pi,p,\delta}(x_0)$ into the sum of the costs of the first k stages and the remaining stages, we have for all $\delta > 0$ and $k > 0$,

$$J_{\pi,p,\delta}(x_0) = \sum_{m=0}^{k-1} g(x_m, \mu_m(x_m)) + \delta \sum_{m=0}^{k-1} p(x_m) + J_{\pi_k,p,\delta}(x_k),$$

where $\{x_k\}$ is the sequence generated starting from x_0 and using π , and π_k is the policy $\{\mu_k, \mu_{k+1}, \dots\}$. By taking the limit as $k \rightarrow \infty$ and using Eq. (4.50), it follows that

$$\lim_{k \rightarrow \infty} J_{\pi_k, p, \delta}(x_k) = 0, \quad \forall \text{ pairs } (\pi, x_0) \text{ with } \pi \in \Pi_{p, x_0}, \delta > 0.$$

Also, since $\hat{J}_p(x_k) \leq \hat{J}_{p, \delta}(x_k) \leq J_{\pi_k, p, \delta}(x_k)$, it follows that

$$\lim_{k \rightarrow \infty} J_{p, \delta}(x_k) = 0, \quad \forall (\pi, x_0) \text{ with } x_0 \in X \text{ and } \pi \in \Pi_{p, x_0}, \delta > 0, \quad (4.54)$$

$$\lim_{k \rightarrow \infty} \hat{J}_p(x_k) = 0, \quad \forall (\pi, x_0) \text{ with } x_0 \in X \text{ and } \pi \in \Pi_{p, x_0}. \quad (4.55)$$

Terminating Policies and Controllability

An important special case is when p is equal to the function

$$p^+(x) = \begin{cases} 0 & \text{if } x = t, \\ 1 & \text{if } x \neq t. \end{cases} \quad (4.56)$$

For $p = p^+$, a policy π is p^+ -stable from x if and only if it is *terminating from* x , i.e., reaches t in a finite number of steps starting from x [cf. Eq. (4.51)]. The set of terminating policies from x is denoted by Π_x^+ and it is contained within every other set of p -stable policies $\Pi_{p, x}$, as can be seen from Eq. (4.51). As a result, the restricted optimal cost function over Π_x^+ ,

$$\hat{J}^+(x) = \inf_{\pi \in \Pi_x^+} J_\pi(x), \quad x \in X,$$

satisfies $J^*(x) \leq \hat{J}_p(x) \leq \hat{J}^+(x)$ for all $x \in X$. A policy π is said to be *terminating* if it is simultaneously terminating from all $x \in X$ such that $\Pi_x^+ \neq \emptyset$. The set of all terminating policies is denoted by Π^+ .

Note that if the state space X is finite, we have for every forcing function p

$$\underline{\beta} p^+(x) \leq p(x) \leq \bar{\beta} p^+(x), \quad \forall x \in X,$$

for some scalars $\underline{\beta}, \bar{\beta} > 0$. As a result it can be seen that $\Pi_{p, x} = \Pi_x^+$ and $\hat{J}_p = \hat{J}^+$, so in effect the case where $p = p^+$ is the only case of interest for finite-state problems.

The notion of a terminating policy is related to the notion of *controllability*. In classical control theory terms, the system $x_{k+1} = f(x_k, u_k)$ is said to be completely controllable if for every $x_0 \in X$, there exists a policy that drives the state x_k to the destination in a finite number of steps. This notion of controllability is equivalent to the existence of a terminating policy from each $x \in X$.

One of our main results, to be shown shortly, is that J^* , \hat{J}_p , and \hat{J}^+ are solutions of Bellman's equation, with J^* being the “smallest” solution and \hat{J}^+ being the “largest” solution within \mathcal{J} . The most favorable situation arises when $J^* = \hat{J}^+$, in which case J^* is the unique solution of Bellman's equation within \mathcal{J} . Moreover, in this case it will be shown that the VI algorithm converges to J^* starting with any $J_0 \in \mathcal{J}$ with $J_0 \geq J^*$, and the PI algorithm converges to J^* as well. Once we prove the fixed point property of \hat{J}_p , we will be able to bring to bear the regularity ideas of the preceding section (cf. Prop. 4.4.2).

4.5.2 Restricted Optimization over Stable Policies

For a given forcing function p , we denote by \hat{X}_p the effective domain of \hat{J}_p , i.e., the set of all x where \hat{J}_p is finite,

$$\hat{X}_p = \{x \in X \mid \hat{J}_p(x) < \infty\}.$$

Since $\hat{J}_p(x) < \infty$ if and only if $\Pi_{p,x} \neq \emptyset$ [cf. Eqs. (4.51)–(4.52)], or equivalently $J_{\pi,p,\delta}(x) < \infty$ for some π and all $\delta > 0$, it follows that \hat{X}_p is also the effective domain of $\hat{J}_{p,\delta}$,

$$\hat{X}_p = \{x \in X \mid \Pi_{p,x} \neq \emptyset\} = \{x \in X \mid \hat{J}_{p,\delta}(x) < \infty\}, \quad \forall \delta > 0.$$

Note that \hat{X}_p may depend on p and may be a strict subset of the effective domain of J^* , which is denoted by

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

(cf. Section 3.5.5). The reason is that there may exist a policy π such that $J_\pi(x) < \infty$, even when there is no p -stable policy from x (for example, no terminating policy from x).

Our first objective is to show that as $\delta \downarrow 0$, the p - δ -perturbed optimal cost function $\hat{J}_{p,\delta}$ converges to the restricted optimal cost function \hat{J}_p .

Proposition 4.5.1 (Approximation Property of $\hat{J}_{p,\delta}$): Let p be a given forcing function and $\delta > 0$.

(a) We have

$$J_{\pi,p,\delta}(x) = J_\pi(x) + w_{\pi,p,\delta}(x), \quad \forall x \in X, \pi \in \Pi_{p,x}, \quad (4.57)$$

where $w_{\pi,p,\delta}$ is a function such that $\lim_{\delta \downarrow 0} w_{\pi,p,\delta}(x) = 0$ for all $x \in X$.

(b) We have

$$\lim_{\delta \downarrow 0} \hat{J}_{p,\delta}(x) = \hat{J}_p(x), \quad \forall x \in X.$$

Proof: (a) Follows by using Eq. (4.53) for $x \in \widehat{X}_p$, and by taking $w_{p,\delta}(x) = 0$ for $x \notin \widehat{X}_p$.

(b) By Prop. 4.4.4(e), there exists an ϵ -optimal policy π_ϵ for the p - δ -perturbed problem, i.e., $J_{\pi_\epsilon,p,\delta}(x) \leq \hat{J}_{p,\delta}(x) + \epsilon$ for all $x \in X$. Moreover, for $x \in \widehat{X}_p$ we have $\hat{J}_{p,\delta}(x) < \infty$, so $J_{\pi_\epsilon,p,\delta}(x) < \infty$. Hence π_ϵ is p -stable from all $x \in \widehat{X}_p$, and we have $\hat{J}_p \leq J_{\pi_\epsilon}$. Using also Eq. (4.57), we have for all $\delta > 0$, $\epsilon > 0$, $x \in X$, and $\pi \in \Pi_{p,x}$,

$$\hat{J}_p(x) - \epsilon \leq J_{\pi_\epsilon}(x) - \epsilon \leq J_{\pi_\epsilon,p,\delta}(x) - \epsilon \leq \hat{J}_{p,\delta}(x) \leq J_{\pi,p,\delta}(x) = J_\pi(x) + w_{\pi,p,\delta}(x),$$

where $\lim_{\delta \downarrow 0} w_{\pi,p,\delta}(x) = 0$ for all $x \in X$. By taking the limit as $\epsilon \downarrow 0$, we obtain for all $\delta > 0$ and $\pi \in \Pi_{p,x}$,

$$\hat{J}_p(x) \leq \hat{J}_{p,\delta}(x) \leq J_\pi(x) + w_{\pi,p,\delta}(x), \quad \forall x \in X.$$

By taking the limit as $\delta \downarrow 0$ and then the infimum over all $\pi \in \Pi_{p,x}$, we have

$$\hat{J}_p(x) \leq \lim_{\delta \downarrow 0} \hat{J}_{p,\delta}(x) \leq \inf_{\pi \in \Pi_{p,x}} J_\pi(x) = \hat{J}_p(x), \quad \forall x \in X,$$

from which the result follows. **Q.E.D.**

We now consider ϵ -optimal policies, setting the stage for our main proof argument. We know that given any $\epsilon > 0$, by Prop. 4.4.4(e), there exists an ϵ -optimal policy for the p - δ -perturbed problem, i.e., a policy π such that $J_\pi(x) \leq J_{\pi,p,\delta}(x) \leq \hat{J}_{p,\delta}(x) + \epsilon$ for all $x \in X$. We address the question whether there exists a p -stable policy π that is ϵ -optimal for the restricted optimization over p -stable policies, i.e., a policy π that is p -stable simultaneously from all $x \in X_p$, (i.e., $\pi \in \Pi_p$) and satisfies

$$J_\pi(x) \leq \hat{J}_p(x) + \epsilon, \quad \forall x \in X.$$

We refer to such a policy as a p - ϵ -optimal policy.

Proposition 4.5.2 (Existence of p - ϵ -Optimal Policy): Let p be a given forcing function and $\delta > 0$. For every $\epsilon > 0$, a policy π that is ϵ -optimal for the p - δ -perturbed problem is p - ϵ -optimal, and hence belongs to Π_p .

Proof: For any ϵ -optimal policy π_ϵ for the p - δ -perturbed problem, we have

$$J_{\pi_\epsilon,p,\delta}(x) \leq \hat{J}_{p,\delta}(x) + \epsilon < \infty, \quad \forall x \in \widehat{X}_p.$$

This implies that $\pi_\epsilon \in \Pi_p$. Moreover, for all sequences $\{x_k\}$ generated from initial state-policy pairs (π, x_0) with $x_0 \in \widehat{X}_p$ and $\pi \in \Pi_{p, x_0}$, we have

$$J_{\pi_\epsilon}(x_0) \leq J_{\pi_\epsilon, p, \delta}(x_0) \leq \hat{J}_{p, \delta}(x_0) + \epsilon \leq J_\pi(x_0) + \delta \sum_{k=0}^{\infty} p(x_k) + \epsilon.$$

Taking the limit as $\delta \downarrow 0$ and using the fact $\sum_{k=0}^{\infty} p(x_k) < \infty$ (since $\pi \in \Pi_{p, x_0}$), we obtain

$$J_{\pi_\epsilon}(x_0) \leq J_\pi(x_0) + \epsilon, \quad \forall x_0 \in \widehat{X}_p, \pi \in \Pi_{p, x_0}.$$

By taking infimum over $\pi \in \Pi_{p, x_0}$, it follows that

$$J_{\pi_\epsilon}(x_0) \leq \hat{J}_p(x_0) + \epsilon, \quad \forall x_0 \in \widehat{X}_p,$$

which in view of the fact $J_{\pi_\epsilon}(x_0) = \hat{J}_p(x_0) = \infty$ for $x_0 \notin \widehat{X}_p$, implies that π_ϵ is p - ϵ -optimal. **Q.E.D.**

Note that the preceding proposition implies that

$$\hat{J}_p(x) = \inf_{\pi \in \Pi_p} J_\pi(x), \quad \forall x \in X, \quad (4.58)$$

which is a stronger statement than the definition $\hat{J}_p(x) = \inf_{\pi \in \Pi_{p, x}} J_\pi(x)$ for all $x \in X$. However, it can be shown through examples that there may not exist a restricted-optimal p -stable policy, i.e., a $\pi \in \Pi_p$ such that $J_\pi = \hat{J}_p$, even if there exists an optimal policy for the original problem. One such example is the one-dimensional linear-quadratic problem of Section 3.1.4 for the case where $p = p^+$. Then, there exists a unique linear stable policy that attains the restricted optimal cost $\hat{J}^+(x)$ for all x , but this policy is not terminating. Note also that there may not exist a *stationary* p - ϵ -optimal policy, since generally in undiscounted nonnegative cost optimal control problems there may not exist a stationary ϵ -optimal policy (an example is given following Prop. 4.4.8).

We now take the first steps for bringing regularity ideas into the analysis. We introduce the set of functions S_p given by

$$S_p = \left\{ J \in \mathcal{J} \mid J(x_k) \rightarrow 0 \text{ for all sequences } \{x_k\} \text{ generated from initial state-policy pairs } (\pi, x_0) \text{ with } x_0 \in X \text{ and } \pi \in \Pi_{p, x_0} \right\}. \quad (4.59)$$

In words, S_p consists of the functions in \mathcal{J} whose value is asymptotically driven to 0 by all the policies that are p -stable starting from some $x_0 \in X$. Similar to the analysis of Section 4.4.2, we can prove that the collection

$\mathcal{C}_p = \{(\pi, x_0) \mid \pi \in \Pi_{p,x_0}\}$ is S_p -regular. Moreover, S_p is the largest set S for which \mathcal{C}_p is S -regular.

Note that S_p contains \hat{J}_p and $\hat{J}_{p,\delta}$ for all $\delta > 0$ [cf. Eq. (4.54), (4.55)]. Moreover, S_p contains all functions J such that

$$0 \leq J \leq c\hat{J}_{p,\delta}$$

for some $c > 0$ and $\delta > 0$.

We summarize the preceding discussion in the following proposition, which also shows that $\hat{J}_{p,\delta}$ is the unique solution (within S_p) of Bellman's equation for the p - δ -perturbed problem. This will be needed to prove that \hat{J}_p solves the Bellman equation of the unperturbed problem, but also shows that the p - δ -perturbed problem can be solved more reliably than the original problem (including by VI methods), and yields a close approximation to \hat{J}_p [cf. Prop. 4.5.1(b)].

Proposition 4.5.3: Let p be a forcing function and $\delta > 0$. The function $\hat{J}_{p,\delta}$ belongs to the set S_p , and is the unique solution within S_p of Bellman's equation for the p - δ -perturbed problem,

$$\hat{J}_{p,\delta}(x) = \inf_{u \in U(x)} \left\{ g(x, u) + \delta p(x) + \hat{J}_{p,\delta}(f(x, u)) \right\}, \quad x \in X. \quad (4.60)$$

Moreover, S_p contains \hat{J}_p and all functions J satisfying

$$0 \leq J \leq c\hat{J}_{p,\delta}$$

for some scalar $c > 0$.

Proof: We have $\hat{J}_{p,\delta} \in S_p$ and $\hat{J}_p \in S_p$ by Eq. (4.54), as noted earlier. We also have that $\hat{J}_{p,\delta}$ is a solution of Bellman's equation (4.60) by Prop. 4.4.4(a). To show that $\hat{J}_{p,\delta}$ is the unique solution within S_p , let $\tilde{J} \in S_p$ be another solution, so that using also Prop. 4.4.4(a), we have

$$\hat{J}_{p,\delta}(x) \leq \tilde{J}(x) \leq g(x, u) + \delta p(x) + \tilde{J}(f(x, u)), \quad \forall x \in X, u \in U(x). \quad (4.61)$$

Fix $\epsilon > 0$, and let $\pi = \{\mu_0, \mu_1, \dots\}$ be an ϵ -optimal policy for the p - δ -perturbed problem. By repeatedly applying the preceding relation, we have for any $x_0 \in \hat{X}_p$,

$$\hat{J}_{p,\delta}(x_0) \leq \tilde{J}(x_0) \leq \tilde{J}(x_k) + \delta \sum_{m=0}^{k-1} p(x_m) + \sum_{m=0}^{k-1} g(x_m, \mu_m(x_m)), \quad \forall k \geq 1, \quad (4.62)$$

where $\{x_k\}$ is the state sequence generated starting from x_0 and using π . We have $\tilde{J}(x_k) \rightarrow 0$ (since $\tilde{J} \in S_p$ and $\pi \in \Pi_p$ by Prop. 4.5.2), so that

$$\lim_{k \rightarrow \infty} \left\{ \tilde{J}(x_k) + \delta \sum_{m=0}^{k-1} p(x_m) + \sum_{m=0}^{k-1} g(x_m, \mu_m(x_m)) \right\} = J_{\pi, \delta}(x_0) \leq \hat{J}_{p, \delta}(x_0) + \epsilon. \quad (4.63)$$

By combining Eqs. (4.62) and (4.63), we obtain

$$\hat{J}_{p, \delta}(x_0) \leq \tilde{J}(x_0) \leq \hat{J}_{p, \delta}(x_0) + \epsilon, \quad \forall x_0 \in \hat{X}_p.$$

By letting $\epsilon \rightarrow 0$, it follows that $\hat{J}_{p, \delta}(x_0) = \tilde{J}(x_0)$ for all $x_0 \in \hat{X}_p$. Also for $x_0 \notin \hat{X}_p$, we have $\hat{J}_{p, \delta}(x_0) = \tilde{J}(x_0) = \infty$ [since $\hat{J}_{p, \delta}(x_0) = \infty$ for $x_0 \notin \hat{X}_p$ and $\hat{J}_{p, \delta} \leq \tilde{J}$, cf. Eq. (4.61)]. Thus $\hat{J}_{p, \delta} = \tilde{J}$, proving that $\hat{J}_{p, \delta}$ is the unique solution of the Bellman Eq. (4.60) within S_p . **Q.E.D.**

We next show our main result in this section, namely that \hat{J}_p is the unique solution of Bellman's equation within the set of functions

$$\mathcal{W}_p = \{J \in S_p \mid \hat{J}_p \leq J\}. \quad (4.64)$$

Moreover, we show that the VI algorithm yields \hat{J}_p in the limit for any initial $J_0 \in \mathcal{W}_p$. This result is intimately connected with the regularity ideas of Section 4.4. The idea is that the collection $\mathcal{C}_p = \{(\pi, x_0) \mid \pi \in \Pi_{p, x_0}\}$ is S_p -regular, as noted earlier. In view of this and the fact that $J_{\mathcal{C}_p}^* = \hat{J}_p$, the result will follow from Prop. 4.4.2 once \hat{J}_p is shown to be a solution of Bellman's equation. This latter property is shown essentially by taking the limit as $\delta \downarrow 0$ in Eq. (4.60).

