3

Semicontractive Models

Contents

We will now consider abstract DP models that are intermediate between the contractive models of Chapter 2, where all stationary policies involve a contraction mapping, and noncontractive models to be discussed in Chapter 4, where there are no contraction-like assumptions. A representative example of such an intermediate model is the stochastic shortest path problem of Example 1.2.6 (SSP for short). In one of the main versions of an SSP theory, there are two types of policies: those that are *proper* where the mapping T_μ is a contraction with respect to a weighted sup-norm, and those that are *improper*, where T_μ is not a contraction with respect to any norm. As noted in Example 1.2.6, results that are comparable to the ones for discounted finite-state MDP have been obtained with appropriate assumptions that guarantee among others that improper policies are "bad" in the sense that they have infinite cost from some initial state.

In this chapter we introduce models where, as in SSP problems, policies are divided 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 contraction of T_μ with a notion of *regularity of μ within an appropriate set S* (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 regular in this sense while others are not, and impose further conditions ensuring that there exist optimal policies that are regular. Under a variety of assumptions, we show results that resemble those available for SSP problems: 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.

We 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. In particular, in this chapter and the next one, we consider several alternative assumptions for different semicontractive models.

The chapter is organized as follows. In Section 3.1, we introduce the notion of a regular policy, and we develop some of the basic results. In Section 3.2, we discuss related lines of analysis to address some major special cases that bear similarity to SSP problems. In Section 3.3, we focus on VI and PI-type algorithms.

3.1 SEMICONTRACTIVE MODELS AND REGULAR POLICIES

Our basic model for this chapter and the next one is similar to the one of Chapter 2, but the assumptions are different. We will maintain the monotonicity assumption, but we will weaken the contraction assumption, and we will introduce some other conditions in its place.

In our analysis of this chapter, the optimal cost function J^* will typically be real-valued. However, the cost function J_μ of some policies μ may take infinite values for some states. To accommodate this, we will use the set of extended real numbers $\Re^* = \Re \cup \{\infty, -\infty\}$, and the set of all extended real-valued functions $J : X \mapsto \Re^*$, which we denote by $E(X)$. We denote by $R(X)$ the set of real-valued functions $J : X \mapsto \Re$, and by $B(X)$ the set of real-valued functions $J : X \mapsto \Re$ that are bounded with respect to a given weighted sup-norm. Throughout this chapter and the next two, when we write lim, lim sup, or lim inf of a sequence of functions we mean it to be pointwise. We also write $J_k \to J$ to mean that $J_k(x) \to J(x)$ for each $x \in X$; see our notational conventions in Appendix A.

As in Chapters 1 and 2, we introduce the set X of states and the set U of controls, and for each $x \in X$, the nonempty control constraint set $U(x) \subset U$. We denote by $\mathcal{M}$ the set of all functions $\mu : X \mapsto U$ with $\mu(x) \in U(x)$, for all $x \in X$, and by Π the set of nonstationary policies $\pi = \{\mu_0, \mu_1, \ldots\}$, with $\mu_k \in \mathcal{M}$ for all k. We refer to a stationary policy $\{\mu, \mu, \ldots\}$ simply as μ. We introduce a mapping $H : X \times U \times E(X) \mapsto \Re^*$, satisfying the following condition.

Assumption 3.1.1: (Monotonicity) If $J, J' \in E(X)$ and $J \leq J'$, then

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

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

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

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

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

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

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

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

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 Section 2.1, following Prop. 2.1.2). We introduce a function $\bar{J} \in E(X)$, 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.

Definition 3.1.1: Given a function $\bar{J} \in E(X)$, for a policy $\pi \in \Pi$ with $\pi = \{\mu_0, \mu_1, \ldots\}$, we define the cost function of π by

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

In the case of a stationary policy $\mu \in \mathcal{M}$, the cost function of μ is denoted by J_μ and is given by

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

The optimal cost function J^* is given by

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

An optimal policy $\pi^* \in \Pi$ is one for which $J_{\pi^*} = J^*$. Note two important differences from Chapter 2:

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

(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 turn out to be "sufficient" in the sense that the optimal cost obtained with nonstationary policies is matched by the one obtained by stationary ones.

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. It would then appear that 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 another 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.

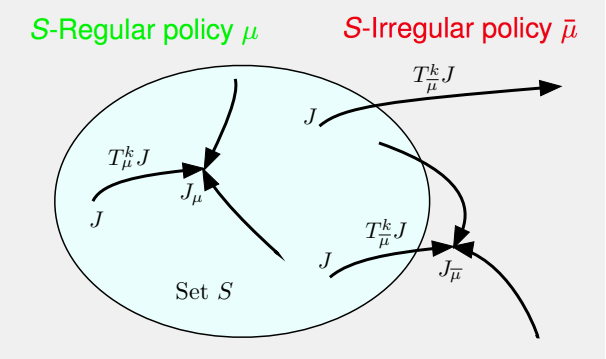

Figure 3.1.1. Illustration of S-regular and S-irregular policies. Policy μ is S-regular because $J_\mu \in S$ and $T_\mu^k J \to J_\mu$ for all $J \in S$. Policy $\overline{\mu}$ is S-irregular.

Definition 3.1.2: Given a set of functions $S \subset E(X)$, we say that a stationary policy μ is *S-regular* if:

(a) $J_\mu \in S$ and $J_\mu = T_\mu J_\mu$.

(b) $T_\mu^k J \to J_\mu$ for all $J \in S$.

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

Thus a policy μ is S-regular if the VI iteration corresponding to μ, $J_{k+1} = T_\mu J_k$, represents a dynamic system that has J_μ as its unique 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. 3.1.1).

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 $R(X)$, $B(X)$, and subsets of $R(X)$, $B(X)$, and $E(X)$ 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

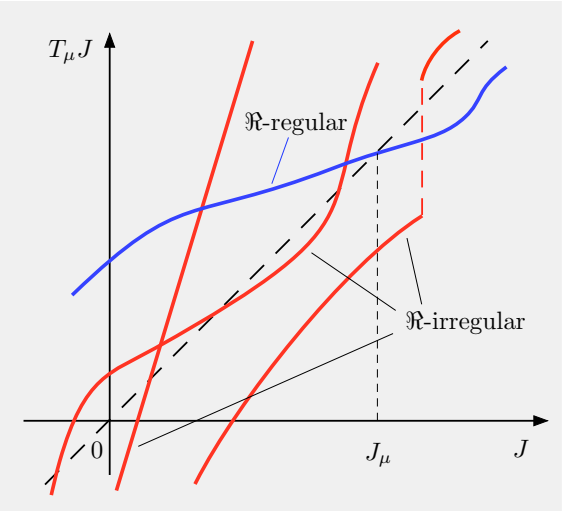

Figure 3.1.2. Illustration of S-regular and S-irregular policies for the case where there is only one state and $S = \Re$. There are three mappings T_μ corresponding to S-irregular policies: one crosses the 45-degree line at multiple points, another crosses at a single point but at an angle greater than 45 degrees, and the third is discontinuous and does not cross at all. The mapping T_μ of the $\Re$-regular policy has J_μ as its unique fixed point and satisfies $T_\mu^k J \to J_\mu$ for all $J \in \Re$.

postpone making the choice of S more specific. Figure 3.1.2 illustrates the mappings T_μ of some S-regular and S-irregular policies for the case where there is a single state and $S = \Re$. Figure 3.1.3 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.

3.1.1 Fixed Points, Optimality Conditions, and Algorithmic Results

We will now introduce an analytical framework where S-regular policies are central. Our focus is reflected by our first assumption in the following proposition, which is that optimal policies can be found among the S-regular policies, i.e., that for some S-regular μ^* we have $J^* = J_{\mu^*}$. This assumption implies that

$$J^* = J_{\mu^*} = T_{\mu^*} J_{\mu^*} \geq T J_{\mu^*} = T J^*,$$

where the second equality follows from the S-regularity of μ^*. Thus the Bellman equation $J^* = TJ^*$ follows if μ^* attains the infimum in the relation $TJ^* = \inf_{\mu \in \mathcal{M}} T_\mu J^*$, which is our second assumption. In addition to existence of solution of the Bellman equation, the regularity of μ^* implies a uniqueness assertion and a convergence result for the VI algorithm, as shown in the following proposition.

