

Semicontractive Models

Contents	
3.1. Pathologies of Noncontractive DP Models	p. 107
3.1.1. Deterministic Shortest Path Problems	p. 111
3.1.2. Stochastic Shortest Path Problems	p. 113
3.1.3. The Blackmailer's Dilemma	p. 115
3.1.4. Linear-Quadratic Problems	p. 118
3.1.5. An Intuitive View of Semicontractive Analysis	p. 123
3.2. Semicontractive Models and Regular Policies	p. 125
3.2.1. S -Regular Policies	p. 128
3.2.2. Restricted Optimization over S -Regular Policies	p. 130
3.2.3. Policy Iteration Analysis of Bellman's Equation	p. 136
3.2.4. Optimistic Policy Iteration and λ -Policy Iteration	p. 144
3.2.5. A Mathematical Programming Approach	p. 148
3.3. Irregular Policies/Infinite Cost Case	p. 149
3.4. Irregular Policies/Finite Cost Case - A Perturbation	p. 155
Approach	p. 155
3.5. Applications in Shortest Path and Other Contexts	p. 161
3.5.1. Stochastic Shortest Path Problems	p. 162
3.5.2. Affine Monotonic Problems	p. 170
3.5.3. Robust Shortest Path Planning	p. 179
3.5.4. Linear-Quadratic Optimal Control	p. 189
3.5.5. Continuous-State Deterministic Optimal Control	p. 191
3.6. Algorithms	p. 195
3.6.1. Asynchronous Value Iteration	p. 195
3.6.2. Asynchronous Policy Iteration	p. 196
3.7. Notes, Sources, and Exercises	p. 203

We will now consider abstract DP models that are intermediate between the contractive models of Chapter 2, where all stationary policies involve a contraction mapping, and noncontractive models to be discussed in Chapter 4, where there are no contraction-like assumptions (although there are some compensating conditions, including monotonicity).

A representative instance of such an intermediate model is the deterministic shortest path problem of Example 1.2.7, where we can distinguish between two types of stationary policies: those that terminate at the destination from every starting node, and those that do not. A more general instance is the stochastic shortest path (SSP for short) problem of Example 1.2.6. In this problem, the analysis revolves around two types of stationary policies μ : those with a mapping T_μ that is a contraction with respect to some norm, and those with a mapping T_μ that is not a contraction with respect to any norm (it can be shown that the former are the ones that terminate with probability 1 starting from any state).

In the models of this chapter, like in SSP problems, we divide policies into two groups, one of which has favorable characteristics. We loosely refer to such models as *semicontractive* to indicate that these favorable characteristics include contraction-like properties of the mapping T_μ . To develop a more broadly applicable theory, we replace the notion of contractiveness of T_μ with a notion of *S-regularity of μ* , where S is an appropriate set of functions of the state (roughly, this is a form of “local stability” of T_μ , which ensures that the cost function J_μ is the unique fixed point of T_μ within S , and that $T_\mu^k J$ converges to J_μ regardless of the choice of J from within S). We allow that some policies are S -regular while others are not.

Note that the term “semicontractive” is not used in a precise mathematical sense here. Rather it refers qualitatively to a collection of models where some policies have a regularity/contraction-like property but others do not. Moreover, regularity is a relative property: the division of policies into “regular” and “irregular” depends on the choice of the set S . On the other hand, typically in practical applications an appropriate choice of S is fairly evident.

Our analysis will involve two types of assumptions:

- (a) *Favorable assumptions*, under which we obtain results that are nearly as strong as those available for the contractive models of Chapter 2. In particular, we show that J^* is a fixed point of T , that the Bellman equation $J = TJ$ has a unique solution, at least within a suitable class of functions, and that variants of the VI and PI algorithms are valid. Some of the VI and PI approaches are suitable for distributed asynchronous computation, similar to their Chapter 2 counterparts for contractive models.
- (a) *Less favorable assumptions*, under which serious difficulties may occur: J^* may not be a fixed point of T , and even when it is, it may not be found using the VI and PI algorithms. These anomalies may ap-

pear in simple problems, such as deterministic and stochastic shortest path problems with some zero length cycles. To address the difficulties, we will consider a restricted problem, where the only admissible policies are the ones that are S -regular. Under reasonable conditions we show that this problem is better-behaved. In particular, J_S^* , the optimal cost function over the S -regular policies only, is the unique solution of Bellman's equation among functions $J \in S$ with $J \geq J_S^*$, while VI converges to J_S^* starting from any $J \in S$ with $J \geq J_S^*$. We will also derive a variety of PI approaches for finding J_S^* and an S -regular policy that is optimal within the class of S -regular policies.

We will illustrate our analysis in Section 3.5, both under favorable and unfavorable assumptions, by means of four classes of practical problems. Some of these problems relate to finding a path to a destination in a graph under stochastic or set membership uncertainty, while others relate to the control of a continuous-state system to a terminal state. In particular, we will consider SSP problems, affine monotonic problems, including problems with multiplicative or risk-sensitive exponential cost function, minimax-type shortest path problems, and continuous-state deterministic problems with nonnegative cost, such as linear-quadratic problems.

The chapter is organized as follows. In Section 3.1, we illustrate the pathologies regarding solutions of Bellman's equation, and the VI and PI algorithms. To this end, we use four simple examples, ranging from finite-state shortest path problems, to continuous-state linear-quadratic problems. These examples provide orientation and motivation for S -regular policies later. In Section 3.2, we formally introduce our abstract DP model, and the notion of an S -regular policy. We then develop some of the basic associated results relating to Bellman's equation, and the convergence of VI and PI, based primarily on the ideas underlying the PI algorithm. In Section 3.3 we refine the results of Section 3.2 under favorable conditions, obtaining results and algorithms that are almost as powerful as the ones for contractive models. In Section 3.4 we develop a complementary analytical approach, which is based on the use of perturbations and applies under less favorable assumptions. In Section 3.5, we discuss in detail the application and refinement of the results of Sections 3.2-3.4 in some important shortest path-type practical contexts. In Section 3.6, we focus on variants of VI and PI-type algorithms for semicontractive DP models, including some that are suitable for asynchronous distributed computation.

3.1 PATHOLOGIES OF NONCONTRACTIVE DP MODELS

In this section we provide a general overview of the analytical and computational difficulties in noncontractive DP models, using for the most part shortest path-type problems. For illustration we will first use two of the simplest and most widely encountered finite-state DP problems: deter-

ministic and SSP problems, whereby we are aiming to reach a destination state at minimum cost.† We will also discuss an example of continuous-state shortest path problem that involves a linear system and a quadratic cost function.

We will adopt the general abstract DP model of Section 1.2. We give a brief description that is adequate for the purposes of this section, and defer a more formal definition to Section 3.2. In particular, we introduce a set of states X , and for each $x \in X$, the nonempty control constraint set $U(x)$. For each policy μ , the mapping T_μ is given by

$$(T_\mu J)(x) = H(x, \mu(x), J), \quad \forall x \in X,$$

where H is a suitable function of (x, u, J) . The mapping T is given by

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

The cost function of a policy $\pi = \{\mu_0, \mu_1, \dots\}$ is

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

where \bar{J} is some function.‡ We want to minimize J_π over π , i.e., to find

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

and a policy that attains the infimum.

For orientation purposes, we recall from Chapter 1 (Examples 1.2.1 and 1.2.2) that for a stochastic optimal control problem involving a finite-state Markov chain with state space $X = \{1, \dots, n\}$, transition probabilities $p_{xy}(u)$, and expected one-stage cost function g , the mapping H is given by

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

and $\bar{J}(x) \equiv 0$. The SSP problem arises when there is an additional termination state that is cost-free, and corresponding transition probabilities $p_{xt}(u)$, $x \in X$.

† These problems are naturally undiscounted, and cannot be readily addressed by introducing a discount factor close to 1, because then the optimal policies may exhibit undesirable behavior. In particular, in the presence of discounting, they may involve moving initially along a small-length cycle in order to postpone the use of an optimal but unavoidably costly path until later, when the discount factor will reduce substantially the cost of that path.

‡ In the contractive models of Chapter 2, the choice of \bar{J} is immaterial, as we discussed in Section 2.1. Here, however, the choice of \bar{J} is important, and affects important characteristics of the model, as we will see later.

A more general undiscounted stochastic optimal control problem involves a stationary discrete-time dynamic system where the state is an element of a space X , and the control is an element of a space U . The control u_k is constrained to take values in a given set $U(x_k) \subset U$, which depends on the current state x_k [$u_k \in U(x_k)$, for all $x_k \in X$]. For a policy $\pi = \{\mu_0, \mu_1, \dots\}$, the state evolves according to a system equation

$$x_{k+1} = f(x_k, \mu_k(x_k), w_k), \quad k = 0, 1, \dots, \quad (3.1)$$

where w_k is a random disturbance that takes values from a space W . We assume that w_k , $k = 0, 1, \dots$, are characterized by probability distributions $P(\cdot | x_k, u_k)$ that are identical for all k , where $P(w_k | x_k, u_k)$ is the probability of occurrence of w_k , when the current state and control are x_k and u_k , respectively. Here, we allow infinite state and control spaces, as well as problems with discrete (finite or countable) state space (in which case the underlying system is a Markov chain). However, for technical reasons that relate to measure-theoretic issues, we assume that W is a countable set. †

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

$$J_\pi(x_0) = \limsup_{k \rightarrow \infty} E \left\{ \sum_{t=0}^k g(x_t, \mu_t(x_t), w_t) \right\}, \quad (3.2)$$

subject to the system equation constraint (3.1), where g is the one-stage cost function. The corresponding mapping of the abstract DP problem is

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

and $\bar{J}(x) \equiv 0$. Again here, $(T_{\mu_0} \cdots T_{\mu_k} \bar{J})(x)$ is the expected cost of the first $k + 1$ periods using π starting from x , and with terminal cost 0.

A discounted version of the problem is defined by the mapping

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

where $\alpha \in (0, 1)$ is the discount factor. It corresponds to minimization of

$$J_\pi(x_0) = \limsup_{k \rightarrow \infty} E \left\{ \sum_{t=0}^k \alpha^t g(x_t, \mu_t(x_t), w_t) \right\}.$$

If the cost per stage g is bounded, then a problem that fits the contractive framework of Chapter 2 is obtained, and can be analyzed using the methods of that chapter. However, there are interesting infinite-state discounted optimal control problems where g is not bounded.

† Measure-theoretic issues are not addressed at all in this second edition of the book. The first edition addressed some of these issues within an abstract DP context in its Chapter 5 and Appendix C (this material is posted at the book's web site); see also the monograph by Bertsekas and Shreve [BeS78], and the paper by Yu and Bertsekas [YuB15].

A Summary of Pathologies

The four examples to be discussed in Sections 3.1.1-3.1.4 are special cases of deterministic and stochastic optimal control problems of the type just described. In each of these examples, we will introduce a subclass of “well-behaved” policies and a restricted optimization problem, which is to minimize the cost over the “well-behaved” subclass (in Section 3.2 the property of being “well-behaved” will be formalized through the notion of S -regularity). The optimal cost function over just the “well-behaved” policies is denoted \hat{J} (we will also use the notation J_S^* later). Here is a summary of the examples and the pathologies that they reveal:

- (a) *A finite-state, finite-control deterministic shortest path problem (Section 3.1.1).* Here the mapping T can have infinitely many fixed points, including J^* and \hat{J} . There exist policies that attain the optimal costs J^* and \hat{J} . Depending on the starting point, the VI algorithm may converge to J^* or to \hat{J} or to a third fixed point of T (for cases where $J^* \neq \hat{J}$, VI converges to \hat{J} starting from any $J \geq \hat{J}$). The PI algorithm can oscillate between two policies that attain J^* and \hat{J} , respectively.
- (b) *A finite-state, finite-control stochastic shortest path problem (Section 3.1.2).* The salient feature of this example is that J^* is not a fixed point of the mapping T . By contrast \hat{J} is a fixed point of T . The VI algorithm converges to \hat{J} starting from any $J \geq \hat{J}$, while it does not converge otherwise.
- (c) *A finite-state, infinite-control stochastic shortest path problem (Section 3.1.3).* We give three variants of this example. In the first variant (a classical problem known as the “blackmailer’s dilemma”), all the policies are “well-behaved,” so $J^* = \hat{J}$, and VI converges to J^* starting from any real-valued initial condition, while PI also succeeds in finding J^* as the limit of the generated sequence $\{J_{\mu^k}\}$. However, PI cannot find an optimal policy, because there is no optimal stationary policy. In a second variant of this example, PI generates a sequence of “well-behaved” policies $\{\mu^k\}$ such that $J_{\mu^k} \downarrow \hat{J}$, but $\{\mu^k\}$ converges to a policy that is either infeasible or is strictly suboptimal. In the third variant of this example, the problem data can strongly affect the multiplicity of the fixed points of T , and the behavior of the VI and PI algorithms.
- (d) *A continuous-state, continuous-control deterministic linear-quadratic problem (Section 3.1.4).* Here the mapping T has exactly two fixed points, J^* and \hat{J} , within the class of positive semidefinite quadratic functions. The VI algorithm converges to \hat{J} starting from all positive initial conditions, and to J^* starting from all other initial conditions. Moreover, starting with a “well-behaved” policy, the PI algorithm

converges to \hat{J} and to an optimal policy within the class of “well-behaved” policies.

It can be seen that the examples exhibit wide-ranging pathological behavior. In Section 3.2, we will aim to construct a theoretical framework that explains this behavior. Moreover, in Section 3.3, we will derive conditions guaranteeing that much of this type of behavior does not occur. These conditions are natural and broadly applicable. They are used to exclude from optimality the policies that are not “well-behaved,” and to obtain results that are nearly as powerful as their counterparts for the contractive models of Chapter 2.

3.1.1 Deterministic Shortest Path Problems

Let us consider the classical deterministic shortest path problem, discussed in Example 1.2.7. Here, we have a graph of n nodes $x = 1, \dots, n$, plus a destination t , and an arc length a_{xy} for each directed arc (x, y) . The objective is to find for each x a directed path that starts at x , ends at t , and has minimum length (the length of a path is defined as the sum of the lengths of its arcs). A standard assumption, which we will adopt here, is that every node x is connected to the destination, i.e., there exists a path from every x to t .

To formulate this shortest path problem as a DP problem, we embed it within a “larger” problem, whereby we view all paths as admissible, including those that do not terminate at t . We also view t as a cost-free and absorbing node. Of course, we need to deal with the presence of policies that do not terminate, and the most common way to do this is to assume that all cycles have strictly positive length, in which case policies that do not terminate cannot be optimal. However, it is not uncommon to encounter shortest path problems with zero length cycles, and even negative length cycles. Thus we will not impose any assumption on the sign of the cycle lengths, particularly since we aim to use the shortest path problem to illustrate behavior that arises in a broader undiscounted/noncontractive DP setting.

As noted in Section 1.2, we can formulate the problem in terms of an abstract DP model where the states are the nodes $x = 1, \dots, n$, and the controls available at x can be identified with the outgoing neighbors of x [the nodes u such that (x, u) is an arc]. The mapping H that defines the corresponding abstract DP problem is

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

A stationary policy μ defines the subgraph whose arcs are $(x, \mu(x))$, $x = 1, \dots, n$. We say that μ is *proper* if this graph is acyclic, i.e., it consists of a tree of paths leading from each node to the destination. If μ is not

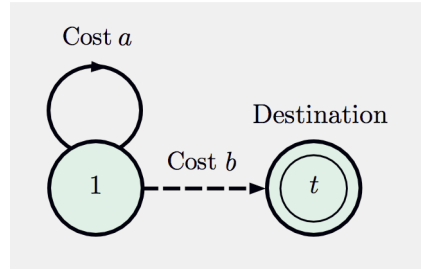


Figure 3.1.1. A deterministic shortest path problem with a single node 1 and a termination node t . At 1 there are two choices; a self-transition, which costs a , and a transition to t , which costs b .

proper, it is called *improper*. Thus there exists a proper policy if and only if each node is connected to t with a path. Furthermore, an improper policy has cost greater than $-\infty$ starting from every initial state if and only if all the cycles of the corresponding subgraph have nonnegative cycle cost.

Let us now get a sense of what may happen by considering the simple one-node example shown in Fig. 3.1.1. Here there is a single state 1 in addition to the termination state t . At state 1 there are two choices: a self-transition, which costs a , and a transition to t , which costs b . The mapping H , abbreviating $J(1)$ with just the scalar J , is

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

There are two policies here: the policy μ that transitions from 1 to t , which is proper, and the policy μ' that self-transitions at state 1, which is improper. We have

$$T_{\mu}J = b, \quad T_{\mu'}J = a + J, \quad J \in \mathfrak{R},$$

and

$$TJ = \min\{b, a + J\}, \quad J \in \mathfrak{R}.$$

Note that for the proper policy μ , the mapping $T_{\mu} : \mathfrak{R} \mapsto \mathfrak{R}$ is a contraction. For the improper policy μ' , the mapping $T_{\mu'} : \mathfrak{R} \mapsto \mathfrak{R}$ is not a contraction, and it has a fixed point within \mathfrak{R} only if $a = 0$, in which case every $J \in \mathfrak{R}$ is a fixed point.

We now consider the optimal cost J^* , the fixed points of T within \mathfrak{R} , and the behavior of the VI and PI methods for different combinations of values of a and b .

- (a) If $a > 0$, the optimal cost, $J^* = b$, is the unique fixed point of T , and the proper policy is optimal.

- (b) If $a = 0$, the set of fixed points of T (within \mathfrak{R}) is the interval $(-\infty, b]$. Here the improper policy is optimal if $b \geq 0$, and the proper policy is optimal if $b \leq 0$ (both policies are optimal if $b = 0$).
- (c) If $a = 0$ and $b > 0$, the proper policy is strictly suboptimal, yet its cost at state 1 (which is b) is a fixed point of T . The optimal cost, $J^* = 0$, lies in the interior of the set of fixed points of T , which is $(-\infty, b]$. Thus the VI method that generates $\{T^k J\}$ starting with $J \neq J^*$ cannot find J^* . In particular if J is a fixed point of T , VI stops at J , while if J is not a fixed point of T (i.e., $J > b$), VI terminates in two iterations at $b \neq J^*$. Moreover, the standard PI method is unreliable in the sense that starting with the suboptimal proper policy μ , it may stop with that policy because $T_\mu J_\mu = b = \min\{b, J_\mu\} = T J_\mu$ (the improper/optimal policy μ' also satisfies $T_{\mu'} J_\mu = T J_\mu$, so a rule for breaking the tie in favor of μ is needed but such a rule may not be obvious in general).
- (d) If $a = 0$ and $b < 0$, the improper policy is strictly suboptimal, and we have $J^* = b$. Here it can be seen that the VI sequence $\{T^k J\}$ converges to J^* for all $J \geq b$, but stops at J for all $J < b$, since the set of fixed points of T is $(-\infty, b]$. Moreover, starting with either the proper policy or the improper policy, the standard form of PI may oscillate, since $T_\mu J_{\mu'} = T J_{\mu'}$ and $T_{\mu'} J_\mu = T J_\mu$, as can be easily verified [the optimal policy μ also satisfies $T_\mu J_\mu = T J_\mu$ but it is not clear how to break the tie; compare also with case (c) above].
- (e) If $a < 0$, the improper policy is optimal and we have $J^* = -\infty$. There are no fixed points of T within \mathfrak{R} , but J^* is the unique fixed point of T within the set $[-\infty, \infty]$. The VI method will converge to J^* starting from any $J \in [-\infty, \infty]$. The PI method will also converge to the optimal policy starting from either policy.

3.1.2 Stochastic Shortest Path Problems

We consider the SSP problem, which was described in Example 1.2.6 and will be revisited in Section 3.5.1. Here a policy is associated with a stationary Markov chain whose states are $1, \dots, n$, plus the cost-free termination state t . The cost of a policy starting at a state x is the sum of the expected cost of its transitions up to reaching t . A policy is said to be *proper*, if in its Markov chain, every state is connected with t with a path of positive probability transitions, and otherwise it is called *improper*. Equivalently, a policy is proper if its Markov chain has t as its unique ergodic state, with all other states being transient.

In deterministic shortest path problems, it turns out that J_μ is always a fixed point of T_μ , and J^* is always a fixed point of T . This is a generic feature of deterministic problems, which was illustrated in Section 1.1 (see Exercise 3.1 for a rigorous proof). However, in SSP problems where the

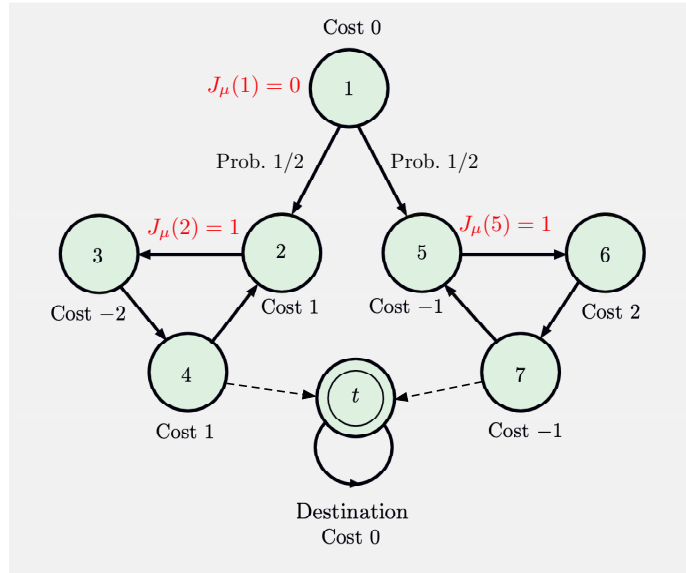


Figure 3.1.2. An example of an improper policy μ , where J_μ is not a fixed point of T_μ . All transitions under μ are shown with solid lines. These transitions are deterministic, except at state 1 where the next state is 2 or 5 with equal probability 1/2. There are additional high cost transitions from nodes 1, 4, and 7 to the destination (shown with broken lines), which create a suboptimal proper policy. We have $J^* = J_\mu$ and J^* is not a fixed point of T .

cost per stage can take both positive and negative values this need not be so, as we will now show with an example due to [BeY16].

Let us consider the problem of Fig. 3.1.2. It involves an improper policy μ , whose transitions are shown with solid lines in the figure, and form the two zero length cycles shown. All the transitions under μ are deterministic, except at state 1 where the successor state is 2 or 5 with equal probability 1/2. The problem has been deliberately constructed so that corresponding costs at the nodes of the two cycles are negatives of each other. As a result, the expected cost at each time period starting from state 1 is 0, implying that the total cost over any number or even infinite number of periods is 0.

Indeed, to verify that $J_\mu(1) = 0$, let c_k denote the cost incurred at time k , starting at state 1, and let $s_N(1) = \sum_{k=0}^{N-1} c_k$ denote the N -step accumulation of c_k starting from state 1. We have

$$s_N(1) = 0 \quad \text{if } N = 1 \text{ or } N = 4 + 3t, t = 0, 1, \dots,$$

$$s_N(1) = 1 \text{ or } s_N(1) = -1 \text{ with probability } 1/2 \text{ each} \\ \text{if } N = 2 + 3t \text{ or } N = 3 + 3t, t = 0, 1, \dots$$

Thus $E\{s_N(1)\} = 0$ for all N , and

$$J_\mu(1) = \limsup_{N \rightarrow \infty} E\{s_N(1)\} = 0.$$

On the other hand, using the definition of J_μ in terms of \limsup , we have

$$J_\mu(2) = J_\mu(5) = 1,$$

(the sequence of N -stage costs undergoes a cycle $\{1, -1, 0, 1, -1, 0, \dots\}$ when starting from state 2, and undergoes a cycle $\{-1, 1, 0, -1, 1, 0, \dots\}$ when starting from state 5). Thus the Bellman equation at state 1,

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

is not satisfied, and J_μ is not a fixed point of T_μ .

The mathematical reason why Bellman's equation $J_\mu = T_\mu J_\mu$ may not hold for stochastic problems is that \limsup may not commute with the expected value that is inherent in T_μ , and the proof argument given for deterministic problems in Section 1.1 breaks down. We can also modify this example so that there are multiple policies. To this end, we can add for $i = 1, 4, 7$, another control that leads from i to t with a cost $c > 1$ (cf. the broken line arcs in Fig. 3.1.2). Then we create a proper policy that is strictly suboptimal, while not affecting J^* , which again is not a fixed point of T .

Let us finally note an anomaly around randomized policies in noncontractive models. The improper policy shown in Fig. 3.1.2 may be viewed as a randomized policy for a deterministic shortest path problem: this is the problem for which at state 1 we must (deterministically) choose one of the two successor states 2 and 5. For this deterministic problem, J^* takes the same values as before for all $i \neq 1$, but it takes the value $J^*(1) = 1$ rather than $J^*(1) = 0$. Thus, remarkably, once we allow randomized policies into the problem, the optimal cost function ceases to be a solution of Bellman's equation and simultaneously the optimal cost at state 1 is improved!

In subsequent sections we will see that favorable results hold in SSP problems where the restricted optimal cost function over just the proper policies is equal to the overall optimal J^* . This can be guaranteed by assumptions that essentially imply that improper policies cannot be optimal (see Sections 3.3 and 3.5.1). We will then see that not only is J^* a fixed point of T , but it is also the unique fixed point (within the class of real-valued functions), and that the VI and PI algorithms yield J^* and an optimal proper policy in the limit.

3.1.3 The Blackmailer's Dilemma

This is a classical example involving a profit maximizing blackmailer. We formulate it as an SSP problem involving cost minimization, with a single state $x = 1$, in addition to the termination state t .

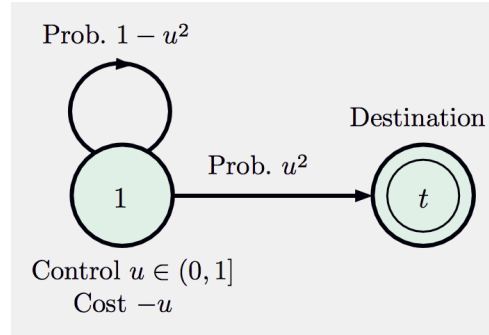


Figure 3.1.3. Transition diagram for the first variant of the blackmailer problem. At state 1, the blackmailer may demand any amount $u \in (0, 1]$. The victim will comply with probability $1 - u^2$ and will not comply with probability u^2 , in which case the process will terminate.

In a first variant of the problem, at state 1, we can choose a control $u \in (0, 1]$, while incurring a cost $-u$; we then move to state t with probability u^2 , and stay in state 1 with probability $1 - u^2$; see Fig. 3.1.3. We may regard u as a demand made by the blackmailer, and state 1 as the situation where the victim complies. State t is arrived at when the victim (permanently) refuses to yield to the blackmailer's demand. The problem then can be viewed as one where the blackmailer tries to maximize his expected total gain by balancing his desire for increased demands (large u) with keeping his victim compliant (small u).

For notational simplicity, let us abbreviate $J(1)$ and $\mu(1)$ with just the scalars J and μ , respectively. Then in terms of abstract DP we have

$$X = \{1\}, \quad U = (0, 1], \quad \bar{J} = 0, \quad H(1, u, J) = -u + (1 - u^2)J,$$

and for every stationary policy μ , we have

$$T_\mu J = -\mu + (1 - \mu^2)J. \quad (3.3)$$

Clearly T_μ , viewed as a mapping from \mathfrak{R} to \mathfrak{R} , is a contraction with modulus $1 - \mu^2$, and its unique fixed point within \mathfrak{R} , J_μ , is the solution of

$$J_\mu = T_\mu J_\mu = -\mu + (1 - \mu^2)J_\mu,$$

which yields

$$J_\mu = -\frac{1}{\mu}.$$

Here all policies are proper in the sense that they lead asymptotically to t with probability 1, and the infimum of J_μ over μ is $-\infty$, implying also

that $J^* = -\infty$. However, there is no optimal stationary policy within the class of proper policies. †

Another interesting fact about this problem is that T_μ is a contraction for all μ . However the theory of contractive models does not apply because there is no uniform modulus of contraction ($\alpha < 1$) that applies simultaneously to all $\mu \in (0, 1]$ [cf. Eq. (3.3)]. As a result, the contraction Assumption 2.1.2 of Section 2.1 does not hold.

Let us now consider Bellman's equation. The mapping T is given by

$$TJ = \inf_{0 < u \leq 1} \{ -u + (1 - u^2)J \},$$

and Bellman's equation is written as

$$J = J - \sup_{0 < u \leq 1} \{ u + u^2J \}.$$

It can be verified that this equation has no real-valued solution. However, $J^* = -\infty$ is a solution within the set of extended real numbers $[-\infty, \infty]$. Moreover the VI method will converge to J^* starting from any $J \in [-\infty, \infty)$. The PI method, starting from any policy μ^0 , will produce the ever improving sequence of policies $\{\mu^k\}$ with $\mu^{k+1} = \mu^k/2$ and $J_{\mu^k} \downarrow J^*$, while μ^k will converge to 0, which is not a feasible policy.

A Second Problem Variant

Consider next a variant of the problem where at state 1, we terminate at no cost with probability u , and stay in state 1 at a cost $-u$ with probability $1 - u$. The control constraint is still $u \in (0, 1]$.

Here we have

$$H(1, u, J) = (1 - u)(-u) + (1 - u)J.$$

It can be seen that for every policy μ , T_μ is again a contraction and we have $J_\mu = \mu - 1$. Thus $J^* = -1$, but again there is no optimal policy, stationary or not. Moreover, T has multiple fixed points: its set of fixed points within \Re is $\{J \mid J \leq -1\}$. Here the VI method will converge to J^* starting from any $J \in [-1, \infty)$. The PI method will produce an ever improving sequence of policies $\{\mu^k\}$ with $J_{\mu^k} \downarrow J^*$, starting from any policy μ^0 , while μ^k will converge to 0, which is not a feasible policy.

† An unusual fact about this problem is that there exists a *nonstationary* policy π^* that is optimal in the sense that $J_{\pi^*} = J^* = -\infty$ (for a proof see [Ber12a], Section 3.2). The underlying intuition is that when the amount demanded u is decreased toward 0, the probability of noncompliance, u^2 , decreases much faster. This fact, however, will not be significant in the context of our analysis.

A Third Problem Variant

Finally, let us again assume that

$$H(1, u, J) = (1 - u)(-u) + (1 - u)J, \quad \forall u \in (0, 1],$$

but also allow, in addition to $u \in (0, 1]$, the choice $u = 0$ that self-transitions to state 1 at a cost c (this is the choice where the blackmailer can forego blackmail for a single period in exchange for a fixed payment $-c$). Here there is the extra (improper) policy μ' that chooses $\mu'(1) = 0$. We have

$$T_{\mu'} J = c + J,$$

and the mapping T is given by

$$TJ = \min \left\{ c + J, \inf_{0 < u \leq 1} \{ -u + u^2 + (1 - u)J \} \right\}. \quad (3.4)$$

Let us consider the optimal policies and the fixed points of T in the two cases where $c \geq 0$ and $c < 0$.

When $c \geq 0$, we have $J^* = -1$, while $J_{\mu'} = \infty$ (if $c > 0$) or $J_{\mu'} = 0$ (if $c = 0$). It can be seen that there is no optimal policy, and that all $J \in (-\infty, -1]$ are fixed points of T , including J^* . Here the VI method will converge to J^* starting from any $J \in [-1, \infty)$. The PI method will produce an ever improving sequence of policies $\{\mu^k\}$, with $J_{\mu^k} \downarrow J^*$. However, μ^k will converge to 0, which is a feasible but strictly suboptimal policy.

When $c < 0$, we have $J_{\mu'} = -\infty$, and the improper policy μ' is optimal. Here the optimal cost over just the proper policies is $\hat{J} = -1$, while $J^* = -\infty$. Moreover \hat{J} is not a fixed point of T , and in fact T has no real-valued fixed points, although J^* is a fixed point. It can be verified that the VI algorithm will converge to J^* starting from any scalar J . Furthermore, starting with a proper policy, the PI method will produce the optimal (improper) policy within a finite number of iterations.

3.1.4 Linear-Quadratic Problems

One of the most important optimal control problems involves a linear system and a cost per stage that is positive semidefinite quadratic in the state and the control. The objective here is roughly to bring the system at or close to the origin, which can be viewed as a cost-free and absorbing state. Thus the problem has a shortest path character, even though the state space is continuous.

Under reasonable assumptions (involving the notions of system controllability and observability; see e.g., [Ber17a], Section 3.1), the problem admits a favorable analysis and an elegant solution: the optimal cost function is positive semidefinite quadratic and the optimal policy is a linear

function of the state. Moreover, Bellman's equation can be equivalently written as an algebraic Riccati equation, which admits a unique solution within the class of nonnegative cost functions.

On the other hand, the favorable results just noted depend on the assumptions and the structure of the linear-quadratic problem. There is no corresponding analysis for more general deterministic continuous-state optimal control problems. Moreover, even for linear-quadratic problems, when the aforementioned controllability and observability assumptions do not hold, the favorable results break down and pathological behavior can occur. This suggests analytical difficulties in more general continuous-state contexts, which we will discuss later in Section 3.5.4.

To illustrate what can happen, consider the scalar system

$$x_{k+1} = \gamma x_k + u_k, \quad x_k \in \mathfrak{R}, u_k \in \mathfrak{R},$$

with $X = U(x) = \mathfrak{R}$, and a cost per stage equal to u^2 . Here we have $J^*(x) = 0$ for all $x \in \mathfrak{R}$, while the policy that applies control $u = 0$ at every state x is optimal. This is reminiscent of the deterministic shortest path problem of Section 3.1.1, for the case where $a = 0$ and there is a zero length cycle. Bellman's equation has the form

$$J(x) = \min_{u \in \mathfrak{R}} \{u^2 + J(\gamma x + u)\}, \quad x \in \mathfrak{R},$$

and it is seen that J^* is a solution. We will now show that there is another solution, which has an interesting interpretation.

Let us assume that $\gamma > 1$ so the system is unstable (the instability of the system is important for the purpose of this example). It is well-known that for linear-quadratic problems the class of quadratic cost functions,

$$S = \{J \mid J(x) = px^2, p \geq 0\},$$

plays a special role. Linear policies of the form

$$\mu(x) = rx,$$

where r is a scalar, also play a special role, particularly the subclass \mathcal{L} of linear policies that are *stable*, in the sense that the closed-loop system

$$x_{k+1} = (\gamma + r)x_k$$

is stable, i.e., $|\gamma + r| < 1$. For such a policy, the generated system trajectory $\{x_k\}$, starting from an initial state x_0 , is $\{(\gamma + r)^k x_0\}$, and the corresponding cost function is quadratic as shown by the following calculation,

$$J_\mu(x_0) = \sum_{k=0}^{\infty} (\mu(x_k))^2 = \sum_{k=0}^{\infty} r^2 x_k^2 = \sum_{k=0}^{\infty} r^2 (\gamma + r)^{2k} x_0^2 = \frac{r^2}{1 - (\gamma + r)^2} x_0^2. \quad (3.5)$$

Note that there is no policy in \mathcal{L} that is optimal, since the optimal policy $\mu^*(x) \equiv 0$ is unstable and does not belong to \mathcal{L} .