Proposition 4.5.4: Let p be a given forcing function. Then:

- (a) \hat{J}_p is the unique solution of Bellman's equation

$$J(x) = \inf_{u \in U(x)} \left\{ g(x, u) + J(f(x, u)) \right\}, \quad x \in X, \quad (4.65)$$

within the set \mathcal{W}_p of Eq. (4.64).

- (b) (*VI Convergence*) If $\{J_k\}$ is the sequence generated by the VI algorithm (4.47) starting with some $J_0 \in \mathcal{W}_p$, then $J_k \rightarrow \hat{J}_p$.
 (c) (*Optimality Condition*) If $\hat{\mu}$ is a p -stable stationary policy and

$$\hat{\mu}(x) \in \arg \min_{u \in U(x)} \left\{ g(x, u) + \hat{J}_p(f(x, u)) \right\}, \quad \forall x \in X, \quad (4.66)$$

then $\hat{\mu}$ is optimal over the set of p -stable policies. Conversely, if $\hat{\mu}$ is optimal within the set of p -stable policies, then it satisfies the preceding condition (4.66).

Proof: (a), (b) We first show that \hat{J}_p is a solution of Bellman's equation. Since $\hat{J}_{p,\delta}$ is a solution of Bellman's equation for the p - δ -perturbed problem (cf. Prop. 4.5.3) and $\hat{J}_{p,\delta} \geq \hat{J}_p$ [cf. Prop. 4.5.1(b)], we have for all $\delta > 0$,

$$\begin{aligned} \hat{J}_{p,\delta}(x) &= \inf_{u \in U(x)} \left\{ g(x, u) + \delta p(x) + \hat{J}_{p,\delta}(f(x, u)) \right\} \\ &\geq \inf_{u \in U(x)} \left\{ g(x, u) + \hat{J}_{p,\delta}(f(x, u)) \right\} \\ &\geq \inf_{u \in U(x)} \left\{ g(x, u) + \hat{J}_p(f(x, u)) \right\}. \end{aligned}$$

By taking the limit as $\delta \downarrow 0$ and using the fact $\lim_{\delta \downarrow 0} \hat{J}_{p,\delta} = \hat{J}_p$ [cf. Prop. 4.5.1(b)], we obtain

$$\hat{J}_p(x) \geq \inf_{u \in U(x)} \left\{ g(x, u) + \hat{J}_p(f(x, u)) \right\}, \quad \forall x \in X. \quad (4.67)$$

For the reverse inequality, let $\{\delta_m\}$ be a sequence with $\delta_m \downarrow 0$. From Prop. 4.5.3, we have for all m , $x \in X$, and $u \in U(x)$,

$$\begin{aligned} g(x, u) + \delta_m p(x) + \hat{J}_{p,\delta_m}(f(x, u)) &\geq \inf_{v \in U(x)} \left\{ g(x, v) + \delta_m p(x) \right. \\ &\quad \left. + \hat{J}_{p,\delta_m}(f(x, v)) \right\} \\ &= \hat{J}_{p,\delta_m}(x). \end{aligned}$$

Taking the limit as $m \rightarrow \infty$, and using the fact $\lim_{\delta_m \downarrow 0} \hat{J}_{p,\delta_m} = \hat{J}_p$ [cf. Prop. 4.5.1(b)], we have

$$g(x, u) + \hat{J}_p(f(x, u)) \geq \hat{J}_p(x), \quad \forall x \in X, u \in U(x),$$

so that

$$\inf_{u \in U(x)} \left\{ g(x, u) + \hat{J}_p(f(x, u)) \right\} \geq \hat{J}_p(x), \quad \forall x \in X. \quad (4.68)$$

By combining Eqs. (4.67) and (4.68), we see that \hat{J}_p is a solution of Bellman's equation. We also have $\hat{J}_p \in S_p$ by Prop. 4.5.3, implying that $\hat{J}_p \in \mathcal{W}_p$ and proving part (a) except for the uniqueness assertion. Part

(b) and the uniqueness part of part (a) follow from Prop. 4.4.2; see the discussion preceding the proposition.

(c) If μ is p -stable and Eq. (4.66) holds, then

$$\hat{J}_p(x) = g(x, \mu(x)) + \hat{J}_p(f(x, \mu(x))), \quad x \in X.$$

By Prop. 4.4.4(b), this implies that $J_\mu \leq \hat{J}_p$, so μ is optimal over the set of p -stable policies. Conversely, assume that μ is p -stable and $J_\mu = \hat{J}_p$. Then by Prop. 4.4.4(b), we have

$$\hat{J}_p(x) = g(x, \mu(x)) + \hat{J}_p(f(x, \mu(x))), \quad x \in X,$$

and since [by part (a)] \hat{J}_p is a solution of Bellman's equation,

$$\hat{J}_p(x) = \inf_{u \in U(x)} \{g(x, u) + \hat{J}_p(f(x, u))\}, \quad x \in X.$$

Combining the last two relations, we obtain Eq. (4.66). **Q.E.D.**

As a supplement to the preceding proposition, we note the specialization of Prop. 4.4.5 that relates to the optimal cost function J^* .

Proposition 4.5.5: Let S^* be the set

$$S^* = \left\{ J \in \mathcal{J} \mid J(x_k) \rightarrow 0 \text{ for all sequences } \{x_k\} \text{ generated from} \right. \\ \left. \text{initial state-policy pairs } (\pi, x_0) \text{ with } J_\pi(x_0) < \infty \right\},$$

and \mathcal{W}^* be the set

$$\mathcal{W}^* = \{J \in S^* \mid J^* \leq J\}.$$

Then J^* belongs to S^* and is the unique solution of Bellman's equation within S^* . Moreover, we have $T^k J \rightarrow J^*$ for all $J \in \mathcal{W}^*$.

Proof: Follows from Prop. 4.4.5 in the deterministic special case where w_k takes a single value. **Q.E.D.**

We now consider the special case where p is equal to the function $p^+(x) = 1$ for all $x \neq t$ [cf. Eq. (4.56)]. Then the set of p^+ -stable policies from x is Π_x^+ , the set of terminating policies from x , and the corresponding restricted optimal cost is $\hat{J}^+(x)$:

$$\hat{J}^+(x) = \hat{J}_{p^+}(x) = \inf_{\pi \in \Pi_x^+} J_\pi(x) = \inf_{\pi \in \Pi^+} J_\pi(x), \quad x \in X,$$

[the last equality follows from Eq. (4.58)]. In this case, the set S_{p^+} of Eq. (4.59) is the entire set \mathcal{J} ,

$$S_{p^+} = \mathcal{J},$$

since for all $J \in \mathcal{J}$ and all sequences $\{x_k\}$ generated from initial state-policy pairs (π, x_0) with $x_0 \in X$ and π terminating from x_0 , we have $J(x_k) = 0$ for k sufficiently large. Thus, the corresponding set of Eq. (4.64) is

$$\mathcal{W}^+ = \{J \in \mathcal{J} \mid \hat{J}^+ \leq J\}.$$

By specializing to the case $p = p^+$ the result of Prop. 4.5.4, we obtain the following proposition, which makes a stronger assertion than Prop. 4.5.4(a), namely that \hat{J}^+ is the largest solution of Bellman's equation within \mathcal{J} (rather than the smallest solution within \mathcal{W}^+).

Proposition 4.5.6:

- (a) \hat{J}^+ is the largest solution of the Bellman equation (4.65) within \mathcal{J} , i.e., \hat{J}^+ is a solution and if $J' \in \mathcal{J}$ is another solution, then $J' \leq \hat{J}^+$.
- (b) (*VI Convergence*) If $\{J_k\}$ is the sequence generated by the VI algorithm (4.47) starting with some $J_0 \in \mathcal{J}$ with $J_0 \geq \hat{J}^+$, then $J_k \rightarrow \hat{J}^+$.
- (c) (*Optimality Condition*) If μ^+ is a terminating stationary policy and

$$\mu^+(x) \in \arg \min_{u \in U(x)} \{g(x, u) + \hat{J}^+(f(x, u))\}, \quad \forall x \in X, \quad (4.69)$$

then μ^+ is optimal over the set of terminating policies. Conversely, if μ^+ is optimal within the set of terminating policies, then it satisfies the preceding condition (4.69).

Proof: In view of Prop. 4.5.4, we only need to show that \hat{J}^+ is the largest solution of the Bellman equation. From Prop. 4.5.4(a), \hat{J}^+ is a solution that belongs to \mathcal{J} . If $J' \in \mathcal{J}$ is another solution, we have $J' \leq \tilde{J}$ for some $\tilde{J} \in \mathcal{W}^+$, so $J' = T^k J' \leq T^k \tilde{J}$ for all k . Since $T^k \tilde{J} \rightarrow \hat{J}^+$, it follows that $J' \leq \hat{J}^+$. **Q.E.D.**

We illustrate Props. 4.5.4 and 4.5.6 in Fig. 4.5.2. In particular, each forcing function p delineates the set of initial functions \mathcal{W}_p from which VI converges to \hat{J}_p . The function \hat{J}_p is the minimal element of \mathcal{W}_p . Moreover, we have $\mathcal{W}_p \cap \mathcal{W}_{p'} = \emptyset$ if $\hat{J}_p \neq \hat{J}_{p'}$, in view of the VI convergence result of Prop. 4.5.4(b).

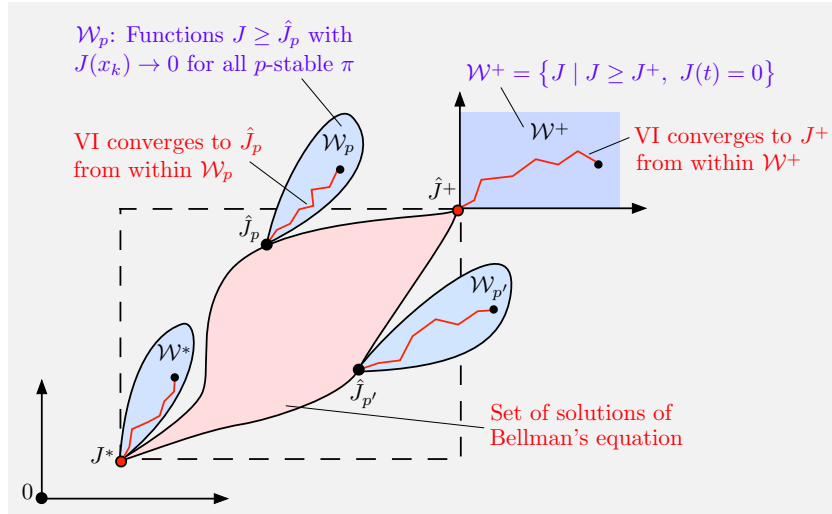


Figure 4.5.2 Schematic two-dimensional illustration of the results of Prop. 4.5.4 and 4.5.6. The functions J^* , \hat{J}^+ , and \hat{J}_p are all solutions of Bellman's equation. Moreover J^* and \hat{J}^+ are the smallest and largest solutions, respectively. Each p defines the set of initial functions \mathcal{W}_p from which VI converges to \hat{J}_p from above. For two forcing functions p and p' , we have $\mathcal{W}_p \cap \mathcal{W}_{p'} = \emptyset$ if $\hat{J}_p \neq \hat{J}_{p'}$. Moreover, \mathcal{W}_p contains no solutions of Bellman's equation other than \hat{J}_p . It is also possible that \mathcal{W}_p consists of just \hat{J}_p .

Note a significant fact: Proposition 4.5.6(b) implies that VI converges to \hat{J}^+ starting from the readily available initial condition

$$J_0(x) = \begin{cases} 0 & \text{if } x = t, \\ \infty & \text{if } x \neq t. \end{cases}$$

For this choice of J_0 , the value $J_k(x)$ generated by VI is the optimal cost that can be achieved starting from x subject to the constraint that t is reached in k steps or less. As we have noted earlier, in shortest-path type problems VI tends to converge faster when started from above.

Consider now the favorable case where terminating policies are sufficient, in the sense that $\hat{J}^+ = J^*$; cf. Fig. 4.5.3. Then, from Prop. 4.5.6, it follows that J^* is the unique solution of Bellman's equation within \mathcal{J} , and the VI algorithm converges to J^* from above, i.e., starting from any $J_0 \in \mathcal{J}$ with $J_0 \geq J^*$. Under additional conditions, such as finiteness of $U(x)$ for all $x \in X$ [cf. Prop. 4.4.4(d)], VI converges to J^* starting from any $J_0 \in \mathcal{E}^+(X)$ with $J_0(t) = 0$. These results are consistent with our analysis of Section 3.5.5.

Examples of problems where terminating policies are sufficient include linear-quadratic problems under the classical conditions of controllability and observability, and finite-node deterministic shortest path prob-

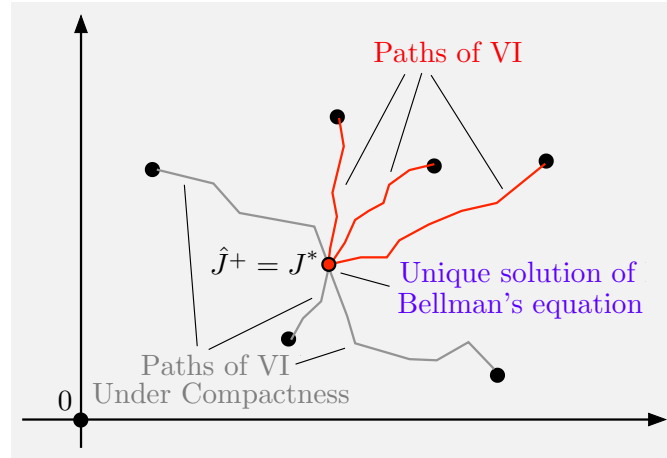


Figure 4.5.3 Schematic two-dimensional illustration of the favorable case where $\hat{J}^+ = J^*$. Then J^* is the unique solution of Bellman's equation within \mathcal{J} , and the VI algorithm converges to J^* from above [and also starting from any $J_0 \geq 0$ under a suitable compactness condition; cf. Prop. 4.4.4(d)].

lems with all cycles having positive length. Note that in the former case, despite the fact $\hat{J}^+ = J^*$, there is no optimal terminating policy, since the only optimal policy is a linear policy that drives the system to the origin asymptotically, but not in finite time.

Let us illustrate the results of this section with two examples.

Example 4.5.1 (Minimum Energy Stable Control of Linear Systems)

Consider the linear-quadratic problem of Section 3.5.4. We assume that there exists at least one linear stable policy, so that J^* is real-valued. However, we are making no assumptions on the state weighting matrix Q other than positive semidefiniteness. This includes the case $Q = 0$, when $J^*(x) \equiv 0$. In this case an optimal policy is $\mu^*(x) \equiv 0$, which may not be stable, yet the problem of finding a stable policy that minimizes the “control energy” (a cost that is positive definite quadratic on the control with no penalty on the state) among all stable policies is meaningful.

We consider the forcing function

$$p(x) = \|x\|^2,$$

so the p - δ -perturbed problem includes a positive definite state penalty and from the classical linear-quadratic results, $\hat{J}_{p,\delta}$ is a positive definite quadratic function $x'P_\delta x$, where P_δ is the unique solution of the δ -perturbed Riccati equation

$$P_\delta = A'(P_\delta - P_\delta B(B'P_\delta B + R)^{-1}B'P_\delta)A + Q + \delta I, \quad (4.70)$$

within the class of positive semidefinite matrices. By Prop. 4.5.1, we have $\hat{J}_p(x) = x' \hat{P}x$, where $\hat{P} = \lim_{\delta \downarrow 0} P_\delta$ is positive semidefinite, and solves the (unperturbed) Riccati equation

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

Moreover, by Prop. 4.5.4(a), \hat{P} is the largest solution among positive semidefinite matrices, since all positive semidefinite quadratic functions belong to the set S_p of Eq. (4.59). By Prop. 4.5.4(c), any stable stationary policy $\hat{\mu}$ that is optimal among the set of stable policies must satisfy the optimality condition

$$\hat{\mu}(x) \in \arg \min_{u \in \mathbb{R}^m} \{u'Ru + (Ax + Bu)'\hat{P}(Ax + Bu)\}, \quad \forall x \in \mathbb{R}^n,$$

[cf. Eq. (4.66)], or equivalently, by setting the gradient of the minimized expression to 0,

$$(R + B'\hat{P}B)\hat{\mu}(x) = -B'\hat{P}Ax. \quad (4.71)$$

We may solve Eq. (4.71), and check if any of its solutions $\hat{\mu}$ is p -stable; if this is so, $\hat{\mu}$ is optimal within the class of p -stable policies. Note, however, that in the absence of additional conditions, it is possible that some policies $\hat{\mu}$ that solve Eq. (4.71) are p -unstable.

In the case where there is no linear stable policy, the p - δ -perturbed cost function $\hat{J}_{p,\delta}$ need not be real-valued, and the δ -perturbed Riccati equation (4.70) may not have any solution (consider for example the case where $n = 1$, $m = 1$, $A = 2$, $B = 0$, and $Q = R = 1$). Then, Prop. 4.5.6 still applies, but the preceding analytical approach needs to be modified.

As noted earlier, the Bellman equation may have multiple solutions corresponding to different forcing functions p , with each solution being unique within the corresponding set \mathcal{W}_p of Eq. (4.64), consistently with Prop. 4.5.4(a). The following is an illustrative example.

Example 4.5.2 (An Optimal Stopping Problem)

Consider an optimal stopping problem where the state space X is \mathbb{R}^n . We identify the destination with the origin of \mathbb{R}^n , i.e., $t = 0$. At each $x \neq 0$, we may either stop (move to the origin) at a cost $c > 0$, or move to state γx at cost $\|x\|$, where γ is a scalar with $0 < \gamma < 1$; see Fig. 4.5.4.† Thus the Bellman equation has the form

$$J(x) = \begin{cases} \min \{c, \|x\| + J(\gamma x)\} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

† In this example, the salient feature of the policy that never stops at an $x \neq 0$ is that it drives the system asymptotically to the destination according to an equation of the form $x_{k+1} = f(x_k)$, where f is a contraction mapping. The example admits generalization to the broader class of optimal stopping problems that have this property. For simplicity in illustrating our main point, we consider here the special case where $f(x) = \gamma x$ with $\gamma \in (0, 1)$.