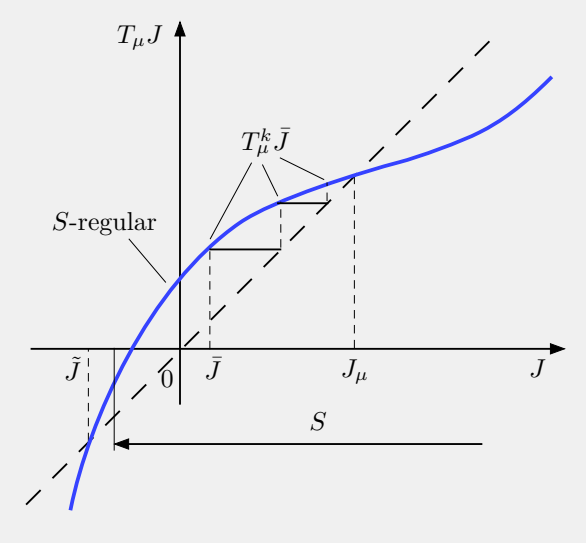

Figure 3.1.3. Illustration of a mapping T_μ where there is only one state and S is a subset of the real line. Here T_μ has two fixed points, J_μ and $\tilde{J}$. If S is as shown, μ is S-regular. If S is enlarged to include $\tilde{J}$, μ becomes S-irregular.

Proposition 3.1.1: Let S be a given subset of $E(X)$. Assume that:

(1) There exists an S-regular policy μ^* that is optimal, i.e., $J_{\mu^*} = J^*$.

(2) The policy μ^* satisfies $T_{\mu^*} J^* = TJ^*$.

Then the following hold:

(a) The optimal cost function J^* is the unique fixed point of T within the set $\{J \in S \mid J \geq J^*\}$.

(b) We have $T^k J \to J^*$ for every $J \in S$ with $J \geq J^*$.

(c) An S-regular policy μ that satisfies $T_\mu J^* = TJ^*$ is optimal. Conversely if μ is an S-regular optimal policy, it satisfies $T_\mu J^* = TJ^*$.

Proof: (a) The proof uses a more refined version of the argument preceding the statement of the proposition. We first show that any fixed point J of T that lies in S satisfies $J \leq J^*$. Indeed, if $J = TJ$, then for the optimal S-regular policy μ^*, we have $J \leq T_{\mu^*} J$, so in view of the monotonicity of T_{μ^*} and the S-regularity of μ^*,

$$J \leq \lim_{k \to \infty} T_{\mu^*}^k J = J_{\mu^*} = J^*.$$

Thus the only function within $\{J \in S \mid J \geq J^*\}$ that can be a fixed point

of T is J^*. Using the optimality and S-regularity of μ^*, and condition (2), we have

$$J^* = J_{\mu^*} = T_{\mu^*} J_{\mu^*} = T_{\mu^*} J^* = T J^*,$$

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

$$T^k_{\mu^*} J \geq T^k J \geq T^k J^* = J^*, \qquad k = 0, 1, \ldots.$$

Taking the limit as $k \to \infty$, and using the fact

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

which holds since μ^* is S-regular and optimal, we see that $T^k J \to J^*$.

(c) If μ satisfies $T_\mu J^* = T J^*$, then using part (a), we have $T_\mu J^* = J^*$ and hence $\lim_{k\to\infty} T^k_\mu J^* = J^*$. If μ is in addition S-regular, then $J_\mu = \lim_{k\to\infty} T^k_\mu J^* = J^*$ 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 $J^* = T J^*$ [cf. part (a)], yields $T_\mu J^* = T J^*$. **Q.E.D.**

Note that given condition (1) of the proposition, condition (2) is equivalent to the seemingly weaker assumption that *some* S-regular μ satisfies $T_\mu J^* = T J^*$. To see this note that if this latter condition holds together with condition (1), we have

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

where the first two equalities follow from the S-regularity and optimality of μ^*, the second inequality follows from the monotonicity of T_μ, and the last equality follows from the S-regularity of μ. Since μ^* is optimal, it follows that μ is also optimal, so equality holds in the above relation, and we have $T_{\mu^*} J^* = T J^*$, implying condition (2) as stated in the proposition.

Let us also show an equivalent variation of the preceding proposition, for problems where the validity of Bellman's equation $J^* = T J^*$ can be independently verified. We will later encounter models where this can be done (e.g., the perturbation model of Section 3.2.2, and the monotone increasing and monotone decreasing models of Section 4.3).

Proposition 3.1.2: Let S be a given subset of $E(X)$. Assume that:

(1) There exists an S-regular policy μ^* that is optimal, i.e., $J_{\mu^*} = J^*$.

(2) We have $J^* = TJ^*$.

Then the assumptions and the conclusions of Prop. 3.1.1 hold.

Proof: We have

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

so, using also the assumption $J^* = TJ^*$, we obtain $T_{\mu^*} J^* = TJ^*$. Hence condition (2) of Prop. 3.1.1 holds. **Q.E.D.**

The following proposition is a special case of Prop. 3.1.1. It applies when functions in S are real-valued, and through its condition (1), it requires that S-irregular policies have a certain "infinite cost-type" property. Conditions of this type will appear prominently in Section 3.2 [see Assumptions 3.2.1(c) and 3.2.3(b)].

Proposition 3.1.3: Let S be a given subset of $R(X)$. Assume that:

(1) There exists an optimal S-regular policy, and for every S-irregular policy μ, there is at least one state $x \in X$ such that

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

(2) There exists a policy μ such that $T_\mu J^* = TJ^*$.

Then the assumptions and the conclusions of Prop. 3.1.1 hold.

Proof: In view of the remark following the proof of Prop. 3.1.1, it will suffice to show that the policy μ of condition (2) is S-regular. Let μ satisfy $T_\mu J^* = TJ^*$, and let μ^* be an optimal S-regular policy. Then for all $k \geq 1$,

$$J_{\mu^*} = T_{\mu^*} J_{\mu^*} \geq TJ_{\mu^*} = T_\mu J_{\mu^*} \geq T_\mu^k J_{\mu^*},$$

where the first equality follows from the definition of an S-regular policy, and the second inequality follows from the monotonicity of T_μ. If μ is S-irregular, by taking the limit as $k \to \infty$ in the preceding relation, the right-hand side tends to ∞ for some $x \in X$, while the left-hand side is finite since $J_{\mu^*} \in S \subset R(X)$ - a contradiction. Thus μ is S-regular. **Q.E.D.**

The examples of Fig. 3.1.2 show how Bellman's equation may fail in the absence of existence of an optimal S-regular policy [cf. condition (1) of Props. 3.1.1-3.1.3]. Consider for instance a problem where there is only one policy μ that is S-irregular and T_μ has no fixed point.

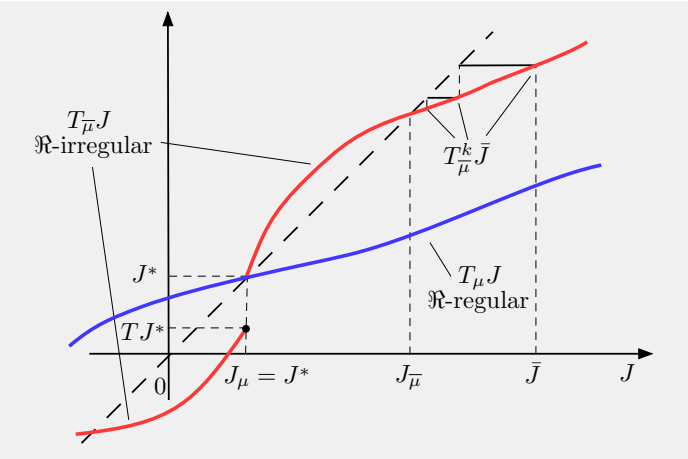

Figure 3.1.4. Illustration of why condition (2) is essential in Prop. 3.1.1. Here there is only one state and $S = \Re$. There are two stationary policies: μ for which T_μ is a contraction, so μ is $\Re$-regular, and $\overline{\mu}$ for which $T_{\overline{\mu}}$ has multiple fixed points, so $\overline{\mu}$ is $\Re$-irregular. Moreover, $T_{\overline{\mu}}$ is discontinuous from above at J_μ as shown. Here, it can be verified that $T_{\mu_0} \cdots T_{\mu_k} \bar{J} \geq J_\mu$ for all $\mu_0, \ldots, \mu_k$ and k, so that $J_\pi \geq J_\mu$ for all π and the S-regular policy μ is optimal. However, μ does not satisfy $T_\mu J^* = TJ^*$ [cf. condition (2) of Prop. 3.1.1] and we have $J^* \neq TJ^*$. Here the conclusions (a) and (c) of Prop. 3.1.1 are violated.