Let us consider fixed points of the mapping T ,

$$(TJ)(x) = \inf_{u \in \mathfrak{R}} \{u^2 + J(\gamma x + u)\},$$

within the class of nonnegative quadratic functions S . For $J(x) = px^2$ with $p \geq 0$, we have

$$(TJ)(x) = \inf_{u \in \mathfrak{R}} \{u^2 + p(\gamma x + u)^2\},$$

and by setting to 0 the derivative with respect to u , we see that the infimum is attained at

$$u^* = -\frac{p\gamma}{1+p}x.$$

By substitution into the formula for TJ , we obtain

$$(TJ)(x) = \frac{p\gamma^2}{1+p}x^2. \quad (3.6)$$

Thus the function $J(x) = px^2$ is a fixed point of T if and only if p solves the equation

$$p = \frac{p\gamma^2}{1+p}.$$

This equation has two solutions:

$$p = 0 \quad \text{and} \quad p = \gamma^2 - 1,$$

as shown in Fig. 3.1.4. Thus there are exactly two fixed points of T within S : the functions

$$J^*(x) \equiv 0 \quad \text{and} \quad \hat{J}(x) = (\gamma^2 - 1)x^2.$$

The fixed point \hat{J} has some significance. It turns out to be *the optimal cost function within the subclass \mathcal{L} of linear policies that are stable*. This can be verified by minimizing the expression (3.5) over the parameter r . In particular, by setting to 0 the derivative with respect to r of

$$\frac{r^2}{1 - (\gamma + r)^2},$$

we obtain after a straightforward calculation that it is minimized for $r = (1 - \gamma^2)/\gamma$, which corresponds to the policy

$$\hat{\mu}(x) = \frac{(1 - \gamma^2)}{\gamma}x,$$

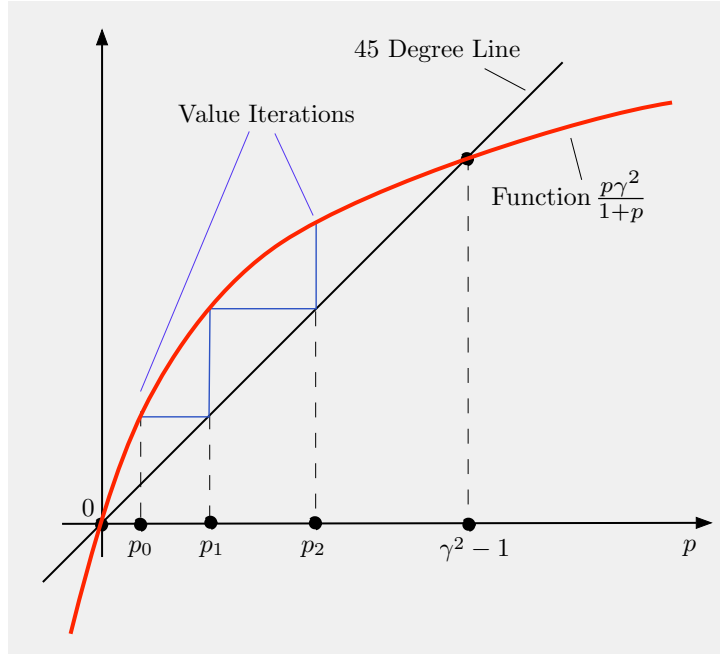


Figure 3.1.4. Illustrating the fixed points of T , and the convergence of the VI algorithm for the one-dimensional linear-quadratic problem.

while from Eq. (3.5), we can verify that

$$J_{\hat{\mu}}(x) = (\gamma^2 - 1)x^2.$$

Thus, we have

$$J_{\hat{\mu}}(x) = \inf_{\mu \in \mathcal{L}} J_{\mu}(x) = \hat{J}(x), \quad x \in \mathfrak{R}.$$

Let us turn now to the VI algorithm starting from a function in S . Using Eq. (3.6), we see that it generates a sequence of functions $J_k \in S$ of the form

$$J_k(x) = p_k x^2,$$

where the sequence $\{p_k\}$ is generated by

$$p_{k+1} = \frac{p_k \gamma^2}{1 + p_k}, \quad k = 0, 1, \dots$$

From Fig. 3.1.4 it can be seen that starting with $p_0 > 0$, the sequence $\{p_k\}$ converges to

$$\hat{p} = \gamma^2 - 1,$$

which corresponds to \hat{J} . In summary, starting from any nonzero function in S , the VI algorithm converges to the optimal cost function \hat{J} over the linear stable policies \mathcal{L} , while starting from the zero function, it converges to the optimal cost function J^* .

Finally, let us consider the PI algorithm starting from a linear stable policy. We first note that given any $\mu \in \mathcal{L}$, i.e.,

$$\mu(x) = rx \quad \text{with} \quad |\gamma + r| < 1,$$

we can compute J_μ as the limit of the VI sequence $\{T_\mu^k J\}$, where J is any function in S , i.e.,

$$J(x) = px^2 \quad \text{with} \quad p \geq 0.$$

This can be verified by writing

$$(T_\mu J)(x) = (r^2 + p(\gamma + r)^2)x^2,$$

and noting that the iteration that maps p to $r^2 + p(\gamma + r)^2$ converges to

$$p_\mu = \frac{r^2}{1 - (\gamma + r)^2},$$

in view of $|\gamma + r| < 1$. Thus,

$$T_\mu^k J \rightarrow J_\mu, \quad \forall \mu \in \mathcal{L}, J \in S.$$

Moreover, we have $J_\mu = T_\mu J_\mu$.

We now use a standard proof argument to show that PI generates a sequence of linear stable policies starting from a linear stable policy. Indeed, we have for all k ,

$$J_{\mu^0} = T_{\mu^0} J_{\mu^0} \geq T J_{\mu^0} = T_{\mu^1} J_{\mu^0} \geq T_{\mu^1}^k J_{\mu^0} \geq T^k \hat{J} = \hat{J},$$

where the second inequality follows by the monotonicity of T_{μ^1} and the third inequality follows from the fact $J_{\mu^0} \geq \hat{J}$. By taking the limit as $k \rightarrow \infty$, we obtain

$$J_{\mu^0} \geq T J_{\mu^0} \geq J_{\mu^1} \geq \hat{J}.$$

It can be verified that μ_1 is a nonzero linear policy, so the preceding relation implies that μ^1 is linear stable. Continuing similarly, it follows that the policies μ^k generated by PI are linear stable and satisfy for all k ,

$$J_{\mu^k} \geq T J_{\mu^k} \geq J_{\mu^{k+1}} \geq \hat{J}.$$

By taking the limit as $k \rightarrow \infty$, we see that the sequence of quadratic functions $\{J_{\mu^k}\}$ converges monotonically to a quadratic function J_∞ , which

is a fixed point of T and satisfies $J_\infty \geq \hat{J}$. Since we have shown that \hat{J} is the only fixed point of T in the range $[\hat{J}, \infty)$, it follows that $J_\infty = \hat{J}$. In summary, the PI algorithm starting from a linear stable policy converges to \hat{J} , the optimal cost function over linear stable policies.

In Section 3.5.4, we will consider a more general multidimensional version of the linear-quadratic problem, using in part the analysis of Section 3.4. We will then explain the phenomena described in this section within a more general setting. We will also see there that the unusual behavior in the present example is due to the fact that there is no penalty for a nonzero state. For example, if the cost per stage is $\delta x^2 + u^2$, where $\delta > 0$, rather than u^2 , then the corresponding Bellman equation has a unique solution with the class of positive semidefinite quadratic functions. We will analyze this case within a more general setting of deterministic optimal control problems in Section 3.5.5.

3.1.5 An Intuitive View of Semicontractive Analysis

In the preceding sections we have demonstrated various aspects of the character of semicontractive analysis in the context of several examples. The salient feature is a class of “well-behaved” policies (e.g., proper policies in shortest path problems, stable policies in linear-quadratic problems), and the restricted optimal cost function \hat{J} over just these policies. The main results we typically derived were that \hat{J} is a fixed point of T , and that the VI and PI algorithms are attracted to \hat{J} , at least from within some suitable class of initial conditions. In the favorable case where $\hat{J} = J^*$, these results hold also for J^* , but in general J^* need not be a fixed point of T .

The central issue of semicontractive analysis is *the choice of a class of “well-behaved” policies $\widehat{\mathcal{M}} \subset \mathcal{M}$ such that the corresponding restricted optimal cost function \hat{J} is a fixed point of T* . Such a choice is often fairly evident, but there are also several systematic approaches to identify a suitable class $\widehat{\mathcal{M}}$ and to show its fixed point property; see the end of Section 3.2.2 for a discussion of various alternatives. As an example, let us introduce a class of policies $\widehat{\mathcal{M}} \subset \mathcal{M}$ for which we assume the following:

- (a) $\widehat{\mathcal{M}}$ is well-behaved with respect to VI: For all $\mu \in \widehat{\mathcal{M}}$ and real-valued functions J , we have

$$J_\mu = T_\mu J_\mu, \quad J_\mu = \lim_{k \rightarrow \infty} T_\mu^k J. \quad (3.7)$$

Moreover J_μ is real-valued.

- (b) $\widehat{\mathcal{M}}$ is well-behaved with respect to PI: For each $\mu \in \widehat{\mathcal{M}}$, any policy μ' such that

$$T_{\mu'} J_\mu = T J_\mu$$

belongs to $\widehat{\mathcal{M}}$, and there exists at least one such μ' .

We can show that \hat{J} is a fixed point of T and obtain our main results with the following line of argument. The first step in this argument is to show that the cost functions of a PI-generated sequence $\{\mu^k\} \subset \widehat{\mathcal{M}}$ (starting from a $\mu^0 \in \widehat{\mathcal{W}}$) are monotonically nonincreasing. Indeed, we have using Eq. (3.7),

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

Using the monotonicity property of $T_{\mu^{k+1}}$, it follows that

$$J_{\mu^k} \geq T J_{\mu^k} \geq \lim_{k \rightarrow \infty} T_{\mu^{k+1}}^k J_{\mu^k} = J_{\mu^{k+1}} \geq \hat{J}, \quad (3.8)$$

where the equality holds by Eq. (3.7), and the rightmost inequality holds since $\mu^{k+1} \in \widehat{\mathcal{M}}$. Thus we obtain

$$J_{\mu^k} \downarrow J_\infty \geq \hat{J},$$

for some function J_∞ .

Now by taking the limit as $k \rightarrow \infty$ in the relation $J_{\mu^k} \geq T J_{\mu^k} \geq J_{\mu^{k+1}}$ [cf. Eq. (3.8)], it follows (under a mild continuity assumption) that J_∞ is a fixed point of T with $J_\infty \geq \hat{J}$.[†] We claim that $J_\infty = \hat{J}$. Indeed we have

$$\hat{J} \leq J_\infty = T^k J_\infty \leq T_\mu^k J_\infty \leq T_\mu^k J_{\mu^0}, \quad \forall \mu \in \widehat{\mathcal{M}}, k = 0, 1, \dots$$

By taking the limit as $k \rightarrow \infty$, and using the fact $\mu \in \widehat{\mathcal{M}}$ [cf. Eq. (3.7)], we obtain $\hat{J} \leq J_\infty \leq J_\mu$ for all $\mu \in \widehat{\mathcal{M}}$. By taking the infimum over $\mu \in \widehat{\mathcal{M}}$, it follows that $J_\infty = \hat{J}$.

Finally, let J be real-valued and satisfy $J \geq \hat{J}$. We claim that $T^k J \rightarrow \hat{J}$. Indeed, since \hat{J} is a fixed point of T , we have

$$T_\mu^k J \geq T^k J \geq T^k \hat{J} = \hat{J}, \quad \forall \mu \in \widehat{\mathcal{M}}, k \geq 0,$$

[†] We elaborate on this argument; see also the proof of Prop. 3.2.4 in the next section. From Eq. (3.8), we have $J_{\mu^k} \geq T J_{\mu^k} \geq T J_\infty$, so by letting $k \rightarrow \infty$, we obtain $J_\infty \geq T J_\infty$. To prove the reverse inequality, we assume that T has the continuity property

$$H(x, u, J_\infty) = \lim_{k \rightarrow \infty} H(x, u, J_{\mu^k}) \geq \lim_{k \rightarrow \infty} (T J_{\mu^k})(x), \quad x \in X, u \in U(x).$$

By taking the limit in Eq. (3.8), we obtain

$$\lim_{k \rightarrow \infty} (T J_{\mu^k})(x) \geq \lim_{k \rightarrow \infty} J_{\mu^{k+1}}(x) = J_\infty(x), \quad x \in X,$$

and from the preceding two relations we have $H(x, u, J_\infty) \geq J_\infty(x)$. By taking the infimum over $u \in U(x)$, it follows that $T J_\infty \geq J_\infty$. Combined with the relation $J_\infty \geq T J_\infty$ shown earlier, this implies that J_∞ is a fixed point of T .

so by taking the limit as $k \rightarrow \infty$ and using Eq. (3.7), we obtain

$$J_\mu \geq \lim_{k \rightarrow \infty} T^k J \geq \hat{J}, \quad \forall \mu \in \widehat{\mathcal{M}}.$$

By taking the infimum over $\mu \in \widehat{\mathcal{M}}$, it follows that $T^k J \rightarrow \hat{J}$, i.e., that VI converges to \hat{J} starting from all initial conditions $J \geq \hat{J}$.

The analysis of the following two sections will be based to a large extent on refinements of the preceding argument. Note that in this argument we have not assumed that $\hat{J} = J^*$, which leaves open the possibility that J^* is not a fixed point of T . Indeed this can happen, as we have seen in the SSP example of Section 3.1.2. Moreover, we have not assumed that \hat{J} is real-valued. In fact \hat{J} may not be real-valued even though all J_μ , $\mu \in \widehat{\mathcal{M}}$, are; see the first variant of the blackmailer problem of Section 3.1.3.

An alternative analytical approach, which does not rely on $\widehat{\mathcal{M}}$ being well-behaved with respect to PI, is given in Section 3.4. The idea there is to introduce a small δ -perturbation to the mapping H and a corresponding “ δ -perturbed” problem. The perturbation is chosen so that the cost function of some policies, the “well-behaved” ones, is minimally affected [say by $O(\delta)$], while the cost function of the policies that are not “well-behaved” is driven to ∞ for some initial states, thereby excluding these policies from optimality. Thus as $\delta \downarrow 0$, the optimal cost function J_δ of the δ -perturbed problem approaches \hat{J} (not J^*). Assuming that J_δ is a solution of the δ -perturbed Bellman equation, and we can then use a limiting argument to show that \hat{J} is a fixed point of T , as well as other results relating to the VI and PI algorithms. The perturbation approach will become more prominent in our semicontractive analysis of Chapter 4 (Sections 4.5 and 4.6), where we will consider “well-behaved” policies that are nonstationary, and thus do not lend themselves to a PI-based analysis.

3.2 SEMICONTRACTIVE MODELS AND REGULAR POLICIES

In the preceding section we illustrated a general pattern of pathologies in noncontractive models, involving the solutions of Bellman’s equation, and the convergence of the VI and PI algorithms. To summarize:

- (a) Bellman’s equation may have multiple solutions (equivalently, T may have multiple fixed points). Often but not always, J^* is a fixed point of T . Moreover, a restricted problem, involving policies that are “well-behaved” (proper in shortest path problems, or linear stable in the linear-quadratic case), may be meaningful and play an important role.
- (b) The optimal cost function over all policies, J^* , may differ from \hat{J} , the optimal cost function over the “well-behaved” policies. Furthermore, it may be that \hat{J} (not J^*) is “well-behaved” from the algorithmic point of view. In particular, \hat{J} is often a fixed point of T , in which

case it is the likely limit of the VI and the PI algorithms, starting from an appropriate set of initial conditions.

In this section we will provide an analytical framework that explains this type of phenomena, and develops the kind of assumptions needed in order to avoid them. We will introduce a concept of regularity that formalizes mathematically the notion of “well-behaved” policy, and we will consider a restricted optimization problem that involves regular policies only. We will show that the optimal cost function of the restricted problem is a fixed point of T under several types of fairly natural assumptions. Moreover, we will show that it can be computed by versions of VI and PI, starting from suitable initial conditions.

Problem Formulation

Let us first introduce formally the model that we will use in this chapter. Compared to the contractive model of Chapter 2, it maintains the monotonicity assumption, but not the contraction assumption.

We introduce the set X of states and the set U of controls, and for each $x \in X$, the nonempty control constraint set $U(x) \subset U$. Since in the absence of the contraction assumption, the cost function J_μ of some policies μ may take infinite values for some states, we will use the set of extended real numbers $\mathfrak{R}^* = \mathfrak{R} \cup \{\infty, -\infty\} = [-\infty, \infty]$. The mathematical operations with ∞ and $-\infty$ are standard and are summarized in Appendix A. We consider the set of all extended real-valued functions $J : X \mapsto \mathfrak{R}^*$, which we denote by $\mathcal{E}(X)$. We also denote by $\mathcal{R}(X)$ the set of real-valued functions $J : X \mapsto \mathfrak{R}$.

As earlier, when we write \lim , \limsup , or \liminf of a sequence of functions we mean it to be pointwise. We also write $J_k \rightarrow J$ to mean that $J_k(x) \rightarrow J(x)$ for each $x \in X$; see Appendix A.

We denote by \mathcal{M} the set of all functions $\mu : X \mapsto U$ with $\mu(x) \in U(x)$, for all $x \in X$, and by Π the set of policies $\pi = \{\mu_0, \mu_1, \dots\}$, where $\mu_k \in \mathcal{M}$ for all k . We refer to a stationary policy $\{\mu, \mu, \dots\}$ simply as μ . We introduce a mapping $H : X \times U \times \mathcal{E}(X) \mapsto \mathfrak{R}^*$ that satisfies the following.

Assumption 3.2.1: (Monotonicity) If $J, J' \in \mathcal{E}(X)$ and $J \leq J'$, then

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

The preceding monotonicity assumption will be in effect throughout this chapter. Consequently, *we will not mention it explicitly in various propositions*. We define the mapping $T : \mathcal{E}(X) \mapsto \mathcal{E}(X)$ by

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

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

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

The monotonicity assumption implies the following properties for all $J, J' \in \mathcal{E}(X)$ and $k = 0, 1, \dots$,

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

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

which we will use extensively in various proof arguments.

We now define cost functions associated with T_μ and T . In Chapter 2 our starting point was to define J_μ and J^* as the unique fixed points of T_μ and T , respectively, based on the contraction assumption used there. However, under our assumptions in this chapter this is not possible, so we use a different definition, which nonetheless is consistent with the one of Chapter 2 (see the discussion of Section 2.1, following Prop. 2.1.2). We introduce a function $\bar{J} \in \mathcal{E}(X)$, and we define the infinite horizon cost of a policy in terms of the limit of its finite horizon costs with \bar{J} being the cost function at the end of the horizon. Note that in the case of the optimal control problems of the preceding section we have taken \bar{J} to be the zero function, $\bar{J}(x) \equiv 0$ [cf. Eq. (3.2)].

Definition 3.2.1: Given a function $\bar{J} \in \mathcal{E}(X)$, for a policy $\pi \in \Pi$ with $\pi = \{\mu_0, \mu_1, \dots\}$, we define the cost function of π by

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

In the case of a stationary policy μ , the cost function of μ is denoted by J_μ and is given by

$$J_\mu(x) = \limsup_{k \rightarrow \infty} (T_\mu^k \bar{J})(x), \quad \forall x \in X.$$

The optimal cost function J^* is given by

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

An optimal policy $\pi^* \in \Pi$ is one for which $J_{\pi^*} = J^*$.

Note two important differences from Chapter 2:

- (1) J_μ is defined in terms of a pointwise lim sup rather than lim, since we don't know whether the limit exists.

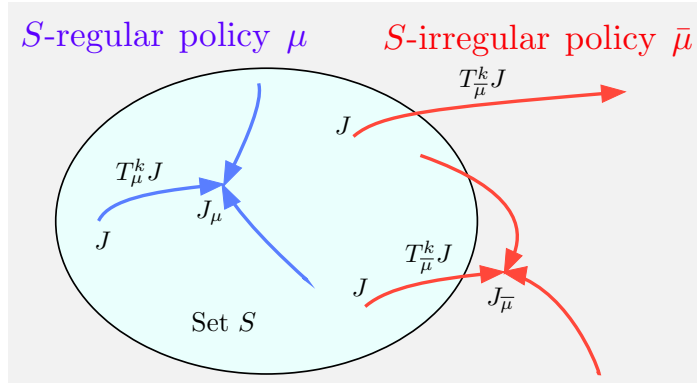


Figure 3.2.1. Illustration of S -regular and S -irregular policies. Policy μ is S -regular because $J_\mu \in S$ and $T_\mu^k J \rightarrow J_\mu$ for all $J \in S$. Policy $\bar{\mu}$ is S -irregular.

- (2) J_π and J_μ in general depend on \bar{J} , so \bar{J} becomes an important part of the problem definition.

Similar to Chapter 2, under the assumptions to be introduced in this chapter, stationary policies will typically turn out to be “sufficient” in the sense that the optimal cost obtained with nonstationary policies that depend on the initial state is matched by the one obtained by stationary ones.

3.2.1 S -Regular Policies

Our objective in this chapter is to construct an analytical framework with a strong connection to fixed point theory, based on the idea of separating policies into those that have “favorable” characteristics and those that do not. Clearly, a favorable property for a policy μ is that J_μ is a fixed point of T_μ . However, J_μ may depend on \bar{J} , even though T_μ does not depend on \bar{J} . It would thus appear that a related favorable property for μ is that J_μ stays the same if \bar{J} is changed arbitrarily within some set S . We express these two properties with the following definition.

Definition 3.2.2: Given a set of functions $S \subset \mathcal{E}(X)$, we say that a stationary policy μ is S -regular if:

- (a) $J_\mu \in S$ and $J_\mu = T_\mu J_\mu$.
- (b) $T_\mu^k J \rightarrow J_\mu$ for all $J \in S$.

A policy that is not S -regular is called S -irregular.

Thus a policy μ is S -regular if the VI algorithm corresponding to μ , $J_{k+1} = T_\mu J_k$, represents a dynamic system that has J_μ as its unique

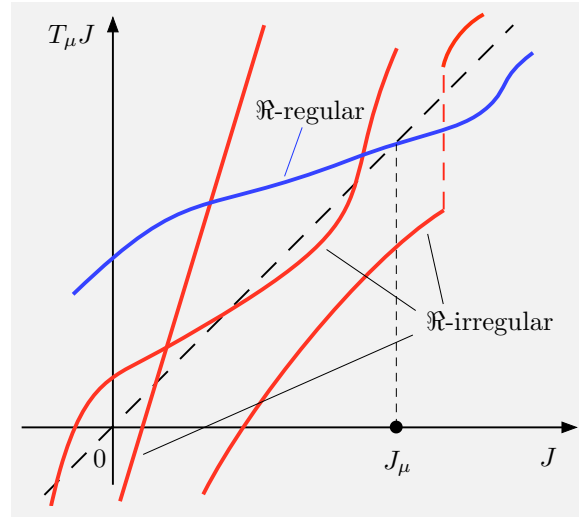


Figure 3.2.2. Illustration of S -regular and S -irregular policies for the case where there is only one state and $S = \mathfrak{R}$. There are three mappings T_μ corresponding to S -irregular policies: one crosses the 45-degree line at multiple points, another crosses at a single point but at an angle greater than 45 degrees, and the third is discontinuous and does not cross at all. The mapping T_μ of the \mathfrak{R} -regular policy has J_μ as its unique fixed point and satisfies $T_\mu^k J \rightarrow J_\mu$ for all $J \in \mathfrak{R}$.

equilibrium within S , and is asymptotically stable in the sense that the iteration converges to J_μ , starting from any $J \in S$ (see Fig. 3.2.1).

For orientation purposes, we note the distinction between the set S and the problem data: S is an analytical device, and is not part of the problem's definition. Its choice, however, can enable analysis and clarify properties of J_μ and J^* . For example, we will later prove local fixed point statements such as

“ J^* is the unique fixed point of T within S ”

or local region of attraction assertions such as

“the VI sequence $\{T^k J\}$ converges to J^* starting from any $J \in S$.”

Results of this type and their proofs depend on the choice of S : they may hold for some choices but not for others.

Generally, with our selection of S we will aim to differentiate between S -regular and S -irregular policies in a manner that produces useful results for the given problem and does not necessitate restrictive assumptions. Examples of sets S that we will use are $\mathcal{R}(X)$, $\mathcal{B}(X)$, $\mathcal{E}(X)$, and subsets of $\mathcal{R}(X)$, $\mathcal{B}(X)$, and $\mathcal{E}(X)$ involving functions J satisfying $J \geq J^*$ or $J \geq \bar{J}$. However, there is a diverse range of other possibilities, so it makes sense to postpone making the choice of S more specific. Figure 3.2.2 illustrates the mappings T_μ of some S -regular and S -irregular policies for the case where

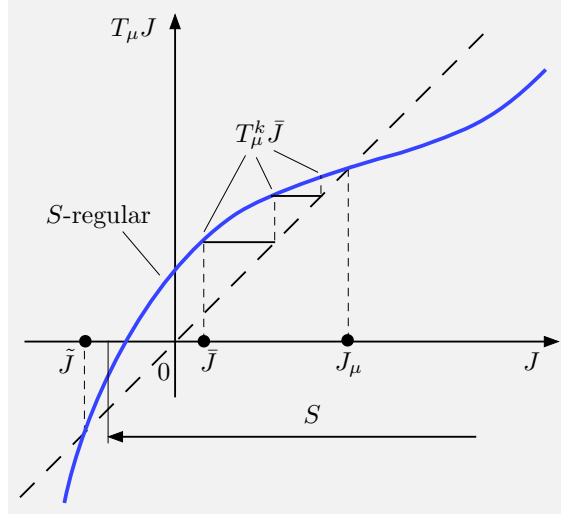


Figure 3.2.3. Illustration of a mapping T_μ where there is only one state and S is a subset of the real line. Here T_μ has two fixed points, J_μ and \tilde{J} . If S is as shown, μ is S -regular. If S is enlarged to include \tilde{J} , μ becomes S -irregular.

there is a single state and $S = \mathfrak{R}$. Figure 3.2.3 illustrates the mapping T_μ of an S -regular policy μ , where T_μ has multiple fixed points, and upon changing S , the policy may become S -irregular.

3.2.2 Restricted Optimization over S -Regular Policies

We will now introduce a restricted optimization framework where S -regular policies are central. Given a nonempty set $S \subset \mathcal{E}(X)$, let \mathcal{M}_S denote the set of policies that are S -regular, and consider optimization over just the set \mathcal{M}_S . The corresponding optimal cost function is denoted J_S^* :

$$J_S^*(x) = \inf_{\mu \in \mathcal{M}_S} J_\mu(x), \quad \forall x \in X. \quad (3.9)$$

We say that μ^* is \mathcal{M}_S -optimal if

$$\mu^* \in \mathcal{M}_S \quad \text{and} \quad J_{\mu^*} = J_S^*.$$

Note that while S is assumed nonempty, it is possible that \mathcal{M}_S is empty. In this case our results will not be useful, but J_S^* is still defined by Eq. (3.9) as $J_S^*(x) \equiv \infty$. This is convenient in various proof arguments.

An important question is whether J_S^* is a fixed point of T and can be obtained by the VI algorithm. Naturally, this depends on the choice of S , but it turns out that reasonable choices can be readily found in several important contexts, so the consequences of J_S^* being a fixed point of T are

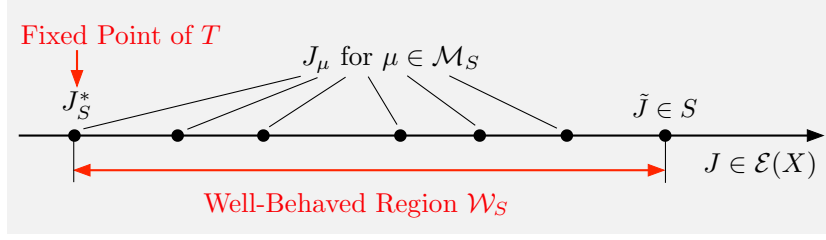


Figure 3.2.4. Interpretation of Prop. 3.2.1, where for illustration purposes, $\mathcal{E}(X)$ is represented by the extended real line. A set $S \subset \mathcal{E}(X)$ such that J_S^* is a fixed point of T , demarcates the well-behaved region \mathcal{W}_S [cf. Eq. (3.10)], within which T has a unique fixed point, and starting from which the VI algorithm converges to J_S^* .

interesting. The next proposition shows that if J_S^* is a fixed point of T , then the VI algorithm is convergent starting from within the set

$$\mathcal{W}_S = \{J \in \mathcal{E}(X) \mid J_S^* \leq J \leq \tilde{J} \text{ for some } \tilde{J} \in S\}, \quad (3.10)$$

which we refer to as the *well-behaved region* (see Fig. 3.2.4). Note that by the definition of S -regularity, the cost functions J_μ , $\mu \in \mathcal{M}_S$, belong to S and hence also to \mathcal{W}_S . The proposition also provides a necessary and sufficient condition for an S -regular policy μ^* to be \mathcal{M}_S -optimal.

Proposition 3.2.1: (Well-Behaved Region Theorem) Given a set $S \subset \mathcal{E}(X)$, assume that J_S^* is a fixed point of T . Then:

- (a) (*Uniqueness of Fixed Point*) If J' is a fixed point of T and there exists $\tilde{J} \in S$ such that $J' \leq \tilde{J}$, then $J' \leq J_S^*$. In particular, if \mathcal{W}_S is nonempty, J_S^* is the unique fixed point of T within \mathcal{W}_S .
- (b) (*VI Convergence*) We have $T^k J \rightarrow J_S^*$ for every $J \in \mathcal{W}_S$.
- (c) (*Optimality Condition*) If μ is S -regular, $J_S^* \in S$, and $T_\mu J_S^* = T J_S^*$, then μ is \mathcal{M}_S -optimal. Conversely, if μ is \mathcal{M}_S -optimal, then $T_\mu J_S^* = T J_S^*$.

Proof: (a) For every $\mu \in \mathcal{M}_S$, we have using the monotonicity of T_μ ,

$$J' = T J' \leq T_\mu J' \leq \dots \leq T_\mu^k J' \leq T_\mu^k \tilde{J}, \quad k = 1, 2, \dots$$

Taking limit as $k \rightarrow \infty$, and using the S -regularity of μ , we obtain $J' \leq J_\mu$ for all $\mu \in \mathcal{M}_S$. Taking the infimum over $\mu \in \mathcal{M}_S$, we have $J' \leq J_S^*$.

Assume that \mathcal{W}_S is nonempty. Then J_S^* is a fixed point of T that belongs to \mathcal{W}_S . To show its uniqueness, let J' be another fixed point that

belongs to \mathcal{W}_S , so that $J_S^* \leq J'$ and there exists $\tilde{J} \in S$ such that $J' \leq \tilde{J}$. By what we have shown so far, $J' \leq J_S^*$, implying that $J' = J_S^*$.

(b) Let $J \in \mathcal{W}_S$, so that $J_S^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$. We have for all $k \geq 1$ and $\mu \in \mathcal{M}_S$,

$$J_S^* = T^k J_S^* \leq T^k J \leq T^k \tilde{J} \leq T_\mu^k \tilde{J},$$

where the equality follows from the fixed point property of J_S^* , while the inequalities follow from the monotonicity and the definition of T . The right-hand side tends to J_μ as $k \rightarrow \infty$, since μ is S -regular and $\tilde{J} \in S$. Hence the infimum over $\mu \in \mathcal{M}_S$ of the limit of the right-hand side tends to the left-hand side J_S^* . It follows that $T^k J \rightarrow J_S^*$.

(c) From the assumptions $T_\mu J_S^* = T J_S^*$ and $T J_S^* = J_S^*$, we have $T_\mu J_S^* = J_S^*$, and since $J_S^* \in S$ and μ is S -regular, we have $J_S^* = J_\mu$. Thus μ is \mathcal{M}_S -optimal. Conversely, if μ is \mathcal{M}_S -optimal, we have $J_\mu = J_S^*$, so that the fixed point property of J_S^* and the S -regularity of μ imply that

$$T J_S^* = J_S^* = J_\mu = T_\mu J_\mu = T_\mu J_S^*.$$

Q.E.D.

Some useful extensions and modified versions of the preceding proposition are given in Exercises 3.2-3.5. Let us illustrate the proposition in the context of the deterministic shortest path example of Section 3.1.1.

Example 3.2.1

Consider the deterministic shortest path example of Section 3.1.1 for the case where there is a zero length cycle ($a = 0$), and let S be the real line \mathfrak{R} . There are two policies: μ which moves from state 1 to the destination at cost b , and μ' which stays at state 1 at cost 0. We use $X = \{1\}$ (i.e., we do not include t in X , since all function values of interest are 0 at t). Then by abbreviating function values $J(1)$ with J , we have

$$J_\mu = b, \quad J_{\mu'} = 0, \quad J^* = \min\{b, 0\};$$

cf. Fig. 3.2.5. The corresponding mappings T_μ , $T_{\mu'}$, and T are

$$T_\mu J = b, \quad T_{\mu'} J = J, \quad J = T J = \min\{b, J\}, \quad J \in \mathcal{E}(X),$$

and the initial function \tilde{J} is taken to be 0. It can be seen from the definition of S -regularity that μ is S -regular, while the policy μ' is not. The cost functions J_μ , $J_{\mu'}$, and J^* are fixed points of the corresponding mappings, but the sets of fixed points of $T_{\mu'}$ and T within S are \mathfrak{R} and $(-\infty, b]$, respectively. Moreover, $J_S^* = J_\mu = b$, so J_S^* is a fixed point of T and Prop. 3.2.1 applies.

The figure also shows the well-behaved regions for the two cases $b > 0$ and $b < 0$. It can be seen that the results of Prop. 3.2.1 are consistent with the discussion of Section 3.1.1. In particular, the VI algorithm fails when

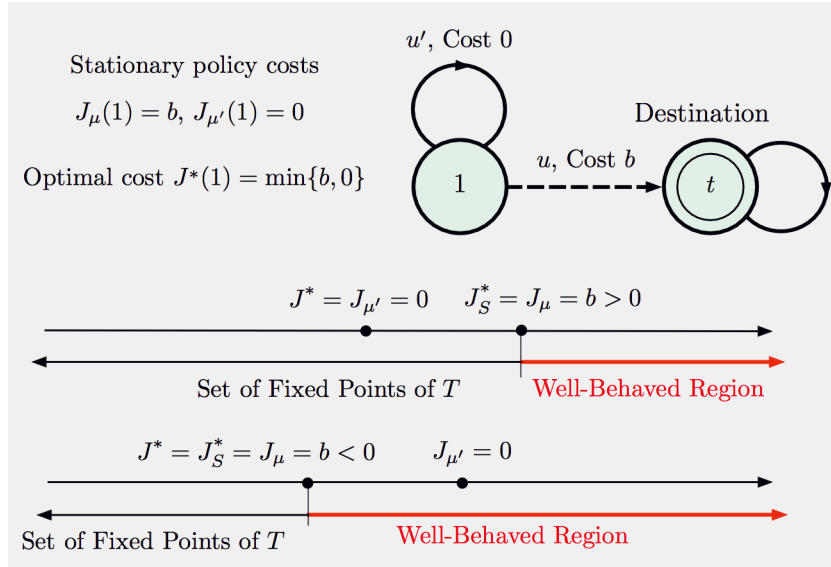


Figure 3.2.5. The well-behaved region of Eq. (3.10) for the deterministic shortest path example of Section 3.1.1 when there is a zero length cycle ($a = 0$). For $S = \mathfrak{R}$, the policy μ is S -regular, while the policy μ' is not. The figure illustrates the two cases where $b > 0$ and $b < 0$.

started outside the well-behaved region, while when started from within the region, it is attracted to J_S^* rather than to J^* .

Let us now discuss some of the fine points of Prop. 3.2.1. The salient assumption of the proposition is that J_S^* is a fixed point of T . Depending on the choice of S , this may or may not be true, and much of the subsequent analysis in this chapter is geared towards the development of approaches to choose S so that J_S^* is a fixed point of T and has some other interesting properties. As an illustration of the range of possibilities, consider the three variants of the blackmailer problem of Section 3.1.3 for the choice $S = \mathfrak{R}$:

- (a) In the first variant, we have $J^* = J_S^* = -\infty$, and J_S^* is a fixed point of T that lies outside S . Here parts (a) and (b) of Prop. 3.2.1 apply. However, part (c) does not apply (even though we have $T_\mu J_S^* = T J_S^*$ for all policies μ) because $J_S^* \notin S$, and in fact there is no \mathcal{M}_S -optimal policy. In the subsequent analysis, we will see that the condition $J_S^* \in S$ plays an important role in being able to assert existence of an \mathcal{M}_S -optimal policy (see the subsequent Props. 3.2.5 and 3.2.6).
- (b) In the second variant, we have $J^* = J_S^* = -1$, and J_S^* is a fixed point of T that lies within S . Here parts (a) and (b) of Prop. 3.2.1 apply, but part (c) still does not apply because there is no S -regular μ such

that $T_\mu J_S^* = T J_S^*$, and in fact there is no M_S -optimal policy.