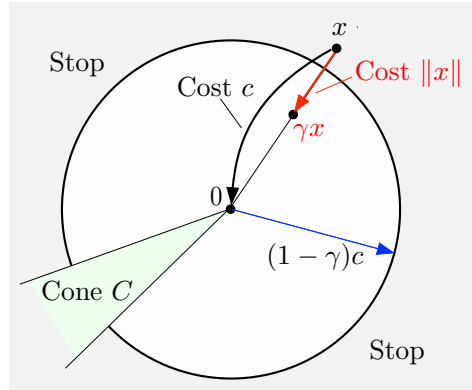


Figure 4.5.4 Illustration of the stopping problem of Example 4.5.2. The optimal policy is to stop outside the sphere of radius $(1 - \gamma)c$ and to continue otherwise. Each cone C of the state space defines a different solution \hat{J}_p of Bellman's equation, with $\hat{J}_p(x) = c$ for all nonzero $x \in C$, and a corresponding region of convergence of the VI algorithm.

Let us consider first the forcing function

$$p(x) = \|x\|.$$

Then it can be verified that all policies are p -stable. We have

$$J^*(x) = \hat{J}_p(x) = \min \left\{ c, \frac{1}{1 - \gamma} \|x\| \right\}, \quad \forall x \in \mathbb{R}^n,$$

and the optimal cost function of the corresponding p - δ -perturbed problem is

$$\hat{J}_{p,\delta}(x) = \min \left\{ c + \delta \|x\|, \frac{1 + \delta}{1 - \gamma} \|x\| \right\}, \quad \forall x \in \mathbb{R}^n.$$

Here the set S_p of Eq. (4.59) is given by

$$S_p = \left\{ J \in \mathcal{J} \mid \lim_{x \rightarrow 0} J(x) = 0 \right\},$$

and the corresponding set \mathcal{W}_p of Eq. (4.64) is given by

$$\mathcal{W}_p = \left\{ J \in \mathcal{J} \mid J^* \leq J, \lim_{x \rightarrow 0} J(x) = 0 \right\}.$$

Let us consider next the forcing function

$$p^+(x) = \begin{cases} 1 & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

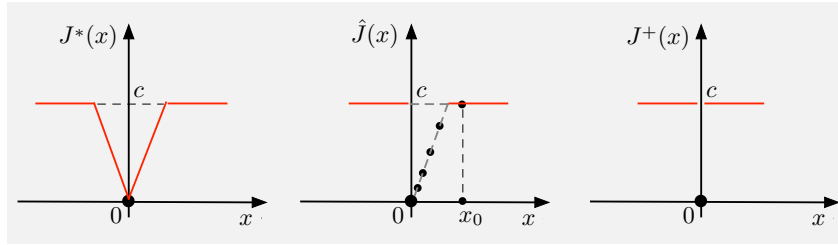


Figure 4.5.5 Illustration of three solutions of Bellman's equation in the one-dimensional case ($n = 1$) of the stopping problem of Example 4.5.2. The solution in the middle is specified by a scalar $x_0 > 0$, and has the form

$$\hat{J}(x) = \begin{cases} 0 & \text{if } x = 0, \\ \frac{1}{1-\gamma}|x| & \text{if } 0 < x < (1-\gamma)c \text{ and } x = \gamma^k x_0 \text{ for some } k \geq 0, \\ c & \text{otherwise.} \end{cases}$$

Then the p^+ -stable policies are the terminating policies. Since stopping at some time and incurring the cost c is a requirement for a p^+ -stable policy, it follows that the optimal p^+ -stable policy is to stop as soon as possible, i.e., stop at every state. The corresponding restricted optimal cost function is

$$\hat{J}^+(x) = \begin{cases} c & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

The optimal cost function of the corresponding p^+ - δ -perturbed problem is

$$\hat{J}_{p^+, \delta}(x) = \begin{cases} c + \delta & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases}$$

since in the p^+ - δ -perturbed problem it is again optimal to stop as soon as possible, at cost $c + \delta$. Here the set S_{p^+} is equal to \mathcal{J} , and the corresponding set \mathcal{W}^+ is equal to $\{J \in \mathcal{J} \mid \hat{J}^+ \leq J\}$.

However, there are infinitely many additional solutions of Bellman's equation between the largest and smallest solutions J^* and \hat{J}^+ . For example, when $n > 1$, functions $J \in \mathcal{J}$ such that $J(x) = J^*(x)$ for x in some cone and $J(x) = \hat{J}^+(x)$ for x in the complementary cone are solutions; see Fig. 4.5.4. There is also a corresponding infinite number of regions of convergence \mathcal{W}_p of VI [cf. Eq. (4.64)]. Also VI converges to J^* starting from any J_0 with $0 \leq J_0 \leq J^*$ [cf. Prop. 4.4.4(d)]. Figure 4.5.5 illustrates additional solutions of Bellman's equation of a different character.

4.5.3 Policy Iteration Methods

Generally, the standard PI algorithm [cf. Eqs. (4.48), (4.49)] produces unclear results under our assumptions. The following example provides an instance where the PI algorithm may converge to either an optimal or a strictly suboptimal policy.

Example 4.5.3 (Counterexample for PI)

Consider the case $X = \{0, 1\}$, $U(0) = U(1) = \{0, 1\}$, and the destination is $t = 0$. Let also

$$f(x, u) = \begin{cases} 0 & \text{if } u = 0, \\ x & \text{if } u = 1, \end{cases} \quad g(x, u) = \begin{cases} 1 & \text{if } u = 0, x = 1, \\ 0 & \text{if } u = 1 \text{ or } x = 0. \end{cases}$$

This is a one-state-plus-destination shortest path problem where the control $u = 0$ moves the state from $x = 1$ to $x = 0$ (the destination) at cost 1, while the control $u = 1$ keeps the state unchanged at cost 0 (cf. the problem of Section 3.1.1). The policy μ^* that keeps the state unchanged is the only optimal policy, with $J_{\mu^*}(x) = J^*(x) = 0$ for both states x . However, under any forcing function p with $p(1) > 0$, the policy $\hat{\mu}$, which moves from state 1 to 0, is the only p -stable policy, and we have $J_{\hat{\mu}}(1) = \hat{J}_p(1) = 1$. The standard PI algorithm (4.48), (4.49) when started with μ^* , it will repeat μ^* . If this algorithm is started with $\hat{\mu}$, it may generate μ^* or it may repeat $\hat{\mu}$, depending on how the policy improvement iteration is implemented. The reason is that for both x we have

$$\hat{\mu}(x) \in \arg \min_{u \in \{0,1\}} \left\{ g(x, u) + \hat{J}_p(f(x, u)) \right\},$$

as can be verified with a straightforward calculation. Thus a rule for breaking a tie in the policy improvement operation is needed, but such a rule may not be obvious in general.

For another illustration, consider the stopping problem of Example 4.5.2. There if PI is started with the policy that stops at every state, it repeats that policy, and this policy is not optimal even within the class of stable policies with respect to the forcing function $p(x) = \|x\|$.

Motivated by the preceding examples, we consider several types of PI methods that bypass the difficulty above either through assumptions or through modifications. We first consider a favorable case where the standard PI algorithm is reliable. This is the case where the terminating policies are sufficient, in the sense that $J^* = \hat{J}^+$, as in Section 3.5.5.

Policy Iteration for the Case $J^* = \hat{J}^+$

The standard PI algorithm starts with a stationary policy μ^0 , and generates a sequence of stationary policies $\{\mu^k\}$ via a sequence of policy evaluations to obtain J_{μ^k} from the equation

$$J_{\mu^k}(x) = g(x, \mu^k(x)) + J_{\mu^k}(f(x, \mu^k(x))), \quad x \in X, \quad (4.72)$$

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

$$\mu^{k+1}(x) \in \arg \min_{u \in U(x)} \left\{ g(x, u) + J_{\mu^k}(f(x, u)) \right\}, \quad x \in X. \quad (4.73)$$

We implicitly assume here that Eq. (4.72) can be solved for J_{μ^k} , and that the minimum in Eq. (4.73) is attained for each $x \in X$, which is true under some compactness condition on either $U(x)$ or the level sets of the function $g(x, \cdot) + J_k(f(x, \cdot))$, or both.

Proposition 4.5.7: (Convergence of PI) Assume that $J^* = \hat{J}^+$. Then the sequence $\{J_{\mu^k}\}$ generated by the PI algorithm (4.72), (4.73), satisfies $J_{\mu^k}(x) \downarrow J^*(x)$ for all $x \in X$.

Proof: For a stationary policy μ , let $\bar{\mu}$ satisfy the policy improvement equation

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

We have shown that

$$J_{\mu}(x) \geq \inf_{u \in U(x)} \{g(x, u) + J_{\mu}(f(x, u))\} \geq J_{\bar{\mu}}(x), \quad x \in X; \quad (4.74)$$

cf. Eq. (4.29). Using μ^k and μ^{k+1} in place of μ and $\bar{\mu}$, we see that the sequence $\{J_{\mu^k}\}$ generated by PI converges monotonically to some function $J_{\infty} \in E^+(X)$, i.e., $J_{\mu^k} \downarrow J_{\infty}$. Moreover, from Eq. (4.74) we have

$$J_{\infty}(x) \geq \inf_{u \in U(x)} \{g(x, u) + J_{\infty}(f(x, u))\}, \quad x \in X,$$

as well as

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

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

$$J_{\infty}(x) \geq \inf_{u \in U(x)} \{g(x, u) + J_{\infty}(f(x, u))\} \geq J_{\infty}(x), \quad x \in X.$$

Thus J_{∞} is a solution of Bellman's equation, satisfying $J_{\infty} \geq J^*$ (since $J_{\mu^k} \geq J^*$ for all k) and $J_{\infty} \in \mathcal{J}$ (since $J_{\mu^k} \in \mathcal{J}$), so by Prop. 4.5.6(a), it must satisfy $J_{\infty} = J^*$. **Q.E.D.**

A Perturbed Version of Policy Iteration for the Case $J^* \neq \hat{J}^+$

We now consider PI algorithms without the condition $J^* = \hat{J}^+$. We provide a version of the PI algorithm, which uses a given forcing function p that is fixed, and generates a sequence $\{\mu^k\}$ of p -stable policies such that

$J_{\mu^k} \rightarrow \hat{J}_p$. Related algorithms were given in Sections 3.4 and 3.5.1. The following assumption requires that the algorithm generates p -stable policies exclusively, which can be quite restrictive. For instance it is not satisfied for the problem of Example 4.5.3.

Assumption 4.5.1: For each $\delta > 0$ there exists at least one p -stable stationary policy μ such that $J_{\mu,p,\delta} \in S_p$. Moreover, given a p -stable stationary policy μ and a scalar $\delta > 0$, every stationary policy $\bar{\mu}$ such that

$$\bar{\mu}(x) \in \arg \min_{u \in U(x)} \left\{ g(x, u) + J_{\mu,p,\delta}(f(x, u)) \right\}, \quad \forall x \in X,$$

is p -stable, and at least one such policy exists.

The perturbed version of the PI algorithm is defined as follows. Let $\{\delta_k\}$ be a positive sequence with $\delta_k \downarrow 0$, and let μ^0 be a p -stable policy that satisfies $J_{\mu^0,p,\delta_0} \in S_p$. One possibility is that μ^0 is an optimal policy for the δ_0 -perturbed problem (cf. the discussion preceding Prop. 4.5.3). At iteration k , we have a p -stable policy μ^k , and we generate a p -stable policy μ^{k+1} according to

$$\mu^{k+1}(x) \in \arg \min_{u \in U(x)} \left\{ g(x, u) + J_{\mu^k,p,\delta_k}(f(x, u)) \right\}, \quad x \in X. \quad (4.75)$$

Note that by Assumption 4.5.1 the algorithm is well-defined, and is guaranteed to generate a sequence of p -stable stationary policies. We have the following proposition.

Proposition 4.5.8: Let Assumption 4.5.1 hold. Then for a sequence of p -stable policies $\{\mu^k\}$ generated by the perturbed PI algorithm (4.75), we have $J_{\mu^k,p,\delta_k} \downarrow \hat{J}_p$ and $J_{\mu^k} \rightarrow \hat{J}_p$.

Proof: Since the forcing function p is kept fixed, to simplify notation, we abbreviate $J_{\mu,p,\delta}$ with $J_{\mu,\delta}$ for all policies μ and scalars $\delta > 0$. Also, we will use the mappings $T_\mu : \mathcal{E}^+(X) \mapsto \mathcal{E}^+(X)$ and $T_{\mu,\delta} : \mathcal{E}^+(X) \mapsto \mathcal{E}^+(X)$ given by

$$(T_\mu J)(x) = g(x, \mu(x)) + J(f(x, \mu(x))), \quad x \in X,$$

$$(T_{\mu,\delta} J)(x) = g(x, \mu(x)) + \delta p(x) + J(f(x, \mu(x))), \quad x \in X.$$

Moreover, we will use the mapping $T : \mathcal{E}^+(X) \mapsto \mathcal{E}^+(X)$ given by

$$(TJ)(x) = \inf_{u \in U(x)} \left\{ g(x, u) + J(f(x, u)) \right\}, \quad x \in X.$$

The algorithm definition (4.75) implies that for all integer $m \geq 1$ we have for all $x_0 \in X$,

$$\begin{aligned} J_{\mu^k, \delta_k}(x_0) &\geq (TJ_{\mu^k, \delta_k})(x_0) + \delta_k p(x_0) \\ &= (T_{\mu^{k+1}, \delta_k} J_{\mu^k, \delta_k})(x_0) \\ &\geq (T_{\mu^{k+1}, \delta_k}^m J_{\mu^k, \delta_k})(x_0) \\ &\geq (T_{\mu^{k+1}, \delta_k}^m \bar{J})(x_0), \end{aligned}$$

where \bar{J} is the identically zero function [$\bar{J}(x) \equiv 0$]. From this relation we obtain

$$\begin{aligned} J_{\mu^k, \delta_k}(x_0) &\geq \lim_{m \rightarrow \infty} (T_{\mu^{k+1}, \delta_k}^m \bar{J})(x_0) \\ &= \lim_{m \rightarrow \infty} \left\{ \sum_{\ell=0}^{m-1} (g(x_\ell, \mu^{k+1}(x_\ell)) + \delta_k p(x_\ell)) \right\} \\ &\geq J_{\mu^{k+1}, \delta_{k+1}}(x_0), \end{aligned}$$

as well as

$$J_{\mu^k, \delta_k}(x_0) \geq (TJ_{\mu^k, \delta_k})(x_0) + \delta_k p(x_0) \geq J_{\mu^{k+1}, \delta_{k+1}}(x_0).$$

It follows that $\{J_{\mu^k, \delta_k}\}$ is monotonically nonincreasing, so that $J_{\mu^k, \delta_k} \downarrow J_\infty$ for some J_∞ , and

$$\lim_{k \rightarrow \infty} TJ_{\mu^k, \delta_k} = J_\infty. \quad (4.76)$$

We also have, using the fact $J_\infty \leq J_{\mu^k, \delta_k}$,

$$\begin{aligned} \inf_{u \in U(x)} \{g(x, u) + J_\infty(f(x, u))\} &\leq \lim_{k \rightarrow \infty} \inf_{u \in U(x)} \{g(x, u) + J_{\mu^k, \delta_k}(f(x, u))\} \\ &\leq \inf_{u \in U(x)} \lim_{k \rightarrow \infty} \{g(x, u) + J_{\mu^k, \delta_k}(f(x, u))\} \\ &= \inf_{u \in U(x)} \left\{ g(x, u) + \lim_{k \rightarrow \infty} J_{\mu^k, \delta_k}(f(x, u)) \right\} \\ &= \inf_{u \in U(x)} \{g(x, u) + J_\infty(f(x, u))\}. \end{aligned}$$

Thus equality holds throughout above, so that

$$\lim_{k \rightarrow \infty} TJ_{\mu^k, \delta_k} = TJ_\infty.$$

Combining this with Eq. (4.76), we obtain $J_\infty = TJ_\infty$, i.e., J_∞ solves Bellman's equation. We also note that $J_\infty \leq J_{\mu^0, \delta_0}$ and that $J_{\mu^0, \delta_0} \in S_p$ by assumption, so that $J_\infty \in S_p$. By Prop. 4.5.4(a), it follows that $J_\infty = \hat{J}_p$.
Q.E.D.

Note that despite the fact $J_{\mu^k} \rightarrow \hat{J}_p$, the generated sequence $\{\mu^k\}$ may exhibit some serious pathologies in the limit. In particular, if U is a metric space and $\{\mu^k\}_{\mathcal{K}}$ is a subsequence of policies that converges to some $\bar{\mu}$, in the sense that

$$\lim_{k \rightarrow \infty, k \in \mathcal{K}} \mu^k(x) = \bar{\mu}(x), \quad \forall x \in X,$$

it does not follow that $\bar{\mu}$ is p -stable. In fact it is possible to construct examples where the generated sequence of p -stable policies $\{\mu^k\}$ satisfies $\lim_{k \rightarrow \infty} J_{\mu^k} = \hat{J}_p = J^*$, yet $\{\mu^k\}$ may converge to a p -unstable policy whose cost function is strictly larger than \hat{J}_p .

An Optimistic Policy Iteration Method

Let us consider an optimistic variant of PI, where policies are evaluated inexactly, with a finite number of VIs. We use a fixed forcing function p . The algorithm aims to compute \hat{J}_p , the restricted optimal cost function over the p -stable policies, and generates a sequence $\{J_k, \mu^k\}$ according to

$$T_{\mu^k} J_k = T J_k, \quad J_{k+1} = T_{\mu^k}^{m_k} J_k, \quad k = 0, 1, \dots, \quad (4.77)$$

where m_k is a positive integer for each k . We assume that a policy μ^k satisfying $T_{\mu^k} J_k = T J_k$ can be found for all k , but it need not be p -stable. However, the algorithm requires that

$$J_0 \in \mathcal{W}_p, \quad J_0 \geq T J_0. \quad (4.78)$$

This may be a restrictive assumption. We have the following proposition.

Proposition 4.5.9: (Convergence of Optimistic PI) Assume that there exists at least one p -stable policy $\pi \in \Pi_p$, and that J_0 satisfies Eq. (4.78). Then a sequence $\{J_k\}$ generated by the optimistic PI algorithm (4.77) belongs to \mathcal{W}_p and satisfies $J_k \downarrow \hat{J}_p$.