Another example that illustrates the need for existence of an optimal S-regular policy is the classical blackmailer problem, described in Exercise 3.1. This is a one-state problem, where T_μ is a contraction for all μ, so all policies are $\Re$-regular, but we have $J^* = -\infty$, so there is no optimal stationary policy. Here Bellman's equation, $J = TJ$, has no solution within $\Re$ (although we do have $J^* = TJ^*$). Moreover, it can be shown that there exists a nonstationary optimal policy for this problem; see [Ber12a] (Example 3.2.1).

Figure 3.1.4 shows what may happen if condition (2) of Prop. 3.1.1 is violated. The figure shows how we can then have $J^* \neq TJ^*$, and underscores the importance of existence of an S-regular μ satisfying the optimality equation $T_\mu J^* = TJ^*$ for a strong connection of our framework with fixed point theory. This condition will be assumed either directly or indirectly, via other conditions, throughout our analysis of semicontractive models.

The two conditions of Props. 3.1.1-3.1.3 may not be easily verified in a given problem. However, they can often be guaranteed through other reasonable conditions, in which case Props. 3.1.1-3.1.3 can be brought to bear on the analysis. We will encounter several such instances in Sections 3.2, 4.4, and 4.5.

Regardless of their assumptions, some of the conclusions (a)-(c) of Props. 3.1.1-3.1.3 are not as strong as one would like. In particular, part (a)

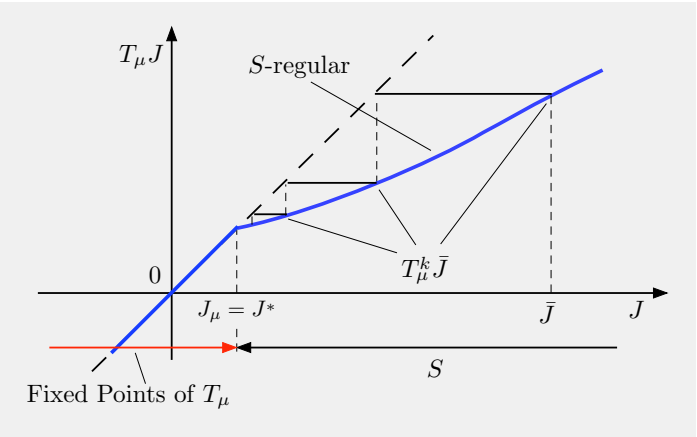

Figure 3.1.5. Illustration of multiplicity of fixed points J satisfying $J \leq J^*$, under the assumptions of Props. 3.1.1-3.1.2. Here there is only one state, so S is a subset of the real line, and there is only one policy μ (so $J_\mu = J^*$), which is S-regular for $S = \{J \mid J \geq J_\mu\}$.

asserts uniqueness of the fixed point of T only within the set $\{J \in S \mid J \geq J^*\}$. One may hope that there would be a unique fixed point, at least within S. Similarly, part (b) asserts the convergence of $T^k J$ to J^* only for J in the set $\{J \in S \mid J \geq J^*\}$, and one would like to assert convergence starting anywhere within S. These results cannot be improved in the absence of additional conditions, as can be illustrated with simple examples involving a single policy; see Fig. 3.1.5. The deterministic shortest path problem with zero length cycles provides an interesting practical context where there may exist additional fixed points J of T that do not satisfy $J \geq J^*$ (see the next subsection). However, with additional conditions, one may be able to show uniqueness of the fixed point of T within S, and demonstrate an enlarged region of initial conditions J from which $T^k J$ converges to J^* (see for example Sections 3.2.1 and 4.4.1).

We finally note a subtle point in part (c) of Props. 3.1.1-3.1.2, which leaves open the possibility that the optimality condition $T_\mu J^* = TJ^*$ is satisfied by a nonoptimal S-irregular μ. Indeed this can happen as can be shown with simple examples; see Fig. 3.1.6. †

† In the important case where $\bar{J} \leq J^*$, we can show that the condition $T_\mu J^* = TJ^*$ implies that μ is optimal, regardless of whether it is S-regular or not. The reason is that in this case we have

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

and taking the lim sup as $k \to \infty$, we obtain $J_\mu \leq J^*$, so μ is optimal. Note that the condition $\bar{J} \leq J^*$ holds for the monotone increasing models of Section 4.3.

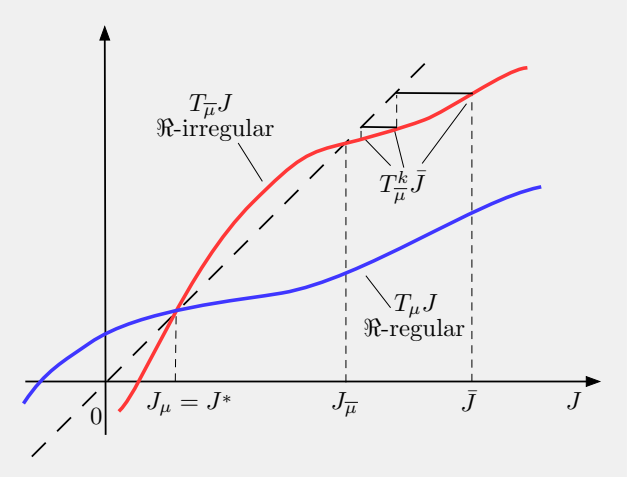

Figure 3.1.6. Illustration of a nonoptimal S-irregular policy $\overline{\mu}$ that satisfies the optimality condition $T_{\overline{\mu}}J^* = TJ^*$. Here there is only one state and $S = \Re$. There are two policies: μ for which T_μ is a contraction, so μ is $\Re$-regular, and $\overline{\mu}$ for which $T_{\overline{\mu}}$ has two fixed points, so $\overline{\mu}$ is $\Re$-irregular. For $\bar{J}$ as shown in the figure, $\overline{\mu}$ is nonoptimal, yet it satisfies $T_{\overline{\mu}}J^* = TJ^*$.

Policy Iteration

We established in Chapter 2 the significance of PI and its variations as a major class of computational methods for abstract DP. However, for semicontractive models, the convergence properties of PI are complicated, and under the conditions of Props. 3.1.1-3.1.2, the sequence of generated policies may not converge to an optimal policy.

What can be proved is that if μ and $\overline{\mu}$ are S-regular policies that satisfy the policy improvement equation $T_{\overline{\mu}}J_\mu = TJ_\mu$, then $J_{\overline{\mu}} \leq J_\mu$. To see this note that by using the S-regularity of μ, we have

$$J_\mu = T_\mu J_\mu \geq TJ_\mu = T_{\bar{\mu}}J_\mu \geq T^k_{\bar{\mu}}J_\mu, \qquad k \geq 1,$$

where the last inequality holds by the monotonicity of $T_{\bar{\mu}}$. By taking the limit as $k \to \infty$ in the preceding relation and using the S-regularity of $\overline{\mu}$, we obtain $J_\mu \geq J_{\bar{\mu}}$.