- (c) In the third variant with $c < 0$, we have $J^* = -\infty$, $J_S^* = -1$, and J_S^* is not a fixed point of T . Thus Prop. 3.2.1 does not apply, and in fact we have $T^k J \rightarrow J^*$ for every $J \in \mathcal{W}_S$ (and not $T^k J \rightarrow J_S^*$).

Another fine point is that Prop. 3.2.1(b) asserts convergence of the VI algorithm to J_S^* only for initial conditions J satisfying $J_S^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$. For an illustrative example of an S -regular μ , where $\{T_\mu^k J\}$ does not converge to J_μ starting from some $J \geq J_\mu$ that lies outside S , consider a case where there is a single state and a single policy μ that is S -regular, so $J_S^* = J_\mu$. Suppose that $T_\mu : \mathfrak{R} \mapsto \mathfrak{R}$ has two fixed points: J_μ and another fixed point $J' > J_\mu$. Let

$$\tilde{J} = (J_\mu + J')/2, \quad S = (-\infty, \tilde{J}],$$

and assume that T_μ is a contraction mapping within S (an example of this type can be easily constructed graphically). Then starting from any $J \in S$, we have $T^k J \rightarrow J_\mu$, so that μ is S -regular. However, since J' is a fixed point of T , the sequence $\{T^k J'\}$ stays at J' and does not converge to J_μ . The difficulty here is that $\mathcal{W}_S = [J_\mu, \tilde{J}]$ and $J' \notin \mathcal{W}_S$.

Still another fine point is that if there exists an \mathcal{M}_S -optimal policy μ , we have $J_S^* = T_\mu J_S^*$ (since $J_S^* = J_\mu$ and μ is S -regular), but this does not guarantee that J_S^* is a fixed point of T , which is essential for Prop. 3.2.1. This can be seen from an example given in Fig. 3.2.6, where there exists an \mathcal{M}_S -optimal policy, but both J_S^* and J^* are not fixed points of T (in this example the \mathcal{M}_S -optimal policy is also overall optimal so $J_S^* = J^*$). In particular, starting from J_S^* , the VI algorithm converges to some $J' \neq J_S^*$ that is a fixed point of T .

Convergence Rate when a Contractive Policy is \mathcal{M}_S -Optimal

In many contexts where Prop. 3.2.1 applies, there exists an \mathcal{M}_S -optimal policy μ such that T_μ is a contraction with respect to a weighted sup-norm. This is true for example in the shortest path problem to be discussed in Section 3.5.1. In such cases, the rate of convergence of VI to J_S^* is linear, as shown in the following proposition.

Proposition 3.2.2: (Convergence Rate of VI) Let S be equal to $\mathcal{B}(X)$, the space of all functions over X that are bounded with respect to a weighted sup-norm $\|\cdot\|_v$ corresponding to a positive function $v : X \mapsto \mathfrak{R}$. Assume that J_S^* is a fixed point of T , and that there exists an \mathcal{M}_S -optimal policy μ such that T_μ is a contraction with respect to $\|\cdot\|_v$, with modulus of contraction β . Then $J_S^* \in \mathcal{B}(X)$, $\mathcal{W}_S \subset \mathcal{B}(X)$, and

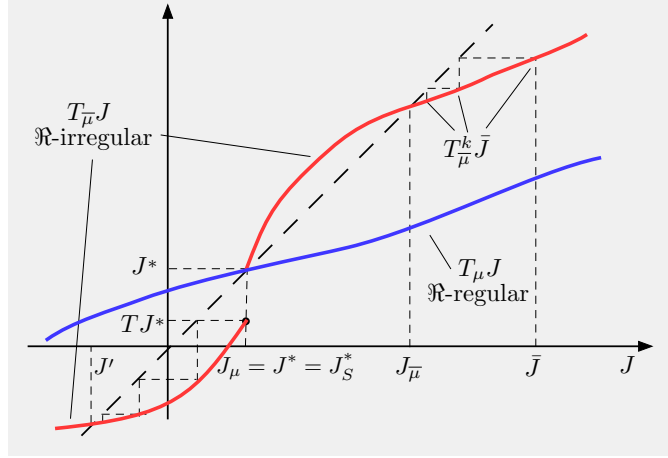


Figure 3.2.6. Illustration of why the assumption that J_S^* is a fixed point of T is essential for Prop. 3.2.1. In this example there is only one state and $S = \mathfrak{R}$. There are two stationary policies: μ for which T_μ is a contraction, so μ is \mathfrak{R} -regular, and $\bar{\mu}$ for which $T_{\bar{\mu}}$ has multiple fixed points, so $\bar{\mu}$ is \mathfrak{R} -irregular. Moreover, $T_{\bar{\mu}}$ is discontinuous from above at J_μ as shown. Here, it can be verified that $T_{\mu_0} \cdots T_{\mu_k} \bar{J} \geq J_\mu$ for all μ_0, \dots, μ_k and k , so that $J_\pi \geq J_\mu$ for all π and the S -regular policy μ is optimal, so $J_S^* = J^*$. However, as can be seen from the figure, we have $J_S^* = J^* \neq T J^* = T J_S^*$. Moreover, starting at J_S^* , the VI sequence $T^k J_S^*$ converges to J' , the fixed point of T shown in the figure, and all parts of Prop. 3.2.1 fail.

$$\|TJ - J_S^*\|_v \leq \beta \|J - J_S^*\|_v, \quad \forall J \in \mathcal{W}_S. \quad (3.11)$$

Moreover, we have

$$\|J - J_S^*\|_v \leq \frac{1}{1 - \beta} \sup_{x \in X} \frac{J(x) - (TJ)(x)}{v(x)}, \quad \forall J \in \mathcal{W}_S. \quad (3.12)$$

Proof: Since μ is S -regular and $S = \mathcal{B}(X)$, we have $J_S^* = J_\mu \in \mathcal{B}(X)$ as well as $\mathcal{W}_S \subset \mathcal{B}(X)$. By using the \mathcal{M}_S -optimality of μ and Prop. 3.2.1(c),

$$J_S^* = T_\mu J_S^* = T J_S^*,$$

so for all $x \in X$ and $J \in \mathcal{W}_S$,

$$\frac{(TJ)(x) - J_S^*(x)}{v(x)} \leq \frac{(T_\mu J)(x) - (T_\mu J_S^*)(x)}{v(x)} \leq \beta \max_{x \in X} \frac{J(x) - J_S^*(x)}{v(x)},$$

where the second inequality holds by the contraction property of T_μ . By taking the supremum of the left-hand side over $x \in X$, and by using the fact $TJ \geq T J_S^* = J_S^*$ for all $J \in \mathcal{W}_S$, we obtain Eq. (3.11).

By using again the relation $T_\mu J_S^* = TJ_S^*$, we have for all $x \in X$ and all $J \in \mathcal{W}_S$,

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

By taking the supremum of both sides over x , we obtain Eq. (3.12). **Q.E.D.**

Approaches to Show that J_S^* is a Fixed Point of T

The critical assumption of Prop. 3.2.1 is that J_S^* is a fixed point of T . For a specific application, this must be proved with a separate analysis after a suitable set S is chosen. To this end, we will provide several approaches that guide the choice of S and facilitate the analysis.

One approach applies to problems where J^* is generically a fixed point of T , in which case for every set S such that $J_S^* = J^*$, Prop. 3.2.1 applies and shows that J^* can be obtained by the VI algorithm starting from any $J \in \mathcal{W}_S$. Exercise 3.1 provides some conditions that guarantee that J^* is a fixed point of T . These conditions can be verified in wide classes of problems such as deterministic models. Sections 3.5.4 and 3.5.5 illustrate this approach. Other important models where J^* is guaranteed to be a fixed point of T are the monotone increasing and monotone decreasing models of Section 4.3. We will discuss the application of Prop. 3.2.1 and other related results to these models in Chapter 4.

In the present chapter the approach for showing that J_S^* is a fixed point of T will be mostly based on the PI algorithm; cf. the discussion of Section 3.1.5. An alternative and complementary approach is the perturbation-based analysis to be given in Section 3.4. This approach will be applied to a variety of problems in Section 3.5, and will also be prominent in Sections 4.5 and 4.6 of the next chapter.

3.2.3 Policy Iteration Analysis of Bellman's Equation

We will develop a PI-based approach for showing that J_S^* is a fixed point of T . The approach is applicable under assumptions that guarantee that there is a sequence $\{\mu^k\}$ of S -regular policies that can be generated by PI. The significance of S -regularity of all μ^k lies in that *the corresponding cost function sequence $\{J_{\mu^k}\}$ belongs to the well-behaved region of Eq. (3.10), and is monotonically nonincreasing* (see the subsequent Prop. 3.2.3). Under an additional mild technical condition, the limit of this sequence is a fixed point of T and is in fact equal to J_S^* (see the subsequent Prop. 3.2.4).

Let us consider the standard form of the PI algorithm, which starts with a policy μ^0 and generates a sequence $\{\mu^k\}$ of stationary policies according to

$$T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k}, \quad k = 0, 1, \dots \quad (3.13)$$

This iteration embodies both the policy evaluation step, which computes J_{μ^k} in some way, and the policy improvement step, which computes $\mu^{k+1}(x)$ as a minimum over $u \in U(x)$ of $H(x, u, J_{\mu^k})$ for each $x \in X$. Of course, to be able to carry out the policy improvement step, there should be enough assumptions to guarantee that the minimum is attained for every x . One such assumption is that $U(x)$ is a finite set for each $x \in X$. A more general assumption, which applies to the case where the constraint sets $U(x)$ are infinite, will be given in Section 3.3.

The evaluation of the cost function J_μ of a policy μ may be done by solving the equation $J_\mu = T_\mu J_\mu$, which holds when μ is an S -regular policy. An important fact is that if the PI algorithm generates a sequence $\{\mu^k\}$ consisting exclusively of S -regular policies, then not only the policy evaluation is facilitated through the equation $J_\mu = T_\mu J_\mu$, but also the sequence of cost functions $\{J_{\mu^k}\}$ is monotonically nonincreasing, as we will show next.

Proposition 3.2.3: (Policy Improvement Under S -Regularity)

Given a set $S \subset \mathcal{E}(X)$, assume that $\{\mu^k\}$ is a sequence generated by the PI algorithm (3.13) that consists of S -regular policies. Then

$$J_{\mu^k} \geq J_{\mu^{k+1}}, \quad k = 0, 1, \dots$$

Proof: Using the S -regularity of μ^k and Eq. (3.13), we have

$$J_{\mu^k} = T_{\mu^k} J_{\mu^k} \geq T J_{\mu^k} = T_{\mu^{k+1}} J_{\mu^k}. \quad (3.14)$$

By using the monotonicity of $T_{\mu^{k+1}}$, we obtain

$$J_{\mu^k} \geq T J_{\mu^k} \geq \lim_{m \rightarrow \infty} T_{\mu^{k+1}}^m J_{\mu^k} = J_{\mu^{k+1}}, \quad (3.15)$$

where the equation on the right holds since μ^{k+1} is S -regular and $J_{\mu^k} \in S$ (in view of the S -regularity of μ^k). **Q.E.D.**

The preceding proposition shows that if a sequence of S -regular policies $\{\mu^k\}$ is generated by PI, the corresponding cost function sequence $\{J_{\mu^k}\}$ is monotonically nonincreasing and hence converges to a limit J_∞ . Under mild conditions, we will show that J_∞ is a fixed point of T and is equal to J_S^* . This is important as it brings to bear Prop. 3.2.1, and the

associated results on VI convergence and optimality conditions. Let us first formalize the property that the PI algorithm can generate a sequence of S -regular policies.

Definition 3.2.3: (Weak PI Property) We say that a set $S \subset \mathcal{E}(X)$ has the *weak PI property* if there exists a sequence of S -regular policies that can be generated by the PI algorithm [i.e., a sequence $\{\mu^k\}$ that satisfies Eq. (3.13) and consists of S -regular policies].

Note a fine point here. For a given starting policy μ^0 , there may be many different sequences $\{\mu^k\}$ that can be generated by PI [i.e., satisfy Eq. (3.13)]. While the weak PI property guarantees that some of these consist of S -regular policies exclusively, there may be some that do not. The policy improvement property shown in Prop. 3.2.3 holds for the former sequences, but not necessarily for the latter. The following proposition provides the basis for showing that J_S^* is a fixed point of T based on the weak PI property.

Proposition 3.2.4: (Weak PI Property Theorem) Given a set $S \subset \mathcal{E}(X)$, assume that:

- (1) S has the weak PI property.
- (2) For each sequence $\{J_m\} \subset S$ with $J_m \downarrow J$ for some $J \in \mathcal{E}(X)$, we have

$$H(x, u, J) = \lim_{m \rightarrow \infty} H(x, u, J_m), \quad \forall x \in X, u \in U(x). \quad (3.16)$$

Then:

- (a) J_S^* is a fixed point of T and the conclusions of Prop. 3.2.1 hold.
- (b) (*PI Convergence*) Every sequence of S -regular policies $\{\mu^k\}$ that can be generated by PI satisfies $J_{\mu^k} \downarrow J_S^*$. If in addition the set of S -regular policies is finite, there exists $\bar{k} \geq 0$ such that $\mu^{\bar{k}}$ is \mathcal{M}_S -optimal.

Proof: (a) Let $\{\mu^k\}$ be a sequence of S -regular policies generated by the PI algorithm (there exists such a sequence by the weak PI property). Then by Prop. 3.2.3, the sequence $\{J_{\mu^k}\}$ is monotonically nonincreasing and must converge to some $J_\infty \geq J_S^*$.

We first show that J_∞ is a fixed point of T . Indeed, from Eq. (3.14),

we have

$$J_{\mu^k} \geq TJ_{\mu^k} \geq TJ_{\infty},$$

so by letting $k \rightarrow \infty$, we obtain $J_{\infty} \geq TJ_{\infty}$. From Eq. (3.15) we also have $TJ_{\mu^k} \geq J_{\mu^{k+1}}$. Taking the limit in this relation as $k \rightarrow \infty$, we obtain

$$\lim_{k \rightarrow \infty} (TJ_{\mu^k})(x) \geq \lim_{k \rightarrow \infty} J_{\mu^{k+1}}(x) = J_{\infty}(x), \quad x \in X.$$

By using Eq. (3.16) we also have

$$H(x, u, J_{\infty}) = \lim_{k \rightarrow \infty} H(x, u, J_{\mu^k}) \geq \lim_{k \rightarrow \infty} (TJ_{\mu^k})(x), \quad x \in X, u \in U(x).$$

By combining the preceding two relations, we obtain

$$H(x, u, J_{\infty}) \geq J_{\infty}(x), \quad x \in X, u \in U(x),$$

and by taking the infimum of the left-hand side over $u \in U(x)$, it follows that $TJ_{\infty} \geq J_{\infty}$. Thus J_{∞} is a fixed point of T .

Finally, we show that $J_{\infty} = J_S^*$. Indeed, since $J_S^* \leq J_{\mu^k}$, we have

$$J_S^* \leq J_{\infty} = T^k J_{\infty} \leq T_{\mu}^k J_{\infty} \leq T_{\mu}^k J_{\mu^0}, \quad \forall \mu \in \mathcal{M}_S, k = 0, 1, \dots$$

By taking the limit as $k \rightarrow \infty$, and using the fact $\mu \in \mathcal{M}_S$ and $J_{\mu^0} \in S$, it follows that $J_S^* \leq J_{\infty} \leq J_{\mu}$, for all $\mu \in \mathcal{M}_S$. By taking the infimum over $\mu \in \mathcal{M}_S$, it follows that $J_{\infty} = J_S^*$, so J_S^* is a fixed point of T .

(b) The limit of $\{J_{\mu^k}\}$ was shown to be equal to J_S^* in the preceding proof. Moreover, the finiteness of \mathcal{M}_S and the policy improvement property of Prop. 3.2.3 imply that some $\mu^{\bar{k}}$ is \mathcal{M}_S -optimal. **Q.E.D.**

Note that under the weak PI property, the preceding proposition shows convergence of the PI-generated cost functions J_{μ^k} to J_S^* but not necessarily to J^* . An example of this type of behavior was seen in the linear-quadratic problem of Section 3.1.4 (where S is the set of nonnegative quadratic functions). Let us describe another example, which shows in addition that under the weak PI property, it is possible for the PI algorithm to generate a nonmonotonic sequence of policy cost functions that includes both optimal and strictly suboptimal policies.

Example 3.2.2: (Weak PI Property and the Deterministic Shortest Path Example)

Consider the deterministic shortest path example of Section 3.1.1 for the case where there is a zero length cycle ($a = 0$), and let S be the real line \mathbb{R} , as in Example 3.2.1. There are two policies: μ which moves from state 1 to the destination at cost b , and μ' which stays at state 1 at cost 0. Starting with the S -regular policy μ , the PI algorithm generates the policy that corresponds

to the minimum in $TJ_\mu = \min\{b, J_\mu\} = \min\{b, b\}$. Thus both the S -regular policy μ and the S -irregular μ' can be generated at the first iteration. This means that the weak PI property holds (although the strong PI property, which will be introduced shortly, does not hold). Indeed, consistent with Prop. 3.2.4, we have that $J_S^* = J_\mu = b$ is a fixed point of T , in fact the only fixed point of T in the well-behaved region $\{J \mid J \geq b\}$.

An interesting fact here is that when $b < 0$, and PI is started with the optimal S -regular policy μ , then it may generate the S -irregular policy μ' , and from that policy, it will generate μ again. Thus the weak PI property does not preclude the PI algorithm from generating a policy sequence that includes S -irregular policies, with corresponding policy cost functions that are oscillating.

Let us also revisit the blackmailer example of Section 3.1.3. In the first variant of that example, when $S = \mathfrak{R}$, all policies are S -regular, the weak PI property holds, and Prop. 3.2.4 applies. In this case, PI will generate a sequence of S -regular policies that converges to $J_S^* = -\infty$, which is a fixed point of T , consistent with Prop. 3.2.4 (even though $J_S^* \notin S$ and there is no \mathcal{M}_S -optimal policy).

Analysis Under the Strong PI Property

Proposition 3.2.4(a) does not guarantee that *every* sequence $\{\mu^k\}$ generated by the PI algorithm satisfies $J_{\mu^k} \downarrow J_S^*$. This is true only for the sequences that consist of S -regular policies. We know that when the weak PI property holds, there exists at least one such sequence, but PI can also generate sequences that contain S -irregular policies, even when started with an S -regular policy, as we have seen in Example 3.2.2. We thus introduce a stronger type of PI property, which will guarantee stronger conclusions.

Definition 3.2.4: (Strong PI Property) We say that a set $S \subset \mathcal{E}(X)$ has the *strong PI property* if:

- (a) There exists at least one S -regular policy.
- (b) For every S -regular policy μ , any policy μ' such that $T_{\mu'}J_\mu = TJ_\mu$ is S -regular, and there exists at least one such μ' .

The strong PI property implies that every sequence that can be generated by PI starting from an S -regular policy consists exclusively of S -regular policies. Moreover, there exists at least one such sequence. Hence the strong PI property implies the weak PI property. Thus if the strong PI property holds together with the mild continuity condition (2) of Prop. 3.2.4, it follows that J_S^* is a fixed point of T and Prop. 3.2.1 applies. We will see that the strong PI property implies additional results, relating to the uniqueness of the fixed point of T .

The following proposition provides conditions guaranteeing that S has the strong PI property. The salient feature of these conditions is that they preclude optimality of an S -irregular policy [see condition (4) of the proposition].

Proposition 3.2.5: (Verifying the Strong PI Property) Given a set $S \subset \mathcal{E}(X)$, assume that:

- (1) $J(x) < \infty$ for all $J \in S$ and $x \in X$.
- (2) There exists at least one S -regular policy.
- (3) For every $J \in S$ there exists a policy μ such that $T_\mu J = TJ$.
- (4) For every $J \in S$ and S -irregular policy μ , there exists a state $x \in X$ such that

$$\limsup_{k \rightarrow \infty} (T_\mu^k J)(x) = \infty. \quad (3.17)$$

Then:

- (a) A policy μ satisfying $T_\mu J \leq J$ for some function $J \in S$ is S -regular.
- (b) S has the strong PI property.

Proof: (a) By the monotonicity of T_μ , we have $\limsup_{k \rightarrow \infty} T_\mu^k J \leq J$, and since by condition (1), $J(x) < \infty$ for all x , it follows from Eq. (3.17) that μ is S -regular.

(b) In view of condition (3), it will suffice to show that for every S -regular policy μ , any policy μ' such that $T_{\mu'} J_\mu = TJ_\mu$ is also S -regular. Indeed we have

$$T_{\mu'} J_\mu = TJ_\mu \leq T_\mu J_\mu = J_\mu,$$

so μ' is S -regular by part (a). **Q.E.D.**

For an example where the assumptions of the preceding proposition fail, consider the linear-quadratic problem of Section 3.1.4. Here S is the set of nonnegative quadratic functions, but the optimal policy μ^* that applies control $u = 0$ at all states is S -irregular, since we do not have $T_{\mu^*}^k J \rightarrow J_{\mu^*} = 0$ for J equal to a positive quadratic function, while condition (4) of the proposition does not hold. Thus we cannot conclude that the strong PI property holds in the absence of additional analysis.

We next derive some of the implications of the strong PI property regarding fixed properties of J_S^* . In particular, we show that if $J_S^* \in S$,

then J_S^* is the unique fixed point of T within S . This result will be the starting point for the analysis of Section 3.3.

Proposition 3.2.6: (Strong PI Property Theorem) Let S satisfy the conditions of Prop. 3.2.5.

- (a) (*Uniqueness of Fixed Point*) If T has a fixed point within S , then this fixed point is equal to J_S^* .
- (b) (*Fixed Point Property and Optimality Condition*) If $J_S^* \in S$, then J_S^* is the unique fixed point of T within S and the conclusions of Prop. 3.2.1 hold. Moreover, every policy μ that satisfies $T_\mu J_S^* = T J_S^*$ is \mathcal{M}_S -optimal and there exists at least one such policy.
- (c) (*PI Convergence*) If for each sequence $\{J_m\} \subset S$ with $J_m \downarrow J$ for some $J \in \mathcal{E}(X)$, we have

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

then J_S^* is a fixed point of T , and every sequence $\{\mu^k\}$ generated by the PI algorithm starting from an S -regular policy μ^0 satisfies $J_{\mu^k} \downarrow J_S^*$. Moreover, if the set of S -regular policies is finite, there exists $\bar{k} \geq 0$ such that $\mu^{\bar{k}}$ is \mathcal{M}_S -optimal.

Proof: (a) Let $J' \in S$ be a fixed point of T . Then for every $\mu \in \mathcal{M}_S$ and $k \geq 1$, we have $J' = T^k J' \leq T_\mu^k J'$. By taking the limit as $k \rightarrow \infty$, we have $J' \leq J_\mu$, and by taking the infimum over $\mu \in \mathcal{M}_S$, we obtain $J' \leq J_S^*$. For the reverse inequality, let μ' be such that $J' = T J' = T_{\mu'} J'$ [cf. condition (3) of Prop. 3.2.5]. Then by Prop. 3.2.5(a), it follows that μ' is S -regular, and since $J' \in S$, by the definition of S -regularity, we have $J' = J_{\mu'} \geq J_S^*$, showing that $J' = J_S^*$.

(b) For every $\mu \in \mathcal{M}_S$ we have $J_\mu \geq J_S^*$, so that

$$J_\mu = T_\mu J_\mu \geq T_\mu J_S^* \geq T J_S^*.$$

Taking the infimum over all $\mu \in \mathcal{M}_S$, we obtain $J_S^* \geq T J_S^*$. Let μ be a policy such that $T J_S^* = T_\mu J_S^*$, [there exists one by condition (3) of Prop. 3.2.5, since we assume that $J_S^* \in S$]. The preceding relations yield $J_S^* \geq T_\mu J_S^*$, so by Prop. 3.2.5(a), μ is S -regular. Therefore, we have

$$J_S^* \geq T J_S^* = T_\mu J_S^* \geq \lim_{k \rightarrow \infty} T_\mu^k J_S^* = J_\mu \geq J_S^*,$$

where the second equality holds since μ was proved to be S -regular, and $J_S^* \in S$ by assumption. Hence equality holds throughout in the above

relation, which proves that J_S^* is a fixed point of T (implying the conclusions of Prop. 3.2.1) and that μ is \mathcal{M}_S -optimal.

(c) Since the strong PI property [which holds by Prop. 3.2.5(b)] implies the weak PI property, the result follows from Prop. 3.2.4(b). **Q.E.D.**

The preceding proposition does not address the question whether J^* is a fixed point of T , and does not guarantee that VI converges to J_S^* or J^* starting from every $J \in S$. We will consider both of these issues in the next section. Note, however, a consequence of part (a): if J^* is known to be a fixed point of T and $J^* \in S$, then $J^* = J_S^*$.

Let us now illustrate with examples some of the fine points of the analysis. For an example where the preceding proposition does not apply, consider the first two variants of the blackmailer problem of Section 3.1.3. Let us take $S = \mathfrak{R}$, so that all policies are S -regular and the strong PI property holds. In the first variant of the problem, we have $J^* = J_S^* = -\infty$, and consistent with Prop. 3.2.4, J_S^* is a fixed point of T . However, $J_S^* \notin S$, and T has no fixed points within S . On the other hand if we change S to be $[-\infty, \infty)$, there are no S -regular policies at all, since for $J = -\infty \in S$, we have $T_\mu^k J = -\infty < J_\mu$ for all μ . As noted earlier, both Props. 3.2.1 and 3.2.4 do apply. In the second variant of the problem, we have $J^* = J_S^* = -1$, while the set of fixed points of T within S is $(-\infty, -1]$, so Prop. 3.2.6(a) fails. The reason is that the condition (3) of Prop. 3.2.5 is violated.

The next example, when compared with Example 3.2.2, illustrates the difference in PI-related results obtained under the weak and the strong PI properties. Moreover it highlights a generic difficulty in applying PI, even if the strong PI property holds, namely that an initial S -regular policy must be available.

Example 3.2.3: (Strong PI Property and the Deterministic Shortest Path Example)

Consider the deterministic shortest path example of Section 3.1.1 for the case where the cycle has positive length ($a > 0$), and let S be the real line \mathfrak{R} , as in Example 3.2.1. The two policies are: μ which moves from state 1 to the destination at cost b and is S -regular, and μ' which stays at state 1 at cost a , which is S -irregular. However, μ' has infinite cost and satisfies Eq (3.17). As a result, Prop. 3.2.5 applies and the strong PI property holds. Consistent with Prop. 3.2.6, J_S^* is the unique fixed point of T within S .

Turning now to the PI algorithm, we see that starting from the S -regular μ , which is optimal, it stops at μ , consistent with Prop. 3.2.6(c). However, starting from the S -irregular policy μ' the policy evaluation portion of the PI algorithm must be able to deal with the infinite cost values associated with μ' . This is a generic difficulty in applying PI to problems where there are irregular policies: we either need to know an initial S -regular policy, or

appropriately modify the PI algorithm. See the discussions in Sections 3.5.1 and 3.6.2.

3.2.4 Optimistic Policy Iteration and λ -Policy Iteration

We have already shown the validity of the VI and PI algorithms for computing J_S^* (subject to various assumptions, and restrictions involving the starting points). In this section and the next one we will consider some additional algorithmic approaches that can be justified based on the preceding analysis.

An Optimistic Form of PI

Let us consider an optimistic variant of PI, where policies are evaluated inexactly, with a finite number of VIs. In particular, this algorithm starts with some $J_0 \in \mathcal{E}(X)$ such that $J_0 \geq TJ_0$, and generates a sequence $\{J_k, \mu^k\}$ according to

$$T_{\mu^k} J_k = TJ_k, \quad J_{k+1} = T_{\mu^k}^{m_k} J_k, \quad k = 0, 1, \dots, \quad (3.18)$$

where m_k is a positive integer for each k .

The following proposition shows that optimistic PI converges under mild assumptions to a fixed point of T , independently of any S -regularity framework. However, when such a framework is introduced, and the sequence generated by optimistic PI generates a sequence of S -regular policies, then the algorithm converges to J_S^* , which is in turn a fixed point of T , similar to the PI convergence result under the weak PI property; cf. Prop. 3.2.4(b).

Proposition 3.2.7: (Convergence of Optimistic PI) Let $J_0 \in \mathcal{E}(X)$ be a function such that $J_0 \geq TJ_0$, and assume that:

- (1) For all $\mu \in \mathcal{M}$, we have $J_\mu = T_\mu J_\mu$, and for all $J \in \mathcal{E}(X)$ with $J \leq J_0$, there exists $\bar{\mu} \in \mathcal{M}$ such that $T_{\bar{\mu}} J = TJ$.
- (2) For each sequence $\{J_m\} \subset \mathcal{E}(X)$ with $J_m \downarrow J$ for some $J \in \mathcal{E}(X)$, we have

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

Then the optimistic PI algorithm (3.18) is well defined and the following hold:

- (a) The sequence $\{J_k\}$ generated by the algorithm satisfies $J_k \downarrow J_\infty$, where J_∞ is a fixed point of T .

(b) If for a set $S \subset \mathcal{E}(X)$, the sequence $\{\mu^k\}$ generated by the algorithm consists of S -regular policies, and we have $J_k \in S$ for all k , then $J_k \downarrow J_S^*$ and J_S^* is a fixed point of T .

Proof: (a) Condition (1) guarantees that the sequence $\{J_k, \mu^k\}$ is well defined in the following argument. We have

$$\begin{aligned} J_0 &\geq TJ_0 = T_{\mu^0}J_0 \geq T_{\mu^0}^{m_0}J_0 = J_1 \\ &\geq T_{\mu^0}^{m_0+1}J_0 = T_{\mu^0}J_1 \geq TJ_1 = T_{\mu^1}J_1 \geq \cdots \geq J_2, \end{aligned} \quad (3.19)$$

and continuing similarly, we obtain

$$J_k \geq TJ_k \geq J_{k+1}, \quad k = 0, 1, \dots \quad (3.20)$$

Thus $J_k \downarrow J_\infty$ for some J_∞ .

The proof that J_∞ is a fixed point of T is similar to the case of the PI algorithm (3.13) in Prop. 3.2.4. In particular, from Eq. (3.20), we have $J_k \geq TJ_\infty$, and by taking the limit as $k \rightarrow \infty$,

$$J_\infty \geq TJ_\infty.$$

For the reverse inequality, we use Eq. (3.20) to write

$$H(x, u, J_k) \geq (TJ_k)(x) \geq J_\infty(x), \quad \forall x \in X, u \in U(x).$$

By taking the limit as $k \rightarrow \infty$ and using condition (2), we have that

$$H(x, u, J_\infty) \geq J_\infty(x), \quad \forall x \in X, u \in U(x).$$

By taking the infimum over $u \in U(x)$, we obtain

$$TJ_\infty \geq J_\infty,$$

thus showing that $TJ_\infty = J_\infty$.

(b) In the case where all the policies μ^k are S -regular and $\{J_k\} \subset S$, from Eq. (3.19), we have $J_{k+1} \geq J_{\mu^k}$ for all k , so it follows that

$$J_\infty = \lim_{k \rightarrow \infty} J_k \geq \liminf_{k \rightarrow \infty} J_{\mu^k} \geq J_S^*.$$

We will also show that the reverse inequality holds, so that $J_\infty = J_S^*$. Indeed, for every S -regular policy μ and all $k \geq 0$, we have

$$J_\infty = T^k J_\infty \leq T_\mu^k J_\infty \leq T_\mu^k J_0,$$

from which by taking limit as $k \rightarrow \infty$ and using the assumption $J_0 \in S$, we obtain

$$J_\infty \leq \lim_{k \rightarrow \infty} T_\mu^k J_0 = J_\mu, \quad \forall \mu \in \mathcal{M}_S.$$

Taking infimum over $\mu \in \mathcal{M}_S$, we have $J_\infty \leq J_S^*$. Thus, $J_\infty = J_S^*$, and by using the properties of J_∞ proved in part (a), the result follows. **Q.E.D.**

Note that, in general, the fixed point J_∞ in Prop. 3.2.7(a) need not be equal to J_S^* or J^* . As an illustration, consider the shortest path Example 3.2.1 with $S = \mathfrak{R}$, and $a = 0$, $b > 0$. Then if $0 < J_0 < b$, it can be seen that $J_k = J_0$ for all k , so $J^* = 0 < J_\infty$ and $J_\infty < J_S^* = b$.

λ -Policy Iteration

We next consider λ -policy iteration (λ -PI for short), which was described in Section 2.5. It involves a scalar $\lambda \in (0, 1)$ and it is defined by

$$T_{\mu^k} J_k = T J_k, \quad J_{k+1} = T_{\mu^k}^{(\lambda)} J_k, \quad (3.21)$$

where for any policy μ and scalar $\lambda \in (0, 1)$, $T_\mu^{(\lambda)}$ is the multistep mapping discussed in Section 1.2.5:

$$(T_\mu^{(\lambda)} J)(x) = (1 - \lambda) \sum_{t=0}^{\infty} \lambda^t (T_\mu^{t+1} J)(x), \quad x \in X. \quad (3.22)$$

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

We will also assume that T_μ and $T_\mu^{(\lambda)}$ commute, i.e.,

$$T_\mu(T_\mu^{(\lambda)} J) = T_\mu^{(\lambda)}(T_\mu J), \quad \forall \mu \in \mathcal{M}, J \in \mathcal{E}(X). \quad (3.23)$$

This assumption is commonly satisfied in DP problems where T_μ is linear, such as the stochastic optimal control problem of Example 1.2.1.

To compare the λ -PI method (3.21) with the exact PI algorithm (3.13), note that by the analysis of Section 1.2.5 (see also Exercise 1.2), the mapping $T_{\mu^k}^{(\lambda)}$ is an extrapolated version of the proximal mapping for solving the fixed point equation $J = T_{\mu^k} J$. *Thus in λ -PI, the policy evaluation phase is done approximately with a single iteration of the (extrapolated) proximal algorithm.*

As noted in Section 2.5, the λ -PI and the optimistic PI methods are related. The reason is that both mappings $T_{\mu^k}^{(\lambda)}$ and $T_{\mu^k}^{m_k}$ involve multiple applications of the VI mapping T_{μ^k} : a fixed number m_k in the latter case, and a geometrically weighted infinite number in the former case [cf. Eq. (3.22)]. *Thus λ -PI and optimistic PI use VI in alternative ways to evaluate J_{μ^k} approximately.*

Since λ -PI and optimistic PI are related, it is not surprising that they have the same type of convergence properties. We have the following proposition, which is similar to Prop. 3.2.7.

Proposition 3.2.8: (Convergence of λ -PI) Let $J_0 \in \mathcal{E}(X)$ be a function such that $J_0 \geq TJ_0$, assume that the limit in the series (3.22) is well defined and Eq. (3.23) holds. Assume further that:

- (1) For all $\mu \in \mathcal{M}$, we have $J_\mu = T_\mu J_\mu$, and for all $J \in \mathcal{E}(X)$ with $J \leq J_0$, there exists $\bar{\mu} \in \mathcal{M}$ such that $T_{\bar{\mu}} J = TJ$.
- (2) For each sequence $\{J_m\} \subset \mathcal{E}(X)$ with $J_m \downarrow J$ for some $J \in \mathcal{E}(X)$, we have

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

Then the λ -PI algorithm (3.21) is well defined and the following hold:

- (a) A sequence $\{J_k\}$ generated by the algorithm satisfies $J_k \downarrow J_\infty$, where J_∞ is a fixed point of T .
- (b) If for a set $S \subset \mathcal{E}(X)$, the sequence $\{\mu^k\}$ generated by the algorithm consists of S -regular policies, and we have $J_k \in S$ for all k , then $J_k \downarrow J_S^*$ and J_S^* is a fixed point of T .