Proof: Since $J_0 \geq \hat{J}_p$ and $\hat{J}_p = T \hat{J}_p$ [cf. Prop. 4.5.6(a)], all operations on any of the functions J_k with T_{μ^k} or T maintain the inequality $J_k \geq \hat{J}_p$ for all k , so that $J_k \in \mathcal{W}_p$ for all k . Also the conditions $J_0 \geq T J_0$ and $T_{\mu^k} J_k = T J_k$ imply that

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

and continuing similarly,

$$J_k \geq T J_k \geq J_{k+1}, \quad k = 0, 1, \dots \quad (4.79)$$

Thus $J_k \downarrow J_\infty$ for some J_∞ , which must satisfy $J_\infty \geq \hat{J}_p$, and hence belong to \mathcal{W}_p . By taking limit as $k \rightarrow \infty$ in Eq. (4.79) and using an argument similar to the one in the proof of Prop. 4.5.8, it follows that $J_\infty = TJ_\infty$. By Prop. 4.5.6(a), this implies that $J_\infty \leq \hat{J}_p$. Together with the inequality $J_\infty \geq \hat{J}_p$ shown earlier, this proves that $J_\infty = \hat{J}_p$. **Q.E.D.**

As an example, for the shortest path problem of Example 4.5.3, the reader may verify that in the case where $p(x) = 1$ for $x = 1$, the optimistic PI algorithm converges in a single iteration to

$$\hat{J}_p(x) = \begin{cases} 1 & \text{if } x = 1, \\ 0 & \text{if } x = 0, \end{cases}$$

provided that $J_0 \in \mathcal{W}_p = \{J \mid J(1) \geq 1, J(0) = 0\}$. For other starting functions J_0 , the algorithm converges in a single iteration to the function

$$J_\infty(1) = \min\{1, J_0(1)\}, \quad J_\infty(0) = 0.$$

All functions J_∞ of the form above are solutions of Bellman's equation, but only \hat{J}_p is restricted optimal.

4.6 INFINITE-SPACES STOCHASTIC SHORTEST PATH PROBLEMS

In this section we consider a stochastic discrete-time infinite horizon optimal control problem involving the system

$$x_{k+1} = f(x_k, u_k, w_k), \quad k = 0, 1, \dots, \quad (4.80)$$

where x_k and u_k are the state and control at stage k , which belong to sets X and U , w_k is a random disturbance that takes values in a countable set W with given probability distribution $P(w_k \mid x_k, u_k)$, and $f : X \times U \times W \mapsto X$ is a given function (cf. Example 1.2.1 in Chapter 1). The state and control spaces X and U are arbitrary, but we assume that W is countable to bypass complex measurability issues in the choice of control (see [BeS78]).

The control u must be chosen from a constraint set $U(x) \subset U$ that may depend on x . The expected cost per stage, $E\{g(x, u, w)\}$, is assumed nonnegative:

$$0 \leq E\{g(x, u, w)\} < \infty, \quad \forall x \in X, u \in U(x), w \in W.$$

We assume that X contains a special cost-free and absorbing state t , referred to as the *destination*:

$$f(t, u, w) = t, \quad g(t, u, w) = 0, \quad \forall u \in U(t), w \in W.$$

This is a special case of an SSP problem, where the cost per stage is nonnegative, but the state and control spaces are arbitrary. It is also a special case of the nonnegative cost stochastic optimal control problem of Section 4.4.2. We adopt the notation and terminology of that section, but we review it here briefly for convenience.

Given an initial state x_0 , a policy $\pi = \{\mu_0, \mu_1, \dots\}$ when applied to the system (4.80), generates a random sequence of state-control pairs $(x_k, \mu_k(x_k))$, $k = 0, 1, \dots$, with cost

$$J_\pi(x_0) = E_{x_0}^\pi \left\{ \sum_{k=0}^{\infty} g(x_k, \mu_k(x_k), w_k) \right\},$$

where $E_{x_0}^\pi\{\cdot\}$ denotes expectation with respect to the probability measure corresponding to initial state x_0 and policy π . For a stationary policy μ , the corresponding cost function is denoted by J_μ . The optimal cost function is

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

and its effective domain is denoted X^* , i.e.,

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

A policy π^* is said to be optimal if $J_{\pi^*}(x) = J^*(x)$ for all $x \in X$.

We denote by $\mathcal{E}^+(X)$ the set of functions $J : X \mapsto [0, \infty]$. In our analysis, we will use the set of functions

$$\mathcal{J} = \{J \in \mathcal{E}^+(X) \mid J(t) = 0\}.$$

Since t is cost-free and absorbing, this set contains the cost functions J_π of all $\pi \in \Pi$, as well as J^* .

Here the results of Section 4.3 under Assumption I apply, and the optimal cost function J^* is a solution of the Bellman equation

$$J(x) = \inf_{u \in U(x)} E \left\{ g(x, u, w) + J(f(x, u, w)) \right\}, \quad x \in X,$$

where the expected value is with respect to the distribution $P(w \mid x, u)$. Moreover, an optimal stationary policy (if it exists) may be obtained through the minimization in the right side of this equation (with J replaced by J^* , cf. Prop. 4.4.4). The VI algorithm starts from some function $J_0 \in \mathcal{J}$, and generates a sequence $\{J_k\} \subset \mathcal{J}$ according to

$$J_{k+1}(x) = \inf_{u \in U(x)} E \left\{ g(x, u, w) + J_k(f(x, u, w)) \right\}, \quad x \in X, \quad k = 0, 1, \dots$$

Proper Policies and the δ -Perturbed Problem

We will now introduce a notion of proper policy with a definition that extends the one used for finite-state SSP in Section 3.5.1. For a given state $x \in X$, a policy π is said to be *proper at x* if

$$J_\pi(x) < \infty, \quad \sum_{k=0}^{\infty} r_k(\pi, x) < \infty, \quad (4.81)$$

where $r_k(\pi, x)$ is the probability that $x_k \neq t$ when using π and starting from $x_0 = x$. We denote by $\widehat{\Pi}_x$ the set of all policies that are proper at x , and we denote by \hat{J} the corresponding restricted optimal cost function,

$$\hat{J}(x) = \inf_{\pi \in \widehat{\Pi}_x} J_\pi(x), \quad x \in X,$$

(with the convention that the infimum over the empty set is ∞). Finally we denote by \widehat{X} the effective domain of \hat{J} , i.e.,

$$\widehat{X} = \{x \in X \mid \hat{J}(x) < \infty\}. \quad (4.82)$$

Note that \widehat{X} is the set of x such that $\widehat{\Pi}_x$ is nonempty and that $t \in \widehat{X}$.

For any $\delta > 0$, let us consider the δ -perturbed optimal control problem. This is the same problem as the original, except that the cost per stage is changed to

$$g(x, u, w) + \delta, \quad \forall x \neq t,$$

while $g(x, u, w)$ is left unchanged at 0 when $x = t$. Thus t is still cost-free as well as absorbing in the δ -perturbed problem. The δ -perturbed cost function of a policy π starting from x is denoted by $J_{\pi, \delta}(x)$ and involves an additional expected cost $\delta r_k(\pi, x)$ for each stage k , so that

$$J_{\pi, \delta}(x) = J_\pi(x) + \delta \sum_{k=0}^{\infty} r_k(\pi, x).$$

Clearly, the sum $\sum_{k=0}^{\infty} r_k(\pi, x)$ is the expected number of steps to reach the destination starting from x and using π , if the sum is finite. We denote by \hat{J}_δ the optimal cost function of the δ -perturbed problem, i.e., $\hat{J}_\delta(x) = \inf_{\pi \in \Pi} J_{\pi, \delta}(x)$. The following proposition provides some characterizations of proper policies in relation to the δ -perturbed problem.

Proposition 4.6.1:

- (a) A policy π is proper at a state $x \in X$ if and only if $J_{\pi, \delta}(x) < \infty$ for all $\delta > 0$.
- (b) We have $\hat{J}_\delta(x) < \infty$ for all $\delta > 0$ if and only if $x \in \widehat{X}$.
- (c) For every $\epsilon > 0$ and $\delta > 0$, a policy π_ϵ that is ϵ -optimal for the δ -perturbed problem is proper at all $x \in \widehat{X}$, and such a policy exists.

Proof: (a) Follows from Eq. (4.50) and the definition (4.81) of a proper policy.

(b) If $x \in \widehat{X}$ there exists a policy π that is proper at x , and by part (a), $\hat{J}_\delta(x) \leq J_{\pi,\delta}(x) < \infty$ for all $\delta > 0$. Conversely, if $\hat{J}_\delta(x) < \infty$, there exists π such that $J_{\pi,\delta}(x) < \infty$, implying [by part (a)] that $\pi \in \widehat{\Pi}_x$, so that $x \in \widehat{X}$.

(c) An ϵ -optimal π_ϵ exists by Prop. 4.4.4(e). We have $J_{\pi_\epsilon,\delta}(x) \leq \hat{J}_\delta(x) + \epsilon$ for all $x \in X$. Hence $J_{\pi_\epsilon,\delta}(x) < \infty$ for all $x \in \widehat{X}$, implying by part (a) that π_ϵ is proper at all $x \in \widehat{X}$. **Q.E.D.**

The next proposition shows that the cost function \hat{J}_δ of the δ -perturbed problem can be used to approximate \hat{J} .

Proposition 4.6.2: We have $\lim_{\delta \downarrow 0} \hat{J}_\delta(x) = \hat{J}(x)$ for all $x \in X$. Moreover, for any $\epsilon > 0$ and $\delta > 0$, a policy π_ϵ that is ϵ -optimal for the δ -perturbed problem is ϵ -optimal within the class of proper policies, i.e., satisfies

$$J_{\pi_\epsilon}(x) \leq \hat{J}(x) + \epsilon, \quad \forall x \in X.$$

Proof: Let us fix $\delta > 0$, and for a given $\epsilon > 0$, let π_ϵ be a policy that is proper at all $x \in \widehat{X}$ and is ϵ -optimal for the δ -perturbed problem [cf. Prop. 4.6.1(c)]. By using Eq. (4.50), we have for all $\epsilon > 0$, $x \in \widehat{X}$, and $\pi \in \widehat{\Pi}_x$,

$$\hat{J}(x) - \epsilon \leq J_{\pi_\epsilon}(x) - \epsilon \leq J_{\pi_\epsilon,\delta}(x) - \epsilon \leq \hat{J}_\delta(x) \leq J_{\pi,\delta}(x) = J_\pi(x) + w_{\pi,\delta}(x),$$

where

$$w_{\pi,\delta}(x) = \delta \sum_{k=0}^{\infty} r_k(\pi, x) < \infty, \quad \forall x \in \widehat{X}, \pi \in \widehat{\Pi}_x.$$

By taking the limit as $\epsilon \downarrow 0$, we obtain for all $\delta > 0$ and $\pi \in \widehat{\Pi}_x$,

$$\hat{J}(x) \leq \hat{J}_\delta(x) \leq J_\pi(x) + w_{\pi,\delta}(x), \quad \forall x \in \widehat{X}, \pi \in \widehat{\Pi}_x.$$

We have $\lim_{\delta \downarrow 0} w_{\pi,\delta}(x) = 0$ for all $x \in \widehat{X}$ and $\pi \in \widehat{\Pi}_x$, so by taking the limit as $\delta \downarrow 0$ and then the infimum over all $\pi \in \widehat{\Pi}_x$,

$$\hat{J}(x) \leq \lim_{\delta \downarrow 0} \hat{J}_\delta(x) \leq \inf_{\pi \in \widehat{\Pi}_x} J_\pi(x) = \hat{J}(x), \quad \forall x \in \widehat{X},$$

from which $\hat{J}(x) = \lim_{\delta \downarrow 0} \hat{J}_\delta(x)$ for all $x \in \widehat{X}$. Moreover, by Prop. 4.6.1(b), $\hat{J}_\delta(x) = \hat{J}(x) = \infty$ for all $x \notin \widehat{X}$, so that $\hat{J}(x) = \lim_{\delta \downarrow 0} \hat{J}_\delta(x)$ for all $x \in X$.

We also have

$$J_{\pi_\epsilon}(x) \leq J_{\pi_\epsilon, \delta}(x) \leq \hat{J}_\delta(x) + \epsilon \leq J_\pi(x) + \delta \sum_{k=0}^{\infty} r(\pi, x) + \epsilon, \quad \forall x \in \hat{X}, \pi \in \hat{\Pi}_x.$$

By taking the limit as $\delta \downarrow 0$, we obtain

$$J_{\pi_\epsilon}(x) \leq J_\pi(x) + \epsilon, \quad \forall x \in \hat{X}, \pi \in \hat{\Pi}_x.$$

By taking the infimum over $\pi \in \hat{\Pi}_x$, it follows that $J_{\pi_\epsilon}(x) \leq \hat{J}(x) + \epsilon$ for all $x \in \hat{X}$, which combined with the fact $J_{\pi_\epsilon}(x) = \hat{J}(x) = \infty$ for all $x \notin \hat{X}$, yields the result. **Q.E.D.**

Main Results

By Prop. 4.4.4(a), \hat{J}_δ solves Bellman's equation for the δ -perturbed problem, while by Prop. 4.6.2, $\lim_{\delta \downarrow 0} \hat{J}_\delta(x) = \hat{J}(x)$. This suggests that \hat{J} solves the unperturbed Bellman equation, which is the "limit" as $\delta \downarrow 0$ of the δ -perturbed version. Indeed we will show a stronger result, namely that \hat{J} is the unique solution of Bellman's equation within the set of functions

$$\widehat{\mathcal{W}} = \{J \in S \mid \hat{J} \leq J\}, \quad (4.83)$$

where

$$S = \left\{ J \in \mathcal{J} \mid E_{x_0}^\pi \{J(x_k)\} \rightarrow 0, \forall (\pi, x_0) \text{ with } \pi \in \hat{\Pi}_{x_0} \right\}. \quad (4.84)$$

Here $E_{x_0}^\pi \{J(x_k)\}$ denotes the expected value of the function J along the sequence $\{x_k\}$ generated starting from x_0 and using π . Similar to earlier proofs in Sections 4.4 and 4.5, we have that the collection

$$\mathcal{C} = \{(\pi, x) \mid \pi \in \hat{\Pi}_x\} \quad (4.85)$$

is S -regular.

We first show a preliminary result. Given a policy $\pi = \{\mu_0, \mu_1, \dots\}$, we denote by π_k the policy

$$\pi_k = \{\mu_k, \mu_{k+1}, \dots\}. \quad (4.86)$$

Proposition 4.6.3:

(a) For all pairs $(\pi, x_0) \in \mathcal{C}$ and $k = 0, 1, \dots$, we have

$$0 \leq E_{x_0}^\pi \{\hat{J}(x_k)\} \leq E_{x_0}^{\pi_k} \{J_{\pi_k}(x_k)\} < \infty,$$

where π_k is the policy given by Eq. (4.86).

(b) The set $\widehat{\mathcal{W}}$ of Eq. (4.83) contains \hat{J} , as well as all functions $J \in S$ satisfying $\hat{J} \leq J \leq c\hat{J}$ for some $c \geq 1$.

Proof: (a) For any pair $(\pi, x_0) \in \mathcal{C}$ and $\delta > 0$, we have

$$J_{\pi, \delta}(x_0) = E_{x_0}^{\pi} \left\{ J_{\pi_k, \delta}(x_k) + \sum_{m=0}^{k-1} g(x_m, \mu_m(x_m), w_m) \right\} + \delta \sum_{m=0}^{k-1} r_m(\pi, x_0).$$

Since $J_{\pi, \delta}(x_0) < \infty$ [cf. Prop. 4.6.1(a)], it follows that $E_{x_0}^{\pi} \{J_{\pi_k, \delta}(x_k)\} < \infty$. Hence for all x_k that can be reached with positive probability using π and starting from x_0 , we have $J_{\pi_k, \delta}(x_k) < \infty$, implying [by Prop. 4.6.1(a)] that $(\pi_k, x_k) \in \mathcal{C}$. Hence $\hat{J}(x_k) \leq J_{\pi_k}(x_k)$ and by applying $E_{x_0}^{\pi} \{\cdot\}$, the result follows.

(b) We have for all $(\pi, x_0) \in \mathcal{C}$,

$$J_{\pi}(x_0) = E_{x_0}^{\pi} \left\{ g(x_0, \mu_0(x_0), w_0) \right\} + E_{x_0}^{\pi} \{J_{\pi_1}(x_1)\}, \quad (4.87)$$

and for $m = 1, 2, \dots$,

$$E_{x_0}^{\pi} \{J_{\pi_m}(x_m)\} = E_{x_0}^{\pi} \left\{ g(x_m, \mu_m(x_m), w_m) \right\} + E_{x_0}^{\pi} \{J_{\pi_{m+1}}(x_{m+1})\}, \quad (4.88)$$

where $\{x_m\}$ is the sequence generated starting from x_0 and using π . By using repeatedly the expression (4.88) for $m = 1, \dots, k-1$, and combining it with Eq. (4.87), we obtain for all $k = 1, 2, \dots$,

$$J_{\pi}(x_0) = E_{x_0}^{\pi} \{J_{\pi_k}(x_k)\} + \sum_{m=0}^{k-1} E_{x_0}^{\pi} \left\{ g(x_m, \mu_m(x_m), w_m) \right\}, \quad \forall (\pi, x_0) \in \mathcal{C}.$$

The rightmost term above tends to $J_{\pi}(x_0)$ as $k \rightarrow \infty$, so by using the fact $J_{\pi}(x_0) < \infty$, we obtain

$$E_{x_0}^{\pi} \{J_{\pi_k}(x_k)\} \rightarrow 0, \quad \forall (\pi, x_0) \in \mathcal{C}.$$

By part (a), it follows that

$$E_{x_0}^{\pi} \{\hat{J}(x_k)\} \rightarrow 0, \quad \forall (\pi, x_0) \in \mathcal{C},$$

so that $\hat{J} \in \widehat{\mathcal{W}}$. This also implies that

$$E_{x_0}^{\pi} \{J(x_k)\} \rightarrow 0, \quad \forall (\pi, x_0) \in \mathcal{C},$$

if $\hat{J} \leq J \leq c\hat{J}$ for some $c \geq 1$. **Q.E.D.**

We can now prove our main result.