The policy improvement relation $J_\mu \geq J_{\bar{\mu}}$ shows that the PI algorithm, when restricted to S-regular policies, generates a nonincreasing sequence $\{J_{\mu^k}\}$, but this does not guarantee that $J_{\mu^k} \downarrow J^*$. Moreover, guaranteeing that the policies μ^k are S-regular may not be easy since the equation $T_{\overline{\mu}}J_\mu = TJ_\mu$ may be satisfied by an S-irregular $\overline{\mu}$, in which case there is no guarantee that $J_{\overline{\mu}} \leq J_\mu$, and an oscillation between policies may occur. This can be seen from the example of Fig. 3.1.7, where there

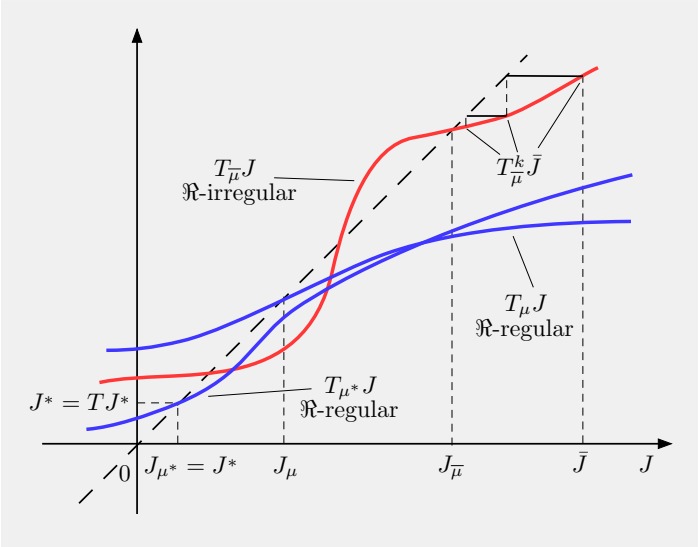

Figure 3.1.7. Oscillation of PI between two nonoptimal policies: an S-regular policy μ and an S-irregular policy $\overline{\mu}$ satisfying

$$T_{\overline{\mu}}J_\mu = TJ_\mu, \qquad T_\mu J_{\overline{\mu}} = TJ_{\overline{\mu}}.$$

Here there is only one state and $S = \Re$. In addition to μ and $\overline{\mu}$, there is a third policy μ^*, which is S-regular and optimal. In this example all the assumptions and conclusions of Props. 3.1.1-3.1.2 are satisfied.

is oscillation between two nonoptimal policies: an $\Re$-regular policy μ and an $\Re$-irregular policy $\overline{\mu}$ satisfying

$$T_{\overline{\mu}}J_\mu = TJ_\mu, \qquad T_\mu J_{\overline{\mu}} = TJ_{\overline{\mu}}.$$

A similar example of PI oscillation will be given for deterministic shortest path problems with zero length cycles in the next subsection. Thus additional assumptions or modifications of the PI algorithm are needed to improve its reliability. We will address this issue in Section 3.3.2, as well as in Sections 4.4 and 4.5 of the next chapter.

3.1.2 Illustrative Example: Deterministic Shortest Path Problems

In this section, we will highlight some of the analytical issues raised in the preceding subsection through the classical deterministic shortest path problem described in Example 1.2.7. We have a graph of n nodes $x = 1, \ldots, n$, plus the destination 0, and an arc length a_{xy} for each directed arc (x, y). Here $X = \{1, \ldots, n\}$. A policy chooses at state/node $x \in X$

an outgoing arc from x. Thus the controls available at x can be identified with the outgoing neighbors of x [the nodes u such that (x, u) is an arc]. The corresponding mapping H is given by

$$H(x, u, J) = \begin{cases} a_{xu} + J(u) & \text{if } u \neq 0, \\ a_{x0} & \text{if } u = 0, \end{cases}$$

and $\bar{J} = 0$.

We will consider S-regularity with $S = \Re^n$. A policy μ defines a graph whose arcs are $\big(x, \mu(x)\big)$, $x = 1, \ldots, n$. If this graph contains a cycle with m arcs, $x_1 \to x_2 \to \cdots \to x_m \to x_1$, with length $L = a_{x_1x_2} + \cdots + a_{x_{m-1}x_m} + a_{x_mx_1}$, then it can be seen that for all $k \geq 1$, we have

$$(T_\mu^{km} J)(x_1) = kL + J(x_1).$$

Thus such a policy cannot be $\Re^n$-regular (and if $L \neq 0$, its cost function J_μ has some infinite entries, so it is outside $\Re^n$). By contrast if a policy defines a graph that is acyclic, it can be verified to be $\Re^n$-regular.

Let us assume now that all cycles have positive cost ($L > 0$ above), and that every node is connected to the destination with some path (this is a common assumption in deterministic shortest path problems, which will be revisited and generalized considerably in Section 3.2.1). Then every $\Re^n$-irregular policy has infinite cost starting from some node/state, and it can be shown that there exists an optimal $\Re^n$-regular policy. Thus Prop. 3.1.3 applies, and guarantees that J^* is the unique fixed point of T within the set $\{J \mid J \geq J^*\}$, and that the VI algorithm converges to J^* starting only from within that set. Actually the uniqueness of the fixed point and the convergence of VI can be shown within the entire space $\Re^n$. This is well-known in shortest path theory, and will be covered by results to be given in Section 3.2.1.

In the other extreme case where there is a cycle of negative cost, there are $\Re^n$-irregular policies that are optimal and no $\Re^n$-regular policy can be optimal. Thus Props. 3.1.1-3.1.3 do not apply in this case.

The case where there is a cycle with zero cost exhibits the most complex behavior and will be illustrated for the example of Fig. 3.1.8. Here

$$X = \{1, 2\}, \qquad U(1) = \{0, 2\}, \qquad U(2) = \{1\}, \qquad \bar{J}(1) = \bar{J}(2) = 0.$$

There are two policies:

μ : where $\mu(1) = 0$, corresponding to the path $2 \to 1 \to 0$,

$\overline{\mu}$: where $\overline{\mu}(1) = 2$, corresponding to the cycle $1 \to 2 \to 1$,

and the corresponding mapping H is

$$H(x, u, J) = \begin{cases} b & \text{if } x = 1,\ u = 0, \\ a + J(2) & \text{if } x = 1,\ u = 2, \\ a + J(1) & \text{if } x = 2,\ u = 1. \end{cases} \tag{3.1}$$

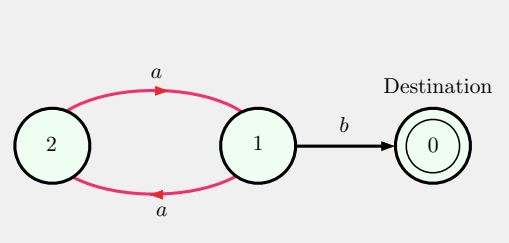

Figure 3.1.8. A deterministic shortest path problem with nodes 1, 2, and destination 0. Arc lengths are shown next to the arcs.

The Bellman equation is given by

$$J(1) = \min\big\{b, a + J(2)\big\}, \qquad J(2) = a + J(1), \tag{3.2}$$

while the VI algorithm takes the form

$$J_{k+1}(1) = \min\big\{b, a + J_k(2)\big\}, \qquad J_{k+1}(2) = a + J_k(1). \tag{3.3}$$

Here the policy μ is $\Re^2$-regular [the VI for μ is $J_{k+1}(1) = b$, $J_{k+1}(2) = a + J_k(1)$], while the policy $\overline{\mu}$ is $\Re^2$-irregular [the VI for $\overline{\mu}$ is $J_{k+1}(1) = a + J_k(2)$, $J_{k+1}(2) = a + J_k(1)$].