Proof: (a) We first note that for all $\mu \in \mathcal{M}$ and $J \in \mathcal{E}(X)$ such that $J \geq T_\mu J$, we have

$$T_\mu J \geq T_\mu^{(\lambda)} J.$$

This follows from the power series expansion (3.22) and the fact that $J \geq T_\mu J$ implies that

$$T_\mu J \geq T_\mu^2 J \geq \dots \geq T_\mu^{m+1} J, \quad \forall m \geq 1.$$

Using also the monotonicity of T_μ and $T_\mu^{(\lambda)}$, and Eq. (3.23), we have that

$$J \geq T_\mu J \quad \Rightarrow \quad T_\mu J \geq T_\mu^{(\lambda)} J \geq T_\mu^{(\lambda)}(T_\mu J) = T_\mu(T_\mu^{(\lambda)} J).$$

The preceding relation and our assumptions imply that

$$\begin{aligned} J_0 \geq TJ_0 = T_{\mu_0} J_0 &\geq T_{\mu_0}^{(\lambda)} J_0 = J_1 \\ &\geq T_{\mu_0}(T_{\mu_0}^{(\lambda)} J_0) = T_{\mu_0} J_1 \geq TJ_1 = T_{\mu_1} J_1 \geq \dots \geq J_2. \end{aligned}$$

Continuing similarly, we obtain $J_k \geq TJ_k \geq J_{k+1}$ for all k . Thus $J_k \downarrow J_\infty$ for some J_∞ . From this point, the proof that J_∞ is a fixed point of T is similar to the one of Prop. 3.2.7(a).

(b) Similar to the proof of Prop. 3.2.7(b). **Q.E.D.**

3.2.5 A Mathematical Programming Approach

Let us finally consider an alternative to the VI and PI approaches. It is based on the fact that J_S^* is an upper bound to all functions $J \in S$ that satisfy $J \leq TJ$, as we will show shortly. We will exploit this fact to obtain a method to compute J_S^* that is based on solution of a related mathematical programming problem. We have the following proposition.

Proposition 3.2.9: Given a set $S \subset \mathcal{E}(X)$, for all functions $J \in S$ satisfying $J \leq TJ$, we have $J \leq J_S^*$.

Proof: If $J \in S$ and $J \leq TJ$, by repeatedly applying T to both sides and using the monotonicity of T , we obtain $J \leq T^k J \leq T_\mu^k J$ for all k and S -regular policies μ . Taking the limit as $k \rightarrow \infty$, we obtain $J \leq J_\mu$, so by taking the infimum over $\mu \in \mathcal{M}_S$, we obtain $J \leq J_S^*$. **Q.E.D.**

Thus if J_S^* is a fixed point of T , it is the “largest” fixed point of T , and we can use the preceding proposition to compute J_S^* by maximizing an appropriate monotonically increasing function of J subject to the constraints $J \in S$ and $J \leq TJ$.[†] This approach, when applied to finite-spaces Markovian decision problems, is usually referred to as the *linear programming solution method*, since then the resulting optimization problem is a linear program (see e.g., see Exercise 2.5 for the case of contractive problems or [Ber12a], Ch. 2).

Suppose now that $X = \{1, \dots, n\}$, $S = \mathfrak{R}^n$, and J_S^* is a fixed point of T . Then Prop. 3.2.9 shows that $J_S^* = (J_S^*(1), \dots, J_S^*(n))$ is the unique solution of the following optimization problem in the vector $J = (J(1), \dots, J(n))$:

$$\begin{aligned} & \text{maximize} && \sum_{i=1}^n \beta_i J(i) \\ & \text{subject to} && J(i) \leq H(i, u, J), \quad i = 1, \dots, n, \quad u \in U(i), \end{aligned}$$

where β_1, \dots, β_n are any positive scalars. If H is linear in J and each $U(i)$ is a finite set, this is a linear program, which can be solved by using standard linear programming methods.

[†] For the mathematical programming approach to apply, it is sufficient that $J_S^* \leq TJ_S^*$. However, we generally have $J_S^* \geq TJ_S^*$ (this follows by writing

$$J_\mu = T_\mu J_\mu \geq TJ_\mu \geq TJ_S^*, \quad \forall \mu \in \mathcal{M}_S,$$

and taking the infimum over all $\mu \in \mathcal{M}_S$), so the condition $J_S^* \leq TJ_S^*$ is equivalent to J_S^* being a fixed point of T .

3.3 IRREGULAR POLICIES/INFINITE COST CASE

The results of the preceding section guarantee (under various conditions) that J_S^* is a fixed point of T and can be found by the VI and PI algorithms, but they do not assert that J^* is a fixed point of T or that $J^* = J_S^*$. In this section we address these issues by carrying the strong PI property analysis further with some additional assumptions. A critical part of the analysis is based on the strong PI property theorem of Prop. 3.2.6. We first collect all of our assumptions. We will verify these assumptions in the context of several applications in Section 3.5.

Assumption 3.3.1: We have a subset $S \subset \mathcal{R}(X)$ satisfying the following:

- (a) S contains \bar{J} , and has the property that if J_1, J_2 are two functions in S , then S contains all functions J with $J_1 \leq J \leq J_2$.
- (b) The function $J_S^* = \inf_{\mu \in \mathcal{M}_S} J_\mu$ belongs to S .
- (c) For each S -irregular policy μ and each $J \in S$, there is at least one state $x \in X$ such that

$$\limsup_{k \rightarrow \infty} (T_\mu^k J)(x) = \infty.$$

- (d) The control set U is a metric space, and the set

$$\{u \in U(x) \mid H(x, u, J) \leq \lambda\}$$

is compact for every $J \in S$, $x \in X$, and $\lambda \in \mathfrak{R}$.

- (e) For each sequence $\{J_m\} \subset S$ with $J_m \uparrow J$ for some $J \in S$,

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

- (f) For each function $J \in S$, there exists a function $J' \in S$ such that $J' \leq J$ and $J' \leq TJ'$.

An important restriction of the preceding assumption is that S consists of real-valued functions. This underlies the mechanism of differentiating between S -regular and S -irregular policies that is embodied in Assumption 3.3.1(c).

The conditions (b) and (c) of the preceding assumption have been introduced in Props. 3.2.5 and 3.2.6 in the context of the strong PI property-related analysis. New conditions, not encountered earlier, are (a), (e), and

(f). They will be used to assert that $J^* = J_S^*$, that J^* is the unique fixed point of T within S , and that the VI and PI algorithms have improved convergence properties compared with the ones of Section 3.2.

Note that in the case where S is the set of real-valued functions $\mathcal{R}(X)$ and $\bar{J} \in \mathcal{R}(X)$, condition (a) is automatically satisfied, while condition (e) is typically verified easily. The verification of condition (f) may be nontrivial in some cases. We postpone the discussion of this issue for later (see the subsequent Prop. 3.3.2).

The main result of this section is the following proposition, which provides results that are almost as strong as the ones for contractive models.

Proposition 3.3.1: Let Assumption 3.3.1 hold. Then:

- (a) The optimal cost function J^* is the unique fixed point of T within the set S .
- (b) We have $T^k J \rightarrow J^*$ for all $J \in S$.
- (c) A policy μ is optimal if and only if $T_\mu J^* = T J^*$. Moreover, there exists an optimal policy that is S -regular.
- (d) For any $J \in S$, if $J \leq T J$ we have $J \leq J^*$, and if $J \geq T J$ we have $J \geq J^*$.
- (e) If in addition for each sequence $\{J_m\} \subset S$ with $J_m \downarrow J$ for some $J \in S$, we have

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

then every sequence $\{\mu^k\}$ generated by the PI algorithm starting from an S -regular policy μ^0 satisfies $J_{\mu^k} \downarrow J^*$. Moreover, if the set of S -regular policies is finite, there exists $\bar{k} \geq 0$ such that $\mu^{\bar{k}}$ is optimal.

We will prove Prop. 3.3.1 through a sequence of lemmas, which delineate the assumptions that are needed for each part of the proof. Our first lemma guarantees that starting from an S -regular policy, the PI algorithm is well defined.

Lemma 3.3.1: Let Assumption 3.3.1(d) hold. For every $J \in S$, there exists a policy μ such that $T_\mu J = T J$.

Proof: For any $x \in X$ with $(T J)(x) < \infty$, let $\{\lambda_m(x)\}$ be a decreasing

scalar sequence with

$$\lambda_m(x) \downarrow \inf_{u \in U(x)} H(x, u, J).$$

The set

$$U_m(x) = \{u \in U(x) \mid H(x, u, J) \leq \lambda_m(x)\},$$

is nonempty, and by assumption it is compact. The set of points attaining the infimum of $H(x, u, J)$ over $U(x)$ is $\bigcap_{m=0}^{\infty} U_m(x)$, and is therefore nonempty. Let u_x be a point in this intersection. Then we have

$$H(x, u_x, J) \leq \lambda_m(x), \quad \forall m \geq 0. \quad (3.24)$$

Consider now a policy μ , which is formed by the point u_x for x with $(TJ)(x) < \infty$, and by any point $u_x \in U(x)$ for x with $(TJ)(x) = \infty$. Taking the limit in Eq. (3.24) as $m \rightarrow \infty$ shows that μ satisfies $(T_\mu J)(x) = (TJ)(x)$ for x with $(TJ)(x) < \infty$. For x with $(TJ)(x) = \infty$, we also have trivially $(T_\mu J)(x) = (TJ)(x)$, so $T_\mu J = TJ$. **Q.E.D.**

The next two lemmas follow from the analysis of the preceding section.

Lemma 3.3.2: Let Assumption 3.3.1(c) hold. A policy μ that satisfies $T_\mu J \leq J$ for some $J \in S$ is S -regular.

Proof: This is Prop. 3.2.5(a). **Q.E.D.**

Lemma 3.3.3: Let Assumption 3.3.1(b),(c),(d) hold. Then:

- (a) The function J_S^* of Assumption 3.3.1(b) is the unique fixed point of T within S .
- (b) Every policy μ satisfying $T_\mu J_S^* = TJ_S^*$ is optimal within the set of S -regular policies, i.e., μ is S -regular and $J_\mu = J_S^*$. Moreover, there exists at least one such policy.

Proof: This is Prop. 3.2.6(b) [Assumption 3.3.1(d) guarantees that for every $J \in S$, there exists a policy μ such that $T_\mu J = TJ$ (cf. Lemma 3.3.1), which is part of the assumptions of Prop. 3.2.6]. **Q.E.D.**

Let us also prove the following technical lemma, which makes use of the additional part (e) of Assumption 3.3.1.

Lemma 3.3.4: Let Assumption 3.3.1(b),(c),(d),(e) hold. Then if $J \in S$, $\{T^k J\} \subset S$, and $T^k J \uparrow J_\infty$ for some $J_\infty \in S$, we have $J_\infty = J_S^*$.

Proof: We fix $x \in X$, and consider the sets

$$U_k(x) = \left\{ u \in U(x) \mid H(x, u, T^k J) \leq J_\infty(x) \right\}, \quad k = 0, 1, \dots, \quad (3.25)$$

which are compact by assumption. Let $u_k \in U(x)$ be such that

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

(such a point exists by Lemma 3.3.1). Then $u_k \in U_k(x)$.

For every k , consider the sequence $\{u_i\}_{i=k}^\infty$. Since $T^k J \uparrow J_\infty$, it follows using the monotonicity of H , that for all $i \geq k$,

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

Therefore from the definition (3.25), we have $\{u_i\}_{i=k}^\infty \subset U_k(x)$. Since $U_k(x)$ is compact, all the limit points of $\{u_i\}_{i=k}^\infty$ belong to $U_k(x)$ and at least one limit point exists. Hence the same is true for the limit points of the whole sequence $\{u_i\}$. Thus if \tilde{u} is a limit point of $\{u_i\}$, we have

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

By Eq. (3.25), this implies that

$$H(x, \tilde{u}, T^k J) \leq J_\infty(x), \quad k = 0, 1, \dots$$

Taking the limit as $k \rightarrow \infty$ and using Assumption 3.3.1(e), we obtain

$$(TJ_\infty)(x) \leq H(x, \tilde{u}, J_\infty) \leq J_\infty(x).$$

Thus, since x was chosen arbitrarily within X , we have $TJ_\infty \leq J_\infty$. To show the reverse inequality, we write $T^k J \leq J_\infty$, apply T to this inequality, and take the limit as $k \rightarrow \infty$, so that $J_\infty = \lim_{k \rightarrow \infty} T^{k+1} J \leq TJ_\infty$. It follows that $J_\infty = TJ_\infty$. Since $J_\infty \in S$ by assumption, by applying Lemma 3.3.3(a) we have $J_\infty = J_S^*$. **Q.E.D.**

We are now ready to prove Prop. 3.3.1 by making use of the additional parts (a) and (f) of Assumption 3.3.1.

Proof of Prop. 3.3.1: (a), (b) We will first prove that $T^k J \rightarrow J_S^*$ for all $J \in S$, and we will use this to prove that $J_S^* = J^*$ and that there exists

an optimal S -regular policy. Thus parts (a) and (b), together with the existence of an optimal S -regular policy, will be shown simultaneously.

We fix $J \in S$, and choose $J' \in S$ such that $J' \leq J$ and $J' \leq TJ'$ [cf. Assumption 3.3.1(f)]. By the monotonicity of T , we have $T^k J' \uparrow J_\infty$ for some $J_\infty \in \mathcal{E}(X)$. Let μ be an S -regular policy such that $J_\mu = J_S^*$ [cf. Lemma 3.3.3(b)]. Then we have, using again the monotonicity of T ,

$$J_\infty = \lim_{k \rightarrow \infty} T^k J' \leq \limsup_{k \rightarrow \infty} T^k J \leq \lim_{k \rightarrow \infty} T_\mu^k J = J_\mu = J_S^*. \quad (3.26)$$

Since J' and J_S^* belong to S , and $J' \leq T^k J' \leq J_\infty \leq J_S^*$, Assumption 3.3.1(a) implies that $\{T^k J'\} \subset S$, and $J_\infty \in S$. From Lemma 3.3.4, it then follows that $J_\infty = J_S^*$. Thus equality holds throughout in Eq. (3.26), proving that $\lim_{k \rightarrow \infty} T^k J = J_S^*$.

There remains to show that $J_S^* = J^*$ and that there exists an optimal S -regular policy. To this end, we note that by the monotonicity Assumption 3.2.1, for any policy $\pi = \{\mu_0, \mu_1, \dots\}$, we have

$$T_{\mu_0} \cdots T_{\mu_{k-1}} \bar{J} \geq T^k \bar{J}.$$

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

$$J_\pi \geq \lim_{k \rightarrow \infty} T^k \bar{J} = J_S^*,$$

where the equality follows since $T^k J \rightarrow J_S^*$ for all $J \in S$ (as shown earlier), and $\bar{J} \in S$ [cf. Assumption 3.3.1(a)]. Thus for all $\pi \in \Pi$, $J_\pi \geq J_S^* = J_\mu$, implying that the policy μ that is optimal within the class of S -regular policies is optimal over all policies, and that $J_S^* = J^*$.

(c) If μ is optimal, then $J_\mu = J^* \in S$, so by Assumption 3.3.1(c), μ is S -regular and therefore $T_\mu J_\mu = J_\mu$. Hence,

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

Conversely, if

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

μ is S -regular (cf. Lemma 3.3.2), so $J^* = \lim_{k \rightarrow \infty} T_\mu^k J^* = J_\mu$. Therefore, μ is optimal.

(d) If $J \in S$ and $J \leq TJ$, by repeatedly applying T to both sides and using the monotonicity of T , we obtain $J \leq T^k J$ for all k . Taking the limit as $k \rightarrow \infty$ and using the fact $T^k J \rightarrow J^*$ [cf. part (b)], we obtain $J \leq J^*$. The proof that $J \geq TJ$ implies $J \geq J^*$ is similar.

(e) As in the proof of Prop. 3.2.4(b), the sequence $\{J_{\mu^k}\}$ converges monotonically to a fixed point of T , call it J_∞ . Since J_∞ lies between $J_{\mu^0} \in S$ and $J_S^* \in S$, it must belong to S , by Assumption 3.3.1(a). Since the only

fixed point of T within S is J^* [cf. part (a)], it follows that $J_\infty = J^*$.
Q.E.D.

Note that Prop. 3.3.1(d) provides the basis for a solution method based on mathematical programming; cf. the discussion following Prop. 3.2.9. Here is an example where Prop. 3.3.1 does not apply, because the compactness condition of Assumption 3.3.1(d) fails.

Example 3.3.1

Consider the third variant of the blackmailer problem (Section 3.1.3) for the case where $c > 0$ and $S = \mathfrak{R}$. Then the (nonoptimal) S -irregular policy $\bar{\mu}$ whereby at each period, the blackmailer may demand no payment ($u = 0$) and pay cost $c > 0$, has infinite cost ($J_{\bar{\mu}} = \infty$). However, T has multiple fixed points within the real line, namely the set $(-\infty, -1]$. By choosing $S = \mathfrak{R}$, we see that the uniqueness of fixed point part (a) of Prop. 3.3.1 fails because the compactness part (d) of Assumption 3.3.1 is violated (all other parts of the assumption are satisfied). In this example, the results of Prop. 3.2.1 apply with $S = \mathfrak{R}$, because J_S^* is a fixed point of T .

In various applications, the verification of part (f) of Assumption 3.3.1 may not be simple. The following proposition is useful in several contexts, including some that we will encounter in Section 3.5.

Proposition 3.3.2: Let S be equal to $R_b(X)$, the subset of $\mathcal{R}(X)$ that consists of functions J that are bounded above and below, in the sense that for some $b \in \mathfrak{R}$, we have $|J(x)| \leq b$ for all $x \in X$. Let parts (b), (c), and (d) of Assumption 3.3.1 hold, and assume further that for all scalars $r > 0$, we have

$$TJ_S^* - re \leq T(J_S^* - re), \quad (3.27)$$

where e is the unit function, $e(x) \equiv 1$. Then part (f) of Assumption 3.3.1 also holds.

Proof: Let $J \in R_b(x)$, and let $r > 0$ be a scalar such that $J_S^* - re \leq J$ [such a scalar exists since $J_S^* \in R_b(x)$ by Assumption 3.3.1(b)]. Define $J' = J_S^* - re$, and note that by Lemma 3.3.3, J_S^* is a fixed point of T . By using Eq. (3.27), we have

$$J' = J_S^* - re = TJ_S^* - re \leq T(J_S^* - re) = TJ',$$

while $J' \in R_b(x)$, thus proving part (f) of Assumption 3.3.1. **Q.E.D.**

The relation (3.27) is satisfied among others in stochastic optimal control problems (cf. Example 1.2.1), where

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

with $\alpha \in (0, 1]$. Note that application of the preceding proposition is facilitated when X is a finite set, in which case $R_b(X) = \mathcal{R}(X)$. This fact will be used in the context of some of the applications of Sections 3.5.1-3.5.4.

3.4 IRREGULAR POLICIES/FINITE COST CASE - A PERTURBATION APPROACH

In this section, we address problems where some S -irregular policies may have finite cost for all states [thus violating Assumption 3.3.1(c)], so Prop. 3.3.1 cannot be used. Our approach instead will be to assert that J_S^* is a fixed point of T , so that Prop. 3.2.1 applies and can be used to guarantee convergence of VI to J_S^* starting from $J_0 \geq J_S^*$.

Our line of analysis is quite different from the one of Sections 3.2.3 and 3.3, which was based on PI ideas. Instead, we *add a perturbation to the mapping* H , designed to provide adequate differentiation between S -regular and S -irregular policies. Using a limiting argument, as the size of the perturbation diminishes to 0, we are able to prove that J_S^* is a fixed point of T . Moreover, we provide a perturbation-based PI algorithm that may be more reliable than the standard PI algorithm, which can fail for problems where irregular policies may have finite cost for all states; cf. Example 3.2.2. We will also use the perturbation approach in Sections 4.5 and 4.6, where we will extend the notion of S -regularity to nonstationary policies that do not lend themselves to a PI-based analysis.

An example where the approach of this section will be shown to apply is an SSP problem where Assumption 3.3.1 is violated while $J^*(x) > -\infty$ for all x (see also Section 3.5.1). Here is a classical problem of this type.

Example 3.4.1 (Search Problem)

Consider a situation where the objective is to move within a finite set of states searching for a state to stop while minimizing the expected cost. We formulate this as a DP problem with finite state space X , and two controls at each $x \in X$: *stop*, which yields an immediate cost $s(x)$, and *continue*, in which case we move to a state $f(x, w)$ at cost $g(x, w)$, where w is a random variable with given distribution that may depend on x . The mapping H is

$$H(x, u, J) = \begin{cases} s(x) & \text{if } u = \text{stop,} \\ E\{g(x, w) + J(f(x, w))\} & \text{if } u = \text{continue,} \end{cases}$$

and the function \bar{J} is identically 0.

Letting $S = \mathcal{R}(X)$, we note that the policy $\bar{\mu}$ that stops nowhere is S -irregular, since $T_{\bar{\mu}}$ cannot have a unique fixed point within S (adding any unit function multiple to J adds to $T_{\bar{\mu}}J$ the same multiple). This policy may violate Assumption 3.3.1(c) of the preceding section, because its cost may be

finite for all states. A special case where this occurs is when $g(x, w) \equiv 0$ for all x . Then the cost function of $\bar{\mu}$ is identically 0.

Note that case (b) of the deterministic shortest path problem of Section 3.1.1, which involves a zero length cycle, is a special case of the search problem just described. Therefore, the anomalous behavior we saw there (nonconvergence of VI to J^* and oscillation of PI; cf. Examples 3.2.1 and 3.2.2) may also arise in the context of the present example. We will see that by adding a small positive constant to the length of the cycle we can rectify the difficulties of VI and PI, at least partially; this is the idea behind the perturbation approach that we will use in this section.

We will address the finite cost issue for irregular policies by introducing a perturbation that makes their cost infinite for some states. We can then use Prop. 3.3.1 of the preceding section. The idea is that with a perturbation, the cost functions of S -irregular policies may increase disproportionately relative to the cost functions of the S -regular policies, thereby making the problem more amenable to analysis.

We introduce a nonnegative “forcing function” $p : X \mapsto [0, \infty)$, and for each $\delta > 0$ and policy μ , we consider the mappings

$$(T_{\mu, \delta} J)(x) = H(x, \mu(x), J) + \delta p(x), \quad x \in X, \quad T_{\delta} J = \inf_{\mu \in \mathcal{M}} T_{\mu, \delta} J.$$

We refer to the problem associated with the mappings $T_{\mu, \delta}$ as the δ -perturbed problem. The cost functions of policies $\pi = \{\mu_0, \mu_1, \dots\} \in \Pi$ and $\mu \in \mathcal{M}$ for this problem are

$$J_{\pi, \delta} = \limsup_{k \rightarrow \infty} T_{\mu_0, \delta} \cdots T_{\mu_k, \delta} \bar{J}, \quad J_{\mu, \delta} = \limsup_{k \rightarrow \infty} T_{\mu, \delta}^k \bar{J},$$

and the optimal cost function is $\hat{J}_{\delta} = \inf_{\pi \in \Pi} J_{\pi, \delta}$.

The following proposition shows that if the δ -perturbed problem is “well-behaved” with respect to a subset of S -regular policies, then its cost function \hat{J}_{δ} can be used to approximate the optimal cost function over this subset of policies only. Moreover J_S^* is a fixed point of T . Note that the unperturbed problem need not be as well-behaved, and indeed J^* need not be a fixed point of T .

Proposition 3.4.1: Given a set $S \subset \mathcal{E}(X)$, let $\widehat{\mathcal{M}}$ be a subset of S -regular policies, and let \hat{J} be the optimal cost function over the policies in $\widehat{\mathcal{M}}$ only, i.e.,

$$\hat{J} = \inf_{\mu \in \widehat{\mathcal{M}}} J_{\mu}.$$

Assume that for every $\delta > 0$:

- (1) The optimal cost function \hat{J}_{δ} of the δ -perturbed problem satisfies the corresponding Bellman equation $\hat{J}_{\delta} = T_{\delta} \hat{J}_{\delta}$.

(2) We have $\inf_{\mu \in \widehat{\mathcal{M}}} J_{\mu, \delta} = \hat{J}_\delta$, i.e., for every $x \in X$ and $\epsilon > 0$, there exists a policy $\mu_{x, \epsilon} \in \widehat{\mathcal{M}}$ such that $J_{\mu_{x, \epsilon}, \delta}(x) \leq \hat{J}_\delta(x) + \epsilon$.

(3) For every $\mu \in \widehat{\mathcal{M}}$, we have

$$J_{\mu, \delta} \leq J_\mu + w_{\mu, \delta},$$

where $w_{\mu, \delta}$ is a function such that $\lim_{\delta \downarrow 0} w_{\mu, \delta} = 0$.

(4) For every sequence $\{J_m\} \subset S$ with $J_m \downarrow J$, we have

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

Then J_S^* is a fixed point of T and the conclusions of Prop. 3.2.1 hold. Moreover, we have

$$J_S^* = \hat{J} = \lim_{\delta \downarrow 0} \hat{J}_\delta.$$

Proof: For every $x \in X$, using conditions (2) and (3), we have for all $\delta > 0$, $\epsilon > 0$, and $\mu \in \widehat{\mathcal{M}}$,

$$\hat{J}(x) - \epsilon \leq J_{\mu_{x, \epsilon}}(x) - \epsilon \leq J_{\mu_{x, \epsilon}, \delta}(x) - \epsilon \leq \hat{J}_\delta(x) \leq J_{\mu, \delta}(x) \leq J_\mu(x) + w_{\mu, \delta}(x).$$

By taking the limit as $\epsilon \downarrow 0$, we obtain for all $\delta > 0$ and $\mu \in \widehat{\mathcal{M}}$,

$$\hat{J} \leq \hat{J}_\delta \leq J_{\mu, \delta} \leq J_\mu + w_{\mu, \delta}.$$

By taking the limit as $\delta \downarrow 0$ and then the infimum over all $\mu \in \widehat{\mathcal{M}}$, it follows [using also condition (3)] that

$$\hat{J} \leq \lim_{\delta \downarrow 0} \hat{J}_\delta \leq \inf_{\mu \in \widehat{\mathcal{M}}} \lim_{\delta \downarrow 0} J_{\mu, \delta} \leq \inf_{\mu \in \widehat{\mathcal{M}}} J_\mu = \hat{J},$$

so that $\hat{J} = \lim_{\delta \downarrow 0} \hat{J}_\delta$.

Next we prove that \hat{J} is a fixed point of T and use this fact to show that $\hat{J} = J_S^*$, thereby concluding the proof. Indeed, from condition (1) and the fact $\hat{J}_\delta \geq \hat{J}$ shown earlier, we have for all $\delta > 0$,

$$\hat{J}_\delta = T_\delta \hat{J}_\delta \geq T \hat{J}_\delta \geq T \hat{J},$$

and by taking the limit as $\delta \downarrow 0$ and using part (a), we obtain $\hat{J} \geq T \hat{J}$. For the reverse inequality, let $\{\delta_m\}$ be a sequence with $\delta_m \downarrow 0$. Using condition (1) we have for all m ,

$$H(x, u, \hat{J}_{\delta_m}) + \delta_m p(x) \geq (T_{\delta_m} \hat{J}_{\delta_m})(x) = \hat{J}_{\delta_m}(x), \quad \forall x \in X, u \in U(x).$$

Taking the limit as $m \rightarrow \infty$, and using condition (4) and the fact $\hat{J}_{\delta_m} \downarrow \hat{J}$ shown earlier, we have

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

so that $T\hat{J} \geq \hat{J}$. Thus \hat{J} is a fixed point of T .

Finally, to show that $\hat{J} = J_S^*$, we first note that $J_S^* \leq \hat{J}$ since every policy in $\widehat{\mathcal{M}}$ is S -regular. For the reverse inequality, let μ be S -regular. We have $\hat{J} = T\hat{J} \leq T_\mu \hat{J} \leq T_\mu^k \hat{J}$ for all $k \geq 1$, so that for all $\mu' \in \widehat{\mathcal{M}}$,

$$\hat{J} \leq \lim_{k \rightarrow \infty} T_\mu^k \hat{J} \leq \lim_{k \rightarrow \infty} T_\mu^k J_{\mu'} = J_\mu,$$

where the equality follows since μ and μ' are S -regular (so $J_{\mu'} \in S$). Taking the infimum over all S -regular μ , we obtain $\hat{J} \leq J_S^*$, so that $J_S^* = \hat{J}$.

Q.E.D.

Aside from S -regularity of the set $\widehat{\mathcal{M}}$, a key assumption of the preceding proposition is that $\inf_{\mu \in \widehat{\mathcal{M}}} J_{\mu, \delta} = \hat{J}_\delta$, i.e., that with a perturbation added, the subset of policies $\widehat{\mathcal{M}}$ is sufficient (the optimal cost of the δ -perturbed problem can be achieved using the policies in $\widehat{\mathcal{M}}$). This is the key insight to apply when selecting $\widehat{\mathcal{M}}$.

Note that the preceding proposition applies even if

$$\lim_{\delta \downarrow 0} \hat{J}_\delta(x) > J^*(x)$$

for some $x \in X$. This is illustrated by the deterministic shortest path example of Section 3.1.1, for the zero-cycle case where $a = 0$ and $b > 0$. Then for $S = \mathfrak{R}$, we have $J_S^* = b > 0 = J^*$, while the proposition applies because its assumptions are satisfied with $p(x) \equiv 1$. Consistently with the conclusions of the proposition, we have $\hat{J}_\delta = b + \delta$, so $J_S^* = \hat{J} = \lim_{\delta \downarrow 0} \hat{J}_\delta$ and J_S^* is a fixed point of T .

Proposition 3.4.1 also applies to Example 3.4.1. In particular, it can be used to assert that J_S^* is a fixed point of T , and hence also that the conclusions of Prop. 3.2.1 hold. These conclusions imply that J_S^* is the unique fixed point of T within the set $\{J \mid J \geq J_S^*\}$ and that the VI algorithm converges to J_S^* starting from within this set.

We finally note that while Props. 3.3.1 and 3.4.1 relate to qualitatively different problems, they can often be used synergistically. In particular, Prop. 3.3.1 may be applied to the δ -perturbed problem in order to verify the assumptions of Prop. 3.4.1.

A Policy Iteration Algorithm with Perturbations

We now consider a subset $\widehat{\mathcal{M}}$ of S -regular policies, and introduce a version of the PI algorithm that uses perturbations and generates a sequence $\{\mu^k\} \subset \widehat{\mathcal{M}}$ such that $J_{\mu^k} \rightarrow J_S^*$. We assume the following.

Assumption 3.4.1: The subset of S -regular policies $\widehat{\mathcal{M}}$ is such that:

- (a) The conditions of Prop. 3.4.1 are satisfied.
- (b) Every policy $\mu \in \widehat{\mathcal{M}}$ is S -regular for all the δ -perturbed problems, $\delta > 0$.
- (c) Given a policy $\mu \in \widehat{\mathcal{M}}$ and a scalar $\delta > 0$, every policy μ' such that

$$T_{\mu'} J_{\mu, \delta} = T J_{\mu, \delta}$$

belongs to $\widehat{\mathcal{M}}$, and at least one such policy exists.

The perturbed version of the PI algorithm is defined as follows. Let $\{\delta_k\}$ be a positive sequence with $\delta_k \downarrow 0$, and let μ^0 be a policy in $\widehat{\mathcal{M}}$. At iteration k , we have a policy $\mu^k \in \widehat{\mathcal{M}}$, and we generate $\mu^{k+1} \in \widehat{\mathcal{M}}$ according to

$$T_{\mu^{k+1}} J_{\mu^k, \delta_k} = T J_{\mu^k, \delta_k}. \quad (3.28)$$

Note that by Assumption 3.4.1(c) the algorithm is well-defined, and is guaranteed to generate a sequence of policies $\{\mu^k\} \subset \widehat{\mathcal{M}}$. We have the following proposition.

Proposition 3.4.2: Let Assumption 3.4.1 hold. Then J_S^* is a fixed point of T and for a sequence of S -regular policies $\{\mu^k\}$ generated by the perturbed PI algorithm (3.28), we have $J_{\mu^k, \delta_k} \downarrow J_S^*$ and $J_{\mu^k} \rightarrow J_S^*$.

Proof: We have that J_S^* is a fixed point of T by Prop. 3.4.1. The algorithm definition (3.28) implies that for all $m \geq 1$ we have

$$T_{\mu^{k+1}, \delta_k}^m J_{\mu^k, \delta_k} \leq T_{\mu^{k+1}, \delta_k} J_{\mu^k, \delta_k} = T J_{\mu^k, \delta_k} + \delta_k p \leq J_{\mu^k, \delta_k}.$$

From this relation it follows that

$$J_{\mu^{k+1}, \delta_{k+1}} \leq J_{\mu^{k+1}, \delta_k} = \lim_{m \rightarrow \infty} T_{\mu^{k+1}, \delta_k}^m J_{\mu^k, \delta_k} \leq J_{\mu^k, \delta_k},$$

where the equality holds because μ^{k+1} and μ^k are S -regular for all the δ -perturbed problems. It follows that $\{J_{\mu^k, \delta_k}\}$ is monotonically nonincreasing, so that $J_{\mu^k, \delta_k} \downarrow J_\infty$ for some J_∞ . Moreover, we must have $J_\infty \geq J_S^*$ since $J_{\mu^k, \delta_k} \geq J_{\mu^k} \geq J_S^*$. Thus

$$J_S^* \leq J_\infty = \lim_{k \rightarrow \infty} T J_{\mu^k, \delta_k}. \quad (3.29)$$

We also have

$$\begin{aligned}
\inf_{u \in U(x)} H(x, u, J_\infty) &\leq \lim_{k \rightarrow \infty} \inf_{u \in U(x)} H(x, u, J_{\mu^k, \delta_k}) \\
&\leq \inf_{u \in U(x)} \lim_{k \rightarrow \infty} H(x, u, J_{\mu^k, \delta_k}) \\
&= \inf_{u \in U(x)} H(x, u, \lim_{k \rightarrow \infty} J_{\mu^k, \delta_k}) \\
&= \inf_{u \in U(x)} H(x, u, J_\infty),
\end{aligned}$$

where the first inequality follows from the fact $J_\infty \leq J_{\mu^k, \delta_k}$, which implies that $H(x, u, J_\infty) \leq H(x, u, J_{\mu^k, \delta_k})$, and the first equality follows from the continuity property that is assumed in Prop. 3.4.1. Thus equality holds throughout above, so that

$$\lim_{k \rightarrow \infty} TJ_{\mu^k, \delta_k} = TJ_\infty. \quad (3.30)$$

Combining Eqs. (3.29) and (3.30), we obtain $J_S^* \leq J_\infty = TJ_\infty$. By replacing \hat{J} with J_∞ in the last part of the proof of Prop. 3.4.1, we obtain $J_S^* = J_\infty$. Thus $J_{\mu^k, \delta_k} \downarrow J_S^*$, which in view of the fact $J_{\mu^k, \delta_k} \geq J_{\mu^k} \geq J_S^*$, implies that $J_{\mu^k} \rightarrow J_S^*$. **Q.E.D.**

When the control space U is finite, Prop. 3.4.2 also implies that the generated policies μ^k will be optimal for all k sufficiently large. The reason is that the set of policies is finite and there exists a sufficiently small $\epsilon > 0$, such that for all nonoptimal μ there is some state x such that $J_\mu(x) \geq \hat{J}(x) + \epsilon$. This convergence behavior should be contrasted with the behavior of PI without perturbations, which may lead to oscillations, as noted earlier.

However, when the control space U is infinite, the generated sequence $\{\mu^k\}$ may exhibit some serious pathologies in the limit. If $\{\mu^k\}_{\mathcal{K}}$ is a subsequence of policies that converges to some $\bar{\mu}$, in the sense that

$$\lim_{k \rightarrow \infty, k \in \mathcal{K}} \mu^k(x) = \bar{\mu}(x), \quad \forall x = 1, \dots, n,$$

it does not follow that $\bar{\mu}$ is S -regular. In fact it is possible that the generated sequence of S -regular policies $\{\mu^k\}$ satisfies $\lim_{k \rightarrow \infty} J_{\mu^k} \rightarrow J_S^* = J^*$, yet $\{\mu^k\}$ may converge to an S -irregular policy whose cost function is strictly larger than J_S^* , as illustrated by the following example.