Proposition 4.6.4: Assume that either W is finite or there exists a $\delta > 0$ such that

$$E\left\{g(x, u, w) + \hat{J}_\delta(f(x, u, w))\right\} < \infty, \quad \forall x \in X^*, u \in U(x).$$

- (a) \hat{J} is the unique solution of the Bellman Eq. (4.65) within the set $\widehat{\mathcal{W}}$ of Eq. (4.83).
- (b) (*VI Convergence*) If $\{J_k\}$ is the sequence generated by the VI algorithm (4.47) starting with some $J_0 \in \widehat{\mathcal{W}}$, then $J_k \rightarrow \hat{J}$.
- (c) (*Optimality Condition*) If μ is a stationary policy that is proper at all $x \in \widehat{X}$ and

$$\mu(x) \in \arg \min_{u \in U(x)} E\left\{g(x, u, w) + \hat{J}(f(x, u, w))\right\}, \quad \forall x \in X, \quad (4.89)$$

then μ is optimal over the set of proper policies, i.e., $J_\mu = \hat{J}$. Conversely, if μ is proper at all $x \in \widehat{X}$ and $J_\mu = \hat{J}$, then μ satisfies the preceding condition (4.89).

Proof: (a), (b) By Prop. 4.6.3(b), $\hat{J} \in \widehat{\mathcal{W}}$. We will first show that \hat{J} is a solution of Bellman's equation. Since \hat{J}_δ solves the Bellman equation for the δ -perturbed problem, and $\hat{J}_\delta \geq \hat{J}$ (cf. Prop. 4.6.2), we have for all $\delta > 0$ and $x \neq t$,

$$\begin{aligned} \hat{J}_\delta(x) &= \inf_{u \in U(x)} E\left\{g(x, u, w) + \delta + \hat{J}_\delta(f(x, u, w))\right\} \\ &\geq \inf_{u \in U(x)} E\left\{g(x, u, w) + \hat{J}_\delta(f(x, u, w))\right\} \\ &\geq \inf_{u \in U(x)} E\left\{g(x, u, w) + \hat{J}(f(x, u, w))\right\}. \end{aligned}$$

By taking the limit as $\delta \downarrow 0$ and using Prop. 4.6.2, we obtain

$$\hat{J}(x) \geq \inf_{u \in U(x)} E\left\{g(x, u, w) + \hat{J}(f(x, u, w))\right\}, \quad \forall x \in X. \quad (4.90)$$

For the reverse inequality, let $\{\delta_m\}$ be a sequence with $\delta_m \downarrow 0$. We have for all m , $x \neq t$, and $u \in U(x)$,

$$\begin{aligned} E\left\{g(x, u, w) + \delta_m + \hat{J}_{\delta_m}(f(x, u, w))\right\} &\geq \inf_{v \in U(x)} E\left\{g(x, v, w) + \delta_m \right. \\ &\quad \left. + \hat{J}_{\delta_m}(f(x, v, w))\right\} \\ &= \hat{J}_{\delta_m}(x). \end{aligned}$$

We now take limit as $m \rightarrow \infty$ in the preceding relation, and we interchange limit and expectation (our assumptions allow the use of the monotone convergence theorem for this purpose; Exercise 4.11 illustrates the need for these assumptions). Using also the fact $\lim_{\delta_m \downarrow 0} \hat{J}_{\delta_m} = \hat{J}$ (cf. Prop. 4.6.2), we have

$$E\left\{g(x, u, w) + \hat{J}(f(x, u, w))\right\} \geq \hat{J}(x), \quad \forall x \in X, u \in U(x),$$

so that

$$\inf_{u \in U(x)} E\left\{g(x, u, w) + \hat{J}(f(x, u, w))\right\} \geq \hat{J}(x), \quad \forall x \in X. \quad (4.91)$$

By combining Eqs. (4.90) and (4.91), we see that \hat{J} is a solution of Bellman's equation.

Part (b) follows by using the S -regularity of the collection (4.85) and Prop. 4.4.2(b). Finally, since $\hat{J} \in \widehat{\mathcal{W}}$ and \hat{J} is a solution of Bellman's equation, part (b) implies the uniqueness assertion of part (a).

(c) If μ is proper at all $x \in \widehat{X}$ and Eq. (4.89) holds, then

$$\hat{J}(x) = E\left\{g(x, \mu(x), w) + \hat{J}(f(x, \mu(x), w))\right\}, \quad x \in X.$$

By Prop. 4.4.4(b), this implies that $J_\mu \leq \hat{J}$, so μ is optimal over the set of proper policies. Conversely, assume that μ is proper at all $x \in \widehat{X}$ and $J_\mu = \hat{J}$. Then by Prop. 4.4.4(b), we have

$$\hat{J}(x) = E\left\{g(x, \mu(x), w) + \hat{J}(f(x, \mu(x), w))\right\}, \quad x \in X,$$

while [by part (a)] \hat{J} is a solution of Bellman's equation,

$$\hat{J}(x) = \inf_{u \in U(x)} E\left\{g(x, u, w) + \hat{J}(f(x, u, w))\right\}, \quad x \in X.$$

Combining the last two relations, we obtain Eq. (4.89). **Q.E.D.**

We illustrate Prop. 4.6.4 in Fig. 4.6.1. Let us consider now the favorable case where the set of proper policies is sufficient in the sense that it can achieve the same optimal cost as the set of all policies, i.e., $\hat{J} = J^*$. This is true for example if all policies are proper at all x such that $J^*(x) < \infty$. Moreover it is true in some of the finite-state formulations of SSP that we discussed in Chapter 3; see also the subsequent Prop. 4.6.5. When $\hat{J} = J^*$, it follows from Prop. 4.6.4 that J^* is the unique solution of Bellman's equation within $\widehat{\mathcal{W}}$, and that the VI algorithm converges to J^* starting from any $J_0 \in \widehat{\mathcal{W}}$. Under an additional compactness condition, such as finiteness

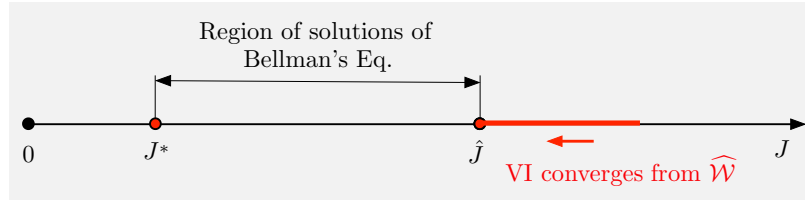


Figure 4.6.1 Illustration of the solutions of Bellman's equation. All solutions either lie between J^* and \hat{J} , or they lie outside the set $\widehat{\mathcal{W}}$. The VI algorithm converges to \hat{J} starting from any $J_0 \in \widehat{\mathcal{W}}$.

of $U(x)$ for all $x \in X$ [cf. Prop. 4.4.4(e)], VI converges to J^* starting from any J_0 in the set S of Eq. (4.84).

Proposition 4.6.4 does not say anything about the existence of a proper policy that is optimal within the class of proper policies. For a simple example where $J^* = \hat{J}$ but the only optimal policy is improper, consider a deterministic shortest path problem with a single state 1 plus the destination t . At state 1 we may choose $u \in [0, 1]$ with cost u , and move to t if $u \neq 0$ and stay at 1 if $u = 0$. Note that here we have $J^*(1) = \hat{J}(1) = 0$, and the minimum over $u \in [0, 1]$ is attained in Bellman's equation, which has the form

$$J^*(1) = \min \left\{ \inf_{u \in (0,1]} u, J^*(1) \right\}.$$

However, the only optimal policy (staying at 1) is improper.

4.6.1 The Multiplicity of Solutions of Bellman's Equation

Let us now discuss the issue of multiplicity of solutions of Bellman's equation within the set of functions

$$\mathcal{J} = \{J \in \mathcal{E}^+(X) \mid J(t) = 0\}.$$

We know from Props. 4.4.4(a) and 4.6.4(a) that J^* and \hat{J} are solutions, and that all other solutions J must satisfy either $J^* \leq J \leq \hat{J}$ or $J \notin \widehat{\mathcal{W}}$.

In the special case of a deterministic problem (one where the disturbance w_k takes a single value), it was shown in Section 4.5 that \hat{J} is the largest solution of Bellman's equation within \mathcal{J} , so all solutions $J' \in \mathcal{J}$ satisfy $J^* \leq J' \leq \hat{J}$. It was also shown through examples that there can be any number of solutions that lie between J^* and \hat{J} : a finite number, an infinite number, or none at all.

In stochastic problems, however, the situation is strikingly different in the following sense: there can be an infinite number of solutions that do not lie below \hat{J} , i.e., solutions $J' \in \mathcal{J}$ that do not satisfy $J' \leq \hat{J}$. Of course, by Prop. 4.6.4(a), these solutions must lie outside $\widehat{\mathcal{W}}$. The following example, which involves a finite set W , is an illustration.

Example 4.6.1

Let $X = \mathfrak{R}$, $t = 0$, and assume that there is only one control at each state, and hence a single policy π . The disturbance w_k takes two values: 1 and 0 with probabilities $\alpha \in (0, 1)$ and $1 - \alpha$, respectively. The system equation is

$$x_{k+1} = \frac{w_k x_k}{\alpha},$$

and there is no cost at each state and stage:

$$g(x, u, w) \equiv 0.$$

Thus from state x_k we move to state x_k/α with probability α and to the termination state $t = 0$ with probability $1 - \alpha$.

Here, the unique policy is stationary and proper at all $x \in X$, and we have

$$J^*(x) = \hat{J}(x) = 0, \quad \forall x \in X.$$

Bellman's equation has the form

$$J(x) = (1 - \alpha)J(0) + \alpha J\left(\frac{x}{\alpha}\right),$$

which within \mathcal{J} reduces to

$$J(x) = \alpha J\left(\frac{x}{\alpha}\right), \quad \forall J \in \mathcal{J}, x \in X. \quad (4.92)$$

It can be seen that Bellman's equation has an infinite number of solutions within \mathcal{J} in addition to J^* and \hat{J} : any positively homogeneous function, such as, for example,

$$J(x) = \gamma|x|, \quad \gamma > 0,$$

is a solution. Consistently with Prop. 4.6.4(a), none of these solutions belongs to $\widehat{\mathcal{W}}$, since x_k is either equal to x_0/α^k (with probability α^k) or equal to 0 (with probability $1 - \alpha^k$). For example, in the case of $J(x) = \gamma|x|$, we have

$$E_{x_0}^\pi \{J(x_k)\} = \alpha^k \gamma \left| \frac{x_0}{\alpha^k} \right| = \gamma|x_0|, \quad \forall k \geq 0,$$

so $J(x_k)$ does not converge to 0, unless $x_0 = 0$. Moreover, none of these additional solutions seems to be significant in some discernible way.

The preceding example illustrates an important structural difference between deterministic and stochastic shortest path problems with infinite state space. For a terminating policy μ in the context of the deterministic problem of Section 4.5, the corresponding Bellman equation $J = T_\mu J$ has a unique solution within \mathcal{J} [to see this, consider the restricted problem for which μ is the only policy, and apply Prop. 4.5.6(a)]. By contrast, for a proper policy in the stochastic context of the present section, the corresponding Bellman equation may have an infinite number of solutions within \mathcal{J} , as Example 4.6.1 shows. This discrepancy does not occur when the state space is finite, as we have seen in Section 3.5.1. We will next elaborate on the preceding observations and refine our analysis regarding multiplicity of solutions of Bellman's equation for problems where the cost per stage is bounded.

4.6.2 The Case of Bounded Cost per Stage

Let us consider the special case where the cost per stage g is bounded over $X \times U \times W$, i.e.,

$$\sup_{(x,u,w) \in X \times U \times W} g(x, u, w) < \infty. \quad (4.93)$$

We will show that \hat{J} is the largest solution of Bellman's equation within the class of functions that are bounded over the effective domain \hat{X} of \hat{J} [cf. Eq. (4.82)].

We say that a policy π is *uniformly proper* if there is a uniform bound on the expected number of steps to reach the destination from states $x \in \hat{X}$ using π :

$$\sup_{x \in \hat{X}} \sum_{k=0}^{\infty} r_k(\pi, x) < \infty.$$

Since we have

$$J_\pi(x_0) \leq \left(\sup_{(x,u,w) \in X \times U \times W} g(x, u, w) \right) \cdot \sum_{k=0}^{\infty} r_k(\pi, x_0) < \infty, \quad \forall \pi \in \hat{\Pi}_{x_0},$$

it follows that the cost function J_π of a uniformly proper π belongs to the set \mathcal{B} , defined by

$$\mathcal{B} = \left\{ J \in \mathcal{J} \mid \sup_{x \in \hat{X}} J(x) < \infty \right\}. \quad (4.94)$$

When $\hat{X} = X$, the notion of a uniformly proper policy coincides with the notion of a transient policy used in [Pli78] and [JaC06], which itself descends from earlier works. However, our definition is somewhat more general, since it also applies to the case where \hat{X} is a strict subset of X .

Let us denote by $\widehat{\mathcal{W}}_b$ the set of functions

$$\widehat{\mathcal{W}}_b = \{ J \in \mathcal{B} \mid \hat{J} \leq J \}.$$

The following proposition, illustrated in Fig. 4.6.2, provides conditions for \hat{J} to be the largest fixed point of T within \mathcal{B} . Its assumptions include the existence of a uniformly proper policy, which implies that \hat{J} belongs to \mathcal{B} . The proposition also uses the earlier Prop. 4.4.6 in order to provide conditions for $J^* = \hat{J}$, in which case J^* is the unique fixed point of T within \mathcal{B} .

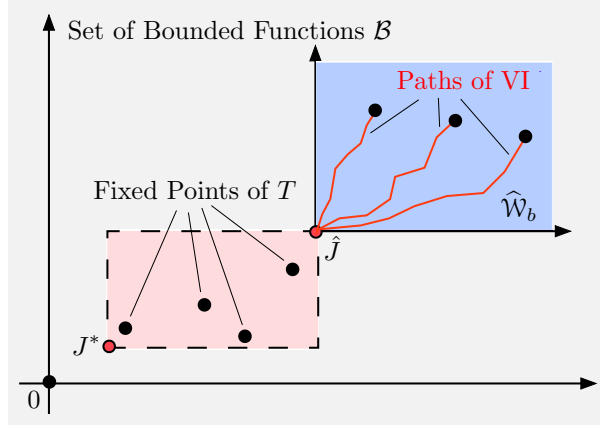


Figure 4.6.2. Schematic illustration of Prop. 4.6.5 for a nonnegative cost SSP problem. The functions J^* and \hat{J} are the smallest and largest solutions, respectively, of Bellman's equation within the set \mathcal{B} . Moreover, the VI algorithm converges to \hat{J} starting from $J_0 \in \widehat{\mathcal{W}}_b = \{J \in \mathcal{B} \mid \hat{J} \leq J\}$.

Proposition 4.6.5: Let the assumptions of Prop. 4.6.4 hold, and assume further that the cost per stage g is bounded over $X \times U \times W$ [cf. Eq. (4.93)], and that there exists a uniformly proper policy. Then:

- (a) \hat{J} is the largest solution of the Bellman Eq. (4.65) within the set \mathcal{B} of Eq. (4.94), i.e., \hat{J} is a solution that belongs to \mathcal{B} and if $J' \in \mathcal{B}$ is another solution, then $J' \leq \hat{J}$. Moreover, if $\hat{J} = J^*$, then J^* is the unique solution of Bellman's equation within \mathcal{B} .
- (b) If $\{J_k\}$ is the sequence generated by the VI algorithm (4.47) starting with some $J_0 \in \mathcal{B}$ with $J_0 \geq \hat{J}$, then $J_k \rightarrow \hat{J}$.
- (c) Assume in addition that X is finite, that $J^*(x) > 0$ for all $x \neq t$, and that $X^* = \hat{X}$. Then $\hat{J} = J^*$.

Proof: (a) Since the cost function of a uniformly proper policy belongs to \mathcal{B} , we have $\hat{J} \in \mathcal{B}$. On the other hand, for all $J \in \mathcal{B}$, we have

$$E_{x_0}^\pi \{J(x_k)\} \leq \left(\sup_{x \in \hat{X}} J(x) \right) \cdot r_k(\pi, x_0) \rightarrow 0, \quad \forall \pi \in \hat{\Pi}_{x_0}.$$

It follows that the set $\widehat{\mathcal{W}}_b$ is contained in $\widehat{\mathcal{W}}$, while the function \hat{J} belongs to $\widehat{\mathcal{W}}_b$. Since $\widehat{\mathcal{W}}_b$ is unbounded above within the set \mathcal{B} , for every solution $J' \in \mathcal{B}$ of Bellman's equation we have $J' \leq J$ for some $J \in \widehat{\mathcal{W}}_b$, and hence also $J' \leq \tilde{J}$ for some \tilde{J} in the set S of Eq. (4.84). It follows from Prop. 4.4.2(a) and the S -regularity of the collection (4.85) that $J' \leq \hat{J}$.

If in addition $\hat{J} = J^*$, from Prop. 4.4.4(a), \hat{J} is also the smallest solution of Bellman's equation within \mathcal{J} . Hence J^* is the unique solution of Bellman's equation within \mathcal{B} .

(b) Follows from Prop. 4.6.4(b), since $\widehat{\mathcal{W}}_b \subset \widehat{\mathcal{W}}$, as shown in the proof of part (a).

(c) We have by assumption

$$0 < J^*(x) \leq \hat{J}(x), \quad \forall x \neq t,$$

while $\hat{J}(x) < \infty$ for all $x \in X^*$ since $X^* = \hat{X}$. In view of the finiteness of X , we can find a sufficiently large c such that $\hat{J} \leq cJ^*$, so by Prop. 4.4.6, it follows that $\hat{J} = J^*$. **Q.E.D.**

The uniqueness of solution of Bellman's equation within \mathcal{B} when $\hat{J} = J^*$ [cf. part (a) of the preceding proposition] is consistent with Example 4.6.1. In that example, J^* and \hat{J} are equal and bounded, and all the additional solutions of Bellman's equation are unbounded, as can be verified by using Eq. (4.92).

Note that without the assumption of existence of a uniformly proper π , \hat{J} and J^* need not belong to \mathcal{B} . As an example, let X be the set of nonnegative integers, let $t = 0$, and let there be a single policy that moves the system deterministically from a state $x \geq 1$ to the state $x - 1$ at cost $g(x, x - 1) = 1$. Then

$$\hat{J}(x) = J^*(x) = x, \quad \forall x \in X,$$

so \hat{J} and J^* do not belong to \mathcal{B} , even though g is bounded. Here the unique policy is proper at all x , but is not uniformly proper.