In the case where the cycle has zero length, so $a = 0$, there are two possibilities, $b \leq 0$ and $b > 0$, which we will consider separately:

(a) $a = 0$, $b \leq 0$: Here the $\Re^2$-regular policy is optimal, and Prop. 3.1.1 applies. Bellman's equation (3.2) has the unique solution

$$J^*(1) = b, \qquad J^*(2) = b,$$

within the set $\Re^2$, and the VI algorithm (3.3) converges to J^* from any starting $J_0 \geq J^*$. However, it can be verified that Bellman's equation has multiple solutions: the set of solutions is

$$\big\{J \mid J(1) = J(2),\ J \leq J^*\big\},$$

and VI starting from one of these solutions, will keep generating that solution. Moreover we can verify that PI may oscillate between the optimal $\Re^2$-regular policy and the $\Re^2$-irregular policy (which is nonoptimal if $b < 0$). Indeed, the $\Re^2$-irregular policy $\overline{\mu}$ is evaluated as $J_{\overline{\mu}}(1) = J_{\overline{\mu}}(2) = 0$, while the $\Re^2$-regular policy μ is evaluated as $J_\mu(1) = J_\mu(2) = b$, so in the policy improvement phase of the algorithm, we have

$$\mu(1) \in \arg\min\big\{b,\ J_{\overline{\mu}}(2)\big\}, \qquad \overline{\mu}(1) \in \arg\min\big\{b,\ J_\mu(2)\big\}.$$

Thus policy improvement starting with $\overline{\mu}$ yields μ, and starting with μ may yield $\overline{\mu}$, with the oscillatory sequence $\{\mu, \overline{\mu}, \mu, \overline{\mu}, \ldots\}$ resulting. Note that here we have $T_{\overline{\mu}} J^* = TJ^*$, so the optimality condition of Prop. 3.1.1(b) is attained by $\overline{\mu}$, which is nonoptimal when $b < 0$.

(b) $a = 0$, $b > 0$: Here the unique optimal policy is the $\Re^2$-irregular $\overline{\mu}$, and Props. 3.1.1-3.1.3 do not apply. The policy generates a sequence of cycles $1 \to 2 \to 1$, rather than a path that leads to the destination. In this case, the abstract DP model based on the mapping H of Eq. (3.1) cannot be used to model the shortest path problem (its optimal solution does not yield paths from nodes 1 and 2 to the destination). However, we may still be interested in finding an optimal policy within the class of $\Re^2$-regular policies. It turns out that we may address this problem by using a perturbation approach, which is described in Section 3.2.2. In particular, we will add a small amount $\delta > 0$ to the length of each arc. This has a strong effect on the problem: the cost function of the $\Re^2$-irregular policy becomes infinite, while the cost function of the $\Re^2$-regular policy changes by an $O(\delta)$ amount. Thus with $\delta > 0$, we obtain the acyclic $\Re^2$-regular policy μ.

3.2 IRREGULAR POLICIES AND A PERTURBATION APPROACH

In this section we will use the model and the results of the preceding section as a starting point and motivation for the analysis of various special cases. In particular, we will introduce various analytical techniques and alternative conditions, in order to strengthen the results of Props. 3.1.1-3.1.3, and to extend the existing theory of the SSP problem of Example 1.2.6.

3.2.1 The Case Where Irregular Policies Have Infinite Cost

A weakness of Props. 3.1.1-3.1.3 is that it is sometimes difficult to verify their assumptions, particularly the existence of an optimal S-regular policy. In this section we will use the following assumption that combines elements of the assumptions of these propositions, indirectly guarantees the existence of an optimal S-regular policy, and yields stronger results.

Assumption 3.2.1: We are given a subset $S \subset R(X)$ such that the following hold:

(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 $\hat{J}$ given by

$$\hat{J}(x) = \inf_{\mu:\, S\text{-regular}} J_\mu(x), \qquad x \in X,$$

belongs to S.

(c) For each S-irregular policy μ and each $J \in S$, there is at least one state $x \in X$ such that

$$\limsup_{k\to\infty} (T_\mu^k J)(x) = \infty. \tag{3.4}$$

(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 $\lambda \in \Re$.

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

$$\lim_{m\to\infty} H(x,u,J_m) = H\left(x,u,J\right), \qquad \forall\ x \in X,\ u \in U(x).$$

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

Note that the preceding assumption requires that S is a set of *real-valued* functions; this is similar to Prop. 3.1.3, and it allows us to make use of Eq. (3.4) in the subsequent analysis. Part (a) holds for some common choices of S, such as when $S = R(X)$ or $S = B(X)$, or when S is a subset of $R(X)$ or $B(X)$ of the form $S = \left\{J \in R(X) \mid J \geq J'\right\}$ or $S = \left\{J \in B(X) \mid J \geq J'\right\}$ for a given function J' in $R(X)$ or $B(X)$, respectively. For a finite number n of states, $R(X)$ and $B(X)$ can both be identified with $\Re^n$, while otherwise $R(X)$ may be simpler as it does not require the use of a norm. On the other hand, for an infinite number of states, the choice between $R(X)$ and $B(X)$ may have a substantial impact on whether a policy μ is S-regular or not. In particular, $T_\mu^k J$ may converge to J_μ for all $J \in B(X)$ but not for all $J \in R(X)$, because $T_\mu J \in B(X)$ for $J \in B(X)$ while $T_\mu J \notin R(X)$ for some $J \in R(X)$. This can happen even if T_μ is a contraction mapping with respect to the norm of $B(X)$. Thus some care is needed in deciding whether S should be a subset of $R(X)$ or a subset of $B(X)$.

Since part (b) requires that $\hat{J}$ belongs to S, and is therefore real-valued, it also implies that *there exists at least one S-regular policy* [otherwise the infimum in part (b) is taken over the empty set, and $\hat{J}(x) = \infty$ for all x]. If S satisfies part (a), then part (b) holds if and only if there exists at least one S-regular policy, and also there exists a function in S that is a lower bound to the cost function of all S-regular policies. The existence of such a lower bound is essential, as can be shown by the blackmailer problem of Exercise 3.1.

Part (c) asserts among others that an S-irregular policy cannot be optimal, and is consistent with our analytical approach for semicontractive models, which relies on the dominance of S-regular policies (cf. Prop. 3.1.3). Part (e) is a mild technical continuity assumption that is needed for the subsequent analysis. Part (f) is also of technical nature, but it may not be satisfied in some problems of interest, a notable case being multiplicative models discussed in Example 1.2.8, where

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

S contains only strictly positive functions, and the function g may take values less than 1; see also the affine monotonic and exponential cost SSP models of Section 4.5.

The compactness part (d) of the semicontraction Assumption 3.2.1 plays a key role for asserting the existence of an optimal S-regular policy, as well as for various proof arguments (see Exercise 3.1 for counterexamples). It implies that *for every $J \in S$, the infimum in the equation*

$$(TJ)(x) = \inf_{u \in U(x)} H(x,u,J), \tag{3.5}$$

is attained for all $x \in X$, and it also implies that *for every $J \in S$, there exists a policy μ such that $T_\mu J = TJ$*. This will be shown as part of the proof of the following proposition.

The compactness condition of Assumption 3.2.1(d) can be verified in a few interesting cases involving both finite and infinite state and control spaces:

(1) The case where U is a finite set.

(2) Cases where for each x, $H(x,u,J)$ depends on J only through its values $J(y)$ for y in a finite set Y_x. For an illustration, consider a mapping like the one of the SSP Example 1.2.6:

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

Then the infimum in Eq. (3.5) is attained if $U(x)$ is compact, $g(x,\cdot)$ is lower semicontinuous, and $p_{xy}(\cdot)$ is continuous for each $y \in Y_x$,

since then $H(x, \cdot, J)$ is lower semicontinuous as a function of u. This covers important cases of finite-state and countable-state MDP with compact control spaces (see [BeT91]).

(3) Cases of stochastic optimal control problems such as those of Example 1.2.1, under conditions involving a continuous state space, a compact constraint set $U(x)$, and a problem structure implying that $H(x, \cdot, J)$ is continuous. However, in such cases one must often require that policies obey additional measurability restrictions, and to this end a more complex mathematical formulation is needed to address these restrictions. Such a formulation and corresponding analysis is given for abstract contractive DP models in Chapter 5; see also [BeS78], Ch. 6, and [JaC06] for SSP problems. The extension to the semicontractive models of this section, while in principle straightforward, has not been worked out.

We will show the following proposition, which is the main result of this section.

Proposition 3.2.1: Let Assumption 3.2.1 hold. Then:

(a) The optimal cost function J^* is the unique fixed point of T within the set S.

(b) We have $T^k J \to J^*$ for all $J \in S$. Moreover, there exists an optimal S-regular policy.

(c) A policy μ is optimal if and only if $T_\mu J^* = TJ^*$.

(d) For any $J \in S$, if $J \leq TJ$ we have $J \leq J^*$, and if $J \geq TJ$ we have $J \geq J^*$.