Example 3.4.2

Consider the third variant of the blackmailer problem (Section 3.1.3) for the case where $c = 0$ (the blackmailer may forgo demanding a payment at cost $c = 0$); see Fig. 3.4.1. Here the mapping T is given by

$$TJ = \min \left\{ J, \inf_{0 < u \leq 1} \left\{ -u + u^2 + (1-u)J \right\} \right\},$$

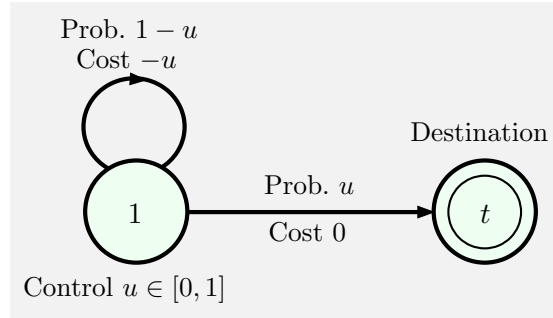


Figure 3.4.1. Transition diagram for a blackmailer problem (the third variant of Section 3.1.3 in the case where $c = 0$). At state 1, the blackmailer may demand any amount $u \in [0, 1]$. The victim will comply with probability $1 - u$ and will not comply with probability u , in which case the process will terminate.

[cf. Eq. (3.4)], and can be written as

$$TJ = \min_{0 \leq u \leq 1} \{ -u + u^2 + (1 - u)J \}.$$

Letting $S = \mathfrak{R}$, it can be seen that the set of fixed points of T within S is $(-\infty, -1]$. Here the policy whereby the blackmailer demands no payment ($u = 0$) and pays no cost at each period, is S -irregular and strictly suboptimal, yet has finite (zero) cost, so part (c) of Assumption 3.3.1 is violated (all other parts of the assumption are satisfied).

It can be seen that

$$J^* = J_S^* = -1,$$

J_S^* is a fixed point of T , Prop. 3.2.1 applies, and VI converges to J^* starting from any $J \geq J^*$. Moreover, starting from any policy (including the S -irregular one that applies $u = 0$), the PI algorithm (3.28) generates a sequence of S -regular policies $\{\mu^k\}$ with $J_{\mu^k} \rightarrow J_S^*$. However, $\{\mu^k\}$ converges to the S -irregular and strictly suboptimal policy that applies $u = 0$.

Here a phenomenon of “oscillation in the limit” is observed: starting with the S -irregular policy that applies $u = 0$, we generate a sequence of S -regular policies that converges to the S -irregular policy we started from! The perturbation-based PI algorithm of this section cannot rectify this type of behavior; it can only guarantee that a sequence of S -regular policies with $J_{\mu^k} \rightarrow J_S^*$ is generated.

3.5 APPLICATIONS IN SHORTEST PATH AND OTHER CONTEXTS

In this section we will apply the results of the preceding sections to various problems with a semicontractive character, including shortest path and deterministic optimal control problems of various types.

As we are about to apply the theory developed so far in this chapter, it may be helpful to summarize our results. Given a suitable set of functions S , we have been dealing with two problems. These are the original problem whose optimal cost function is J^* , and the restricted problem whose optimal cost function is J_S^* , the optimal cost over the S -regular policies. In summary, the aims of our analysis have been the following:

- (a) *To establish the fixed point properties of T .* We have showed under various conditions (cf. Prop. 3.2.1) that J_S^* is the unique fixed point of T within the well-behaved region \mathcal{W}_S , and moreover the VI algorithm converges from above to J_S^* . Related analyses involve the use of infinite cost assumptions for S -irregular policies (Section 3.3), possibly in conjunction with the use of perturbations (Section 3.4). A favorable case is when $J_S^* = J^*$. However, we may also have $J_S^* \neq J^*$. Generally, proving that J^* is a fixed point of T is a separate issue, which may either be addressed in conjunction with the analysis of properties of J_S^* as in Section 3.3 (cf. Prop. 3.3.1), or independently of J_S^* (for example J^* is generically a fixed point of T in deterministic problems, among other classes of problems; see Exercise 3.1).
- (b) *To delineate the initial conditions under which the VI and PI algorithms are guaranteed to converge to J_S^* or to J^* .* This was done in conjunction with the analysis of the fixed point properties of T . For example, a major line of analysis for establishing that J_S^* is a fixed point of T is based on the PI algorithm (cf. Sections 3.2.3 and 3.3). We have also obtained several other results relating to the convergence of variants of PI (the optimistic version, cf. Prop. 3.2.7, the λ -PI version, cf. Prop. 3.2.8, and the perturbation-based version, cf. Prop. 3.4.2), and to the mathematical programming-based solution, cf. Section 3.2.5.
- (c) *To establish the existence of optimal policies for the original or for the restricted problem, and the associated optimality conditions.* This was accomplished in conjunction with the analysis of the fixed points of T , and under special compactness-like conditions (cf. Props. 3.2.1, 3.2.6, and 3.3.1).

As we apply our analysis to various specific contexts in this section, we will make frequent reference to the pathological behavior that we witnessed in the examples of Section 3.1. In particular, we will explain this behavior through our theoretical results, and we will discuss how to preclude this behavior through appropriate assumptions.

3.5.1 Stochastic Shortest Path Problems

Let us consider the SSP problem that we discussed in Section 1.3.2. It involves a directed graph with nodes $x = 1, \dots, n$, plus a destination node

t that is cost-free and absorbing. At each node x , we must select a control $u \in U(x)$, which defines a probability distribution $p_{xy}(u)$ over all possible successor nodes $y = 1, \dots, n, t$, while a cost $g(x, u)$ is incurred. We wish to minimize the expected cost of the traversed path, with cost accumulated up to reaching the destination.

Note that if for every feasible control the corresponding probability distribution assigns probability 1 to a single successor node, we obtain the deterministic shortest path problem of Section 3.1.1. This problem admits a relatively simple analysis, yet exhibits pathological behavior that we have described. The pathologies exhibited by SSP problems are more severe, and were illustrated in Sections 3.1.2 and 3.1.3.

We formulate the SSP problem as an abstract DP problem where:

- (a) The state space is $X = \{1, \dots, n\}$ and the control constraint set is $U(x)$ for all $x \in X$. (For technical reasons, it is convenient to exclude from X the destination t ; we know that the optimal cost starting from t is 0, and including t within X would just complicate the notation and the analysis, with no tangible benefit.)
- (b) The mapping H is given by

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

- (c) The function \bar{J} is identically 0, $\bar{J}(x) = 0$ for all x .

We continue to denote by $\mathcal{E}(X)$ the set of all extended real-valued functions $J : X \mapsto \mathfrak{R}^*$, and by $\mathcal{R}(X)$ the set of real-valued functions $J : X \mapsto \mathfrak{R}$. Note that since $X = \{1, \dots, n\}$, $\mathcal{R}(X)$ is essentially the n -dimensional space R^n .

Here the mapping T_μ corresponding to a policy μ maps $\mathcal{R}(X)$ to $\mathcal{R}(X)$, and is given by

$$(T_\mu J)(x) = g(x, \mu(x)) + \sum_{y=1}^n p_{xy}(\mu(x))J(y), \quad x = 1, \dots, n.$$

The corresponding cost for a given initial state $x_0 \in \{1, \dots, n\}$ is

$$J_\mu(x_0) = \limsup_{k \rightarrow \infty} (T_\mu^k \bar{J})(x_0) = \limsup_{k \rightarrow \infty} \sum_{m=0}^{k-1} E\{g(x_m, \mu(x_m))\},$$

where $\{x_m\}$ is the (random) state trajectory generated under policy μ , starting from initial state x_0 . The expected value $E\{g(x_m, \mu(x_m))\}$ above is defined in the natural way: it is the weighted sum of the numerical values $g(x, \mu(x))$, $x = 1, \dots, n$, weighted by the probabilities $p(x_m = x \mid x_0, \mu)$

that $x_m = x$ given that the initial state is x_0 and policy μ is used. Thus $J_\mu(x_0)$ is the upper limit as $k \rightarrow \infty$ of the cost for the first k steps or up to reaching the destination, whichever comes first.

A stationary policy μ is said to be *proper* if for every initial state there is positive probability that the destination will be reached under that policy after at most n stages. A stationary policy that is not proper is said to be *improper*. The relation between proper policies and S -regularity is given in the following proposition.

Proposition 3.5.1: (Proper Policies and Regularity) A policy is proper if and only if it is $\mathcal{R}(X)$ -regular.

Proof: Clearly μ is $\mathcal{R}(X)$ -regular if and only if the $n \times n$ matrix P_μ , whose components are $p_{ij}(\mu(i))$, $i, j = 1, \dots, n$, is a contraction (since T_μ is a linear mapping with matrix P_μ). If μ is proper then P_μ is a contraction mapping with respect to some weighted sup-norm; this is a classical result, given for example in [BeT89], Section 4.2. Conversely, it can be seen that if μ is improper, P_μ is not a contraction mapping since the Markov chain corresponding to μ has multiple ergodic classes and hence the equilibrium equation $\xi' = \xi' P_\mu$ has multiple solutions. **Q.E.D.**

Looking back to the shortest path examples of Sections 3.1.1-3.1.3, we can make some observations. In deterministic shortest path problems, $\mu(x)$ can be identified with the single successor node of node x . Thus μ is proper if and only if the corresponding graph of arcs $(x, \mu(x))$ is acyclic. Moreover, there exists a proper policy if and only if each node is connected to the destination with a sequence of arcs. Every improper policy involves at least one cycle. Depending on the sign of the length of their cycle(s), improper policies can be strictly suboptimal (if all cycles have positive length), or may be optimal (possibly together with some proper policies, if all cycles have nonnegative length). Moreover, if there are cycles with negative length, no proper policy can be optimal and for the states x that lie on some negative length cycle we have $J^*(x) = -\infty$.

A further characterization of the optimal solution is possible in deterministic shortest path problems. Since the sets $U(x)$ are finite, there exists an optimal policy, which can be separated into a “proper” part consisting of arcs that form an acyclic subgraph, and an “improper” part consisting of cycles that have negative or zero length. These facts can be proved with simple arguments, which will not be given here (deterministic shortest path theory and algorithms are developed in detail in the author’s text [Ber98]).

In SSP problems, the situation is more complicated. In particular, the cost function of an improper policy μ may not be a fixed point of T_μ while J^* may not be a fixed point of T (cf. the example of Section

3.1.2). Moreover, there may not exist an optimal stationary policy even if all policies are proper (cf. the three variants of the blackmailer example of Section 3.1.3).

In this section we will use various assumptions, which we will in turn translate into the conditions and corresponding results of Sections 3.2-3.4. Throughout this section we will assume the following.

Assumption 3.5.1: There exists at least one proper policy.

Depending on the circumstances, we will also consider the use of one or both of the following assumptions.

Assumption 3.5.2: The control space U is a metric space. Moreover, for each state x , the set $U(x)$ is a compact subset of U , the functions $p_{xy}(\cdot)$, $y = 1, \dots, n$, are continuous over $U(x)$, and the function $g(x, \cdot)$ is lower semicontinuous over $U(x)$.

Assumption 3.5.3: For every improper policy μ and function $J \in \mathcal{R}(X)$, there exists at least one state $x \in X$ such that $J_\mu(x) = \infty$.

An important consequence of Assumption 3.5.2 is that it implies the compactness condition (d) of Assumption 3.3.1. We will also see from the proof of the following proposition that Assumption 3.5.3 implies the infinite cost condition (c) of Assumption 3.3.1.

Analysis Under the Strong SSP Conditions

The preceding three assumptions, referred to as the *strong SSP conditions*,[†] were introduced in the paper [BeT91], and they were used to show strong results for the SSP problem. In particular, the following proposition was shown.

Proposition 3.5.2: Let the strong SSP conditions hold. Then:

- (a) The optimal cost function J^* is the unique solution of Bellman's equation $J = TJ$ within $\mathcal{R}(X)$.

[†] The strong SSP conditions and the weak SSP conditions, which will be introduced shortly, connect to the strong and weak PI properties of Section 3.2.

- (b) The VI sequence $\{T^k J\}$ converges to J^* starting from any $J \in \mathcal{R}(X)$.
- (c) A policy μ is optimal if and only if $T_\mu J^* = T J^*$. Moreover, there exists an optimal policy that is proper.
- (d) The PI algorithm, starting from any proper policy, is valid in the sense described by the conclusions of Prop. 3.3.1(e).

We will prove the proposition by using the strong SSP conditions to verify Assumption 3.3.1 for $S = \mathcal{R}(X)$, and then by applying Prop. 3.3.1. To this end, we first state without proof the following result relating to proper policies from [BeT91].

Proposition 3.5.3: Under the strong SSP conditions, the optimal cost function \hat{J} over proper policies only,

$$\hat{J}(x) = \inf_{\mu: \text{proper}} J_\mu(x), \quad x \in X,$$

is real-valued.

The preceding proposition holds trivially if the control space U is finite (since then the set of all policies is finite), or if J^* is somehow known to be real-valued [for example if $g(x, u) \geq 0$ for all (x, u)]. The three variants of the blackmailer problem of Section 3.1.3 provide examples illustrating what can happen if U is infinite. In particular, in the first variant of the blackmailer problem all policies are proper (and hence Assumptions 3.5.1 and 3.5.3 are satisfied), but \hat{J} is not real-valued. The proof of Prop. 3.5.3 in the case of an infinite control space U was given as part of Prop. 2 of the paper [BeT91]. Despite the intuitive nature of Prop. 3.5.3, the proof embodies a fairly complicated argument (see Lemma 3 of [BeT91]).

Another related result is that if all policies are proper, then for all $\mu \in \mathcal{M}$, T_μ is a contraction mapping with respect to a common weighted sup-norm, so the contractive model analysis and algorithms of Chapter 2 apply (see [BeT96], Prop. 2.2). However, this fact will not be useful to us in this section.

Proof of Prop. 3.5.2: In the context of Section 3.3, let us choose $S = \mathcal{R}(X)$, so the proper policies are identified with the S -regular policies by Prop. 3.5.1. We will verify Assumption 3.3.1.

Indeed parts (a) and (e) are trivially satisfied, part (b) is satisfied by Prop. 3.5.3, part (d) can be easily verified by using Assumption 3.5.2. To verify part (f), we use Prop. 3.3.2, which applies because $S = \mathcal{R}(X) =$

$R_b(X)$ (since X is finite) and Eq. (3.27) clearly holds. Finally, to verify part (c) we must show that given an improper policy μ , for every $J \in \mathcal{R}(X)$ there exists an $x \in X$ such that $\limsup_{k \rightarrow \infty} (T_\mu^k J)(x) = \infty$. This follows since by Assumption 3.5.3, $J_\mu(x) = \limsup_{k \rightarrow \infty} (T_\mu^k \bar{J})(x) = \infty$, for some $x \in X$, and $(T_\mu^k J)(x)$ and $(T_\mu^k \bar{J})(x)$ differ by $E\{J(x_k)\}$, an amount that is finite since J is real-valued and has a finite number of components $J(x)$. Thus Assumption 3.3.1 holds and the result follows from Prop. 3.3.1. **Q.E.D.**

Analysis Under the Weak SSP Conditions

Under the strong SSP conditions, we showed in Prop. 3.5.2 that J^* is the unique fixed point of T within $\mathcal{R}(X)$. Moreover, we showed that a policy μ^* is optimal if and only if $T_{\mu^*} J^* = T J^*$, and an optimal proper policy exists (so in particular J^* , being the cost function of a proper policy, is real-valued). In addition, J^* can be computed by the VI algorithm starting with any $J \in \mathfrak{R}^n$.

We will now replace Assumption 3.5.3 (improper policies have cost ∞ for some initial states) with the following weaker assumption:

Assumption 3.5.4: The optimal cost function J^* is real-valued.

We will refer to the Assumptions 3.5.1, 3.5.2, and 3.5.4 as the *weak SSP conditions*. The examples of Sections 3.1.1 and 3.1.2 show that under these assumptions, it is possible that

$$J^* \neq \hat{J} = \inf_{\mu: \text{proper}} J_\mu,$$

while J^* need not be a fixed point of T (Section 3.1.2). The key fact is that under Assumption 3.5.4, we can use the perturbation approach of Section 3.4, whereby adding $\delta > 0$ to the mapping T_μ makes all improper policies have infinite cost for some initial states, so the results of Prop. 3.5.2 can be used for the δ -perturbed problem. In particular, Prop. 3.5.1 implies that $J_S^* = \hat{J}$, so from Prop. 3.4.1 it follows that \hat{J} is a fixed point of T and the conclusions of Prop. 3.2.1 hold. We thus obtain the following proposition, which provides additional results, not implied by Prop. 3.2.1; see Fig. 3.5.1.

Proposition 3.5.4: Let the weak SSP conditions hold. Then:

- (a) The optimal cost function over proper policies, \hat{J} , is the largest solution of Bellman's equation $J = TJ$ within $\mathcal{R}(X)$, i.e., \hat{J} is a solution that belongs to $\mathcal{R}(X)$, and if $J' \in \mathcal{R}(X)$ is another solution, we have $J' \leq \hat{J}$.

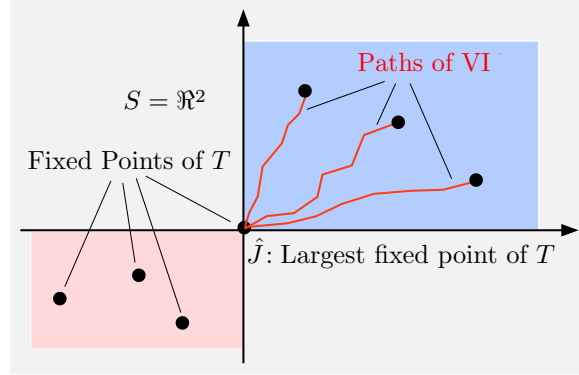


Figure 3.5.1. Schematic illustration of Prop. 3.5.4 for a problem with two states, so $\mathcal{R}(X) = \mathbb{R}^2 = S$. We have that \hat{J} is the largest solution of Bellman's equation, while VI converges to \hat{J} starting from $J \geq \hat{J}$. As shown in Section 3.1.2, J^* need not be a solution of Bellman's equation.

- (b) The VI sequence $\{T^k J\}$ converges linearly to \hat{J} starting from any $J \in \mathcal{R}(X)$ with $J \geq \hat{J}$.
- (c) Let μ be a proper policy. Then μ is optimal within the class of proper policies (i.e., $J_\mu = \hat{J}$) if and only if $T_\mu \hat{J} = T \hat{J}$.
- (d) For every $J \in \mathcal{R}(X)$ such that $J \leq T J$, we have $J \leq \hat{J}$.

Proof: (a), (b) Let $S = \mathcal{R}(X)$, so the proper policies are identified with the S -regular policies by Prop. 3.5.1. We use the perturbation framework of Section 3.4 with forcing function $p(x) \equiv 1$. From Prop. 3.5.2 it follows that Prop. 3.4.1 applies so that \hat{J} is a fixed point of T , and the conclusions of Prop. 3.2.1 hold, so $T^k J \rightarrow \hat{J}$ starting from any $J \in \mathcal{R}(X)$ with $J \geq \hat{J}$. The convergence rate of VI is linear in view of Prop. 3.2.2 and the existence of an optimal proper policy to be shown in part (c). Finally, let $J' \in \mathcal{R}(X)$ be another solution of Bellman's equation, and let $J \in \mathcal{R}(X)$ be such that $J \geq \hat{J}$ and $J \geq J'$. Then $T^k J \rightarrow \hat{J}$, while $T^k J \geq T^k J' = J'$. It follows that $\hat{J} \geq J'$.

(c) If the proper policy μ satisfies $J_\mu = \hat{J}$, we have $\hat{J} = J_\mu = T_\mu J_\mu = T_\mu \hat{J}$, so, using also the relation $\hat{J} = T \hat{J}$ [cf. part (a)], we obtain $T_\mu \hat{J} = T \hat{J}$. Conversely, if μ satisfies $T_\mu \hat{J} = T \hat{J}$, then using part (a), we have $T_\mu \hat{J} = \hat{J}$ and hence $\lim_{k \rightarrow \infty} T_\mu^k \hat{J} = \hat{J}$. Since μ is proper, we have $J_\mu = \lim_{k \rightarrow \infty} T_\mu^k \hat{J}$, so $J_\mu = \hat{J}$.

(d) Let $J \leq T J$ and $\delta > 0$. We have $J \leq T J + \delta e = T_\delta J$, and hence $J \leq T_\delta^k J$ for all k . Since the strong SSP conditions hold for the δ -perturbed

problem, it follows that $T_\delta^k J \rightarrow \hat{J}_\delta$, so $J \leq \hat{J}_\delta$. By taking $\delta \downarrow 0$ and using Prop. 3.4.1, it follows that $J \leq \hat{J}$. **Q.E.D.**

The first variant of the blackmailer Example 3.4.2 shows that under the weak SSP conditions there may not exist an optimal policy or an optimal policy within the class of proper policies if the control space is infinite. This is consistent with Prop. 3.5.4(c). Another interesting fact is provided by the third variant of this example in the case where $c < 0$. Then $J^*(1) = -\infty$ (violating Assumption 3.5.4), but \hat{J} is real-valued and does not solve Bellman's equation, contrary to the conclusion of Prop. 3.5.4(a).

Part (d) of Prop. 3.5.4 shows that \hat{J} is the unique solution of the problem of maximizing $\sum_{i=1}^n \beta_i J(i)$ over all $J = (J(1), \dots, J(n))$ such that $J \leq TJ$, where β_1, \dots, β_n are any positive scalars (cf. Prop. 3.2.9). This problem can be written as

$$\begin{aligned} & \text{maximize} && \sum_{i=1}^n J(i) \\ & \text{subject to} && J(x) \leq g(x, u) + \sum_{y=1}^n p_{ij}(u) J(j), \quad i = 1, \dots, n, \quad u \in U(i), \end{aligned}$$

and is a linear program if each $U(i)$ is a finite set.

Generally, under the weak SSP conditions the strong PI property may not hold, so a sequence generated by PI starting from a proper policy need not have the cost improvement property. An example is the deterministic shortest path problem of Section 3.1.1, when there is a zero length cycle ($a = 0$) and the only optimal policy is proper ($b = 0$). Then the PI algorithm may oscillate between the optimal proper policy and the strictly suboptimal improper policy. We will next consider the modified version of the PI algorithm that is based on the use of perturbations (Section 3.4).

Policy Iteration with Perturbations

To deal with the oscillatory behavior of PI, which was illustrated in the deterministic shortest path Example 3.2.2, we may use the perturbed version of the PI algorithm of Section 3.4, with forcing function $p(x) \equiv 1$. Thus, we have

$$(T_{\mu, \delta} J)(x) = H(x, \mu(x), J) + \delta, \quad x \in X, \quad T_\delta J = \inf_{\mu \in \mathcal{M}} T_{\mu, \delta} J.$$

The algorithm generates the sequence $\{\mu^k\}$ as follows.

Let $\{\delta_k\}$ be a positive sequence with $\delta_k \downarrow 0$, and let μ^0 be any proper policy. At iteration k , we have a proper policy μ^k , and we generate μ^{k+1} according to

$$T_{\mu^{k+1}} J_{\mu^k, \delta_k} = T J_{\mu^k, \delta_k}, \quad (3.31)$$

where J_{μ^k, δ_k} is computed as the unique fixed point of the mapping T_{μ^k, δ_k} given by

$$T_{\mu^k, \delta_k} J = T_{\mu^k} J + \delta_k e.$$

The policy μ^{k+1} of Eq. (3.31) exists by the compactness Assumption 3.5.2. We claim that μ^{k+1} is proper. To see this, note that

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

so that by the monotonicity of T_{μ}^{k+1} ,

$$T_{\mu^{k+1}, \delta_k}^m J_{\mu^k, \delta_k} \leq T_{\mu^{k+1}, \delta_k} J_{\mu^k, \delta_k} = T J_{\mu^k, \delta_k} + \delta_k e \leq J_{\mu^k, \delta_k}, \quad \forall m \geq 1.$$

Since J_{μ^k, δ_k} forms an upper bound to $T_{\mu^{k+1}, \delta_k}^m J_{\mu^k, \delta_k}$, it follows that μ^{k+1} is proper [if it were improper, we would have $(T_{\mu^{k+1}, \delta_k}^m J_{\mu^k, \delta_k})(x) \rightarrow \infty$ for some x , because of the perturbation δ_k]. Thus the sequence $\{\mu^k\}$ generated by the perturbed PI algorithm (3.31) is well-defined and consists of proper policies. We have the following proposition.

Proposition 3.5.5: Let the weak SSP conditions hold. Then the sequence $\{J_{\mu^k}\}$ generated by the perturbed PI algorithm (3.31) satisfies $J_{\mu^k} \rightarrow \hat{J}$.

Proof: We apply the perturbation framework of Section 3.4 with $S = \mathcal{R}(X)$, $\widehat{\mathcal{M}}$ equal to the set of proper policies, and the forcing function $p(x) \equiv 1$. Clearly Assumption 3.4.1 holds, so Prop. 3.4.2 applies. **Q.E.D.**

When the control space U is finite, the generated policies μ^k will be optimal for all k sufficiently large, as noted following Prop. 3.4.2. However, when the control space U is infinite, the generated sequence $\{\mu^k\}$ may exhibit some serious pathologies in the limit, as we have seen in Example 3.4.2.

3.5.2 Affine Monotonic Problems

In this section, we consider a class of semicontractive models, called *affine monotonic*, where the abstract mapping T_{μ} associated with a stationary policy μ is affine and maps nonnegative functions to nonnegative functions. These models include as special cases stochastic undiscounted nonnegative cost problems, and multiplicative cost problems, such as risk-averse problems with exponentiated additive cost and a termination state (see Example 1.2.8). Here we will focus on the special case where the state space is finite and a certain compactness condition holds.

We consider a finite state space $X = \{1, \dots, n\}$ and a (possibly infinite) control constraint set $U(x)$ for each state x . For each $\mu \in \mathcal{M}$ the mapping T_μ is given by

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

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

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

Thus T_μ maps $\mathcal{E}^+(X)$ into $\mathcal{E}^+(X)$, where $\mathcal{E}^+(X)$ denotes the set of nonnegative extended real-valued functions $J : X \mapsto [0, \infty]$. Moreover T maps $\mathcal{R}^+(X)$ to $\mathcal{R}^+(X)$, where $\mathcal{R}^+(X)$ denotes the set of nonnegative real-valued functions $J : X \mapsto [0, \infty)$.

The mapping $T : \mathcal{E}^+(X) \mapsto \mathcal{E}^+(X)$ is given by

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

or equivalently,

$$(TJ)(x) = \inf_{u \in U(x)} \left[b(x, u) + \sum_{y=1}^n A_{xy}(u) J(y) \right], \quad x \in X.$$

Multiplicative and Exponential Cost SSP Problems

Affine monotonic models appear in several contexts. In particular, finite-state sequential stochastic control problems (including SSP problems) with nonnegative cost per stage (see, e.g., [Ber12a], Chapter 3, and Section 4.1) are special cases where \bar{J} is the identically zero function [$\bar{J}(i) \equiv 0$]. We will describe another type of SSP problem, where the cost function of a policy accumulates over time multiplicatively, rather than additively.

As in the SSP problems of the preceding section, we assume that there are n states $i = 1, \dots, n$, and a cost-free and absorbing state t . There are probabilistic state transitions among the states $i = 1, \dots, n$, up to the first time a transition to state t occurs, in which case the state transitions terminate. We denote by $p_{it}(u)$ and $p_{ij}(u)$ the probabilities of transition under u from i to t and to j , respectively, so that

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

We introduce nonnegative scalars $h(i, u, t)$ and $h(i, u, j)$,

$$h(i, u, t) \geq 0, \quad h(i, u, j) \geq 0, \quad \forall i, j = 1, \dots, n, \quad u \in U(i),$$

and we consider the affine monotonic problem where the scalars $A_{ij}(u)$ and $b(i, u)$ are defined by

$$A_{ij}(u) = p_{ij}(u)h(i, u, j), \quad i, j = 1, \dots, n, \quad u \in U(i),$$

and

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

and the vector \bar{J} is the unit vector,

$$\bar{J}(i) = 1, \quad i = 1, \dots, n.$$

The cost function of this problem has a multiplicative character as we show next.

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

$$J_\pi(x_0) = \limsup_{N \rightarrow \infty} (T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(x_0), \quad x_0 = 1, \dots, n,$$

can be written in the multiplicative form

$$J_\pi(x_0) = \limsup_{N \rightarrow \infty} E \left\{ \prod_{k=0}^{N-1} h(x_k, \mu_k(x_k), x_{k+1}) \right\}, \quad x_0 = 1, \dots, n, \quad (3.32)$$

where:

- (a) $\{x_0, x_1, \dots\}$ is the random state trajectory generated starting from x_0 , using π .
- (b) The expected value is with respect to the probability distribution of that trajectory.
- (c) We use the notation

$$h(x_k, \mu_k(x_k), x_{k+1}) = 1, \quad \text{if } x_k = x_{k+1} = t,$$

(so that the multiplicative cost accumulation stops once the state reaches t).

Thus, we claim that $J_\pi(x_0)$ can be viewed as the expected value of cost accumulated multiplicatively, starting from x_0 up to reaching the termination state t (or indefinitely accumulated multiplicatively, if t is never reached).

To verify the formula (3.32) for J_π , we use the definition $T_\mu J = b_\mu + A_\mu J$, to show by induction that for every $\pi = \{\mu_0, \mu_1, \dots\}$, we have

$$T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J} = A_{\mu_0} \cdots A_{\mu_{N-1}} \bar{J} + b_{\mu_0} + \sum_{k=1}^{N-1} A_{\mu_0} \cdots A_{\mu_{k-1}} b_{\mu_k}. \quad (3.33)$$

We then interpret the n components of each vector on the right as conditional expected values of the expression

$$\prod_{k=0}^{N-1} h(x_k, \mu_k(x_k), x_{k+1}) \quad (3.34)$$

multiplied with the appropriate conditional probability. In particular:

- (a) The i th component of the vector $A_{\mu_0} \cdots A_{\mu_{N-1}} \bar{J}$ in Eq. (3.33) is the conditional expected value of the expression (3.34), given that $x_0 = i$ and $x_N \neq t$, multiplied with the conditional probability that $x_N \neq t$, given that $x_0 = i$.
- (b) The i th component of the vector b_{μ_0} in Eq. (3.33) is the conditional expected value of the expression (3.34), given that $x_0 = i$ and $x_1 = t$, multiplied with the conditional probability that $x_1 = t$, given that $x_0 = i$.
- (c) The i th component of the vector $A_{\mu_0} \cdots A_{\mu_{k-1}} b_{\mu_k}$ in Eq. (3.33) is the conditional expected value of the expression (3.34), given that $x_0 = i$, $x_1, \dots, x_{k-1} \neq t$, and $x_k = t$, multiplied with the conditional probability that $x_1, \dots, x_{k-1} \neq t$, and $x_k = t$, given that $x_0 = i$.

By adding these conditional probability expressions, we obtain the i th component of the unconditional expected value

$$E \left\{ \prod_{k=0}^{N-1} h(x_k, \mu_k(x_k), x_{k+1}) \right\},$$

thus verifying the formula (3.32).

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

$$h(i, u, j) = \exp(g(i, u, j)), \quad j = 1, \dots, n, t,$$

and the function g can take both positive and negative values. The mapping T_μ has the form

$$\begin{aligned} (T_\mu J)(i) &= p_{it}(\mu(i)) \exp(g(i, \mu(i), t)) \\ &\quad + \sum_{j=1}^n p_{ij}(\mu(i)) \exp(g(i, \mu(i), j)) J(j), \quad i = 1, \dots, n, \end{aligned} \quad (3.35)$$

where $p_{ij}(u)$ is the probability of transition from i to j under u , and $g(i, u, j)$ is the cost of the transition. The Bellman equation is

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

Based on Eq. (3.32), we have that $J_\pi(x_0)$ is the limit superior of the expected value of the exponential of the N -step additive finite horizon cost up to termination, i.e., $\sum_{k=0}^{\bar{k}} g(x_k, \mu_k(x_k), x_{k+1})$, where \bar{k} is equal to the first index prior to $N - 1$ such that $x_{\bar{k}+1} = t$, or is equal to $N - 1$ if there is no such index. The use of the exponential introduces risk aversion, by assigning a strictly convex increasing penalty for large rather than small cost of a trajectory up to termination (and hence a preference for small variance of the additive cost up to termination).

The deterministic version of the exponential cost problem where for each $u \in U(i)$, one of the transition probabilities $p_{it}(u), p_{i1}(u), \dots, p_{in}(u)$ is equal to 1 and all others are equal to 0, is mathematically equivalent to the classical deterministic shortest path problem (since minimizing the exponential of a deterministic expression is equivalent to minimizing that expression). For this problem a standard assumption is that there are no cycles that have negative total length to ensure that the shortest path length is finite. However, it is interesting that this assumption is not required for the analysis of the present section: when there are paths that travel perpetually around a negative length cycle we simply have $J^*(i) = 0$ for all states i on the cycle, which is permissible within our context.

Assumptions on Policies - Contractive Policies

Let us now derive an expression for the cost function of a policy. By repeatedly applying the mapping T to the equation $T_\mu J = b_\mu + A_\mu J$, we have

$$T_\mu^N J = A_\mu^N J + \sum_{k=0}^{N-1} A_\mu^k b_\mu, \quad \forall J \in \mathcal{E}^+(X), \quad N = 1, 2, \dots,$$

and hence

$$J_\mu = \limsup_{N \rightarrow \infty} T_\mu^N \bar{J} = \limsup_{N \rightarrow \infty} A_\mu^N \bar{J} + \sum_{k=0}^{\infty} A_\mu^k b_\mu \quad (3.36)$$

(the series converges since A_μ and b_μ have nonnegative components).

We say that μ is *contractive* if A_μ has eigenvalues that are strictly within the unit circle. In this case T_μ is a contraction mapping with respect to some weighted sup-norm (see Prop. B.3 in Appendix B). If μ is contractive, then $A_\mu^N \bar{J} \rightarrow 0$ and from Eq. (3.36), it follows that

$$J_\mu = \sum_{k=0}^{\infty} A_\mu^k b_\mu = (I - A_\mu)^{-1} b_\mu,$$

and J_μ is real-valued as well as nonnegative, i.e., $J_\mu \in \mathcal{R}^+(X)$. Moreover, a contractive μ is also $\mathcal{R}^+(X)$ -regular, since J_μ does not depend on the initial function \bar{J} . The reverse is also true as shown by the following proposition.

Proposition 3.5.6: A policy μ is contractive if and only if it is $\mathcal{R}^+(X)$ -regular. Moreover, if μ is noncontractive and all the components of b_μ are strictly positive, there exists a state x such that the corresponding component of the vector $\sum_{k=0}^{\infty} A_\mu^k b_\mu$ is ∞ .

Proof: As noted earlier, if μ is contractive it is $\mathcal{R}^+(X)$ -regular. It will thus suffice to show that for a noncontractive μ and strictly positive components of b_μ , some component of $\sum_{k=0}^{\infty} A_\mu^k b_\mu$ is ∞ . Indeed, according to the Perron-Frobenius Theorem, the nonnegative matrix A_μ has a real eigenvalue λ , which is equal to its spectral radius, and an associated nonnegative eigenvector $\xi \neq 0$ [see Prop. B.3(a) in Appendix B]. Choose $\gamma > 0$ to be such that $b_\mu \geq \gamma\xi$, so that

$$\sum_{k=0}^{\infty} A_\mu^k b_\mu \geq \gamma \sum_{k=0}^{\infty} A_\mu^k \xi = \gamma \left(\sum_{k=0}^{\infty} \lambda^k \right) \xi.$$

Since some component of ξ is positive while $\lambda \geq 1$ (since μ is noncontractive), the corresponding component of the infinite sum on the right is infinite, and the same is true for the corresponding component of the vector $\sum_{k=0}^{\infty} A_\mu^k b_\mu$ on the left. **Q.E.D.**

Let us introduce some assumptions that are similar to the ones of the preceding section.