In a given practical application, we may be interested in computing either J^* or \hat{J} . If the cost per stage is bounded, we may compute \hat{J} with the VI algorithm, assuming that an initial function in the set $\widehat{\mathcal{W}}_b$ can be found. The computation of J^* is also possible by using the VI algorithm and starting from the zero initial condition, assuming that the conditions of Prop. 4.4.4(d) are satisfied.

An alternative possibility for the case of a finite spaces SSP is to approximate the problem with a sequence of α_k -discounted problems where the discount factors α_k tend to 1. This approach, developed in some detail in Exercise 5.28 of the book [Ber17a], has the advantage that the discounted problems can be solved more reliably and with a broader variety of methods than the original undiscounted SSP.

Another technique, developed in the paper [BeY16], is to transform a finite-state SSP problem such that $J^*(x) = 0$ for some $x \neq t$ into an equivalent SSP problem that satisfies the conditions of Prop. 4.6.5(c), and thus allow the computation of J^* by a VI or PI algorithm. The idea is to lump t together with the states x for which $J^*(x) = 0$ into a single

state, which is the termination state for the equivalent SSP problem. This technique is strictly limited to finite-state problems, since in general the conditions $J^*(x) > 0$ for all $x \neq t$ and $X^* = \widehat{X}$ do not imply that $\hat{J} = J^*$, even under the bounded cost and uniform properness assumptions of this section (see the deterministic stopping Example 4.5.2).

4.7 NOTES, SOURCES, AND EXERCISES

Sections 4.1: The use of monotonicity as the foundational property of abstract DP models was initiated in the author's papers [Ber75], [Ber77].

Section 4.2: The finite horizon analysis of Section 4.2 was given in Chapter 3 of the monograph by Bertsekas and Shreve [BeS78].

Section 4.3: The analysis of the monotone increasing and decreasing abstract DP models of Section 4.3 is due to the author's papers [Ber75], [Ber77]. This analysis was also presented in Chapter 5 of [BeS78].

Important examples of noncontractive infinite horizon models are the classical negative cost DP problems, analyzed by Blackwell [Bla65], and by Dubins and Savage [DuS65], and the positive cost DP problems analyzed in Strauch [Str66] (and also in Strauch's Ph.D. thesis, written under the supervision of Blackwell). The monograph by Bertsekas and Shreve [BeS78] provides a detailed treatment of these two models, which also resolves the associated measurability questions using the notion of universally measurable policies. The paper by Yu and Bertsekas [YuB15] provides a more recent analysis that addresses some issues regarding the convergence of the VI and PI algorithms that were left unresolved in the monograph [BeS78]. A simpler textbook treatment, which bypasses the measurability questions, is given in the author's [Ber12a], Chapter 4.

The compactness condition that guarantees convergence of VI to J^* starting with the initial condition $J_0 = \bar{J}$ under Assumption I (cf. Prop. 4.3.14) was obtained by the author in [Ber72] for reachability problems (see Exercise 4.5), and in [Ber75], [Ber77] for positive cost DP models; see also Schal [Sch75] and Whittle [Whi80]. A more refined analysis of the question of convergence of VI to J^* is possible. This analysis provides a necessary and sufficient condition for convergence, and improves over the compactness condition of Prop. 4.3.14. In particular, the following characterization is shown in [Ber77], Prop. 11 (see also [BeS78], Prop. 5.9):

For a set $C \subset X \times U \times \mathfrak{R}$, let $\Pi(C)$ be the projection of C onto $X \times \mathfrak{R}$:

$$\Pi(C) = \{(x, \lambda) \mid (x, u, \lambda) \in C \text{ for some } u \in U(x)\},$$

and denote also

$$\overline{\Pi(C)} = \{(x, \lambda) \mid \lambda_m \rightarrow \lambda \text{ for some sequence } \{\lambda_m\} \text{ with } \{(x, \lambda_m)\} \subset C\}.$$

Consider the sets $C_k \subset X \times U \times \mathfrak{R}$ given by

$$C_k = \{(x, u, \lambda) \mid H(x, u, T^k \bar{J}) \leq \lambda, x \in X, u \in U(x)\}, \quad k = 0, 1, \dots$$

Then under Assumption I we have $T^k \bar{J} \rightarrow J^*$ if and only if

$$\overline{\Pi(\cap_{k=0}^{\infty} C_k)} = \cap_{k=0}^{\infty} \overline{\Pi(C_k)}.$$

Moreover we have $T^k \bar{J} \rightarrow J^*$ and in addition there exists an optimal stationary policy if and only if

$$\Pi(\cap_{k=0}^{\infty} C_k) = \cap_{k=0}^{\infty} \overline{\Pi(C_k)}. \quad (4.95)$$

For a connection with Prop. 4.3.14, it can be shown that compactness of

$$U_k(x, \lambda) = \{u \in U(x) \mid H(x, u, T^k \bar{J}) \leq \lambda\}$$

implies Eq. (4.95) (see [Ber77], Prop. 12, or [BeS78], Prop. 5.10).

The analysis of convergence of VI to J^* under Assumption I and starting with an initial condition $J_0 \geq J^*$ is far more complicated than for the initial condition $J_0 = \bar{J}$. A principal reason for this is the multiplicity of solutions of Bellman's equation within the set $\{J \in \mathcal{E}^+(X) \mid J \geq \bar{J}\}$. We know that J^* is the smallest solution (cf. Prop. 4.4.9), and an interesting issue is the characterization of the largest solution and other solutions within some restricted class of functions of interest. We substantially resolved this question in Sections 4.5 and 4.6 for infinite-spaces deterministic and stochastic shortest path problems, respectively (as well in Sections 3.5.1 and 3.5.2 for finite-state stochastic shortest path and affine monotonic problems). Generally, optimal control problems with nonnegative cost per stage can typically be reduced to problems with a cost-free and absorbing termination state (see [BeY16] for an analysis of the finite-state case). However, the fuller characterization of the set of solutions of Bellman's equation for general abstract DP models under Assumption I requires further investigation.

Optimistic PI and λ -PI under Assumption D have not been considered prior to the 2013 edition of this book, and the corresponding analysis of Section 4.3.3 is new. See [Bei96], [ThS10a], [ThS10b], [Ber11b], [Sch11], [Ber16b] for analyses of λ -PI for discounted and SSP problems.

Section 4.4: The definition and analysis of regularity for nonstationary policies was introduced in the author's paper [Ber15]. We have primarily used regularity in this book to analyze the structure of the solution set of Bellman's equation, and to identify the region of attraction of value and policy iteration algorithms. This analysis is multifaceted, so it is worth summarizing here:

- (a) We have characterized the fixed point properties of the optimal cost function J^* and the restricted optimal cost function J_C^* over S -regular

collections \mathcal{C} , for various sets S . While J^* and $J_{\mathcal{C}}^*$ need not be fixed points of T , they are fixed points in a large variety of interesting contexts (Sections 3.3-3.5 and 4.4-4.6).

- (b) We have shown that when $J^* = J_{\mathcal{C}}^*$, then J^* is the unique solution of Bellman's equation in several interesting noncontractive contexts. In particular, Section 3.3 deals with an important case that covers among others, the most common type of stochastic shortest path problems. However, even when $J^* \neq J_{\mathcal{C}}^*$, the functions J^* and $J_{\mathcal{C}}^*$ often bound the set of solutions from below and/or from above (see Sections 3.5.1, 3.5.2, 4.5, 4.6).
- (c) Simultaneously with the analysis of the fixed point properties of J^* and $J_{\mathcal{C}}^*$, we have used regularity to identify the region of convergence of value iteration. Often convergence to $J_{\mathcal{C}}^*$ can be shown from starting functions $J \geq J_{\mathcal{C}}^*$, assuming that $J_{\mathcal{C}}^*$ is a fixed point of T . In the favorable case where $J^* = J_{\mathcal{C}}^*$, convergence to J^* can often be shown from every starting function of interest. In addition regularity has been used to guarantee the validity of policy iteration algorithms that generate exclusively regular policies, and are guaranteed to converge to J^* or $J_{\mathcal{C}}^*$.
- (d) We have been able to characterize some of the solutions of Bellman's equation, but not the entire set. Generally, there may exist an infinite number of solutions, and some of them may not be associated with an S -regular collection for any set S , unless we change the starting function \bar{J} that is part of the definition of the cost function J_{π} of the policies. There is a fundamental difficulty here: the solutions of the Bellman equation $J = TJ$ do not depend on \bar{J} , but S -regularity of a collection of policy-state pairs depends strongly on \bar{J} . A sharper characterization of the solution set of Bellman's equation remains an open interesting question, in both specific problem contexts as well as in generality.

The use of regularity in the analysis of undiscounted and discounted stochastic optimal control in Sections 4.4.2 and 4.4.3 is new, and was presented in the author's paper [Ber15]. The analysis of convergent models in Section 4.4.4, under the condition

$$J^*(x) \geq \bar{J}(x) > -\infty, \quad \forall x \in X,$$

is also new. A survey of stochastic optimal control problems under convergence conditions that are more general than the ones considered here is given by Feinberg [Fei02]. An analysis of convergent models for stochastic optimal control, which illustrates the broad range of pathological behaviors that can occur without the condition $J^* \geq \bar{J}$, is given in the paper by Yu [Yu15].

Section 4.5: This section follows the author’s paper [Ber17a]. The issue of the connection of optimality with stability (and also with controllability and observability) was raised in the classic paper by Kalman [Kal60] in the context of linear-quadratic problems.

The set of solutions of the Riccati equation has been extensively investigated starting with the papers by Willems [Wil71] and Kucera [Kuc72], [Kuc73], which were followed up by several other works; see the book by Lancaster and Rodman [LaR95] for a comprehensive treatment. In these works, the “largest” solution of the Riccati equation is referred to as the “stabilizing” solution, and the stability of the corresponding policy is shown, although the author could not find an explicit statement in the literature regarding the optimality of this policy within the class of all linear stable policies. Also the lines of analysis of these works are tied to the structure of the linear-quadratic problem and are unrelated to our analysis of Section 4.5, which is based on semicontractive ideas.

Section 4.6: Proper policies for infinite-state SSP problems have been considered earlier in the works of Pliska [Pli78], and James and Collins [JaC06], where they are called “transient.” There are a few differences between the frameworks of [Pli78], [JaC06] and Section 4.6, which impact on the results obtained. In particular, the papers [Pli78] and [JaC06] use a related (but not identical) definition of properness to the one of Section 4.6, while the notion of a transient policy used in [JaC06] coincides with the notion of a uniformly proper policy of Section 4.6.2 when $\hat{X} = X$. Furthermore, [Pli78] and [JaC06] do not consider the notion of policy that is “proper at a state.” The paper [Pli78] assumes that all policies are transient, that g is bounded, and that J^* is real-valued. The paper [JaC06] allows for notransient policies that have infinite cost from some initial states, and extends the analysis of Bertsekas and Tsitsiklis [BeT91] from finite state space to infinite state space (addressing also measurability issues). Also, [JaC06] allows the cost per stage g to take both positive and negative values, and uses assumptions that guarantee that $J^* = \hat{J}$, that J^* is real-valued, and that improper policies cannot be optimal. Instead, in Section 4.6 we allow that $J^* \neq \hat{J}$ and that J^* can take the value ∞ , while requiring that g is nonnegative and that the disturbance space W is countable.

The analysis of Section 4.6 comes from the author’s paper [Ber17b], and is most closely related to the SSP analysis under the weak conditions of Section 3.5.1, where we assumed that the state space is finite, but allowed g to take both positive and negative values. The extension of some of our results of Section 4.6 to SSP problems where g takes both positive and negative values may be possible; Exercises 4.8 and 4.9 suggest some research directions. However, our analysis of infinite-spaces SSP problems in this chapter relies strongly on the nonnegativity of g and cannot be extended without major modifications. In this connection, it is worth mentioning the example of Section 3.1.2, which shows that J^* may not be a solution

of Bellman's equation when g can take negative values.

E X E R C I S E S

4.1 (Example of Nonexistence of an Optimal Policy Under D)

This is an example of a deterministic stopping problem where Assumption D holds, and an optimal policy does not exist, even though only two controls are available at each state (stop and continue). The state space is $X = \{1, 2, \dots\}$. Continuation from state x leads to state $x + 1$ with certainty and no cost, while the stopping cost is $-1 + (1/x)$, so that there is an incentive to delay stopping at every state. Here for all x , $\bar{J}(x) = 0$, and

$$H(x, u, J) = \begin{cases} J(x+1) & \text{if } u = \text{continue,} \\ -1 + (1/x) & \text{if } u = \text{stop.} \end{cases}$$

Show that $J^*(x) = -1$ for all x , but there is no policy (stationary or not) that attains the optimal cost starting from x .

Solution: Since a cost is incurred only upon stopping, and the stopping cost is greater than -1 , we have $J_\mu(x) > -1$ for all x and μ . On the other hand, starting from any state x and stopping at $x + n$ yields a cost $-1 + \frac{1}{x+n}$, so by taking n sufficiently large, we can attain a cost arbitrarily close to -1 . Thus $J^*(x) = -1$ for all x , but no policy can attain this optimal cost.

4.2 (Counterexample for Optimality Condition Under D)

For the problem of Exercise 4.1, show that the policy μ that never stops is not optimal but satisfies $T_\mu J^* = T J^*$.

Solution: We have $J^*(x) = -1$ and $J_\mu(x) = 0$ for all $x \in X$. Thus μ is nonoptimal, yet attains the minimum in Bellman's equation

$$J^*(x) = \min \left\{ J^*(x+1), -1 + \frac{1}{x} \right\}$$

for all x .

4.3 (Counterexample for Optimality Condition Under I)

Let

$$\begin{aligned} X &= \mathfrak{R}, & U(x) &\equiv (0, 1], & \bar{J}(x) &\equiv 0, \\ H(x, u, J) &= |x| + J(ux), & \forall x \in X, u \in U(x). \end{aligned}$$

Let $\mu(x) = 1$ for all $x \in X$. Then $J_\mu(x) = \infty$ if $x \neq 0$ and $J_\mu(0) = 0$. Verify that $T_\mu J_\mu = T J_\mu$. Verify also that $J^*(x) = |x|$, and hence μ is not optimal.

Solution: The verification of $T_\mu J_\mu = T J_\mu$ is straightforward. To show that $J^*(x) = |x|$, we first note that $|x|$ is a fixed point of T , so by Prop. 4.3.2, $J^*(x) \leq |x|$. Also $(T\bar{J})(x) = |x|$ for all x , while under Assumption I, we have $J^* \geq T\bar{J}$, so $J^*(x) \geq |x|$. Hence $J^*(x) = |x|$.

4.4 (Solution by Mathematical Programming)

This exercise shows that under Assumptions I and D, it is possible to use a computational method based on mathematical programming when $X = \{1, \dots, n\}$.

- (a) Under Assumption I, show that J^* is the unique solution of the following optimization problem in $z = (z_1, \dots, z_n)$:

$$\begin{aligned} & \text{minimize} && \sum_{i=1}^n z_i \\ & \text{subject to} && z_i \geq \bar{J}(i), \quad z_i \geq \inf_{u \in U(i)} H(i, u, z), \quad i = 1, \dots, n. \end{aligned}$$

- (b) Under Assumption D, show that J^* is the unique solution of the following optimization problem in $z = (z_1, \dots, z_n)$:

$$\begin{aligned} & \text{maximize} && \sum_{i=1}^n z_i \\ & \text{subject to} && z_i \leq \bar{J}(i), \quad z_i \leq H(i, u, z), \quad i = 1, \dots, n, \quad u \in U(i). \end{aligned}$$

Note: Generally, these programs may not be linear or even convex.

Solution: (a) Any feasible solution z of the given optimization problem satisfies $z \geq \bar{J}$ as well as $z_i \geq \inf_{u \in U(i)} H(i, u, z)$ for all $i = 1, \dots, n$, so that $z \geq Tz$. It follows from Prop. 4.4.9 that $z \geq J^*$, which implies that J^* is an optimal solution of the given optimization problem. Also J^* is the unique optimal solution since if z is feasible and $z \neq J^*$, the inequality $z \geq J^*$ implies that $\sum_i z_i > \sum_i J^*(i)$, so z cannot be optimal.

(b) Any feasible solution z of the given optimization problem satisfies $z \leq \bar{J}$ as well as $z_i \leq H(i, u, z)$ for all $i = 1, \dots, n$ and $u \in U(i)$, so that $z \leq Tz$. It follows from Prop. 4.3.6 that $z \leq J^*$, which implies that J^* is an optimal solution of the given optimization problem. Similar to part (a), J^* is the unique optimal solution.

4.5 (Infinite Time Reachability [Ber71], [Ber72])

This exercise provides an instance of an interesting problem where the mapping H is naturally extended real-valued. Consider a dynamic system

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

where w_k is viewed as an uncertain disturbance that may be any point in a set $W(x_k, u_k)$ (this is known in the literature as an “unknown but bounded” disturbance, and is the basis for a worst case/minimax treatment of uncertainty in the control of uncertain dynamic systems). We introduce an abstract DP model where the objective is to find a policy that keeps the state x_k of the system within a given set X at all times, for all possible values of the sequence $\{w_k\}$. This is a common objective, which arises in a variety of control theory contexts, including model predictive control (see [Ber17a], Section 6.4.3).

Let

$$\bar{J}(x) = \begin{cases} 0 & \text{if } x \in X, \\ \infty & \text{otherwise,} \end{cases}$$

and

$$H(x, u, J) = \begin{cases} 0 & \text{if } J(x) = 0, u \in U(x), \text{ and } J(f(x, u, w)) = 0, \forall w \in W(x, u), \\ \infty & \text{otherwise.} \end{cases}$$

- (a) Show that Assumption I holds, and that the optimal cost function has the form

$$J^*(x) = \begin{cases} 0 & \text{if } x \in X^*, \\ \infty & \text{otherwise,} \end{cases}$$

where X^* is some subset of X .