In comparing this proposition with Props. 3.1.1-3.1.3 of Section 3.1, we see that it requires more conditions (cf. Assumption 3.2.1). However, the two assumptions of Props. 3.1.1-3.1.3 require some prior analysis for verification. Assumption 3.2.1 provides a reasonably convenient way to verify these two assumptions in the case where S consists of real-valued functions, but goes further with additional technical conditions, and in doing so it leads to stronger conclusions. These are the uniqueness of the fixed point of T within S (not just within the set $\{J \in S \mid J \geq J^*\}$), and the convergence of the VI sequence $\{T^k J\}$ starting from any $J \in S$ (not just starting from $J \in S$ with $J \geq J^*$).

The proof of Prop. 3.2.1 is long and is developed through several lemmas. These lemmas will also help to illuminate the implications of the various parts of Assumption 3.2.1, and to identify the roles of these parts in the major steps of the proof. The lemmas culminate with showing that the function $\hat{J}$ of Assumption 3.2.1(b) is the unique fixed point of T, and that

any policy μ satisfying $T_\mu \hat{J} = T\hat{J}$ is optimal within the set of S-regular policies. Then the proposition is proved by first showing that $T^k J \to \hat{J}$ for all $J \in S$, and then by using this to prove that $\hat{J} = J^*$ and that there exists an optimal S-regular policy.

Lemma 3.2.1: Let Assumption 3.2.1(d) hold. For every $J \in S$, there exists a policy μ such that $T_\mu J = TJ$.

Proof: Since H is in general extended real-valued, for a given $x \in X$, we need to consider separately the cases $(TJ)(x) < \infty$ and $(TJ)(x) = \infty$. Consider any $x \in X$ with $(TJ)(x) < \infty$, and 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 3.2.1(d), it is compact. The set of points attaining the infimum of $H(x, u, J)$ over $U(x)$ is $\cap_{m=0}^{\infty} U_m(x)$, and is therefore nonempty. Let u_x be a point in this intersection. Then we have

$$H(x, u_x, J) \leq \lambda_m(x), \qquad \forall\ m \geq 0. \tag{3.6}$$

Consider the policy μ 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. (3.6) as $m \to \infty$ shows that μ satisfies $(T_\mu J)(x) = (TJ)(x)$ for x with $(TJ)(x) < \infty$. For x with $(TJ)(x) = \infty$, we also have trivially $(T_\mu J)(x) = (TJ)(x)$, so $T_\mu J = TJ$. **Q.E.D.**

Lemma 3.2.2: Let Assumption 3.2.1(c) hold. A policy μ that satisfies $T_\mu J \leq J$ for some $J \in S$ is S-regular.

Proof: By the monotonicity of T_μ, we have $T_\mu^k J \leq J$, for all $k \geq 1$. Thus $\limsup_{k\to\infty} T_\mu^k J \leq J$ and since J is real-valued, from Assumption 3.2.1(c) it follows that μ cannot be S-irregular. **Q.E.D.**

Lemma 3.2.3: Let Assumption 3.2.1(a),(d),(e) hold. Then if $J' \in S$ and $T^k J' \uparrow J$ for some $J \in S$, we have $J = TJ$.

Proof: We first note that since $J' \in S$ and $J \in S$, from Assumption 3.2.1(a) it follows that $T^k J' \in S$. We fix $x \in X$, and consider the sets

$$U_k(x) = \Big\{ u \in U(x) \mid H(x, u, T^k J') \leq J(x) \Big\}, \qquad k = 0, 1, \ldots, \tag{3.7}$$

which are compact by Assumption 3.2.1(d). Let $u_k \in U(x)$ be such that

$$H(x, u_k, T^k J') = \inf_{u \in U(x)} H(x, u, T^k J') = (T^{k+1} J')(x) \leq J(x);$$

(such a point exists by Lemma 3.2.1). Then $u_k \in U_k(x)$.

For every k, consider the sequence $\{u_i\}_{i=k}^{\infty}$. Since $T^k J' \uparrow J$, it follows using the monotonicity of H, that for all $i \geq k$,

$$H(x, u_i, T^k J') \leq H(x, u_i, T^i J') \leq J(x).$$

Therefore from the definition (3.7), we have $\{u_i\}_{i=k}^{\infty} \subset U_k(x)$. Since $U_k(x)$ is compact, all the limit points of $\{u_i\}_{i=k}^{\infty}$ belong to $U_k(x)$ and at least one limit point exists. Hence the same is true for the limit points of the whole sequence $\{u_i\}$. Thus if $\tilde{u}$ is a limit point of $\{u_i\}$, we have

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

By Eq. (3.7), this implies that

$$H\big(x, \tilde{u}, T^k J'\big) \leq J(x), \qquad k = 0, 1, \ldots.$$

Taking the limit as $k \to \infty$ and using Assumption 3.2.1(e), we obtain

$$(TJ)(x) \leq H(x, \tilde{u}, J) \leq J(x).$$

Thus, since x was chosen arbitrarily within X, we have $TJ \leq J$. To show the reverse inequality, we write $T^k J' \leq J$, apply T to this inequality, and take the limit as $k \to \infty$, so that $J = \lim_{k \to \infty} T^{k+1} J' \leq TJ$. It follows that $J = TJ$. **Q.E.D.**

Lemma 3.2.4: Let Assumption 3.2.1(b),(c),(d) hold. Then:

(a) The function $\hat{J}$ of Assumption 3.2.1(b),

$$\hat{J}(x) = \inf_{\mu:\, S\text{-regular}} J_\mu(x), \qquad x \in X,$$

is the unique fixed point of T within S.

(b) Every policy μ satisfying $T_\mu \hat{J} = T\hat{J}$ is optimal within the set of S-regular policies, i.e., μ is S-regular and $J_\mu = \hat{J}$. Moreover, there exists at least one such policy.

Proof: For all S-regular policies μ, we have $J_\mu \geq \hat{J}$, and by applying T_μ to this relation, we have

$$J_\mu = T_\mu J_\mu \geq T_\mu \hat{J} \geq T\hat{J},$$

where the first equality follows from the S-regularity of μ. Taking the infimum in this relation over all S-regular policies μ and using the definition of $\hat{J}$, we obtain $\hat{J} \geq T\hat{J}$.

To prove the reverse relation, let μ be any policy such that $T_\mu \hat{J} = T\hat{J}$ (there exists one by Lemma 3.2.1). In view of the inequality $\hat{J} \geq T\hat{J}$ just shown, we have $\hat{J} \geq T_\mu \hat{J}$, so μ is S-regular by Lemma 3.2.2. Thus we have, using also the monotonicity of T_μ,

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

From the definition of $\hat{J}$, it follows that equality holds throughout in the preceding relation, so μ is optimal within the class of S-regular policies, and $\hat{J}$ is a fixed point of T.

Next we show that $\hat{J}$ is the unique fixed point of T within S. Indeed if $J' \in S$ is another fixed point, we choose an S-regular μ such that $J_\mu = \hat{J}$ (there exists one by the preceding argument), and we have

$$J' = TJ' \leq T_\mu J' \leq \lim_{k\to\infty} T_\mu^k J' = J_\mu = \hat{J}.$$

Let μ' be such that $J' = TJ' = T_{\mu'}J'$ (cf. Lemma 3.2.1). Then μ' is S-regular (cf. Lemma 3.2.2), and we have

$$J' = \lim_{k\to\infty} T_{\mu'}^k J' = J_{\mu'}.$$

Combining the preceding two relations, we have $J' = J_{\mu'} \leq \hat{J}$, which in view of the definition of $\hat{J}$, implies that $J' = \hat{J}$. **Q.E.D.**

Proof of Prop. 3.2.1: (a), (b) We will first prove that $T^k J \to \hat{J}$ for all $J \in S$, and we will use this to prove that $\hat{J} = J^*$ and that there exists an optimal S-regular policy. Thus both parts (a) and (b) 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 3.2.1(f)]. By the monotonicity of T, we have $T^k J' \uparrow \tilde{J}$ for some $\tilde{J} \in E(X)$. Let μ be an S-regular policy such that $J_\mu = \hat{J}$ [cf. Lemma 3.2.4(b)]. Then we have, using again the monotonicity of T,

$$\tilde{J} = \lim_{k\to\infty} T^k J' \leq \limsup_{k\to\infty} T^k J \leq \lim_{k\to\infty} T_\mu^k J = J_\mu = \hat{J}. \tag{3.8}$$

Since J' and $\hat{J}$ belong to S, and $J' \leq \tilde{J} \leq \hat{J}$, Assumption 3.2.1(a) implies that $\tilde{J} \in S$. From Lemma 3.2.3, it then follows that $\tilde{J} = T\tilde{J}$. Since $\hat{J}$

is the unique fixed point of T within S [cf. Lemma 3.2.4(a)], it follows that $\tilde{J} = \hat{J}$. Thus equality holds throughout in Eq. (3.8), proving that $\lim_{k\to\infty} T^k J = \hat{J}$.