Assumption 3.5.5: There exists at least one contractive policy.

Assumption 3.5.6: (Compactness and Continuity) The control space U is a metric space, and $p_{xy}(\cdot)$ and $b(x, \cdot)$ are continuous functions of u over $U(x)$, for all x and y . Moreover, for each state x , the sets

$$\left\{ u \in U(x) \mid b(x, u) + \sum_{y=1}^n A_{xy}(u)J(y) \leq \lambda \right\}$$

are compact subsets of U for all scalars $\lambda \in \Re$ and $J \in \mathcal{R}^+(X)$.

Case of Infinite Cost Noncontractive Policies

We now turn to questions relating to Bellman's equation, the convergence of the VI and PI algorithms, as well as conditions for optimality of a stationary

policy. We first consider the following assumption, which parallels the infinite cost Assumption 3.5.3 for SSP problems.

Assumption 3.5.7: (Infinite Cost Condition) For every noncontractive policy μ , there is at least one state such that the corresponding component of the vector $\sum_{k=0}^{\infty} A_{\mu}^k b_{\mu}$ is equal to ∞ .

We will now show that for $S = \mathcal{R}^+(X)$, Assumptions 3.5.5, 3.5.6, and 3.5.7 imply all the parts of Assumption 3.3.1 of Section 3.3, so Prop. 3.3.1 can be applied to the affine monotonic model. Indeed parts (a), (e) of Assumption 3.3.1 clearly hold. Part (b) also holds, since by Assumption 3.5.5 there exists a contractive and hence S -regular policy, so we have $J_S^* \in \mathcal{R}^+(X)$. Moreover Assumption 3.5.6 implies part (d), while Assumption 3.5.7 implies part (c). Finally part (f) holds since for every $J \in \mathcal{R}^+(X)$, the zero function, $J'(x) \equiv 0$, lies in $\mathcal{R}^+(X)$, and satisfies $J' \leq J$ and $J' \leq TJ'$. Thus Prop. 3.3.1 yields the following result.

Proposition 3.5.7: (Bellman's Equation, Policy Iteration, Value Iteration, and Optimality Conditions) Let Assumptions 3.5.5, 3.5.6, and 3.5.7 hold.

- (a) The optimal cost vector J^* is the unique fixed point of T within $\mathcal{R}^+(X)$.
- (b) We have $T^k J \rightarrow J^*$ for all $J \in \mathcal{R}^+(X)$.
- (c) A policy μ is optimal if and only if $T_{\mu} J^* = TJ^*$. Moreover there exists an optimal policy that is contractive.
- (d) For any $J \in \mathcal{R}^+(X)$, if $J \leq TJ$ we have $J \leq J^*$, and if $J \geq TJ$ we have $J \geq J^*$.
- (e) Every sequence $\{\mu^k\}$ generated by the PI algorithm starting from a contractive policy μ^0 satisfies $J_{\mu^k} \downarrow J^*$. Moreover, if the set of contractive policies is finite, there exists $\bar{k} \geq 0$ such that $\mu^{\bar{k}}$ is optimal.

Example 3.5.1 (Exponential Cost Shortest Path Problem)

Consider the deterministic shortest path example of Section 3.1.1, but with the exponential cost function of the present subsection; cf. Eq. (3.35). There are two policies denoted μ and μ' ; see Fig. 3.5.2. The corresponding mappings and costs are shown in the figure, and Bellman's equation is given by

$$J(1) = (TJ)(1) = \min \{ \exp(b), \exp(a)J(1) \}.$$

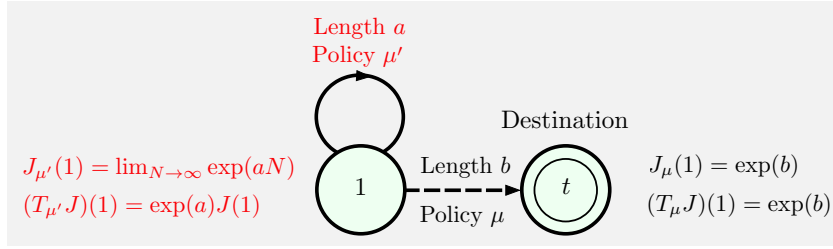


Figure 3.5.2. Shortest path problem with exponential cost function.

We consider three cases:

- (a) $a > 0$: Here the proper policy μ is optimal, and the improper policy μ' is $\mathcal{R}^+(X)$ -irregular (noncontractive) and has infinite cost, $J_{\mu'}(1) = \infty$. The assumptions of Prop. 3.5.7 hold, and consistently with the conclusions of the proposition, $J^*(1) = \exp(b)$ is the unique solution of Bellman's equation.
- (b) $a = 0$: Here the improper policy μ' is $\mathcal{R}^+(X)$ -irregular (noncontractive) and has finite cost, $J_{\mu'}(1) = 1$, so the assumptions of Prop. 3.5.7 are violated. The set of solutions of Bellman's equation within $S = \mathcal{R}^+(X)$ is the interval $[0, \exp(b)]$.
- (c) $a < 0$: Here both policies are contractive, including the improper policy μ' . The assumptions of Prop. 3.5.7 hold, and consistently with the conclusions of the proposition, $J^*(1) = 0$ is the unique solution of Bellman's equation.

The reader may also verify that in the cases where $a \neq 0$, the assumptions and the results of Prop. 3.5.7 hold.

Case of Finite Cost Noncontractive Policies

We will now eliminate Assumption 3.5.7, thus allowing noncontractive policies with real-valued cost functions, similar to the corresponding case of the preceding section, under the weak SSP conditions. Let us denote by \hat{J} the optimal cost function that can be achieved with contractive policies only,

$$\hat{J}(x) = \inf_{\mu: \text{contractive}} J_{\mu}(x), \quad x = 1, \dots, n. \quad (3.37)$$

We use the perturbation approach of Section 3.4 and Prop. 3.4.1 to show that \hat{J} is a solution of Bellman's equation. In particular, we add a constant $\delta > 0$ to all components of b_{μ} . By using arguments that are entirely analogous to the ones for the SSP case of Section 3.5.1, we obtain the following proposition, which is illustrated in Fig. 3.5.3. A detailed analysis and proof is given in the exercises.

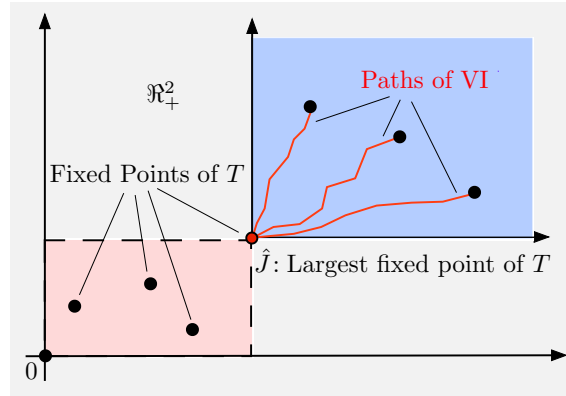


Figure 3.5.3. Schematic illustration of Prop. 3.5.8 for a problem with two states. The optimal cost function over contractive policies, \hat{J} , is the largest solution of Bellman's equation, while VI converges to \hat{J} starting from $J \geq \hat{J}$.

Proposition 3.5.8: (Bellman's Equation, Value Iteration, and Optimality Conditions) Let Assumptions 3.5.5 and 3.5.6 hold. Then:

- (a) The optimal cost function over contractive policies, \hat{J} , is the largest solution of Bellman's equation $J = TJ$ within $\mathcal{R}^+(X)$, i.e., \hat{J} is a solution that belongs to $\mathcal{R}^+(X)$, and if $J' \in \mathcal{R}^+(X)$ is another solution, we have $J' \leq \hat{J}$.
- (b) We have $T^k J \rightarrow \hat{J}$ for every $J \in \mathcal{R}^+(X)$ with $J \geq \hat{J}$.
- (c) Let μ be a contractive policy. Then μ is optimal within the class of contractive policies (i.e., $J_\mu = \hat{J}$) if and only if $T_\mu \hat{J} = T \hat{J}$.
- (d) For every $J \in \mathcal{R}^+(X)$ such that $J \leq TJ$, we have $J \leq \hat{J}$.

The other results of Section 3.5.1 for SSP problems also have straightforward analogs. Moreover, there is an adaptation of the example of Section 3.1.2, which provides an affine monotonic model for which J^* is not a fixed point of T (see the author's paper [Ber16a], to which we refer for further discussion).

Example 3.5.2 (Deterministic Shortest Path Problem with Exponential Cost - Continued)

Consider the problem of Fig. 3.5.2, for the case $a = 0$. This is the case where the noncontractive policy μ' has finite cost, so Assumption 3.5.7 is violated and Prop. 3.5.7 does not apply. However, it can be seen that the assumptions of Prop. 3.5.8 hold. Consistent with part (a) of the proposition, the optimal

cost over contractive policies, $\hat{J}(1) = \exp(b)$, is the largest of the fixed points of T . The other parts of Prop. 3.5.8 may also be easily verified.

We note that in the absence of the infinite cost Assumption 3.5.7, it is possible that the only optimal policy is noncontractive, even if the compactness Assumption 3.5.6 holds and $\hat{J} = J^*$. This is shown in the following example.

Example 3.5.3 (A Counterexample on the Existence of an Optimal Contractive Policy)

Consider the exponential cost version of the blackmailer problem of Example 3.4.2 (cf. Fig. 3.4.1). Here there is a single state 1, at which we must choose $u \in [0, 1]$. Then, we terminate at no cost [$g(1, u, t) = 0$ in Eq. (3.35)] with probability u , and we stay at state 1 at cost $-u$ [i.e., $g(1, u, 1) = -u$ in Eq. (3.35)] with probability $1 - u$. We have

$$b(i, u) = u \exp(0) = u, \quad A_{11}(u) = (1 - u) \exp(-u),$$

so that

$$H(1, u, J) = u + (1 - u) \exp(-u)J.$$

Here there is a unique noncontractive policy μ' : it chooses $u = 0$ at state 1, and has cost $J_{\mu'}(1) = 1$. Every policy μ with $\mu(1) \in (0, 1]$ is contractive, and J_{μ} can be obtained by solving the equation $J_{\mu} = T_{\mu}J_{\mu}$, i.e.,

$$J_{\mu}(1) = \mu(1) + (1 - \mu(1)) \exp(-\mu(1))J_{\mu}(1).$$

We thus obtain

$$J_{\mu}(1) = \frac{\mu(1)}{1 - (1 - \mu(1)) \exp(-\mu(1))}.$$

By minimizing over $\mu(1) \in (0, 1]$ this expression, it can be seen that $\hat{J}(1) = J^*(1) = \frac{1}{2}$, but there exists no optimal policy, and no optimal policy within the class of contractive policies [$J_{\mu}(1)$ decreases monotonically to $\frac{1}{2}$ as $\mu(1) \rightarrow 0$].

3.5.3 Robust Shortest Path Planning

We will now discuss how the analysis of Sections 3.3 and 3.4 applies to minimax shortest path-type problems, following the author's paper [Ber14], to which we refer for further discussion. To formally describe the problem, we consider a graph with a finite set of nodes $X \cup \{t\}$ and a finite set of directed arcs $\mathcal{A} \subset \{(x, y) \mid x, y \in X \cup \{t\}\}$, where t is a special node called the *destination*. At each node $x \in X$ we may choose a control u from a nonempty set $U(x)$, which is a subset of a finite set U . Then a

successor node y is selected by an antagonistic opponent from a nonempty set $Y(x, u) \subset X \cup \{t\}$ and a cost $g(x, u, y)$ is incurred. The destination node t is absorbing and cost-free, in the sense that the only outgoing arc from t is (t, t) , and we have $Y(t, u) = \{t\}$ and $g(t, u, t) = 0$ for all $u \in U(t)$.

As earlier, we denote the set of all policies by Π , and the finite set of all stationary policies by \mathcal{M} . Also, we denote the set of functions $J : X \mapsto [-\infty, \infty]$ by $\mathcal{E}(X)$, and the set of functions $J : X \mapsto (-\infty, \infty)$ by $\mathcal{R}(X)$. We introduce the mapping $H : X \times U \times \mathcal{E}(X) \mapsto [-\infty, \infty]$ given by

$$H(x, u, J) = \max_{y \in Y(x, u)} [g(x, u, y) + \tilde{J}(y)], \quad x \in X, \quad (3.38)$$

where for any $J \in \mathcal{E}(X)$ we denote by \tilde{J} the function given by

$$\tilde{J}(y) = \begin{cases} J(y) & \text{if } y \in X, \\ 0 & \text{if } y = t. \end{cases} \quad (3.39)$$

We consider the mapping $T : \mathcal{E}(X) \mapsto \mathcal{E}(X)$ defined by

$$(TJ)(x) = \min_{u \in U(x)} H(x, u, J), \quad x \in X, \quad (3.40)$$

and for each policy μ , the mapping $T_\mu : \mathcal{E}(X) \mapsto \mathcal{E}(X)$, defined by

$$(T_\mu J)(x) = H(x, \mu(x), J), \quad x \in X. \quad (3.41)$$

We let \bar{J} be the zero function,

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

The cost function of a policy $\pi = \{\mu_0, \mu_1, \dots\}$ is

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

and $J^*(x) = \inf_{\pi \in \Pi} J_\pi(x)$, cf. Definition 3.2.1.

For a policy $\mu \in \mathcal{M}$, we define a *possible path under μ starting at node $x_0 \in X$* to be an arc sequence of the form

$$p = \{(x_0, x_1), (x_1, x_2), \dots\},$$

such that $x_{k+1} \in Y(x_k, \mu(x_k))$ for all $k \geq 0$. The set of all possible paths under μ starting at x_0 is denoted by $P(x_0, \mu)$. The length of a path $p \in P(x_0, \mu)$ is defined by

$$L_\mu(p) = \limsup_{m \rightarrow \infty} \sum_{k=0}^m g(x_k, \mu(x_k), x_{k+1}).$$

Using Eqs. (3.38)-(3.41), we see that for any $\mu \in \mathcal{M}$ and $x \in X$, $(T_\mu^k \bar{J})(x)$ is the result of the k -stage DP algorithm that computes the length of the *longest path* under μ that starts at x and consists of k arcs.

For completeness, we also define the length of a portion

$$\{(x_i, x_{i+1}), (x_{i+1}, x_{i+2}), \dots, (x_m, x_{m+1})\}$$

of a path $p \in P(x_0, \mu)$, consisting of a finite number of consecutive arcs, by

$$\sum_{k=i}^m g(x_k, \mu(x_k), x_{k+1}).$$

When confusion cannot arise we will also refer to such a finite-arc portion as a path. Of special interest are *cycles*, i.e., paths of the form $\{(x_i, x_{i+1}), (x_{i+1}, x_{i+2}), \dots, (x_{i+m}, x_i)\}$. Paths that do not contain any cycle other than the self-cycle (t, t) are called *simple*.

For a given policy $\mu \in \mathcal{M}$ and $x_0 \neq t$, a path $p \in P(x_0, \mu)$ is said to be *terminating* if it has the form

$$p = \{(x_0, x_1), (x_1, x_2), \dots, (x_m, t), (t, t), \dots\}, \quad (3.42)$$

where m is a positive integer, and x_0, \dots, x_m are distinct nondestination nodes. Since $g(t, u, t) = 0$ for all $u \in U(t)$, the length of a terminating path p of the form (3.42), corresponding to μ , is given by

$$L_\mu(p) = g(x_m, \mu(x_m), t) + \sum_{k=0}^{m-1} g(x_k, \mu(x_k), x_{k+1}),$$

and is equal to the finite length of its initial portion that consists of the first $m + 1$ arcs.

An important characterization of a policy $\mu \in \mathcal{M}$ is provided by the subset of arcs

$$\mathcal{A}_\mu = \cup_{x \in X} \{(x, y) \mid y \in Y(x, \mu(x))\}.$$

Thus $\mathcal{A}_\mu \cup (t, t)$ can be viewed as the set of all possible paths under μ , $\cup_{x \in X} P(x, \mu)$, in the sense that it contains this set of paths and no other paths. We refer to \mathcal{A}_μ as the *characteristic graph of μ* . We say that \mathcal{A}_μ is *destination-connected* if for each $x \in X$ there exists a terminating path in $P(x, \mu)$.

We say that μ is *proper* if the characteristic graph \mathcal{A}_μ is acyclic (i.e., contains no cycles). Thus μ is proper if and only if all the paths in $\cup_{x \in X} P(x, \mu)$ are simple and hence terminating (equivalently μ is proper if and only if \mathcal{A}_μ is destination-connected and has no cycles). The term “proper” is consistent with the one used in Section 3.5.1 for SSP problems, where it indicates a policy under which the destination is reached

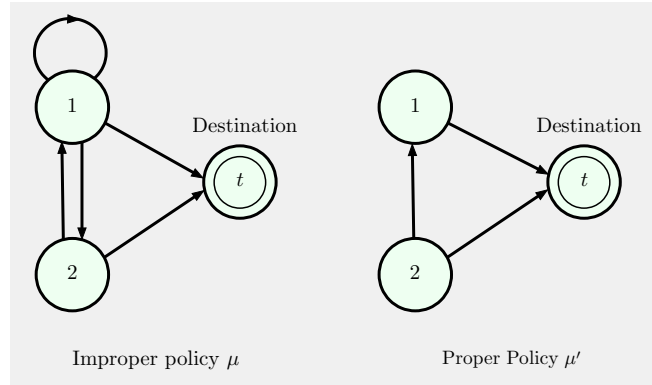


Figure 3.5.4. A robust shortest path problem with $X = \{1, 2\}$, two controls at node 1, and one control at node 2. The two policies, μ and μ' , correspond to the two controls at node 1. The figure shows the characteristic graphs \mathcal{A}_μ and $\mathcal{A}_{\mu'}$.

with probability 1. If μ is not proper, it is called *improper*, in which case the characteristic graph \mathcal{A}_μ must contain a cycle; see the examples of Fig. 3.5.4. Intuitively, a policy is improper, if and only if under that policy there are initial states such that the antagonistic opponent can force movement along a cycle without ever reaching the destination.

The following proposition clarifies the properties of J_μ when μ is improper.

Proposition 3.5.9: Let μ be an improper policy.

- (a) If all cycles in the characteristic graph \mathcal{A}_μ have nonpositive length, $J_\mu(x) < \infty$ for all $x \in X$.
- (b) If all cycles in the characteristic graph \mathcal{A}_μ have nonnegative length, $J_\mu(x) > -\infty$ for all $x \in X$.
- (c) If all cycles in the characteristic graph \mathcal{A}_μ have zero length, J_μ is real-valued.
- (d) If there is a positive length cycle in the characteristic graph \mathcal{A}_μ , we have $J_\mu(x) = \infty$ for at least one node $x \in X$. More generally, for each $J \in \mathcal{R}(X)$, we have $\limsup_{k \rightarrow \infty} (T_\mu^k J)(x) = \infty$ for at least one $x \in X$.

Proof: Any path with a finite number of arcs, can be decomposed into a simple path, and a finite number of cycles (see e.g., the path decomposition theorem of [Ber98], Prop. 1.1, and Exercise 1.4). Since there is only a finite number of simple paths under μ , their length is bounded above and below. Thus in part (a) the length of all paths with a finite number of

arcs is bounded above, and in part (b) it is bounded below, implying that $J_\mu(x) < \infty$ for all $x \in X$ or $J_\mu(x) > -\infty$ for all $x \in X$, respectively. Part (c) follows by combining parts (a) and (b).

To show part (d), consider a path p , which consists of an infinite repetition of the positive length cycle that is assumed to exist. Let $C_\mu^k(p)$ be the length of the path that consists of the first k cycles in p . Then $C_\mu^k(p) \rightarrow \infty$ and $C_\mu^k(p) \leq J_\mu(x)$ for all k , where x is the first node in the cycle, thus implying that $J_\mu(x) = \infty$. Moreover for every $J \in \mathcal{R}(X)$ and all k , $(T_\mu^k J)(x)$ is the maximum over the lengths of the k -arc paths that start at x , plus a terminal cost that is equal to either $J(y)$ (if the terminal node of the k -arc path is $y \in X$), or 0 (if the terminal node of the k -arc path is the destination). Thus we have,

$$(T_\mu^k \bar{J})(x) + \min \left\{ 0, \min_{x \in X} J(x) \right\} \leq (T_\mu^k J)(x).$$

Since $\limsup_{k \rightarrow \infty} (T_\mu^k \bar{J})(x) = J_\mu(x) = \infty$ as shown earlier, it follows that $\limsup_{k \rightarrow \infty} (T_\mu^k J)(x) = \infty$ for all $J \in \mathcal{R}(X)$. **Q.E.D.**

Note that if there is a negative length cycle in the characteristic graph \mathcal{A}_μ , it is not necessarily true that for some $x \in X$ we have $J_\mu(x) = -\infty$. Even for x on the negative length cycle, the value of $J_\mu(x)$ is determined by the *longest* path in $P(x, \mu)$, which may be simple in which case $J_\mu(x)$ is a real number, or contain an infinite repetition of a positive length cycle in which case $J_\mu(x) = \infty$.

Properness and Regularity

We will now make a formal connection between the notions of properness and $\mathcal{R}(X)$ -regularity. We recall that μ is $\mathcal{R}(X)$ -regular if $J_\mu \in \mathcal{R}(X)$, $J_\mu = T_\mu J_\mu$, and $T_\mu^k J \rightarrow J_\mu$ for all $J \in \mathcal{R}(X)$ (cf. Definition 3.2.2). Clearly if μ is proper, we have $J_\mu \in \mathcal{R}(X)$ and the equation $J_\mu = T_\mu J_\mu$ holds (this is Bellman's equation for the longest path problem involving the acyclic graph \mathcal{A}_μ). We will also show that $T_\mu^k J \rightarrow J_\mu$ for all $J \in \mathcal{R}(X)$, so that a proper policy is $\mathcal{R}(X)$ -regular. However, the following proposition shows that there may be some $\mathcal{R}(X)$ -regular policies that are improper, depending on the sign of the lengths of their associated cycles.

Proposition 3.5.10: The following are equivalent for a policy μ :

- (i) μ is $\mathcal{R}(X)$ -regular.
- (ii) The characteristic graph \mathcal{A}_μ is destination-connected and all its cycles have negative length.
- (iii) μ is either proper or else it is improper, all the cycles of the characteristic graph \mathcal{A}_μ have negative length, and $J_\mu \in \mathcal{R}(X)$.

Proof: To show that (i) implies (ii), let μ be $\mathcal{R}(X)$ -regular and to arrive at a contradiction, assume that \mathcal{A}_μ contains a nonnegative length cycle. Let x be a node on the cycle, consider the path p that starts at x and consists of an infinite repetition of this cycle, and let $L_\mu^k(p)$ be the length of the first k arcs of that path. Let also J be a constant function, $J(x) \equiv r$, where r is a scalar. Then we have

$$L_\mu^k(p) + r \leq (T_\mu^k J)(x),$$

since from the definition of T_μ , we have that $(T_\mu^k J)(x)$ is the maximum over the lengths of all k -arc paths under μ starting at x , plus r , if the last node in the path is not the destination. Since μ is $\mathcal{R}(X)$ -regular, we have $J_\mu \in \mathcal{R}(X)$ and $\limsup_{k \rightarrow \infty} (T_\mu^k J)(x) = J_\mu(x) < \infty$, so that for all scalars r ,

$$\limsup_{k \rightarrow \infty} (L_\mu^k(p) + r) \leq J_\mu(x) < \infty.$$

Taking supremum over $r \in \mathfrak{R}$, it follows that $\limsup_{k \rightarrow \infty} L_\mu^k(p) = -\infty$, which contradicts the nonnegativity of the cycle of p . Thus all cycles of \mathcal{A}_μ have negative length. To show that \mathcal{A}_μ is destination-connected, assume the contrary. Then there exists some node $x \in X$ such that all paths in $P(x, \mu)$ contain an infinite number of cycles. Since the length of all cycles is negative, as just shown, it follows that $J_\mu(x) = -\infty$, which contradicts the $\mathcal{R}(X)$ -regularity of μ .

To show that (ii) implies (iii), we assume that μ is improper and show that $J_\mu \in \mathcal{R}(X)$. By (ii) \mathcal{A}_μ is destination-connected, so the set $P(x, \mu)$ contains a simple path for all $x \in X$. Moreover, since by (ii) the cycles of \mathcal{A}_μ have negative length, each path in $P(x, \mu)$ that is not simple has smaller length than some simple path in $P(x, \mu)$. This implies that $J_\mu(x)$ is equal to the largest path length among simple paths in $P(x, \mu)$, so $J_\mu(x)$ is a real number for all $x \in X$.

To show that (iii) implies (i), we note that if μ is proper, it is $\mathcal{R}(X)$ -regular, so we focus on the case where μ is improper. Then by (iii), $J_\mu \in \mathcal{R}(X)$, so to show $\mathcal{R}(X)$ -regularity of μ , we must show that $(T_\mu^k J)(x) \rightarrow J_\mu(x)$ for all $x \in X$ and $J \in \mathcal{R}(X)$, and that $J_\mu = T_\mu J_\mu$. Indeed, from the definition of T_μ , we have

$$(T_\mu^k J)(x) = \sup_{p \in P(x, \mu)} [L_\mu^k(p) + J(x_p^k)], \quad (3.43)$$

where $L_\mu^k(p)$ is the length of the first k arcs of path p , x_p^k is the node reached after k arcs along the path p , and $J(t)$ is defined to be equal to 0. Thus as $k \rightarrow \infty$, for every path p that contains an infinite number of cycles (each necessarily having negative length), the sequence $L_\mu^k(p) + J(x_p^k)$ approaches $-\infty$. It follows that for sufficiently large k , the supremum in Eq. (3.43) is attained by one of the simple paths in $P(x, \mu)$, so $x_p^k = t$ and $J(x_p^k) = 0$. Thus the limit of $(T_\mu^k J)(x)$ does not depend on J , and is equal to the limit

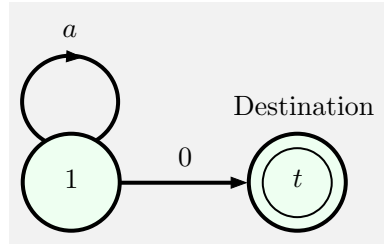


Figure 3.5.5. The characteristic graph \mathcal{A}_μ corresponding to an improper policy, for the case of a single node 1 and a destination node t . The arcs lengths are shown in the figure.

of $(T_\mu^k \bar{J})(x)$, i.e., $J_\mu(x)$. To show that $J_\mu = T_\mu J_\mu$, we note that by the preceding argument, $J_\mu(x)$ is the length of the longest path among paths that start at x and terminate at t . Moreover, we have

$$(T_\mu J_\mu)(x) = \max_{y \in Y(x, \mu(x))} [g(x, \mu(x), y) + J_\mu(y)],$$

where we denote $J_\mu(t) = 0$. Thus $(T_\mu J_\mu)(x)$ is also the length of the longest path among paths that start at x and terminate at t , and hence it is equal to $J_\mu(x)$. **Q.E.D.**

We illustrate the preceding proposition, in relation to the infinite cost condition of Assumption 3.3.1, with a two-node example involving an improper policy with a cycle that may have positive, zero, or negative length.

Example 3.5.4:

Let $X = \{1\}$, and consider the policy μ where at state 1, the antagonistic opponent may force either staying at 1 or terminating, i.e., $Y(1, \mu(1)) = \{1, t\}$; cf. Fig. 3.5.5. Then μ is improper since its characteristic graph \mathcal{A}_μ contains the self-cycle $(1, 1)$. Let

$$g(1, \mu(1), 1) = a, \quad g(1, \mu(1), t) = 0.$$

Then,

$$(T_\mu J_\mu)(1) = \max [0, a + J_\mu(1)],$$

and

$$J_\mu(1) = \begin{cases} \infty & \text{if } a > 0, \\ 0 & \text{if } a \leq 0. \end{cases}$$

Consistently with Prop. 3.5.10, the following hold:

- (a) For $a > 0$, the cycle $(1, 1)$ has positive length, and μ is $\mathcal{R}(X)$ -irregular. Here we have $J_\mu(1) = \infty$, and the infinite cost condition of Assumption 3.3.1 is satisfied.

- (b) For $a = 0$, the cycle $(1, 1)$ has zero length, and μ is $\mathcal{R}(X)$ -irregular. Here we have $J_\mu(1) = 0$, and the infinite cost condition of Assumption 3.3.1 is violated because for a function $J \in \mathcal{R}(X)$ with $J(1) > 0$,

$$\limsup_{k \rightarrow \infty} (T_\mu^k J)(x) = J(1) > 0 = J_\mu(1).$$

- (c) For $a < 0$, the cycle $(1, 1)$ has negative length, and μ is $\mathcal{R}(X)$ -regular. Here we have $J_\mu \in \mathcal{R}(X)$, $J_\mu(1) = \max [0, a + J_\mu(1)] = (T_\mu J_\mu)(1)$, and for all $J \in \mathcal{R}(X)$,

$$\lim_{k \rightarrow \infty} (T_\mu^k J)(1) = 0 = J_\mu(1).$$

We will now apply the regularity results of Sections 3.2-3.4 with $S = \mathcal{R}(X)$. To this end, we introduce assumptions that will allow the use of Prop. 3.3.1.

Assumption 3.5.8:

- (a) There exists at least one $\mathcal{R}(X)$ -regular policy.
- (b) For every $\mathcal{R}(X)$ -irregular policy μ , some cycle in the characteristic graph \mathcal{A}_μ has positive length.

Assumption 3.5.8 is implied by the weaker conditions given in the following proposition. These conditions may be more easily verifiable in some contexts.

Proposition 3.5.11: Assumption 3.5.8 holds if anyone of the following two conditions is satisfied.

- (1) There exists at least one proper policy, and for every improper policy μ , all cycles in the characteristic graph \mathcal{A}_μ have positive length.
- (2) Every policy μ is either proper or else it is improper and its characteristic graph \mathcal{A}_μ is destination-connected with all cycles having negative length, and $J_\mu \in \mathcal{R}(X)$.

Proof: Under condition (1), by Prop. 3.5.10, a policy is $\mathcal{R}(X)$ -regular if and only if it is proper. Moreover, since each $\mathcal{R}(X)$ -irregular and hence improper policy μ has cycles with positive length, it follows that for all $J \in \mathcal{R}(X)$, we have

$$\limsup_{k \rightarrow \infty} (T_\mu^k J)(x) = \infty$$

for some $x \in X$. The proof under condition (2) is similar, using Prop. 3.5.10. **Q.E.D.**

We now show our main result for the problem of this section.

Proposition 3.5.12: Let Assumption 3.5.8 hold. Then:

- (a) The optimal cost function J^* is the unique fixed point of T within $\mathcal{R}(X)$.
- (b) We have $T^k J \rightarrow J^*$ for all $J \in \mathcal{R}(X)$.
- (c) A policy μ^* is optimal if and only if $T_{\mu^*} J^* = T J^*$. Moreover, there exists an optimal proper policy.
- (d) For any $J \in \mathcal{R}(X)$, if $J \leq T J$ we have $J \leq J^*$, and if $J \geq T J$ we have $J \geq J^*$.

Proof: We verify the parts (a)-(f) of Assumption 3.3.1 with $S = \mathcal{R}(X)$, and we then use Prop. 3.3.1. To this end we argue as follows:

- (1) Part (a) is satisfied since $S = \mathcal{R}(X)$.
- (2) Part (b) is satisfied since by Assumption 3.5.8(a), there exists at least one $\mathcal{R}(X)$ -regular policy. Moreover, for each $\mathcal{R}(X)$ -regular policy μ , we have $J_\mu \in \mathcal{R}(X)$. Since the number of all policies is finite, it follows that $J_S^* \in \mathcal{R}(X)$.
- (3) To show that part (c) is satisfied, note that by Prop. 3.5.10 every $\mathcal{R}(X)$ -irregular policy μ must be improper, so by Assumption 3.5.8(b), the characteristic graph \mathcal{A}_μ contains a cycle of positive length. By Prop. 3.5.9(d), this implies that for each $J \in \mathcal{R}(X)$ and for at least one $x \in X$, we have $\limsup_{k \rightarrow \infty} (T_\mu^k J)(x) = \infty$.
- (4) Part (d) is satisfied since $U(x)$ is a finite set.
- (5) Part (e) is satisfied since X is finite and T_μ is a continuous function that maps the finite-dimensional space $\mathcal{R}(X)$ into itself.
- (6) Part (f) follows from Prop. 3.3.2, which applies because $S = \mathcal{R}(X) = R_b(X)$ (since X is finite) and Eq. (3.27) clearly holds.

Thus all parts of Assumption 3.3.1 are satisfied, and Prop. 3.3.1 applies with $S = \mathcal{R}(X)$. The conclusions of this proposition are precisely the results we want to prove [since improper policies have infinite cost for some initial states, as argued earlier, optimal S -regular policies must be proper; cf. the conclusion of part (c)]. **Q.E.D.**

The following example illustrates what may happen in the absence of Assumption 3.5.8(b), when there may exist improper policies that involve

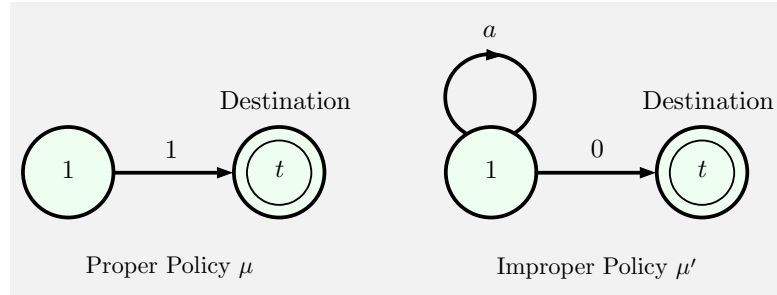


Figure 3.5.6. A counterexample involving a single node 1 in addition to the destination t . There are two policies, μ and μ' , with corresponding characteristic graphs \mathcal{A}_μ and $\mathcal{A}_{\mu'}$, and arc lengths shown in the figure. The improper policy μ' is optimal when $a \leq 0$. It is $\mathcal{R}(X)$ -irregular if $a = 0$, and it is $\mathcal{R}(X)$ -regular if $a < 0$.

a nonpositive length cycle.

Example 3.5.5:

Let $X = \{1\}$, and consider the proper policy μ with $Y(1, \mu(1)) = \{t\}$ and the improper policy μ' with $Y(1, \mu'(1)) = \{1, t\}$ (cf. Fig. 3.5.6). Let

$$g(1, \mu(1), t) = 1, \quad g(1, \mu'(1), 1) = a \leq 0, \quad g(1, \mu'(1), t) = 0.$$

The improper policy is the same as the one of Example 3.5.4. It can be seen that under both policies, the longest path from 1 to t consists of the arc $(1, t)$. Thus,

$$J_\mu(1) = 1, \quad J_{\mu'}(1) = 0,$$

so the improper policy μ' is optimal, and strictly dominates the proper policy μ . To explain what is happening here, we consider two different cases:

- (1) $a = 0$: In this case, the optimal policy μ' is both improper and $\mathcal{R}(X)$ -irregular, but with finite cost $J_{\mu'}(1) < \infty$. Thus the conditions of Props. 3.3.1 and 3.5.12 do not hold because Assumptions 3.3.1(c) and 3.5.9(b) are violated.
- (2) $a < 0$: In this case, μ' is improper but $\mathcal{R}(X)$ -regular, so there are no $\mathcal{R}(X)$ -irregular policies. Then all the conditions of Assumption 3.5.8 are satisfied, and Prop. 3.5.12 applies. Consistent with this proposition, there exists an optimal $\mathcal{R}(X)$ -regular policy (i.e., optimal over both proper and improper policies), which however is improper.