- (b) Consider the sequence of sets $\{X_k\}$, where

$$X_k = \{x \in X \mid (T^k \bar{J})(x) = 0\}.$$

Show that $X_{k+1} \subset X_k$ for all k , and that $X^* \subset \bigcap_{k=0}^{\infty} X_k$. Show also that convergence of VI (i.e., $T^k \bar{J} \rightarrow J^*$) is equivalent to $X^* = \bigcap_{k=0}^{\infty} X_k$.

- (c) Show that $X^* = \bigcap_{k=0}^{\infty} X_k$ and there exists an optimal stationary policy if the sets

$$\hat{U}_k(x) = \{u \in U(x) \mid f(x, u, w) \in X_k, \forall w \in W(x, u)\}$$

are compact for all k greater than some index \bar{k} . *Hint:* Use Prop. 4.3.14.

Solution: Let $\hat{\mathcal{E}}(X)$ be the subset of $\mathcal{E}(X)$ that consists of functions that take only the two values 0 and ∞ , and for all $J \in \hat{\mathcal{E}}(X)$ denote

$$D(J) = \{x \in X \mid J(x) = 0\}.$$

Note that for all $J \in \hat{\mathcal{E}}(X)$ we have $T_\mu J \in \hat{\mathcal{E}}(X)$, $TJ \in \hat{\mathcal{E}}(X)$, and that

$$D(T_\mu J) = \{x \in X \mid x \in D(J), f(x, \mu(x), w) \in D(J), \forall w \in W(x, \mu(x))\},$$

$$D(TJ) = \bigcup_{\mu \in \mathcal{M}} D(T_\mu J).$$

- (a) For all $J \in \hat{\mathcal{E}}(X)$, we have $D(T_\mu J) \subset D(J)$ and $T_\mu J \geq J$, so condition (1) of Assumption I holds, and it is easily verified that the remaining two conditions of

Assumption I also hold. We have $\bar{J} \in \hat{\mathcal{E}}(X)$, so for any policy $\pi = \{\mu_0, \mu_1, \dots\}$, we have $T_{\mu_0} \cdots T_{\mu_k} \bar{J} \in \hat{\mathcal{E}}(X)$. It follows that J_π , given by

$$J_\pi = \lim_{k \rightarrow \infty} T_{\mu_0} \cdots T_{\mu_k} \bar{J},$$

also belongs to $\hat{\mathcal{E}}(X)$, and the same is true for $J^* = \inf_{\pi \in \Pi} J_\pi$. Thus J^* has the given form with $D(J^*) = X^*$.

(b) Since $\{T^k \bar{J}\}$ is monotonically nondecreasing we have $D(T^{k+1} \bar{J}) \subset D(T^k \bar{J})$, or equivalently $X_{k+1} \subset X_k$ for all k . Generally for a sequence $\{J_k\} \subset \hat{\mathcal{E}}(X)$, if $J_k \uparrow J$, we have $J \in \hat{\mathcal{E}}(X)$ and $D(J) = \bigcap_{k=0}^{\infty} D(J_k)$. Thus convergence of VI (i.e., $T^k \bar{J} \uparrow J^*$) is equivalent to $D(J^*) = \bigcap_{k=0}^{\infty} D(J_k)$ or $X^* = \bigcap_{k=0}^{\infty} X_k$.

(c) The compactness condition of Prop. 4.3.14 guarantees that $T^k \bar{J} \uparrow J^*$, or equivalently by part (b), $X^* = \bigcap_{k=0}^{\infty} X_k$. This condition requires that the sets

$$U_k(x, \lambda) = \{u \in U(x) \mid H(x, u, T^k \bar{J}) \leq \lambda\}$$

are compact for every $x \in X$, $\lambda \in \mathfrak{R}$, and for all k greater than some integer \bar{k} . It can be seen that $U_k(x, \lambda)$ is equal to the set

$$\hat{U}_k(x) = \{u \in U(x) \mid f(x, u, w) \in X_k, \forall w \in W(x, u)\}$$

given in the statement of the exercise.

4.6 (Exceptional Linear-Quadratic Problems)

Consider the deterministic linear-quadratic problem of Section 3.5.4 and Example 4.5.1. Assume that there is a single control variable u_k , and two state variables, x_k^1 and x_k^2 , which evolve according to

$$x_{k+1}^1 = \gamma x_k^1 + b u_k, \quad x_{k+1}^2 = x_k^1 + x_k^2 + u_k,$$

where $\gamma > 1$. The cost of stage k is quadratic of the form

$$q((x_k^1)^2 + (x_k^2)^2) + (u_k)^2.$$

Consider the four cases of pairs of values (b, q) where $b \in \{0, 1\}$ and $q \in \{0, 1\}$. For each case, use the theory of Section 4.5 to find the optimal cost function J^* and the optimal cost function over stable policies \hat{J}^+ , and to describe the convergence behavior of VI.

Solution: When $b = 1$ and $q = 1$, the classical controllability and observability conditions are satisfied, and we have $J^* = \hat{J}^+$, while there exists an optimal policy that is linear and stable (so J^* and \hat{J}^+ are real-valued and positive definite quadratic). Moreover, the VI algorithm converges to J^* starting from any $J_0 \geq 0$ (even extended real-valued J_0) with $J_0(0) = 0$.

When $b = 0$ and $q = 0$, we clearly have $J^*(x) \equiv 0$. Also $\hat{J}^+(x^1, x^2) = \infty$ for $x^1 \neq 0$, while $\hat{J}^+(0, x^2)$ is finite for all x^2 , but positive for $x^2 \neq 0$ (since for

$x^1 = 0$, the problem becomes essentially one-dimensional, and similar to the one of Section 3.5.4). The VI algorithm converges to \hat{J}^+ starting from any positive semidefinite quadratic initial condition J_0 with $J_0(0, x^2) = 0$ and $J_0 \neq J^*$.

When $b = 0$ and $q = 1$, we have $J^* = \hat{J}^+$, but J^* and \hat{J}^+ are not real-valued. In particular, since x_k^1 stays constant under all policies when $b = 0$, we have $J^*(x^1, x^2) = \hat{J}^+(x^1, x^2) = \infty$ for $x^1 \neq 0$. Moreover, for an initial state with $x_0^1 = 0$, the problem becomes essentially a one-dimensional problem that satisfies the classical controllability and observability conditions, and we have $J^*(0, x^2) = \hat{J}^+(0, x^2)$ for all x^2 . The VI algorithm takes the form

$$J_{k+1}(0, x^2) = \min_u \{(x^2)^2 + (u)^2 + J_k(0, x^2 + u)\},$$

$$J_{k+1}(x^1, x^2) = \min_u \{(x^1)^2 + (x^2)^2 + (u)^2 + J_k(\gamma x^1, x^1 + x^2 + u)\}, \quad \text{if } x^1 \neq 0.$$

It can be seen that the VI iterates $J_k(0, x^2)$ evolve as in the case of a single state variable problem, where x^1 is fixed at 0. For $x^1 \neq 0$, the VI iterates $J_k(x^1, x^2)$ diverge to ∞ .

When $b = 1$ and $q = 0$, we have $J^*(x) \equiv 0$, while $0 < \hat{J}^+(x) < \infty$ for all $x \neq 0$. Similar to Example 4.5.1, the VI algorithm converges to \hat{J}^+ starting from any initial condition $J_0 \geq \hat{J}^+$. The functions J^* and \hat{J}^+ are real-valued and satisfy Bellman's equation, which has the form

$$J(x^1, x^2) = \min_u \{(u)^2 + J(\gamma x^1 + u, x^1 + x^2 + u)\}.$$

However, Bellman's equation has additional solutions, other than J^* and \hat{J}^+ . One of these is

$$\hat{J}(x^1, x^2) = P(x^1)^2,$$

where $P = \gamma^2 - 1$ (cf. the example of Section 3.5.4).

4.7 (Discontinuities in Infinite-State Shortest Path Problems)

The purpose of this exercise is to show that different types of perturbations in infinite-state shortest path problems, may yield different solutions of Bellman's equation. Consider the optimal stopping problem of Example 4.5.2, and introduce a perturbed version by modifying the effect of the action that moves the state from $x \neq 0$ to γx . Instead, this action stops the system with probability $\delta > 0$ at cost $\beta \geq 0$, and moves the state from x to γx with probability $1 - \delta$ at cost $\|x\|$. Note that with this modification, all policies become uniformly proper. Show that:

- (a) The optimal cost function of the (δ, β) -perturbed version of the problem, denoted $\hat{J}_{\delta, \beta}$, is the unique solution of the corresponding Bellman equation within the class of bounded functions \mathcal{B} of Eq. (4.94).
- (b) For $\beta = 0$, we have $\lim_{\delta \downarrow 0} \hat{J}_{\delta, 0} = J^*$, where J^* is the optimal cost function of the deterministic problem of Example 4.5.2.
- (c) For $\beta = c$, we have $\hat{J}_{\delta, c} = \hat{J}^+$ for all $\delta > 0$, where \hat{J}^+ is the largest solution of Bellman's equation in the deterministic problem of Example

4.5.2 [$\hat{J}^+(x) = c$ for all $x \neq 0$, which corresponds to the policy that stops at all states].

Solution: (a) It can be seen that the Bellman equation for the (δ, β) -perturbed version of the problem is

$$J(x) = \begin{cases} \min \{c, \delta\beta + (1 - \delta)(\|x\| + J(\gamma x))\} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases}$$

and has exactly the same solutions as the equation

$$J(x) = \begin{cases} \min \{c, \delta\beta + (1 - \delta)(\min \{c/(1 - \delta), \|x\|\} + J(\gamma x))\} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

The latter equation involves a bounded cost per stage, and hence according to the theory of Section 4.6, has a unique solution within \mathcal{B} , when all policies are proper.

(b) Evident since the effect of δ on the cost of the optimal policy of the problem of Example 4.5.2 diminishes as $\delta \rightarrow 0$.

(c) Since termination at cost c is inevitable (with probability 1) under every policy, the optimal policy for the (δ, β) -perturbed version of the problem is to stop as soon as possible.

4.8 (A Perturbation Approach for Semicontractive Models)

The purpose of this exercise is to adapt the perturbation approach of Section 3.4 so that it can be used in conjunction with the regularity notion for nonstationary policies of Definition 4.4.1. Given a set of functions $S \subset \mathcal{E}(X)$ and a collection \mathcal{C} of policy-state pairs (π, x) that is S -regular, let $J_{\mathcal{C}}^*$ be the restricted optimal cost function defined by

$$J_{\mathcal{C}}^*(x) = \inf_{(\pi, x) \in \mathcal{C}} J_{\pi}(x), \quad x \in X.$$

Consider also a nonnegative forcing function $p : X \mapsto [0, \infty)$, and for each $\delta > 0$ and stationary policy μ , the mappings $T_{\mu, \delta}$ and T_{δ} given by

$$(T_{\mu, \delta}J)(x) = H(x, \mu(x), J) + \delta p(x), \quad (T_{\delta}J)(x) = \inf_{\mu \in \mathcal{M}} (T_{\mu, \delta}J)(x), \quad x \in X.$$

We refer to the problem associated with the mappings $T_{\mu, \delta}$ as the δ -perturbed problem. The cost function of a policy $\pi = \{\mu_0, \mu_1, \dots\} \in \Pi$ for this problem is

$$J_{\pi, \delta} = \limsup_{k \rightarrow \infty} T_{\mu_0, \delta} \cdots T_{\mu_k, \delta} \bar{J},$$

and the optimal cost function is $\hat{J}_{\delta} = \inf_{\pi \in \Pi} J_{\pi, \delta}$. Assume that for every $\delta > 0$:

- (1) \hat{J}_{δ} satisfies the Bellman equation of the δ -perturbed problem, $\hat{J}_{\delta} = T_{\delta} \hat{J}_{\delta}$.

(2) For every $x \in X$, we have $\inf_{(\pi,x) \in \mathcal{C}} J_{\pi,\delta}(x) = \hat{J}_\delta(x)$.

(3) For all $x \in X$ and $(\pi, x) \in \mathcal{C}$, we have

$$J_{\pi,\delta}(x) \leq J_\pi(x) + w_{\pi,\delta}(x),$$

where $w_{\pi,\delta}$ is a function such that $\lim_{\delta \downarrow 0} w_{\pi,\delta} = 0$.

(4) For every sequence $\{J_m\} \subset S$ with $J_m \downarrow J$, we have

$$\lim_{m \rightarrow \infty} H(x, u, J_m) = H(x, u, J), \quad \forall x \in X, u \in U(x).$$

Then $J_{\mathcal{C}}^*$ is a fixed point of T and the conclusions of Prop. 4.4.2 hold. Moreover, we have

$$J_{\mathcal{C}}^* = \lim_{\delta \downarrow 0} \hat{J}_\delta.$$

Solution: The proof is very similar to the one of Prop. 3.4.1. Condition (2) implies that for every $x \in X$ and $\epsilon > 0$, there exists a policy $\pi_{x,\epsilon}$ such that $(\pi_{x,\epsilon}, x) \in \mathcal{C}$ and $J_{\pi_{x,\epsilon},\delta}(x) \leq \hat{J}_\delta(x) + \epsilon$. Thus, using conditions (2) and (3), we have for all $x \in X$, $\delta > 0$, $\epsilon > 0$, and π with $(\pi, x) \in \mathcal{C}$,

$$J_{\mathcal{C}}^*(x) - \epsilon \leq J_{\pi_{x,\epsilon}}(x) - \epsilon \leq J_{\pi_{x,\epsilon},\delta}(x) - \epsilon \leq \hat{J}_\delta(x) \leq J_{\pi,\delta}(x) \leq J_\pi(x) + w_{\pi,\delta}(x).$$

By taking the limit as $\epsilon \downarrow 0$, we obtain for all $x \in X$, $\delta > 0$, and π with $(\pi, x) \in \mathcal{C}$,

$$J_{\mathcal{C}}^*(x) \leq \hat{J}_\delta(x) \leq J_{\pi,\delta}(x) \leq J_\pi(x) + w_{\pi,\delta}(x).$$

By taking the limit as $\delta \downarrow 0$ and then the infimum over all π with $(\pi, x) \in \mathcal{C}$, it follows [using also condition (3)] that for all $x \in X$,

$$J_{\mathcal{C}}^*(x) \leq \lim_{\delta \downarrow 0} \hat{J}_\delta(x) \leq \inf_{\{\pi | (\pi,x) \in \mathcal{C}\}} \lim_{\delta \downarrow 0} J_{\pi,\delta}(x) \leq \inf_{\{\pi | (\pi,x) \in \mathcal{C}\}} J_\pi(x) = J_{\mathcal{C}}^*(x),$$

so that $J_{\mathcal{C}}^* = \lim_{\delta \downarrow 0} \hat{J}_\delta$.

To prove that $J_{\mathcal{C}}^*$ is a fixed point of T , we prove that both $J_{\mathcal{C}}^* \geq TJ_{\mathcal{C}}^*$ and $J_{\mathcal{C}}^* \leq TJ_{\mathcal{C}}^*$ hold. Indeed, from condition (1) and the fact $\hat{J}_\delta \geq J_{\mathcal{C}}^*$ shown earlier, we have for all $\delta > 0$,

$$\hat{J}_\delta = T_\delta \hat{J}_\delta \geq T \hat{J}_\delta \geq TJ_{\mathcal{C}}^*,$$

and by taking the limit as $\delta \downarrow 0$ and using the fact $J_{\mathcal{C}}^* = \lim_{\delta \downarrow 0} \hat{J}_\delta$ shown earlier, we obtain $J_{\mathcal{C}}^* \geq TJ_{\mathcal{C}}^*$. For the reverse inequality, let $\{\delta_m\}$ be a sequence with $\delta_m \downarrow 0$. Using condition (1) we have for all m ,

$$H(x, u, \hat{J}_{\delta_m}) + \delta_m p(x) \geq (T_{\delta_m} \hat{J}_{\delta_m})(x) = \hat{J}_{\delta_m}(x), \quad \forall x \in X, u \in U(x).$$

Taking the limit as $m \rightarrow \infty$, and using condition (4) and the fact $\hat{J}_{\delta_m} \downarrow J_{\mathcal{C}}^*$ shown earlier, we have

$$H(x, u, J_{\mathcal{C}}^*) \geq J_{\mathcal{C}}^*(x), \quad \forall x \in X, u \in U(x),$$

so that by minimizing over $u \in U(x)$, we obtain $TJ_{\mathcal{C}}^* \geq J_{\mathcal{C}}^*$.

4.9 (Deterministic Optimal Control with Positive and Negative Costs per Stage)

In this exercise, we consider the infinite-spaces optimal control problem of Section 4.5 and its notation, but without the assumption $g \geq 0$ [cf. Eq. (4.46)]. Instead, we assume that

$$-\infty < g(x, u) \leq \infty, \quad \forall x \in X, u \in U(x), k = 0, 1, \dots,$$

and that $J^*(x) > -\infty$ for all $x \in X$. The latter assumption was also made in Section 3.5.5, but in the present exercise, we will not assume the additional near-optimal termination Assumption 3.5.9 of that section, and we will use instead the perturbation framework of Exercise 4.8.

We say that a policy π is *terminating from state* $x_0 \in X$ if the sequence $\{x_k\}$ generated by π starting from x_0 terminates finitely (i.e., satisfies $x_{\bar{k}} = t$ for some index \bar{k}). We denote by Π_x the set of all policies that are terminating from x , and we consider the collection

$$\mathcal{C} = \{(\pi, x) \mid \pi \in \Pi_x\}.$$

Let $J_{\mathcal{C}}^*$ be the corresponding restricted optimal cost function,

$$J_{\mathcal{C}}^*(x) = \inf_{(\pi, x) \in \mathcal{C}} J_{\pi}(x) = \inf_{\pi \in \Pi_x} J_{\pi}(x), \quad x \in X,$$

and let S be the set of functions

$$S = \{J \in \mathcal{E}(X) \mid J(t) = 0, J(x) > -\infty, x \in X\}.$$

Clearly \mathcal{C} is S -regular, so we may consider the perturbation framework of Exercise 4.8 with $p(x) = 1$ for all $x \neq t$ and $p(t) = 0$. Apply the results of that exercise to show that:

(a) We have

$$J_{\mathcal{C}}^* = \lim_{\delta \downarrow 0} \hat{J}_{\delta}.$$

(b) $J_{\mathcal{C}}^*$ is the only fixed point of T within the set

$$\mathcal{W} = \{J \in \mathcal{E}(X) \mid J(t) = 0, J \geq J_{\mathcal{C}}^*\}.$$

(c) We have $T^k J \rightarrow J_{\mathcal{C}}^*$ for all $J \in \mathcal{W}$.