There remains to show that $\hat{J} = J^*$ and that there exists an optimal S-regular policy. To this end, we note that by the monotonicity Assumption 3.1.1, for any policy $\pi = \{\mu_0, \mu_1, \ldots\}$, we have

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

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

$$J_\pi \geq \lim_{k\to\infty} T^k \bar{J} = \hat{J},$$

where the equality follows since $T^k J \to \hat{J}$ for all $J \in S$ (as shown earlier), and $\bar{J} \in S$ [cf. Assumption 3.2.1(a)]. Thus for all $\pi \in \Pi$, $J_\pi \geq \hat{J} = J_\mu$, implying that the policy μ that is optimal within the class of S-regular policies is optimal over all policies, and that $\hat{J} = J^*$.

(c) If μ is optimal, then $J_\mu = J^* \in S$, so by Assumption 3.2.1(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 3.2.2), so $J^* = \lim_{k\to\infty} T_\mu^k J^* = J_\mu$. Therefore, μ is optimal.

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

3.2.2 The Case Where Irregular Policies Have Finite Cost - Perturbations

In this section, we consider problems where some S-irregular policies may have finite cost for all states, so Prop. 3.2.1 does not apply. An example is SSP problems where all one-stage costs are nonpositive and $J^*(x) > -\infty$ for all x. The following example describes a classical problem of this type.

Example 3.2.1 (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 has the form

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

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

Letting $S = R(X)$, we note that the policy $\overline{\mu}$ that stops nowhere is S-irregular, since $T_{\overline{\mu}}$ cannot have a unique fixed point (adding any unit function multiple to J adds to $T_{\overline{\mu}}J$ the same multiple). This policy may violate Assumption 3.2.1(c) of the preceding subsection, because its cost may be finite for all states. A special case where this occurs is when $g(x, w) \equiv 0$ for all x. Then the cost function of $\overline{\mu}$ is identically 0.

Note that case (a) of the three-node shortest path problem given in Section 3.1.2, 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 starting from $J_0 < J^*$ and oscillation of PI) may also arise in the context of the present example.

In this section, we show that the results of Props. 3.1.1-3.1.3 (uniqueness of fixed point of T within the set $\{J \in S \mid J \geq J^*\}$ and convergence of VI starting from within that set) hold, provided that there exists an optimal S-regular policy, and the assumptions are suitably modified by introducing a positive perturbation into T_μ. 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.

In particular, for each $\delta \geq 0$ and policy μ, we consider the mappings $T_{\mu,\delta}$ and T_δ given by

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

We define the corresponding cost functions of policies $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi$ and $\mu \in \mathcal{M}$, and optimal cost function J^*_δ by

$$J_{\pi,\delta}(x) = \limsup_{k\to\infty} T_{\mu_0,\delta} \cdots T_{\mu_k,\delta}\bar{J}, \qquad J_{\mu,\delta}(x) = \limsup_{k\to\infty} T^k_{\mu,\delta}\bar{J},$$

$$J^*_\delta = \inf_{\pi \in \Pi} J_{\pi,\delta}.$$

We refer to the problem associated with the mappings $T_{\mu,\delta}$ as the *δ-perturbed problem*.

The following proposition shows that if the δ-perturbed problem is "well-behaved" with respect to the S-regular policies, then its cost function J^*_δ can be used to approximate the optimal cost function *over the S-regular policies only*.

Proposition 3.2.2: Given a set $S \subset E(X)$, assume that:

(1) For every $\delta > 0$, there exists an optimal S-regular policy for the δ-perturbed problem.

(2) If μ is an S-regular policy, we have

$$J_{\mu,\delta} \leq J_\mu + w_\mu(\delta), \qquad \forall\ \delta > 0,$$

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

Then

$$\lim_{\delta \downarrow 0} J_\delta^* = \inf_{\mu:\, S\text{-regular}} J_\mu,$$

where J_δ^* is the optimal cost function of the δ-perturbed problem.

Proof: For all $\delta > 0$, we have by using condition (2),

$$\inf_{\mu:\, S\text{-regular}} J_\mu \leq J_{\mu_\delta^*} \leq J_{\mu_\delta^*,\delta} = J_\delta^* \leq J_{\mu',\delta} \leq J_{\mu'} + w_{\mu'}(\delta), \quad \forall\ \mu' : S\text{-regular},$$

where μ_δ^* is an optimal S-regular policy of the δ-perturbed problem [cf. condition (1)]. By taking the limit as $\delta \downarrow 0$ and then the infimum over all μ' that are S-regular, it follows that

$$\inf_{\mu:\, S\text{-regular}} J_\mu \leq \lim_{\delta \downarrow 0} J_\delta^* \leq \inf_{\mu:\, S\text{-regular}} J_\mu.$$

Q.E.D.

The preceding proposition does not require that existence of an optimal S-regular policy for the original problem. It applies even if the optimal cost function J^* does not belong to S and we may have $\lim_{\delta \downarrow 0} J_\delta^*(x) > J^*(x)$ for some $x \in X$. This is illustrated by the following example.

Example 3.2.2

Consider the case of a single state where $\bar{J} = 0$, and there are two policies, μ^* and μ, with

$$T_{\mu^*} J = J, \qquad T_\mu J = 1, \qquad \forall\ J \in \Re.$$

Here we have $J_{\mu^*} = 0$ and $J_\mu = 1$. Moreover, it can be verified that for any set $S \subset \Re$ that contains the point 1, the optimal policy μ^* is not S-regular while the suboptimal policy μ is S-regular. For $\delta > 0$, the δ-perturbed problem has optimal cost $J_\delta^* = 1 + \delta$, the unique solution of the Bellman equation

$$J = T_\delta J = \min\{1, J\} + \delta,$$

and its optimal policy is the S-regular policy μ (see Fig. 3.2.1). We also have

$$\lim_{\delta \downarrow 0} J_\delta^* = J_\mu = 1 > 0 = J^*,$$

consistent with Prop. 3.2.2.

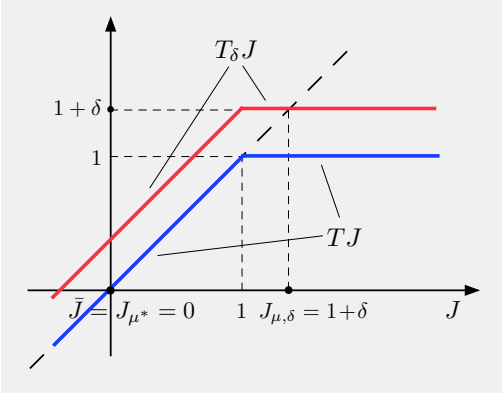

Figure 3.2.1: The mapping T and its perturbed version T_δ in Example 3.2.2.

A simple way to guarantee that $\lim_{\delta \downarrow 0} J_\delta^* = J^*$ is to assume that there exists an optimal S-regular policy for the unperturbed problem. This will also guarantee Bellman's equation $J^* = TJ^*$, under some additional conditions which are collected in the following assumption.

Assumption 3.2.2: We are given a set $S \subset E(X)$ such that the following hold:

(a) There exists an S-regular policy μ^* that is optimal, i.e., $J_{\mu^*} = J^*$, and satisfies

$$J_{\mu^*,\delta} \leq J_{\mu^*} + w(\delta), \qquad \forall\ \delta > 0,$$

where w is a function such that $\lim_{\delta \downarrow 0} w(\delta) = 0$.