For further analysis and algorithms for the robust shortest path planning problem, we refer to the paper [Ber14]. In particular, this paper applies the perturbation approach of Section 3.4 to the case where it may be easier to guarantee nonnegativity rather than positivity of the lengths

of cycles corresponding to improper policies, which is required by Assumption 3.5.8(b). The paper shows that the VI algorithm terminates in a finite number of iterations starting from the initial function J with $J(x) = \infty$ for all $x \in X$. Moreover the paper provides a Dijkstra-like algorithm for problems with nonnegative arc lengths.

3.5.4 Linear-Quadratic Optimal Control

In this subsection, we consider a classical problem from control theory, which involves the deterministic linear system

$$x_{k+1} = Ax_k + Bu_k, \quad k = 0, 1, \dots,$$

where $x_k \in \mathfrak{R}^n$, $u_k \in \mathfrak{R}^m$ for all k , and A and B are given matrices. The cost function of a policy $\pi = \{\mu_0, \mu_1, \dots\}$ has the form

$$J_\pi(x_0) = \lim_{N \rightarrow \infty} \sum_{k=0}^{N-1} (x_k' Q x_k + \mu_k(x_k)' R \mu_k(x_k)),$$

where x' denotes the transpose of a column vector x , Q is a positive semidefinite symmetric $n \times n$ matrix, and R is a positive definite symmetric $m \times m$ matrix. This is a special case of the deterministic optimal control problem of Section 1.1, and was discussed briefly in the context of the one-dimensional example of Section 3.1.4.

The theory of this problem is well-known and is discussed in various forms in many sources, including the textbooks [AnM79] and [Ber17a] (Section 3.1). The solution revolves around stationary policies μ that are *linear*, in the sense that

$$\mu(x) = Lx,$$

where L is some $m \times n$ matrix, and *stable*, in the sense that the matrix $A + BL$ has eigenvalues that are strictly within the unit circle. Thus for a linear stable policy, the closed loop system

$$x_{k+1} = (A + BL)x_k$$

is stable. We assume that *there exists at least one linear stable policy*. Among others, this guarantees that the optimal cost function J^* is real-valued (it is bounded above by the real-valued cost function of every linear stable policy).

The solution also revolves around the *algebraic matrix Riccati equation*, which is given by

$$P = A'(P - PB(B'PB + R)^{-1}B'P)A + Q,$$

where the unknown is P , a symmetric $n \times n$ matrix. It is well-known that if Q is positive definite, then the Riccati equation has a unique solution P^*

within the class of positive semidefinite symmetric matrices, and that the optimal cost function has the form

$$J^*(x) = x'P^*x.$$

Moreover, there is a unique optimal policy, and this policy is linear stable of the form

$$\mu^*(x) = Lx, \quad L = -(B'P^*B + R)^{-1}B'P^*A.$$

The existence of an optimal linear stable policy can be extended to the case where Q is instead positive semidefinite, but satisfies a certain “detectability” condition; see the textbooks cited earlier.

However, in the general case where Q is positive semidefinite without further assumptions (e.g., $Q = 0$), the example of Section 3.1.4 shows that the optimal policy need not be stable, and in fact the optimal cost function over just the linear stable policies may be different than J^* .[†] We will discuss this case by using the perturbation-based approach of Section 3.4, and provide results that are consistent with the behavior observed in the example of Section 3.1.4.

To convert the problem to our abstract format, we let

$$X = \mathfrak{R}^n, \quad U(x) = \mathfrak{R}^m, \quad \bar{J}(x) = 0, \quad \forall x \in X,$$

$$H(x, u, J) = x'Qx + u'Ru + J(Ax + Bu).$$

Let S be the set of positive semidefinite quadratic functions, i.e.,

$$S = \{J \mid J(x) = x'Px, P : \text{positive semidefinite symmetric}\}.$$

Let $\widehat{\mathcal{M}}$ be the set of linear stable policies, and note that every linear stable policy is S -regular. This is due to the fact that for every quadratic function $J(x) = x'Px$ and linear stable policy $\mu(x) = Lx$, the k -stage costs $(T_\mu^k J)(x)$ and $(T_\mu^k \bar{J})(x)$ differ by the term

$$x'(A + BL)^k P (A + BL)^k x,$$

which vanishes in the limit as $k \rightarrow \infty$, since μ is stable.

Consider the perturbation framework of Section 3.4, with forcing function

$$p(x) = \|x\|^2.$$

[†] This is also true in the discounted version of the example of Section 3.1.4, where there is a discount factor $\alpha \in (0, 1)$. The Riccati equation then takes the form $P = A'(\alpha P - \alpha^2 PB(\alpha B'PB + R)^{-1}B'P)A + Q$, and for the given system and cost per stage, it has two solutions, $P^* = 0$ and $\hat{P} = \frac{\alpha\gamma^2 - 1}{\alpha}$. The VI algorithm converges to \hat{P} starting from any $P > 0$.

Then for $\delta > 0$, the mapping $T_{\mu,\delta}$ has the form

$$(T_{\mu,\delta}J)(x) = x'(Q + \delta I)x + \mu(x)'R\mu(x) + J(Ax + B\mu(x)),$$

where I is the identity, and corresponds to the linear-quadratic problem where Q is replaced by the positive definite matrix $Q + \delta I$. This problem admits a quadratic positive definite optimal cost $\hat{J}_\delta(x) = x'P_\delta^*x$, and an optimal linear stable policy. Moreover, all the conditions of Prop. 3.4.1 can be verified. It follows that J_S^* is equal to the optimal cost over just the linear stable policies \hat{J} , and is obtained as $\lim_{\delta \rightarrow 0} \hat{J}_\delta$, which also implies that $\hat{J}(x) = x'\hat{P}x$ where $\hat{P} = \lim_{\delta \rightarrow 0} P_\delta^*$.

The perturbation line of analysis of the linear-quadratic problem will be generalized in Section 4.5. This generalization will address a deterministic discrete-time infinite horizon optimal control problem involving the system

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

a nonnegative cost per stage $g(x, u)$, and a cost-free termination state. We will introduce there a notion of stability, and we will show that the optimal cost function over the stable policies is the largest solution of Bellman's equation. Moreover, we will show that the VI algorithm and several versions of the PI algorithm are valid for suitable initial conditions.

3.5.5 Continuous-State Deterministic Optimal Control

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

$$x_{k+1} = f(x_k, u_k), \quad k = 0, 1, \dots, \quad (3.44)$$

where x_k and u_k are the state and control at stage k , belonging to sets X and U , respectively, and f is a function mapping $X \times U$ to X . The control u_k must be chosen from a constraint set $U(x_k)$. No restrictions are placed on the nature of X and U : for example, they may be finite sets as in deterministic shortest path problems, or they may be continuous spaces as in classical problems of control to the origin or some other terminal set, including the linear-quadratic problem of Section 3.5.4. The cost per stage is denoted by $g(x, u)$, and is assumed to be a real number. †

Because the system is deterministic, given an initial state x_0 , a policy $\pi = \{\mu_0, \mu_1, \dots\}$ when applied to the system (3.44), generates a unique sequence of state-control pairs $(x_k, \mu_k(x_k))$, $k = 0, 1, \dots$. The corresponding

† In Section 4.5, we will consider a similar problem where the cost per stage will be assumed to be nonnegative, but some other assumptions from the present section (e.g., the subsequent Assumption 3.5.9) will be relaxed.

cost function is

$$J_\pi(x_0) = \limsup_{N \rightarrow \infty} \sum_{k=0}^{N-1} g(x_k, \mu_k(x_k)), \quad x_0 \in X.$$

We assume that there is a nonempty *stopping set* $X_0 \subset X$, which consists of cost-free and absorbing states in the sense that

$$g(x, u) = 0, \quad x = f(x, u), \quad \forall x \in X_0, u \in U(x). \quad (3.45)$$

Based on our assumptions to be introduced shortly, the objective will be roughly to reach or asymptotically approach the set X_0 at minimum cost.

To formulate a corresponding abstract DP problem, we introduce the mapping $T_\mu : \mathcal{R}(X) \mapsto \mathcal{R}(X)$ by

$$(T_\mu J)(x) = g(x, \mu(x)) + J(f(x, \mu(x))), \quad x \in X,$$

and the mapping $T : \mathcal{E}(X) \mapsto \mathcal{E}(X)$ given by

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

Here as earlier, we denote by $\mathcal{R}(X)$ the set of real-valued functions over X , and by $\mathcal{E}(X)$ the set of extended real-valued functions over X . The initial function \bar{J} is the zero function [$\bar{J}(x) \equiv 0$]. An important fact is that because the problem is deterministic, J^* is a *fixed point of T* (cf. Exercise 3.1).

The analysis of the linear-quadratic problem of the preceding section has revealed two distinct types of behavior for the case where $g \geq 0$:

- (a) J^* is the unique fixed point of T within the set S (the set of nonnegative definite quadratic functions).
- (b) J^* and the optimal cost function \hat{J} over a restricted subset of S -regular policies (the linear stable policies) are both fixed points of T within the set S , but $J^* \neq \hat{J}$, and the VI algorithm converges to \hat{J} when started with a function $J \geq \hat{J}$.

In what follows we will introduce assumptions that preclude case (b); we will postpone the discussion of this case for Section 4.5, where we will use a perturbation-based line of analysis. Similar to the linear-quadratic problem, the restricted set of policies that we will consider have some “stability” property: they are either terminating (reach X_0 in a finite number of step), or else they asymptotically approach X_0 in a manner to be made precise later.

As a first step in the analysis, let us introduce the effective domain of J^* , i.e., the set

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

Ordinarily, in practical applications, the states in X^* are those from which one can reach the stopping set X_0 , at least asymptotically. We say that a policy μ is *terminating* if starting from any $x_0 \in X^*$, the state sequence $\{x_k\}$ generated using μ reaches X_0 in finite time, i.e., satisfies $x_{\bar{k}} \in X_0$ for some index \bar{k} . The set of terminating policies is denoted by $\widehat{\mathcal{M}}$.

Our key assumption in this section is that for all $x \in X^*$, the optimal cost $J^*(x)$ can be approximated arbitrarily closely by using terminating policies. In Section 4.5 we will relax this assumption.

Assumption 3.5.9: (Near-Optimal Termination) For every pair (x, ϵ) with $x \in X^*$ and $\epsilon > 0$, there exists a terminating policy μ [possibly dependent on (x, ϵ)] that satisfies $J_\mu(x) \leq J^*(x) + \epsilon$.

This assumption implies in particular that the optimal cost function over terminating policies,

$$\hat{J}(x) = \inf_{\mu \in \widehat{\mathcal{M}}} J_\mu(x), \quad x \in X,$$

is equal to J^* . Note that Assumption 3.5.9 is equivalent to a seemingly weaker assumption where nonstationary policies can be used for termination (see Exercise 3.7).

Specific and easily verifiable conditions that imply Assumption 3.5.9 are given in the exercises. A prominent case is when X and U are finite, so the problem becomes a deterministic shortest path problem. If all cycles of the state transition graph have positive length, then for every π and x with $J_\pi(x) < \infty$ the generated path starting from x and using π must reach the destination, and this implies that there exists an optimal policy that terminates from all $x \in X^*$. Thus, in this case Assumption 3.5.9 is naturally satisfied.

Another interesting case arises when $g(x, u) = 0$ for all (x, u) except if $x \notin X_0$ and the next state $f(x, u)$ is a termination state, in which case the cost of the stage is strictly negative, i.e., $g(x, u) < 0$ only when $f(x, u) \in X_0$. Thus no cost is incurred except for a negative cost upon termination. Intuitively, this is the problem of trying to find the best state from which to terminate, out of all states that are reachable from the initial state x_0 . Then, assuming that X_0 can be reached from all states, Assumption 3.5.9 is satisfied.

When X is the n -dimensional Euclidean space \mathfrak{R}^n , it may easily happen that the optimal policies are not terminating from some $x \in X^*$, but instead the optimal state trajectories may approach X_0 asymptotically. This is true for example in the linear-quadratic problem of the preceding section, where $X = \mathfrak{R}^n$, $X_0 = \{0\}$, $U = \mathfrak{R}^m$, the system is linear of the form $x_{k+1} = Ax_k + Bu_k$, where A and B are given matrices, and the optimal cost

function is positive definite quadratic. There the optimal policy is linear stable of the form $\mu^*(x) = Lx$, where L is some matrix obtained through the steady-state solution of the Riccati equation. Since the optimal closed-loop system has the form $x_{k+1} = (A + BL)x_k$, the state will typically never reach the termination set $X_0 = \{0\}$ in finite time, although it will approach it asymptotically. However, the Assumption 3.5.9 is satisfied under some natural and easily verifiable conditions (see Exercise 3.8).

Let us consider the set of functions

$$S = \{J \in \mathcal{E}(X) \mid J(x) = 0, \forall x \in X_0, J(x) \in \mathfrak{R}, \forall x \in X^*\}.$$

Since X_0 consists of cost-free and absorbing states [cf. Eq. (3.45)], and $J^*(x) > -\infty$ for all $x \in X$ (by Assumption 3.5.9), the set S contains the cost functions J_μ of all terminating policies μ , as well as J^* . Moreover it can be seen that every terminating policy is S -regular, i.e., $\widehat{\mathcal{M}} \subset \mathcal{M}_S$, implying that $J_S^* = J^*$. The reason is that the terminal cost is zero after termination for any terminal cost function $J \in S$, i.e.,

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

for $\mu \in \widehat{\mathcal{M}}$, $x \in X^*$, and k sufficiently large.

The following proposition is a consequence of the well-behaved region theorem (Prop. 3.2.1), the deterministic character of the problem (which guarantees that J^* is a fixed point of T), and Assumption 3.5.9 (which guarantees that $J_S^* = J^*$).

Proposition 3.5.13: Let Assumption 3.5.9 hold. Then:

- (a) J^* is the unique solution of the Bellman equation $J = TJ$ within the set of all $J \in S$ such that $J \geq J^*$.
- (b) We have $T^k J \rightarrow J^*$ for every $J \in S$ such that $J \geq J^*$.
- (c) If μ^* is terminating and $T_{\mu^*} J^* = TJ^*$, then μ^* is optimal. Conversely, if μ^* is terminating and is optimal, then $T_{\mu^*} J^* = TJ^*$.

To see what may happen in the absence of Assumption 3.5.9, consider the deterministic shortest path example of Section 3.1.1 with $a = 0$, $b > 0$, and $S = \mathfrak{R}$. Here Assumption 3.5.9 is violated and we have $0 = J^* < \hat{J} = b$, while the set of fixed points of T is the interval $(-\infty, b]$. However, for the same example, but with $b \leq 0$ instead of $b > 0$, Assumption 3.5.9 is satisfied and Prop. 3.5.13 applies. Consider also the linear-quadratic example of Section 3.1.4. Here Assumption 3.5.9 is violated. This results in multiple fixed points of T within S : the functions $J^*(x) \equiv 0$ and $\hat{J}(x) = (\gamma^2 - 1)x^2$. In Section 4.5, we will reconsider this example, as well as the

problem of this section for the case $g(x, u) \geq 0$ for all (x, u) , but under assumptions that are much weaker than Assumption 3.5.9. There, we will make a connection between regularity, perturbations like the ones of Section 3.4, and traditional notions of stability.

Another interesting fact is that when the model of this section is extended in the natural way to a stochastic model with infinite state space, then under the analog of Assumption 3.5.9, J^* need not be the unique solution of Bellman's equation within the set of all $J \in S$ such that $J \geq J^*$. Indeed, we will show this in Section 4.6.1 with a stochastic example that involves a single control per state and nonnegative but unbounded cost per stage (if the cost per stage is nonnegative and bounded, and the optimal cost over the proper policies only is equal to J^* , then J^* will be proved to be the unique solution of Bellman's equation within the set of all bounded J such that $J \geq 0$). This is a striking difference between deterministic and stochastic optimal control problems with infinite state space. Another striking difference is that J^* is always a solution of Bellman's equation in deterministic problems (cf. Exercise 3.1), but this is not so in stochastic problems, even when the state space is finite (cf. Section 3.1.2).

3.6 ALGORITHMS

We have already discussed some VI and PI algorithms for finding J^* and an optimal policy as part of our analysis under the weak and strong PI properties in Section 3.2. Moreover, we have shown that the VI algorithm converges to the optimal cost function J^* for any starting function $J \in S$ in the case of Assumption 3.3.1 (cf. Prop. 3.3.1), or to the restricted optimal cost function J_S^* under the assumptions of Prop. 3.4.1(b).

In this section, we will introduce additional algorithms. In Section 3.6.1, we will discuss asynchronous versions of VI and will prove satisfactory convergence properties under reasonable assumptions. In Section 3.6.2, we will focus on a modified version of PI that is unaffected by the presence of S -irregular policies. This algorithm is similar to the optimistic PI algorithm with uniform fixed point (cf. Section 2.6.3), and can also be implemented in a distributed asynchronous computing environment.

3.6.1 Asynchronous Value Iteration

Let us consider the model of Section 2.6.1 for asynchronous distributed computation of the fixed point of a mapping T , and the asynchronous distributed VI method described there. The model involves a partition of X into disjoint nonempty subsets X_1, \dots, X_m , and a corresponding partition of J as $J = (J_1, \dots, J_m)$, where J_ℓ is the restriction of J on the set X_ℓ .

We consider a network of m processors, each updating asynchronously corresponding components of J . In particular, we assume that J_ℓ is updated only by processor ℓ , and only for times t in a selected subset \mathcal{R}_ℓ of it-

erations. Moreover, as in Section 2.6.1, processor ℓ uses components J_j supplied by other processors $j \neq \ell$ with communication “delays” $t - \tau_{\ell j}(t) \geq 0$:

$$J_\ell^{t+1}(x) = \begin{cases} T\left(J_1^{\tau_{\ell 1}(t)}, \dots, J_m^{\tau_{\ell m}(t)}\right)(x) & \text{if } t \in \mathcal{R}_\ell, x \in X_\ell, \\ J_\ell^t(x) & \text{if } t \notin \mathcal{R}_\ell, x \in X_\ell. \end{cases} \quad (3.46)$$

We can prove convergence within the frameworks of Sections 3.3 and 3.4 by using the asynchronous convergence theorem (cf. Prop. 2.6.1), and the fact that T is monotone and has J^* as its unique fixed point within the appropriate set. We assume that the continuous updating and information renewal Assumption 2.6.1 holds. For simplicity we restrict attention to the framework of Section 3.3, under Assumption 3.3.1 with $S = \mathcal{B}(X)$. Assume further that we have two functions $\underline{V}, \overline{V} \in S$ such that

$$\underline{V} \leq T\underline{V} \leq T\overline{V} \leq \overline{V}, \quad (3.47)$$

so that, by Prop. 3.3.1, $T^k\underline{V} \leq J^* \leq T^k\overline{V}$ for all k , and

$$T^k\underline{V} \uparrow J^*, \quad T^k\overline{V} \downarrow J^*.$$

Then we can show asynchronous convergence of the VI algorithm (3.46), starting from any function J^0 with $\underline{V} \leq J^0 \leq \overline{V}$.

Indeed, let us apply Prop. 2.6.1 with the sets $S(k)$ given by

$$S(k) = \{J \in S \mid T^k\underline{V} \leq J \leq T^k\overline{V}\}, \quad k = 0, 1, \dots$$

The sets $S(k)$ satisfy $S(k+1) \subset S(k)$ in view of Eq. (3.47) and the monotonicity of T . Using Prop. 3.3.1, we also see that $S(k)$ satisfy the synchronous convergence and box conditions of Prop. 2.6.1. Thus, together with Assumption 2.6.1, all the conditions of Prop. 2.6.1 are satisfied, and the convergence of the algorithm follows starting from any $J^0 \in S(0)$.

3.6.2 Asynchronous Policy Iteration

In this section, we focus on PI methods, under Assumption 3.3.1 and some additional assumptions to be introduced shortly. We first discuss briefly a natural form of PI algorithm, which generates S -regular policies exclusively. Let μ^0 be an initial S -regular policy [there exists one by Assumption 3.3.1(b)]. At the typical iteration k , we have an S -regular policy μ^k , and we compute a policy μ^{k+1} such that $T_{\mu^{k+1}}J_{\mu^k} = TJ_{\mu^k}$ (this is possible by Lemma 3.3.1). Then μ^{k+1} is S -regular, by Lemma 3.3.2, and we have

$$J_{\mu^k} = T_{\mu^k}J_{\mu^k} \geq TJ_{\mu^k} = T_{\mu^{k+1}}J_{\mu^k} \geq \lim_{m \rightarrow \infty} T_{\mu^{k+1}}^m J_{\mu^k} = J_{\mu^{k+1}}.$$

We can thus construct a sequence of S -regular policies $\{\mu^k\}$ and a corresponding nonincreasing sequence $\{J_{\mu^k}\}$. Under some additional mild conditions it is then possible to show that $J_{\mu^k} \downarrow J^*$, cf. Prop. 3.3.1(e).

Unfortunately, when there are S -irregular policies, the preceding PI algorithm is somewhat limited, because an initial S -regular policy may not be known. Moreover, when asynchronous versions of the algorithm are implemented, it is difficult to guarantee that all the generated policies are S -regular.

In what follows in this section, we will discuss a PI algorithm that works in the presence of S -irregular policies, and can operate in a distributed asynchronous environment, like the PI algorithm for contractive models of Section 2.6.3. The main assumption is that J^* is the unique fixed point of T within $\mathcal{R}(X)$, the set of real-valued functions over X . This assumption holds under Assumption 3.3.1 with $S = \mathcal{R}(X)$, but it also holds under weaker conditions. Our assumptions also include finiteness of U , which among others facilitates the policy evaluation and policy improvement operations, and ensures that the algorithm generates iterates that lie in $\mathcal{R}(X)$. The algorithm and its analysis also go through if $\mathcal{R}(X)$ is replaced by $\mathcal{R}^+(X)$ (the set of all nonnegative real-valued functions) in the following assumptions, arguments, and propositions.

Assumption 3.6.1: In addition to the monotonicity Assumption 3.2.1, the following hold.

- (a) $H(x, u, J)$ is real-valued for all $J \in \mathcal{R}(X)$, $x \in X$, and $u \in U(x)$.
- (b) U is a finite set.
- (c) For each sequence $\{J_m\} \subset \mathcal{R}(X)$ with either $J_m \uparrow J$ or $J_m \downarrow J$ for some $J \in \mathcal{R}(X)$, we have

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

- (d) For all scalars $r > 0$ and functions $J \in \mathcal{R}(X)$, we have

$$H(x, u, J + r e) \leq H(x, u, J) + r e, \quad \forall x \in X, u \in U(x), \quad (3.48)$$

where e is the unit function.

- (e) J^* is the unique fixed point of T within $\mathcal{R}(X)$.

Part (d) of the preceding assumption is a nonexpansiveness condition for $H(x, u, \cdot)$, and can be easily verified in many DP models, including deterministic, minimax, and stochastic optimal control problems. It is not readily satisfied, however, in the affine monotonic model of Section 3.5.2.

Similar to Section 2.6.3, we introduce a new mapping that is parametrized by μ and can be shown to have a common fixed point for all μ . It operates on a pair (V, Q) where:

- V is a real-valued function with a component denoted $V(x)$ for each $x \in X$.
- Q is a real-valued function with a component denoted $Q(x, u)$ for each pair (x, u) with $x \in X, u \in U(x)$.

The mapping produces a pair

$$(MF_\mu(V, Q), F_\mu(V, Q)),$$

where

- $F_\mu(V, Q)$ is a function with a component $F_\mu(V, Q)(x, u)$ for each (x, u) , defined by

$$F_\mu(V, Q)(x, u) = H(x, u, \min\{V, Q_\mu\}), \quad (3.49)$$

where for any Q and μ , we denote by Q_μ the function of x defined by

$$Q_\mu(x) = Q(x, \mu(x)), \quad x \in X,$$

and for any two functions V_1 and V_2 , we denote by $\min\{V_1, V_2\}$ the function of x given by

$$\min\{V_1, V_2\}(x) = \min\{V_1(x), V_2(x)\}, \quad x \in X.$$

- $MF_\mu(V, Q)$ is a function with a component $(MF_\mu(V, Q))(x)$ for each x , where M is the operator of pointwise minimization over u :

$$(MQ)(x) = \min_{u \in U(x)} Q(x, u),$$

so that

$$(MF_\mu(V, Q))(x) = \min_{u \in U(x)} F_\mu(V, Q)(x, u).$$

Note that under Assumption 3.6.1, M maps real-valued functions to real-valued functions, since by part (b) of that assumption, U is assumed finite.

We consider an algorithm that is similar to the asynchronous PI algorithm given in Section 2.6.3 for contractive models. It applies asynchronously the mapping $MF_\mu(V, Q)$ for local policy improvement and update of V and μ , and the mapping $F_\mu(V, Q)$ for local policy evaluation and update of Q . The algorithm involves a partition of the state space into sets X_1, \dots, X_m , and assignment of each subset X_ℓ to a processor $\ell \in \{1, \dots, m\}$. For each ℓ , there are two infinite disjoint subsets of times $\mathcal{R}_\ell, \overline{\mathcal{R}}_\ell \subset \{0, 1, \dots\}$, corresponding to policy improvement and policy evaluation iterations, respectively. At time t , each processor ℓ operates on

$V^t(x)$, $Q^t(x, u)$, and $\mu^t(x)$, only for x in its “local” state space X_ℓ . In particular, at each time t , each processor ℓ does one of the following:

- (a) *Local policy improvement*: If $t \in \mathcal{R}_\ell$, processor ℓ sets for all $x \in X_\ell$,

$$V^{t+1}(x) = \min_{u \in U(x)} H(x, u, \min\{V^t, Q_{\mu^t}^t\}) = (MF_{\mu^t}(V^t, Q^t))(x), \quad (3.50)$$

sets $\mu^{t+1}(x)$ to a u that attains the minimum, and leaves Q unchanged, i.e., $Q^{t+1}(x, u) = Q^t(x, u)$ for all $x \in X_\ell$ and $u \in U(x)$.

- (b) *Local policy evaluation*: If $t \in \overline{\mathcal{R}}_\ell$, processor ℓ sets for all $x \in X_\ell$ and $u \in U(x)$,

$$Q^{t+1}(x, u) = H(x, u, \min\{V^t, Q_{\mu^t}^t\}) = F_{\mu^t}(V^t, Q^t)(x, u), \quad (3.51)$$

and leaves V and μ unchanged, i.e., $V^{t+1}(x) = V^t(x)$ and $\mu^{t+1}(x) = \mu^t(x)$ for all $x \in X_\ell$.

- (c) *No local change*: If $t \notin \mathcal{R}_\ell \cup \overline{\mathcal{R}}_\ell$, processor ℓ leaves Q , V , and μ unchanged, i.e., $Q^{t+1}(x, u) = Q^t(x, u)$ for all $x \in X_\ell$ and $u \in U(x)$, $V^{t+1}(x) = V^t(x)$, and $\mu^{t+1}(x) = \mu^t(x)$ for all $x \in X_\ell$.

Under Assumption 3.6.1, the algorithm generates real-valued functions if started with real-valued V^0 and Q^0 . We will prove that it converges to (J^*, Q^*) , where J^* is the unique fixed point of T within $\mathcal{R}(X)$ [cf. Assumption 3.6.1(e)], and Q^* is defined by

$$Q^*(x, u) = H(x, u, J^*), \quad x \in X, u \in U(x). \quad (3.52)$$

To this end, we introduce the mapping F defined by

$$(FQ)(x, u) = H(x, u, MQ), \quad x \in X, u \in U(x), \quad (3.53)$$

and we show the following proposition.

Proposition 3.6.1: Let Assumption 3.6.1 hold. Then Q^* is the unique fixed point of F within the class of real-valued functions.

Proof: By minimizing over $u \in U(x)$ in Eq. (3.52) and noting that J^* is a fixed point of T , we have $MQ^* = TJ^* = J^*$. Thus, by applying Eq. (3.53) and then Eq. (3.52), we obtain

$$(FQ^*)(x, u) = H(x, u, J^*) = Q^*(x, u), \quad \forall x \in X, u \in U(x).$$

Thus Q^* is a fixed point of F , and it is real-valued since J^* is real-valued and H is real-valued.

To show uniqueness, let Q' be any real-valued fixed point of F . Then $Q'(x, u) = H(x, u, MQ')$ for all $x \in X$, $u \in U(x)$, and by minimization over $u \in U(x)$, we have $MQ' = T(MQ')$. Hence MQ' is equal to the unique fixed point J^* of T , so that the equation $Q' = FQ'$ yields $Q'(x, u) = H(x, u, MQ') = H(x, u, J^*)$, for all (x, u) . From the definition (3.52) of Q^* , it then follows that $Q' = Q^*$. **Q.E.D.**

We introduce the μ -dependent mapping

$$L_\mu(V, Q) = (MQ, F_\mu(V, Q)), \quad (3.54)$$

where $F_\mu(V, Q)$ is given by Eq. (3.49). For this mapping and other related mappings to be defined shortly, we implicitly assume that it operates on real-valued functions, so by Assumption 3.6.1(a),(b), it produces real-valued functions. Note that the policy evaluation part of the algorithm [cf. Eq. (3.51)] amounts to applying the second component of L_μ , while the policy improvement part of the algorithm [cf. Eq. (3.50)] amounts to applying the second component of L_μ , *and* then applying the first component of L_μ . The following proposition shows that (J^*, Q^*) is the common fixed point of the mappings L_μ , for all μ .

Proposition 3.6.2: Let Assumption 3.6.1 hold. Then for all $\mu \in \mathcal{M}$, the mapping L_μ of Eq. (3.54) is monotone, and (J^*, Q^*) is its unique fixed point within the class of real-valued functions.

Proof: Monotonicity of L_μ follows from the monotonicity of the operators M and F_μ . To show that L_μ has (J^*, Q^*) as its unique fixed point, we first note that $J^* = MQ^*$ and $Q^* = FQ^*$; cf. Prop. 3.6.1. Then, using also the definition of F_μ , we have

$$J^* = MQ^*, \quad Q^* = FQ^* = F_\mu(J^*, Q^*),$$

which shows that (J^*, Q^*) is a fixed point of L_μ .

To show uniqueness, let (V', Q') be a real-valued fixed point of L_μ , i.e., $V' = MQ'$ and $Q' = F_\mu(V', Q')$. Then

$$Q' = F_\mu(V', Q') = FQ',$$

where the last equality follows from $V' = MQ'$. Thus Q' is a fixed point of F , and since Q^* is the unique fixed point of F (cf. Prop. 3.6.1), we have $Q' = Q^*$. It follows that $V' = MQ^* = J^*$, so (J^*, Q^*) is the unique fixed point of L_μ within the class of real-valued functions. **Q.E.D.**

The uniform fixed point property of L_μ just shown is, however, insufficient for the convergence proof of the asynchronous algorithm, in the

absence of a contraction property. For this reason, we introduce two mappings \underline{L} and \overline{L} that are associated with the mappings L_μ and satisfy

$$\underline{L}(V, Q) \leq L_\mu(V, Q) \leq \overline{L}(V, Q), \quad \forall \mu \in \mathcal{M}. \quad (3.55)$$

These are the mappings defined by

$$\underline{L}(V, Q) = \left(MQ, \min_{\mu \in \mathcal{M}} F_\mu(V, Q) \right), \quad \overline{L}(V, Q) = \left(MQ, \max_{\mu \in \mathcal{M}} F_\mu(V, Q) \right), \quad (3.56)$$

where the min and max over μ are attained in view of the finiteness of \mathcal{M} [cf. Assumption 3.6.1(b)]. We will show that \underline{L} and \overline{L} also have (J^*, Q^*) as their unique fixed point. Note that there exists $\bar{\mu}$ that attains the maximum in Eq. (3.56), uniformly for all V and (x, u) , namely a policy $\bar{\mu}$ for which

$$Q(x, \bar{\mu}(x)) = \max_{u \in U(x)} Q(x, u), \quad \forall x \in X,$$

[cf. Eq. (3.49)]. Similarly, there exists $\underline{\mu}$ that attains the minimum in Eq. (3.56), uniformly for all V and (x, u) . Thus for any given (V, Q) , we have

$$\underline{L}(V, Q) = L_{\underline{\mu}}(V, Q), \quad \overline{L}(V, Q) = L_{\bar{\mu}}(V, Q), \quad (3.57)$$

where $\underline{\mu}$ and $\bar{\mu}$ are some policies. The following proposition shows that (J^*, Q^*) , the common fixed point of the mappings L_μ , for all μ , is also the unique fixed point of \underline{L} and \overline{L} .

Proposition 3.6.3: Let Assumption 3.6.1 hold. Then the mappings \underline{L} and \overline{L} of Eq. (3.56) are monotone, and have (J^*, Q^*) as their unique fixed point within the class of real-valued functions.

Proof: Monotonicity of \underline{L} and \overline{L} follows from the monotonicity of the operators M and F_μ . Since (J^*, Q^*) is the common fixed point of L_μ for all μ (cf. Prop. 3.6.2), and there exists $\underline{\mu}$ such that $\underline{L}(J^*, Q^*) = L_{\underline{\mu}}(J^*, Q^*)$ [cf. Eq. (3.57)], it follows that (J^*, Q^*) is a fixed point of \underline{L} . To show uniqueness, suppose that (V, Q) is a fixed point, so $(V, Q) = \underline{L}(V, Q)$. Then by Eq. (3.57), we have

$$(V, Q) = \underline{L}(V, Q) = L_{\underline{\mu}}(V, Q)$$

for some $\underline{\mu} \in \mathcal{M}$. Since by Prop. 3.6.2, (J^*, Q^*) is the only fixed point of $L_{\underline{\mu}}$, it follows that $(V, Q) = (J^*, Q^*)$, so (J^*, Q^*) is the only fixed point of \underline{L} . Similarly, we show that (J^*, Q^*) is the unique fixed point of \overline{L} . **Q.E.D.**

We are now ready to construct a sequence of sets needed to apply Prop. 2.6.1 and prove convergence. For a scalar $c \geq 0$, we denote

$$\begin{aligned} J_c^- &= J^* - ce, & Q_c^- &= Q^* - ce_Q, \\ J_c^+ &= J^* + ce, & Q_c^+ &= Q^* + ce_Q, \end{aligned}$$

with e and e_Q are the unit functions in the spaces of J and Q , respectively.

Proposition 3.6.4: Let Assumption 3.6.1 hold. Then for all $c > 0$,

$$\underline{L}^k(J_c^-, Q_c^-) \uparrow (J^*, Q^*), \quad \overline{L}^k(J_c^+, Q_c^+) \downarrow (J^*, Q^*), \quad (3.58)$$

where \underline{L}^k (or \overline{L}^k) denotes the k -fold composition of \underline{L} (or \overline{L} , respectively).