Solution: Part (a) follows from Exercise 4.8, and parts (b), (c) follow from Exercise 4.8 and Prop. 4.4.2.

4.10 (On Proper Policies for Stochastic Shortest Paths)

Consider the infinite-spaces SSP problem of Section 4.6 under the assumptions of Prop. 4.6.4, and assume that g is bounded over $X \times U \times W$.

- (a) Show that if μ is a uniformly proper policy, then J_μ is the unique solution of the equation $J = T_\mu J$ within \mathcal{B} and that $T_\mu^k J \rightarrow J_\mu$ for all $J \in \mathcal{B}$.
- (b) Let J' be a fixed point of T such that $J' \in \mathcal{B}$ and $J' \neq \hat{J}$. Show that a policy μ satisfying $T_\mu J' = T J'$ cannot be uniformly proper.

Solution: (a) Consider the problem where the only policy is μ , i.e., with control constraint set $\tilde{U}(x) = \{\mu(x)\}$, $x \in X$, and apply Props. 4.6.5 and 4.4.4.

(b) Assume to come to a contradiction that μ is uniformly proper. We have $T_\mu J' = T J' = J'$, so by part (a) we have $J' = J_\mu$, while $J_\mu \geq \hat{J}$ since μ is uniformly proper. Thus $J' \geq \hat{J}$ while $J' \neq \hat{J}$ by assumption. This contradicts the largest fixed point property of \hat{J} [cf. Prop. 4.6.5(a)].

4.11 (Example where \hat{J} is not a Fixed Point of T in Infinite Spaces SSP)

We noted in Section 4.6 that some additional assumption, like

$$E\left\{g(x, u, w) + \hat{J}_\delta(f(x, u, w))\right\} < \infty, \quad \forall x \in X^*, u \in U(x), \quad (4.96)$$

or the finiteness of W , is necessary to prove that \hat{J} is a fixed point for SSP problems (cf. Prop. 4.6.4). [The condition (4.96) is satisfied for example if there exists a policy π (necessarily proper at all $x \in X^*$) such that $J_{\pi, \delta}$ is bounded over X^* .] To see what can happen without such an assumption, consider the following example, which was constructed by Yi Zhang (private communication).

Let $X = \{t, 0, 1, 2, \dots\}$, where t is the termination state, and let $g(x, u, w) \equiv 0$, so that $J^*(x) \equiv 0$. There is only one control at each state, and hence only one policy. The transitions are as follows:

From each state $x = 2, 3, \dots$, we move deterministically to state $x - 1$, from state 1 we move deterministically to state t , and from state 0 we move to state $x = 1, 2, \dots$, with probability p_x such that $\sum_{x=1}^{\infty} x p_x = \infty$.

Verify that the unique policy is proper at all $x = 1, 2, \dots$, and we have $\hat{J}(x) = J^*(x) = 0$. However, the policy is not proper at $x = 0$, since the expected number of transitions from $x = 0$ to termination is $\sum_{x=1}^{\infty} x p_x = \infty$. As a result the set $\hat{\Pi}_0$ is empty and we have $\hat{J}(0) = \infty$. Thus \hat{J} does not satisfy the Bellman equation for $x = 0$, since

$$\infty = \hat{J}(0) \neq E\left\{g(0, u, w) + \hat{J}(f(0, u, w))\right\} = \sum_{x=1}^{\infty} p_x \hat{J}(x) = 0.$$

4.12 (Convergence of Nonexpansive Monotone Fixed Point Iterations with a Unique Fixed Point)

Consider the mapping H of Section 2.1 under the monotonicity Assumption 2.1.1. Assume that instead of the contraction Assumption 2.1.2, the following hold:

- (1) For every $J \in \mathcal{B}(X)$, the function TJ belongs to $\mathcal{B}(X)$, the space of functions on X that are bounded with respect to the weighted sup-norm corresponding to a positive weighting function v .
- (2) T is nonexpansive, i.e., $\|TJ - TJ'\| \leq \|J - J'\|$ for all $J, J' \in \mathcal{B}(X)$.
- (3) T has a unique fixed point within $\mathcal{B}(X)$, denoted J^* .
- (4) If X is infinite the following continuity property holds: For each $J \in \mathcal{B}(X)$ and $\{J_m\} \subset \mathcal{B}(X)$ with either $J_m \downarrow J$ or $J_m \uparrow J$,

$$H(x, u, J) = \lim_{m \rightarrow \infty} H(x, u, J_m), \quad \forall x \in X, u \in U(x).$$

Show the following:

- (a) For every $J \in \mathcal{B}(X)$, we have $\|T^k J - J^*\| \rightarrow 0$ if X is finite, and $T^k J \rightarrow J^*$ if X is infinite.
- (b) Part (a) holds if $\mathcal{B}(X)$ is replaced by $\{J \in \mathcal{B}(X) \mid J \geq 0\}$, or by $\{J \in \mathcal{B}(X) \mid J(t) = 0\}$, or by $\{J \in \mathcal{B}(X) \mid J(t) = 0, J \geq 0\}$, where t is a special cost-free and absorbing destination state t .

(Unpublished joint work of the author with H. Yu.)

Solution: (a) Assume first that X is finite. For any $c > 0$, let $V_0 = J^* + cv$ and consider the sequence $\{V_k\}$ defined by $V_{k+1} = TV_k$ for $k \geq 0$. Note that $\{V_k\} \subset \mathcal{B}(X)$, since $\|V_0\| \leq \|J^*\| + c$ so that $V_0 \in \mathcal{B}(X)$, and we have $V_{k+1} = TV_k$, so that property (1) applies. From the nonexpansiveness property (2), we have

$$H(x, u, J^* + cv) \leq H(x, u, J^*) + cv(x), \quad x \in X, u \in U(x),$$

and by taking the infimum over $u \in U(x)$, we obtain $J^* \leq T(J^* + cv) \leq J^* + cv$, i.e., $J^* \leq V_1 \leq V_0$. From this and the monotonicity of T it follows that $J^* \leq V_{k+1} \leq V_k$ for all k , so that for each $x \in X$, $V_k(x) \downarrow \bar{V}(x)$ where $\bar{V}(x) \geq J^*(x)$. Moreover, \bar{V} lies in $\mathcal{B}(X)$ (since $J^* \leq \bar{V} \leq V_k$), and also satisfies $\|V_k - \bar{V}\| \rightarrow 0$ (since X is finite). From property (2), we have $\|TV_k - T\bar{V}\| \leq \|V_k - \bar{V}\|$, so that $\|TV_k - T\bar{V}\| \rightarrow 0$, which together with the fact $TV_k = V_{k+1} \rightarrow \bar{V}$, implies that $\bar{V} = T\bar{V}$. Thus $\bar{V} = J^*$ by the uniqueness property (3), and it follows that $V_k \downarrow J^*$.

Similarly, define $W_k = T^k(J^* - cv)$, and by an argument symmetric to the above, $W_k \uparrow J^*$. Now for any $J \in \mathcal{B}(X)$, let $c = \|J - J^*\|$ in the definition of V_k and W_k . Then $J^* - cv \leq J \leq J^* + cv$, so by the monotonicity of T , we have $W_k \leq T^k J \leq V_k$ as well as $W_k \leq J^* \leq V_k$ for all k . Therefore $\|T^k J - J^*\| \leq \|W_k - V_k\|$ for all $k \geq 0$. Since $\|W_k - V_k\| \leq \|W_k - J^*\| + \|V_k - J^*\| \rightarrow 0$, the conclusion follows.

If X is infinite and property (4) holds, the preceding proof goes through, except for the part that shows that $\|V_k - \bar{V}\| \rightarrow 0$. Instead we use a different

argument to prove that $\bar{V} = T\bar{V}$. Indeed, since $V_k \geq V_{k+1} = TV_k \geq T\bar{V}$, it follows that $\bar{V} \geq T\bar{V}$. For the reverse inequality we write

$$T\bar{V} = \inf_{u \in U(x)} \lim_{k \rightarrow \infty} H(x, u, V_k) \geq \lim_{k \rightarrow \infty} \inf_{u \in U(x)} H(x, u, V_k) = \lim_{k \rightarrow \infty} TV_k = \bar{V},$$

where the first equality follows from the continuity property (4), and the inequality follows from the generic relation $\inf \lim H \geq \lim \inf H$. Thus we have $\bar{V} = T\bar{V}$, which by the uniqueness property (3), implies that $\bar{V} = J^*$ and $V_k \downarrow J^*$. With a similar argument we obtain $W_k \uparrow J^*$, implying that $T^k J \rightarrow J^*$.

(b) The proof of part (a) applies with simple modifications.

4.13 (Convergence of Nonexpansive Monotone Fixed Point Iterations with Multiple Fixed Points)

Consider the mapping H of Section 2.1 under the monotonicity Assumption 2.1.1. Assume that instead of the contraction Assumption 2.1.2, the following hold:

- (1) For every $J \in \mathcal{B}(X)$, the function TJ belongs to $\mathcal{B}(X)$, the space of functions on X that are bounded with respect to the weighted sup-norm corresponding to a positive weighting function v .
- (2) T is nonexpansive, i.e., $\|TJ - TJ'\| \leq \|J - J'\|$ for all $J, J' \in \mathcal{B}(X)$.
- (3) T has a largest fixed point within $\mathcal{B}(X)$, denoted \hat{J} , i.e., $\hat{J} \in \mathcal{B}(X)$, \hat{J} is a fixed point of T , and for every other fixed point $J' \in \mathcal{B}(X)$ we have $J' \leq \hat{J}$.
- (4) If X is infinite the following continuity property holds: For each $J \in \mathcal{B}(X)$ and $\{J_m\} \subset \mathcal{B}(X)$ with either $J_m \downarrow J$ or $J_m \uparrow J$,

$$H(x, u, J) = \lim_{m \rightarrow \infty} H(x, u, J_m), \quad \forall x \in X, u \in U(x).$$

Show the following:

- (a) For every $J \in \mathcal{B}(X)$ such that $\hat{J} \leq J \leq \hat{J} + cv$ for some $c > 0$, we have $\|T^k J - \hat{J}\| \rightarrow 0$ if X is finite, and $T^k J \rightarrow \hat{J}$ if X is infinite.
- (b) Part (a) holds if $\mathcal{B}(X)$ is replaced by $\{J \in \mathcal{B}(X) \mid J \geq 0\}$, or by $\{J \in \mathcal{B}(X) \mid J(t) = 0\}$, or by $\{J \in \mathcal{B}(X) \mid J(t) = 0, J \geq 0\}$, where t is a special cost-free and absorbing destination state t .

(Note the similarity with the preceding exercise.)

Solution: (a) The proof follows the line of proof of the preceding exercise. Assume first that X is finite. For any $c > 0$, let $V_0 = \hat{J} + cv$ and consider the sequence $\{V_k\}$ defined by $V_{k+1} = TV_k$ for $k \geq 0$. Note that $\{V_k\} \subset \mathcal{B}(X)$, since $\|V_0\| \leq \|\hat{J}\| + c$ so that $V_0 \in \mathcal{B}(X)$, and we have $V_{k+1} = TV_k$, so that property (1) applies. From the nonexpansiveness property (2), we have

$$H(x, u, \hat{J} + cv) \leq H(x, u, \hat{J}) + cv(x), \quad x \in X, u \in U(x),$$

and by taking the infimum over $u \in U(x)$, we obtain $\hat{J} \leq T(\hat{J} + cv) \leq \hat{J} + cv$, i.e., $\hat{J} \leq V_1 \leq V_0$. From this and the monotonicity of T it follows that $\hat{J} \leq V_{k+1} \leq V_k$

for all k , so that for each $x \in X$, $V_k(x) \downarrow \bar{V}(x)$ where $\bar{V}(x) \geq \hat{J}(x)$. Moreover, \bar{V} lies in $\mathcal{B}(X)$ (since $\hat{J} \leq \bar{V} \leq V_k$), and also satisfies $\|V_k - \bar{V}\| \rightarrow 0$ (since X is finite). From property (2), we have $\|TV_k - T\bar{V}\| \leq \|V_k - \bar{V}\|$, so that $\|TV_k - T\bar{V}\| \rightarrow 0$, which together with the fact $TV_k = V_{k+1} \rightarrow \bar{V}$, implies that $\bar{V} = T\bar{V}$. Thus $\bar{V} = \hat{J}$ by property (3), and it follows that $V_k \downarrow \hat{J}$.

If X is infinite and property (4) holds, the preceding proof goes through, except for the part that shows that $\|V_k - \bar{V}\| \rightarrow 0$. Instead we use a different argument to prove that $\bar{V} = T\bar{V}$. Indeed, since $V_k \geq V_{k+1} = TV_k \geq T\bar{V}$, it follows that $\bar{V} \geq T\bar{V}$. For the reverse inequality we write

$$T\bar{V} = \inf_{u \in U(x)} \lim_{k \rightarrow \infty} H(x, u, V_k) \geq \lim_{k \rightarrow \infty} \inf_{u \in U(x)} H(x, u, V_k) = \lim_{k \rightarrow \infty} TV_k = \bar{V},$$

where the first equality follows from the continuity property (4). Thus we have $\bar{V} = T\bar{V}$, which by property (3), implies that $\bar{V} = \hat{J}$ and $V_k \downarrow \hat{J}$.

(b) The proof of part (a) applies with simple modifications.

4.14 (Necessary and Sufficient Condition for an Interpolated Nonexpansive Mapping to be a Contraction)

This exercise (due to unpublished joint work with H. Yu) considers a nonexpansive mapping $G : \mathfrak{R}^n \mapsto \mathfrak{R}^n$, and derives conditions under which the interpolated mapping G_γ defined by

$$G_\gamma(x) = (1 - \gamma)x + \gamma G(x), \quad x \in \mathfrak{R}^n,$$

is a contraction for all $\gamma \in (0, 1)$. Consider \mathfrak{R}^n equipped with a strictly convex norm $\|\cdot\|$, and the set

$$C = \left\{ \left(\frac{x - y}{\|x - y\|}, \frac{G(x) - G(y)}{\|x - y\|} \right) \mid x, y \in \mathfrak{R}^n, x \neq y \right\},$$

which can be viewed as a set of “slopes” of G along all directions. Show that the mapping G_γ defined by

$$G_\gamma(x) = (1 - \gamma)x + \gamma G(x), \quad x \in \mathfrak{R}^n,$$

is a contraction for all $\gamma \in (0, 1)$ if and only if there is no closure point (z, w) of C such that $z = w$. *Note:* To illustrate with some one-dimensional examples what can happen if this closure condition is violated, let $G : \mathfrak{R} \mapsto \mathfrak{R}$ be continuously differentiable, monotonically nondecreasing, and satisfying $0 \leq \frac{dG(x)}{dx} \leq 1$. Note that G is nonexpansive. We consider two cases.

- (1) $G(0) = 0$, $\frac{dG(0)}{dx} = 1$, $0 \leq \frac{dG(x)}{dx} < 1$ for $x \neq 0$, $\lim_{x \rightarrow \infty} \frac{dG(x)}{dx} < 1$ and $\lim_{x \rightarrow -\infty} \frac{dG(x)}{dx} < 1$. Here $(z, w) = (1, 1)$ is a closure point of C and satisfies $z = w$. Note that G_γ is not a contraction for any $\gamma \in (0, 1)$, although it has 0 as its unique fixed point.

- (2) $\lim_{x \rightarrow \infty} \frac{dG(x)}{dx} = 1$. Here we have $\lim_{x \rightarrow \infty} (G(x) - G(y)) = x - y$ for $x = y + 1$, so $(1, 1)$ is a closure point of C . It can also be seen that because $\lim_{x \rightarrow \infty} \frac{dG_\gamma(x)}{dx} = 1$, G_γ is not a contraction for any $\gamma \in (0, 1)$, and may have one, more than one, or no fixed points.

Solution: Assume there is no closure point (z, w) of C such that $z = w$, and for $\gamma \in (0, 1)$, let

$$\rho = \sup_{(z, w) \in C} \|(1 - \gamma)z + \gamma w\|.$$

The set C is bounded since for all $(z, w) \in C$, we have $\|z\| = 1$, and $\|w\| \leq 1$ by the nonexpansiveness of G . Hence, there exists a sequence $\{(z_k, w_k)\} \subset C$ that converges to some (\bar{z}, \bar{w}) , and is such that

$$\|(1 - \gamma)z_k + \gamma w_k\| \rightarrow \rho.$$

Since (\bar{z}, \bar{w}) is a closure point of C , we have $\bar{z} \neq \bar{w}$. Using the continuity of the norm, we have

$$\rho = \|(1 - \gamma)\bar{z} + \gamma\bar{w}\| < (1 - \gamma)\|\bar{z}\| + \gamma\|\bar{w}\| \leq 1,$$

where for the strict inequality we use the strict convexity of the norm, and for the last inequality we use the fact $\|\bar{z}\| = 1$ and $\|\bar{w}\| \leq 1$. Thus $\rho < 1$, and since

$$\begin{aligned} \left\| (1 - \gamma) \frac{x - y}{\|x - y\|} + \gamma \frac{G(x) - G(y)}{\|x - y\|} \right\| &= \frac{\|G_\gamma(x) - G_\gamma(y)\|}{\|x - y\|} \\ &\leq \sup_{(z, w) \in C} \|(1 - \gamma)z + \gamma w\| \\ &= \rho, \quad \forall x \neq y, \end{aligned}$$

it follows that G_γ is a contraction of modulus ρ .

Conversely, if G_γ is a contraction, we have

$$\begin{aligned} \sup_{(z, w) \in C} \|(1 - \gamma)z + \gamma w\| &= \sup_{x \neq y} \left\| (1 - \gamma) \frac{x - y}{\|x - y\|} + \gamma \frac{G(x) - G(y)}{\|x - y\|} \right\| \\ &\leq \sup_{x \neq y} \frac{\|G_\gamma(x) - G_\gamma(y)\|}{\|x - y\|} \\ &< 1. \end{aligned}$$

Thus for every closure point (z, w) of C ,

$$\|(1 - \gamma)z + \gamma w\| < 1,$$

which implies that we cannot have $z = w$.