Proof: For any $\mu \in \mathcal{M}$, using the assumption (3.48), we have for all (x, u) ,

$$\begin{aligned} F_\mu(J_c^+, Q_c^+)(x, u) &= H(x, u, \min\{J_c^+, Q_c^+\}) \\ &= H(x, u, \min\{J^*, Q^*\} + ce) \\ &\leq H(x, u, \min\{J^*, Q^*\}) + c \\ &= Q^*(x, u) + c \\ &= Q_c^+(x, u), \end{aligned}$$

and similarly

$$Q_c^-(x, u) \leq F_\mu(J_c^-, Q_c^-)(x, u).$$

We also have $MQ_c^+ = J_c^+$ and $MQ_c^- = J_c^-$. From these relations, the definition of L_μ , and the fact $L_\mu(J^*, Q^*) = (J^*, Q^*)$ (cf. Prop. 3.6.2), we have

$$(J_c^-, Q_c^-) \leq L_\mu(J_c^-, Q_c^-) \leq (J^*, Q^*) \leq L_\mu(J_c^+, Q_c^+) \leq (J_c^+, Q_c^+).$$

Using this relation and Eqs. (3.55) and (3.57), we obtain

$$(J_c^-, Q_c^-) \leq \underline{L}(J_c^-, Q_c^-) \leq (J^*, Q^*) \leq \overline{L}(J_c^+, Q_c^+) \leq (J_c^+, Q_c^+). \quad (3.59)$$

Denote for $k = 0, 1, \dots$,

$$(\overline{V}_k, \overline{Q}_k) = \overline{L}^k(J_c^+, Q_c^+), \quad (\underline{V}_k, \underline{Q}_k) = \underline{L}^k(J_c^-, Q_c^-).$$

From the monotonicity of \overline{L} and \underline{L} and Eq. (3.59), we have that $(\overline{V}_k, \overline{Q}_k)$ converges monotonically from above to some pair

$$(\overline{V}, \overline{Q}) \geq (J^*, Q^*),$$

while $(\underline{V}_k, \underline{Q}_k)$ converges monotonically from below to some pair

$$(\underline{V}, \underline{Q}) \leq (J^*, Q^*).$$

By taking the limit in the equation

$$(\overline{V}_{k+1}, \overline{Q}_{k+1}) = \overline{L}(\overline{V}_k, \overline{Q}_k),$$

and using the continuity from above and below property of \overline{L} , implied by Assumption 3.6.1(c), it follows that $(\overline{V}, \overline{Q}) = \overline{L}(\overline{V}, \overline{Q})$, so $(\overline{V}, \overline{Q})$ must be equal to (J^*, Q^*) , the unique fixed point of \overline{L} . Thus, $\overline{L}^k(J_c^+, Q_c^+) \downarrow (J^*, Q^*)$. Similarly, $\underline{L}^k(J_c^-, Q_c^-) \uparrow (J^*, Q^*)$. **Q.E.D.**

To show asynchronous convergence of the algorithm (3.50)-(3.51), consider the sets

$$S(k) = \{(V, Q) \mid \underline{L}^k(J_c^-, Q_c^-) \leq (V, Q) \leq \overline{L}^k(J_c^+, Q_c^+)\}, \quad k = 0, 1, \dots,$$

whose intersection is (J^*, Q^*) [cf. Eq. (3.58)]. By Prop. 3.6.4 and Eq. (3.55), this set sequence together with the mappings L_μ satisfy the synchronous convergence and box conditions of the asynchronous convergence theorem of Prop. 2.6.1 (more precisely, its time-varying version of Exercise 2.2). This proves the convergence of the algorithm (3.50)-(3.51) for starting points $(V, Q) \in S(0)$. Since c can be chosen arbitrarily large, it follows that the algorithm is convergent from an arbitrary starting point.

Finally, let us note some variations of the asynchronous PI algorithm. One such variation is to allow “communication delays” $t - \tau_{\ell_j}(t)$. Another variation, for the case where we want to calculate just J^* , is to use a reduced space implementation similar to the one discussed in Section 2.6.3. There is also a variant with interpolation, cf. Section 2.6.3.

3.7 NOTES, SOURCES, AND EXERCISES

The semicontractive model framework of this chapter was first given in the 2013 edition of the book, and it was subsequently expanded in a series of papers and reports by the author: [Ber14], [Ber15], [Ber16a], [BeY16], [Ber17c], [Ber17d]. The framework is inspired from the analysis of the SSP problem of Example 1.2.6, which involves finite state and control spaces, as well as a termination state. In the absence of a termination state, a key idea has been to generalize the notion of a proper policy from one that leads to termination with probability 1, to one that is S -regular for an appropriate set of functions S .

Section 3.1: The counterexample showing that J^* may fail to solve Bellman’s equation in SSP problems is due to Bertsekas and Yu [BeY16]. The

blackmailer's dilemma is a folklore example in the DP literature. The book by Whittle [Whi82] has a substantial discussion. The set of solutions of the Riccati equation in continuous-time linear-quadratic optimal control problems (cf. Section 3.1.4) has been described in the paper by Willems [Wil71], which stimulated considerable further work on the subject (see the book by Lancaster and Rodman [LaR95] for an extensive account). The pathologies of infinite horizon linear-quadratic optimal control problems can be largely eliminated under some well-studied controllability and observability conditions (see, e.g., [Ber17a], Section 3.1).

Section 3.2: The PI-based analysis of Section 3.2 was developed in the author's paper [Ber15] after the 2013 edition of the book was published. The author's joint work with H. Yu [BeY16] was also influential. In particular, the SSP example of Section 3.1.2 where J^* does not satisfy Bellman's equation, and the perturbation analysis of Section 3.4 come from the paper [BeY16]. The same is true for the rate of convergence result of Prop. 3.2.2. The λ -PI method was introduced by Bertsekas and Ioffe [Bei96] in the context of discounted and SSP problems, and subsequent work was referenced in Section 2.7. The analysis of λ -PI in Section 3.2.4 is new and is related to an analysis of a linearized form of the proximal algorithm given in the author's papers [Ber16b], [Ber17e].

Section 3.3: The central result of Section 3.3, Prop. 3.3.1, was given in the 2013 edition of the book. It is patterned after a result of Bertsekas and Tsitsiklis [BeT91] for SSP problems with finite state space and compact control constraint sets, which is reproduced in Section 3.5.1. The proof given there contains an intricate part used to demonstrate a real-valued lower bound on the cost functions of proper policies (Lemma 3 of [BeT91], which implies Prop. 3.5.3).

Section 3.4: The perturbation approach of Section 3.4 was introduced in the 2013 edition of the book. It is given here in somewhat stronger form, which will also be applied to nonstationary S -regular policies in the next chapter.

Section 3.5: The SSP problem analysis of Section 3.5.1 for the case of the strong SSP conditions is due to the paper by Bertsekas and Tsitsiklis [BeT91]. For the case of the weak SSP conditions it is due to the paper by Bertsekas and Yu [BeY16]. The perturbation-based PI algorithm was given in Section 3.3.3 of the 2013 edition of the book. A different PI algorithm that embodies a mechanism for breaking ties in the policy improvement step was given by Guillot and Stauffer [GuS17] for the case of finite state and control spaces.

The affine monotonic model of Section 3.5.2 was first formulated and analyzed in the 2013 edition of the book, in a more general setting where the state space can be an infinite set. The analysis of Section 3.5.2 of the finite-state case comes from the author's paper [Ber16a], which contains

more details. The exponentiated cost version of the SSP problem was analyzed in the papers by Denardo and Rothblum [DeR79], and by Patek [Pat01]. The paper [DeR79] assumes that the state and control spaces are finite, that there exists at least one contractive policy (a transient policy in the terminology of [DeR79]), and that every improper policy is noncontractive and has infinite cost from some initial state. These assumptions bypass the pathologies around infinite control spaces and multiple solutions or no solution of Bellman's equation. Also the approach of [DeR79] is based on linear programming (relying on the finite control space), and is thus quite different from ours. The paper [Pat01] assumes that the state space is finite, that the control constraint set is compact, and that the expected one-stage cost is strictly positive for all state-control pairs, which is much stronger than what we have assumed. Our results of Section 3.5.2, when specialized to the exponential cost problem, are consistent with and subsume the results of [DeR79] and Patek [Pat01].

The robust shortest path planning discussion of Section 3.5.3 follows the author's paper [Ber14]. This paper contains further analysis and algorithms, including a Dijkstra-like finitely terminating algorithms for problems with nonnegative arc lengths.

The deterministic optimal control model of Section 3.5.5 is discussed in more detail in the author's paper [Ber17b] under Assumption 3.5.9 for the case where $g \geq 0$; see also Section 4.5 and the paper [Ber17c]. The analysis under the more general assumptions given here is new. Deterministic and minimax infinite-spaces optimal control problems have also been discussed by Reissig [Rei16] under different assumptions than ours.

Section 3.6: The asynchronous VI algorithm of Section 3.6.1 was first given in the author's paper on distributed DP [Ber82]. It was further formalized in the paper [Ber83], where the solution of a DP problem was viewed as a special case of a fixed point problem.

The asynchronous PI algorithm and analysis of Section 3.6.2, parallels the corresponding algorithm of Section 2.6.3, and is due to joint work of the author with H. Yu, presented in the papers [BeY12] and [YuB13a]. In particular, the algorithm of Section 3.6.2 is one of the optimistic PI algorithms in [YuB13a], which was applied to the SSP problem of Section 3.5.1 under the strong SSP conditions. We have followed the line of analysis of that paper and the related paper [BeY12], which focuses on discounted problems. These papers also analyzed asynchronous stochastic iterative versions of PI, and proved convergence results that parallel those for classical Q-learning for SSP, given in Tsitsiklis [Tsi94], and Yu and Bertsekas [YuB13b]. An earlier paper, which deals with a slightly different asynchronous abstract PI algorithm without a contraction structure, is Bertsekas and Yu [BeY10].

By allowing an infinite state space, the analysis of the present chapter applies among others to SSP problems with a countable state space. Such problems often arise in queueing control problems where the termination

state corresponds to an empty queue. The problem then is to empty the queue with minimum expected cost. Generalized forms of SSP problems, which involve an infinite (uncountable) number of states, in addition to the termination state, were analyzed by Pliska [Pli78], Hernandez-Lerma et al. [HCP99], and James and Collins [JaC06]. The latter paper allows improper policies, assumes that g is bounded and J^* is bounded below, and generalizes the results of [BeT91] to infinite (Borel) state spaces, using a similar line of proof. Infinite spaces SSP problems will also be discussed in Section 4.6.

An important case of an SSP problem where the state space is infinite arises under imperfect state information. There the problem is converted to a perfect state information problem whose states are the belief states, i.e., the posterior probability distributions of the original state given the observations thus far. Patek [Pat07] addresses SSP problems with imperfect state information and proves results that are similar to the ones for their perfect state information counterparts. These results can also be derived using the line of analysis of this chapter. In particular, the critical condition that the cost functions of proper policies are bounded below by some real-valued function [cf. Assumption 3.3.1(b)] is proved as Lemma 5 in [Pat07], using the fact that the cost functions of the proper policies are bounded below by the optimal cost function of a corresponding perfect state information problem.

E X E R C I S E S

3.1 (Conditions for J^* to be a Fixed Point of T)

The purpose of this exercise is to show that the optimal cost function J^* is a fixed point of T under some assumptions, which among others, are satisfied generically in deterministic optimal control problems. Let $\hat{\Pi}$ be a subset of policies such that:

- (1) We have

$$(\mu, \pi) \in \hat{\Pi} \quad \text{if and only if} \quad \mu \in \mathcal{M}, \pi \in \hat{\Pi},$$

where for $\mu \in \mathcal{M}$ and $\pi = \{\mu_0, \mu_1, \dots\}$, we denote by (μ, π) the policy $\{\mu, \mu_0, \mu_1, \dots\}$. *Note:* This condition precludes the possibility that $\hat{\Pi}$ is the set of all stationary policies (unless there is only one stationary policy).

- (2) For every $\pi = \{\mu_0, \mu_1, \dots\} \in \hat{\Pi}$, we have

$$J_\pi = T_{\mu_0} J_{\pi_1},$$

where π_1 is the policy $\pi_1 = \{\mu_1, \mu_2, \dots\}$.

(3) We have

$$\inf_{\mu \in \mathcal{M}, \pi \in \hat{\Pi}} T_{\mu} J_{\pi} = \inf_{\mu \in \mathcal{M}} T_{\mu} \hat{J},$$

where the function \hat{J} is given by

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

Show that:

- (a) \hat{J} is a fixed point of T . In particular, if $\hat{\Pi} = \Pi$, then J^* is a fixed point of T .
- (b) The assumptions (1)-(3) hold with $\hat{\Pi} = \Pi$ in the case of the deterministic mapping

$$H(x, u, J) = g(x, u) + J(f(x, u)), \quad x \in X, u \in u(x), J \in \mathcal{E}(X). \quad (3.60)$$

- (c) Consider the SSP example of Section 3.1.2, where J^* is not a fixed point of T . Which of the conditions (1)-(3) is violated?

Solution: (a) For every $x \in X$, we have

$$\hat{J}(x) = \inf_{\pi \in \hat{\Pi}} J_{\pi}(x) = \inf_{\mu \in \mathcal{M}, \pi \in \hat{\Pi}} (T_{\mu} J_{\pi})(x) = \inf_{\mu \in \mathcal{M}} (T_{\mu} \hat{J})(x) = (T \hat{J})(x),$$

where the second equality holds by conditions (1) and (2), and the third equality holds by condition (3).

(b) This is evident in the case of the deterministic mapping (3.60). *Notes:* (i) If $\hat{\Pi} = \Pi$, parts (a) and (b) show that J^* , which is equal to \hat{J} , is a fixed point of T . Moreover, if we choose a set S such that J_S^* can be shown to be equal to J^* , then Prop. 3.2.1 applies and shows that J^* is the unique fixed point of T with the set $\{J \in \mathcal{E}(X) \mid J_S^* \leq J \leq \bar{J}\}$ for some $\bar{J} \in S$. In addition the VI sequence $\{T^k J\}$ converges to J^* starting from every J within that set. (ii) The assumptions (1)-(3) of this exercise also hold for other choices of $\hat{\Pi}$. For example, when $\hat{\Pi}$ is the set of all *eventually stationary* policies, i.e., policies of the form $\{\mu_0, \dots, \mu_k, \mu, \mu, \dots\}$, where $\mu_0, \dots, \mu_k, \mu \in \mathcal{M}$ and k is some positive integer.

(c) For the SSP problem of Section 3.1.1, condition (2) of the preceding proposition need not be satisfied (because the expected value operation need not commute with $\lim \sup$).

3.2 (Alternative Semicontractive Conditions I)

This exercise provides a different starting point for the semicontractive analysis of Section 3.2. In particular, the results of Prop. 3.2.1 are shown without assuming that J_S^* is a fixed point of T , but by making different assumptions, which include the existence of an S -regular policy that is optimal. Let S be a given subset of $\mathcal{E}(X)$. Assume that:

- (1) There exists an S -regular policy μ^* that is optimal, i.e., $J_{\mu^*} = J^*$.
- (2) The policy μ^* satisfies $T_{\mu^*}J^* = TJ^*$.

Show that the following hold:

- (a) The optimal cost function J^* is the unique fixed point of T within the set $\{J \in S \mid J \geq J^*\}$.
- (b) We have $T^k J \rightarrow J^*$ for every $J \in S$ with $J \geq J^*$.
- (c) An S -regular policy μ that satisfies $T_{\mu}J^* = TJ^*$ is optimal. Conversely if μ is an S -regular optimal policy, it satisfies $T_{\mu}J^* = TJ^*$.

Note: Part (a) and the assumptions show that J_S^* is a fixed point of T (as well as that $J_S^* = J^* \in S$), so parts (b) and (c) also follow from Prop. 3.2.1.

Solution: (a) We first show that any fixed point J of T that lies in S satisfies $J \leq J^*$. Indeed, if $J = TJ$, then for the optimal S -regular policy μ^* , we have $J \leq T_{\mu^*}J$, so in view of the monotonicity of T_{μ^*} and the S -regularity of μ^* ,

$$J \leq \lim_{k \rightarrow \infty} T_{\mu^*}^k J = J_{\mu^*} = J^*.$$

Thus the only function within $\{J \in S \mid J \geq J^*\}$ that can be a fixed point of T is J^* . Using the optimality and S -regularity of μ^* , and condition (2), we have

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

so J^* is a fixed point of T . Finally, $J^* \in S$ since $J^* = J_{\mu^*}$ and μ^* is S -regular, so J^* is the unique fixed point of T within $\{J \in S \mid J \geq J^*\}$.

(b) For the optimal S -regular policy μ^* and any $J \in S$ with $J \geq J^*$, we have

$$T_{\mu^*}^k J \geq T^k J \geq T^k J^* = J^*, \quad k = 0, 1, \dots$$

Taking the limit as $k \rightarrow \infty$, and using the fact $\lim_{k \rightarrow \infty} T_{\mu^*}^k J = J_{\mu^*} = J^*$, which holds since μ^* is S -regular and optimal, we see that $T^k J \rightarrow J^*$.

(c) If μ satisfies $T_{\mu}J^* = TJ^*$, then using part (a), we have $T_{\mu}J^* = J^*$ and hence $\lim_{k \rightarrow \infty} T_{\mu}^k J^* = J^*$. If μ is in addition S -regular, then $J_{\mu} = \lim_{k \rightarrow \infty} T_{\mu}^k J^* = J^*$ and μ is optimal. Conversely, if μ is optimal and S -regular, then $J_{\mu} = J^*$ and $J_{\mu} = T_{\mu}J_{\mu}$, which combined with $J^* = TJ^*$ [cf. part (a)], yields $T_{\mu}J^* = TJ^*$.

3.3 (Alternative Semicontractive Conditions II)

Let S be a given subset of $\mathcal{E}(X)$. Show that the assumptions of Exercise 3.2 hold if and only if $J^* \in S$, $TJ^* \leq J^*$, and there exists an S -regular policy μ such that $T_{\mu}J^* = TJ^*$.

Solution: Let the conditions (1) and (2) of Exercise 3.2 hold, and let μ^* be the S -regular policy that is optimal. Then condition (1) implies that $J^* = J_{\mu^*} \in S$ and $J^* = T_{\mu^*}J^* \geq TJ^*$, while condition (2) implies that there exists an S -regular policy μ such that $T_{\mu}J^* = TJ^*$.

Conversely, assume that $J^* \in S$, $TJ^* \leq J^*$, and there exists an S -regular policy μ such that $T_\mu J^* = TJ^*$. Then we have $T_\mu J^* = TJ^* \leq J^*$. Hence $T_\mu^k J^* \leq J^*$ for all k , and by taking the limit as $k \rightarrow \infty$, we obtain $J_\mu \leq J^*$. Hence the S -regular policy μ is optimal, and the conditions of Exercise 3.2 hold.

3.4 (Alternative Semicontractive Conditions III)

Let S be a given subset of $\mathcal{E}(X)$. Assume that:

- (1) There exists an optimal S -regular policy.
- (2) For every S -irregular policy $\bar{\mu}$, we have $T_{\bar{\mu}} J^* \geq J^*$.

Show that the assumptions of Exercise 3.2 hold.

Solution: It will be sufficient to show that conditions (1) and (2) imply that $J^* = TJ^*$. Assume to obtain a contradiction, that $J^* \neq TJ^*$. Then $J^* \geq TJ^*$, as can be seen from the relations

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

where μ^* is an optimal S -regular policy. Thus the relation $J^* \neq TJ^*$ implies that there exists μ' and $x \in X$ such that

$$J^*(x) \geq (T_{\mu'} J^*)(x), \quad \forall x \in X,$$

with strict inequality for some x [note here that we can choose $\bar{\mu}(x) = \mu^*(x)$ for all x such that $J^*(x) = (TJ^*)(x)$, and we can choose $\bar{\mu}(x)$ to satisfy $J^*(x) > (T_{\bar{\mu}} J^*)(x)$ for all other x]. If $\bar{\mu}$ were S -regular, we would have

$$J^* \geq T_{\bar{\mu}} J^* \geq \lim_{k \rightarrow \infty} T_{\bar{\mu}}^k J^* = J_{\mu'},$$

with strict inequality for some $x \in X$, which is impossible. Hence μ' is S -irregular, which contradicts condition (2).

3.5 (Restricted Optimization over a Subset of S -Regular Policies)

This exercise provides a useful extension of Prop. 3.2.1. Given a set S , it may be more convenient to work with a subset $\widehat{\mathcal{M}} \subset \mathcal{M}_S$. Let \hat{J} denote the corresponding restricted optimal value:

$$\hat{J}(x) = \inf_{\mu \in \widehat{\mathcal{M}}} J_\mu(x),$$

and assume that \hat{J} is a fixed point of T . Show that the following analogs of the conclusions of Prop. 3.2.1 hold:

- (a) (*Uniqueness of Fixed Point*) If J' is a fixed point of T and there exists $\tilde{J} \in S$ such that $J' \leq \tilde{J}$, then $J' \leq \hat{J}$. In particular, if the set $\widehat{\mathcal{W}}$ given by

$$\widehat{\mathcal{W}} = \{J \in \mathcal{E}(X) \mid \hat{J} \leq J \leq \tilde{J} \text{ for some } \tilde{J} \in S\},$$

is nonempty, then \hat{J} is the unique fixed point of T within $\widehat{\mathcal{W}}$.

(b) (*VI Convergence*) We have $T^k J \rightarrow \hat{J}$ for every $J \in \widehat{\mathcal{W}}$.

Solution: The proof is nearly identical to the one of Prop. 3.2.1. Let $J \in \widehat{\mathcal{W}}$, so that

$$\hat{J} \leq J \leq \tilde{J}$$

for some $\tilde{J} \in S$. We have for all $k \geq 1$ and $\mu \in \widehat{\mathcal{M}}$,

$$\hat{J} = T^k \hat{J} \leq T^k J \leq T^k \tilde{J} \leq T_\mu^k \tilde{J},$$

where the equality follows from the fixed point property of \hat{J} , while the inequalities follow by using the monotonicity and the definition of T . The right-hand side tends to J_μ as $k \rightarrow \infty$, since μ is S -regular and $\tilde{J} \in S$. Hence the infimum over $\mu \in \widehat{\mathcal{M}}$ of the limit of the right-hand side tends to the left-hand side \hat{J} . It follows that $T^k J \rightarrow \hat{J}$, proving part (b). To prove part (a), let J' be a fixed point of T that belongs to $\widehat{\mathcal{W}}$. Then J' is equal to $\lim_{k \rightarrow \infty} T^k J'$, which has been proved to be equal to \hat{J} .

3.6 (The Case $J_S^* \leq \bar{J}$)

Within the framework of Section 3.2, assume that $J_S^* \leq \bar{J}$. (This occurs in particular in the monotone decreasing model where $\bar{J} \geq T_\mu \bar{J}$ for all $\mu \in \mathcal{M}$; see Section 4.3.) Show that if J_S^* is a fixed point of T , then we have $J_S^* = J^*$. *Note:* This result manifests itself in the shortest path Example 3.2.1 for the case where $b < 0$.

Solution: For all k and policies $\pi = \{\mu_0, \mu_1, \dots\}$, we have

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

and by taking the infimum over $\pi \in \Pi$, we obtain $J_S^* \leq J^*$. Since generically we have $J_S^* \geq J^*$, it follows that $J_S^* = J^*$.

3.7 (Weakening the Near-Optimal Termination Assumption)

Consider the deterministic optimal control problem of Section 3.5.5. The purpose of this exercise is to show that the Assumption 3.5.9 is equivalent to a seemingly weaker assumption where nonstationary policies can be used for termination. Given a state $x \in X^*$, we say that a (possibly nonstationary) policy $\pi \in \Pi$ *terminates from* x if the sequence $\{x_k\}$, which is generated starting from x and using π , reaches X_0 in the sense that $x_{\bar{k}} \in X_0$ for some index \bar{k} . Assume that for every $x \in X^*$, there exists a policy $\pi \in \Pi$ that terminates from x . Show that:

(a) The set $\widehat{\mathcal{M}}$ of terminating stationary policies is nonempty, i.e., there exists a stationary policy that terminates from every $x \in X^*$.

- (b) Assumption 3.5.9 is satisfied if for every pair (x, ϵ) with $x \in X^*$ and $\epsilon > 0$, there exists a policy $\pi \in \Pi$ that terminates from x and satisfies $J_\pi(x) \leq J^*(x) + \epsilon$.

Solution: (a) Consider the sequence of subsets of X defined for $k = 0, 1, \dots$, by

$$X_k = \{x \in X^* \mid \text{there exists } \pi \in \Pi \text{ that terminates from } x \text{ in } k \text{ steps or less}\},$$

starting with the stopping set X_0 . Note that $\cup_{k=0}^\infty X_k = X^*$. Define a stationary policy $\bar{\mu}$ as follows: For each $x \in X_k$ with $x \notin X_{k-1}$, let $\{\mu_0, \mu_1, \dots\}$ be a policy that terminates from x in the minimum possible number of steps (which is k), and let $\bar{\mu} = \mu_0$. For each $x \notin X^*$, let $\bar{\mu}(x)$ be an arbitrary control in $U(x)$. It can be seen that $\bar{\mu}$ is a terminating stationary policy.

(b) Given any state $\bar{x} \in X^*$ with $\bar{x} \notin X_0$, and a nonstationary policy $\pi = \{\mu_0, \mu_1, \dots\}$ that terminates from \bar{x} , we construct a stationary policy μ that terminates from every $x \in X^*$ and generates essentially the same trajectory as π starting from \bar{x} (i.e., after cycles are subtracted). To construct such a μ , we consider the sequence generated by π starting from \bar{x} . If this sequence contains cycles, we shorten the sequence by eliminating the cycles, and we redefine π so that starting from \bar{x} it generates a terminating trajectory without cycles. This redefined version of π , denoted $\pi' = \{\mu'_0, \mu'_1, \dots\}$, terminates from \bar{x} and has cost $J_{\pi'}(\bar{x}) \leq J_\pi(\bar{x})$ [since all the eliminated transitions that belonged to cycles have nonnegative cost, in view of the fact $J^*(x) > -\infty$ for all x , which is implied by Assumption 3.5.9]. We now consider the sequence of subsets of X defined by

$$X_k = \{x \in X \mid \pi' \text{ terminates from } x \text{ in } k \text{ steps or less}\}, \quad k = 0, 1, \dots,$$

where X_0 is the stopping set. Let \bar{k} be the first $k \geq 1$ such that $\bar{x} \in X_k$. Construct the stationary policy μ as follows: for $x \in \cup_{k=1}^{\bar{k}} X_k$, let

$$\mu(x) = \mu'_{\bar{k}-k}(x), \quad \text{if } x \in X_k \text{ and } x \notin X_{k-1}, \quad k = 1, 2, \dots,$$

and for $x \notin \cup_{k=1}^{\bar{k}} X_k$, let $\mu(x) = \bar{\mu}(x)$, where $\bar{\mu}$ is a stationary policy that terminates from every $x \in X^*$ [and was shown to exist in part (a)]. Then it is seen that μ terminates from every $x \in X^*$, and generates the same sequence as π' starting from the state \bar{x} , so it satisfies $J_\mu(\bar{x}) = J_{\pi'}(\bar{x}) \leq J_\pi(\bar{x})$.

3.8 (Verifying the Near-Optimal Termination Assumption)

In the context of the deterministic optimal control problem of Section 3.5.5, assume that X is a normed space with norm denoted $\|\cdot\|$. We say that π *asymptotically terminates from* x if the sequence $\{x_k\}$ generated starting from x and using π converges to X_0 in the sense that

$$\lim_{k \rightarrow \infty} \text{dist}(x_k, X_0) = 0,$$

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

$$\text{dist}(x, X_0) = \inf_{y \in X_0} \|x - y\|, \quad x \in X.$$

The purpose of this exercise is to provide a readily verifiable condition that guarantees Assumption 3.5.9. Assume that

$$0 \leq g(x, u), \quad x \in X, u \in U(x),$$

and that

$$J^*(x) > 0, \quad \forall x \notin X_0.$$

Assume further the following:

- (1) For every $x \in X^* = \{x \in X \mid J^*(x) < \infty\}$ and $\epsilon > 0$, there exists a policy π that asymptotically terminates from x and satisfies $J_\pi(x) \leq J^*(x) + \epsilon$.
- (2) For every $\epsilon > 0$, there exists a $\delta_\epsilon > 0$ such that for each $x \in X^*$ with

$$\text{dist}(x, X_0) \leq \delta_\epsilon,$$

there is a policy π that terminates from x and satisfies $J_\pi(x) \leq \epsilon$.

Then:

- (a) Show that Assumption 3.5.9 holds.
- (b) Show that condition (1) holds if for each $\delta > 0$ there exists $\epsilon > 0$ such that

$$\inf_{u \in U(x)} g(x, u) \geq \epsilon, \quad \forall x \in X \text{ such that } \text{dist}(x, X_0) \geq \delta.$$

Note: For further discussion, analysis, and application to the case of a linear system, see the author's paper [Ber17b].

Solution: (a) Fix $x \in X^*$ and $\epsilon > 0$. Let π be a policy that asymptotically terminates from x , and satisfies $J_\pi(x) \leq J^*(x) + \epsilon$, as per condition (1). Starting from x , this policy will generate a sequence $\{x_k\}$ such that for some index \bar{k} we have $\text{dist}(x_{\bar{k}}, X_0) \leq \delta_\epsilon$, so by condition (2), there exists a policy $\bar{\pi}$ that terminates from $x_{\bar{k}}$ and is such that $J_{\bar{\pi}}(x_{\bar{k}}) \leq \epsilon$. Consider the policy π' that follows π up to index \bar{k} and follows $\bar{\pi}$ afterwards. This policy terminates from x and satisfies

$$J_{\pi'}(x) = J_{\pi, \bar{k}}(x) + J_{\bar{\pi}}(x_{\bar{k}}) \leq J_\pi(x) + J_{\bar{\pi}}(x_{\bar{k}}) \leq J^*(x) + 2\epsilon,$$

where $J_{\pi, \bar{k}}(x)$ is the cost incurred by π starting from x up to reaching $x_{\bar{k}}$. From Exercise 3.7 it follows that Assumption 3.5.9 holds.

(b) For any x and policy π that does not asymptotically terminate from x , we will have $J_\pi(x) = \infty$, so that if $x \in X^*$, all policies π with $J_\pi(x) < \infty$ must be asymptotically terminating from x .

3.9 (Perturbations and S -Regular Policies)

The purpose of this exercise is to illustrate that the set of S -regular policies may be different in the perturbed and unperturbed problems of Section 3.4. Consider a single-state problem with $\bar{J} = 0$ and two policies μ and μ' , where

$$T_\mu J = \min\{1, J\}, \quad T_{\mu'} J = \beta > 0.$$

Let $S = \mathfrak{R}$.

- (a) Verify that μ is S -irregular and $J_\mu = J^* = 0$.
- (b) Verify that μ' is S -regular and $J_{\mu'} = J_S^* = \beta$.
- (c) For $\delta > 0$ consider the δ -perturbed problem with $p(x) = 1$, where x is the only state. Show that both μ and μ' are S -regular for this problem. Moreover, we have $\hat{J}_\delta = \min\{1, \beta\} + \delta$.
- (d) Verify that Prop. 3.4.1 applies for $\widehat{\mathcal{M}} = \{\mu'\}$ and $\beta \leq 1$, but does not apply if $\widehat{\mathcal{M}} = \{\mu, \mu'\}$ or $\beta > 1$. Which assumptions of the proposition are violated in the latter case?

Solution: Parts (a) and (b) are straightforward. It is also straightforward to verify the definition of S -regularity for both policies in the δ -perturbed problem, and that $J_{\mu, \delta} = 1 + \delta$ and $J_{\mu', \delta} = \beta + \delta$. If $\beta \leq 1$, the policy μ' is optimal for the δ -perturbed problem, and Prop. 3.4.1 applies for $\widehat{\mathcal{M}} = \{\mu'\}$ because all its assumptions are satisfied. However, when $\beta > 1$ and $\widehat{\mathcal{M}} = \{\mu'\}$ there is no ϵ -optimal policy in $\widehat{\mathcal{M}}$ for the δ -perturbed problem (contrary to the assumption of Prop. 3.4.1), and indeed we have $\beta = J_S^* > \lim_{\delta \downarrow 0} \hat{J}_\delta = 1$. Also when $\widehat{\mathcal{M}} = \{\mu, \mu'\}$, the policy μ is not S -regular, contrary to the assumption of Prop. 3.4.1.

3.10 (Perturbations in Affine Monotonic Models [Ber16a])

Consider the affine monotonic model of Section 3.5.2, and let Assumptions 3.5.5 and 3.5.6 hold. In a perturbed version of this model we add a constant $\delta > 0$ to all components of b_μ , thus obtaining what we call the δ -perturbed affine monotonic problem. We denote by \hat{J}_δ and $J_{\mu, \delta}$ the corresponding optimal cost function and policy cost functions, respectively.

- (a) Show that for all $\delta > 0$, \hat{J}_δ is the unique solution within \mathfrak{R}_+^n of the equation

$$J(i) = (TJ)(i) + \delta, \quad i = 1, \dots, n.$$

- (b) Show that for all $\delta > 0$, a policy μ is optimal for the δ -perturbed problem (i.e., $J_{\mu, \delta} = \hat{J}_\delta$) if and only if $T_\mu \hat{J}_\delta = T \hat{J}_\delta$. Moreover, for the δ -perturbed problem, all optimal policies are contractive and there exists at least one contractive policy that is optimal.
- (c) The optimal cost function over contractive policies \hat{J} [cf. Eq. (3.37)] satisfies

$$\hat{J}(i) = \lim_{\delta \downarrow 0} \hat{J}_\delta(i), \quad i = 1, \dots, n.$$

- (d) If the control constraint set $U(i)$ is finite for all states $i = 1, \dots, n$, there exists a contractive policy $\hat{\mu}$ that attains the minimum over all contractive policies, i.e., $J_{\hat{\mu}} = \hat{J}$.
- (e) Show Prop. 3.5.8.

Solution: (a), (b) By Prop. 3.5.6, we have that Assumption 3.3.1 holds for the δ -perturbed problem. The results follow by applying Prop. 3.5.7 [the equation of part (a) is Bellman's equation for the δ -perturbed problem].

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

$$\hat{J} = \inf_{\mu: \text{contractive}} J_\mu \leq J_{\mu_\delta^*} \leq J_{\mu_\delta^*, \delta} = \hat{J}_\delta \leq J_{\mu', \delta}, \quad \forall \mu' : \text{contractive}.$$

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

$$\hat{J} \leq \lim_{\delta \downarrow 0} \hat{J}_\delta \leq J_{\mu'}, \quad \forall \mu' : \text{contractive}.$$

By taking the infimum over all μ' that are contractive, the result follows.

(d) Let $\{\delta_k\}$ be a positive sequence with $\delta_k \downarrow 0$, and consider a corresponding sequence $\{\mu_k\}$ of optimal contractive policies for the δ_k -perturbed problems. Since the set of contractive policies is finite [in view of the finiteness of $U(i)$], some policy $\hat{\mu}$ will be repeated infinitely often within the sequence $\{\mu_k\}$, and since $\{J_{\delta_k}^*\}$ is monotonically nonincreasing, we will have

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

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

(e) For all contractive μ , we have $J_\mu = T_\mu J_\mu \geq T_\mu \hat{J} \geq T\hat{J}$. Taking the infimum over contractive μ , we obtain $\hat{J} \geq T\hat{J}$. Conversely, for all $\delta > 0$ and $\mu \in \mathcal{M}$, we have

$$\hat{J}_\delta = T\hat{J}_\delta + \delta e \leq T_\mu \hat{J}_\delta + \delta e.$$

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

For all $J \in \mathfrak{R}^n$ with $J \geq \hat{J}$ and contractive μ , we have by using the relation $\hat{J} = T\hat{J}$ just shown,

$$\hat{J} = \lim_{k \rightarrow \infty} T^k \hat{J} \leq \lim_{k \rightarrow \infty} T^k J \leq \lim_{k \rightarrow \infty} T_\mu^k J = J_\mu.$$

Taking the infimum over all contractive μ , we obtain

$$\hat{J} \leq \lim_{k \rightarrow \infty} T^k J \leq \hat{J}, \quad \forall J \geq \hat{J}.$$

This proves that $T^k J \rightarrow \hat{J}$. Finally, let $J' \in \mathfrak{R}(X)$ be another solution of Bellman's equation, and let $J \in \mathfrak{R}(X)$ be such that $J \geq \hat{J}$ and $J \geq J'$. Then $T^k J \rightarrow \hat{J}$, while $T^k J \geq T^k J' = J'$. It follows that $\hat{J} \geq J'$.

To prove Prop. 3.5.8(c) note that if μ is a contractive policy with $J_\mu = \hat{J}$, we have $\hat{J} = J_\mu = T_\mu J_\mu = T_\mu \hat{J}$, so, using also the relation $\hat{J} = T\hat{J}$ [cf. part (a)], we obtain $T_\mu \hat{J} = T\hat{J}$. Conversely, if μ satisfies $T_\mu \hat{J} = T\hat{J}$, then from part (a), we have $T_\mu \hat{J} = \hat{J}$ and hence $\lim_{k \rightarrow \infty} T_\mu^k \hat{J} = \hat{J}$. Since μ is contractive, we obtain $J_\mu = \lim_{k \rightarrow \infty} T_\mu^k \hat{J}$, so $J_\mu = \hat{J}$.

The proof of Prop. 3.5.8(d) is nearly identical to the one of Prop. 3.5.4(d).