(b) The optimal cost function J_δ^* of the δ-perturbed problem belongs to S and satisfies the Bellman equation $J_\delta^* = T_\delta J_\delta^*$ for each $\delta > 0$.

(c) For each sequence $\{J_m\} \subset S$ with $J_m \downarrow J$ for some $J \geq J^*$, we have

$$T_\mu J_m \downarrow T_\mu J, \qquad \forall\ \mu \in \mathcal{M}.$$

Under the preceding assumption we will show that $J^* = TJ^*$. This will allow us to use Prop. 3.1.2 and yield the results of Props. 3.1.1-3.1.3.

Proposition 3.2.3: Let Assumption 3.2.2 hold. Then:

(a) The optimal cost function J^* is the unique fixed point of T within the set $\{J \in S \mid J \geq J^*\}$.

(b) We have $T^k J \to J^*$ for every $J \in S$ with $J \geq J^*$.

(c) An S-regular policy μ that satisfies $T_\mu J^* = TJ^*$ is optimal. Conversely if μ is an S-regular optimal policy, it satisfies $T_\mu J^* = TJ^*$.

Proof: By the monotonicity of T_μ, we clearly have $J^* \leq J^*_\delta$, so using Assumption 3.2.2(a), we obtain for all $\delta > 0$,

$$J^* \leq J^*_\delta \leq J_{\mu^*,\delta} \leq J_{\mu^*} + w(\delta) = J^* + w(\delta),$$

and $\lim_{\delta \downarrow 0} J^*_\delta = J^*$. The Bellman equation $J^*_\delta = T_\delta J^*_\delta$ is written as

$$J^*_\delta = \inf_{\mu \in \mathcal{M}} T_\mu J^*_\delta + \delta\, e. \tag{3.9}$$

From this relation, the fact $\lim_{\delta \downarrow 0} J^*_\delta = J^*$ just shown, and Assumption 3.2.2(c), we have

$$J^* = \lim_{\delta \downarrow 0} \inf_{\mu \in \mathcal{M}} T_\mu J^*_\delta \leq \inf_{\mu \in \mathcal{M}} \lim_{\delta \downarrow 0} T_\mu J^*_\delta = \inf_{\mu \in \mathcal{M}} T_\mu J^* = TJ^*. \tag{3.10}$$

We also have

$$TJ^* \leq TJ^*_\delta = \inf_{\mu \in \mathcal{M}} T_\mu J^*_\delta = J^*_\delta - \delta\, e, \qquad \forall\ \delta > 0,$$

where the last equality follows from Eq. (3.9). By taking the limit as $\delta \downarrow 0$, we obtain $TJ^* \leq J^*$, which combined with Eq. (3.10), shows that $J^* = TJ^*$. Thus the assumptions of Prop. 3.1.2 are satisfied and the conclusions follow from this proposition. **Q.E.D.**

The following example illustrates the preceding line of analysis. For another application, see Section 4.5.3 on exponential cost models.

Example 3.2.3 (Search Problem Continued)

Consider the search problem of Example 3.2.1, assuming that the expected costs for not stopping are nonnegative, $E\big\{g(x,w)\big\} \geq 0$ for all x. Then for all policies μ that don't stop with probability 1 starting from state x we have $J_{\mu,\delta}(x) = \infty$ for all $\delta > 0$, since an expected cost of at least δ is incurred at each transition in the δ-perturbed problem.

If the costs $s(x)$ for stopping are nonpositive for all x, then from known results on SSP problems (cf. Example 1.2.6 and [BeT91]), it follows that there exists an optimal $R(X)$-regular policy, which stops with probability 1 starting from every state. In this case, Assumption 3.2.2 holds, and Prop. 3.2.3 applies. If some stopping costs $s(x)$ are positive, it may happen that each optimal policy is S-irregular, and there is no optimal $R(X)$-regular policy. In this case, however, there is an optimal $R(X)$-regular policy for the δ-perturbed problem, for all $\delta > 0$, and Prop. 3.2.2 applies. Thus $\lim_{\delta \downarrow 0} J^*_\delta$ yields the best that can be achieved when restricted to policies that stop with probability 1.

An Alternative Line of Analysis

A weakness of Assumption 3.2.2(b) is that it requires the verification of Bellman's equation $J^*_\delta = T_\delta J^*_\delta$ for the δ-perturbed problem. An alternative line of analysis makes instead regularity assumptions on the policies of this problem, and is based on the following definition.

Definition 3.2.1: Given a set $S \subset E(X)$ and a scalar $\delta \geq 0$, we say that a stationary policy μ is δ-S-*regular* if $J_{\mu,\delta}$ is the unique fixed point of $T_{\mu,\delta}$ within S, and

$$T^k_{\mu,\delta} J \to J_{\mu,\delta}, \qquad \forall\ J \in S.$$

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

Thus μ is δ-S-regular if and only if it is 0-S-regular for the δ-perturbed problem. Our earlier notions of S-regular and S-irregular policies are equivalent to 0-S-regular and 0-S-irregular policies, respectively. To illustrate, consider the case where T_μ is a weighted sup-norm contraction (e.g., a proper μ in a finite-state SSP problem). Then assuming also that $T_{\mu,\delta}$ maps $B(X)$ into $B(X)$ [e.g., when $e \in B(X)$], $T_{\mu,\delta}$ is a weighted sup-norm contraction for all $\delta \geq 0$, μ is δ-$B(X)$-regular for all $\delta \geq 0$.

Note that δ-S-regularity for all $\delta > 0$ does not imply 0-S-regularity, nor does it imply that $J_{\mu,\delta} \downarrow J_\mu$ as $\delta \downarrow 0$. This can be illustrated by a single-state example, a variation of Example 3.2.2, where $\bar{J} = 0$, and there is a single policy μ with $T_\mu J = \min\{1, J\}$. Here for $S = \Re$, μ is δ-S-regular for all $\delta > 0$, but it is not 0-S-regular, and in fact it can be verified that $J_\mu = 0$ while $J_{\mu,\delta} = 1 + \delta$ (cf. Fig. 3.2.1).

We introduce the following assumption, whose conditions resemble the ones of Prop. 3.1.3. In particular, parts (a)-(c) of the assumption are patterned after conditions (1) and (2) of Prop. 3.1.3.

Assumption 3.2.3: We are given a set $S \subset R(X)$ that contains the function $\bar{J}$, and is such that the following hold:

(a) There exists an S-regular policy μ^* that is optimal, i.e., $J_{\mu^*} = J^*$.

(b) Each S-regular policy is δ-S-regular for every $\delta > 0$. Moreover, if μ is a δ-S-irregular policy for a given $\delta > 0$, then there is at least one state $x \in X$ such that

$$\limsup_{k\to\infty} (T^k_{\mu,\delta} J_{\mu^*,\delta})(x) = \infty, \tag{3.11}$$

where μ^* is the optimal S-regular policy of (a).

(c) For each $\delta > 0$, there exists a policy μ such that $T_\mu J_{\mu^*,\delta} = TJ_{\mu^*,\delta}$, where μ^* is the optimal S-regular policy of (a).

(d) For each sequence $\{J_{\mu^k,\delta_k}\}$, where for all k, μ^k is S-regular, $\delta_k > 0$, $\delta_k \downarrow 0$, and $J_{\mu^k,\delta_k} \downarrow J$,

$$\lim_{m\to\infty} H(x,u,J_{\mu^k,\delta_k}) = H(x,u,J), \qquad \forall\ x \in X,\ u \in U(x).$$

Note that similar to Assumption 3.2.1, the preceding assumption requires that S is a set of *real-valued* functions; this allows us to take advantage of Eq. (3.11). We first show a preliminary lemma.

Lemma 3.2.5: Let Assumption 3.2.3 hold. Then:

(a) If $0 \le \delta \le \delta'$, then for all policies μ, $k \ge 1$, and functions $J, J' \in E(X)$, with $J \le J'$, we have $T^k_{\mu,\delta} J \le T^k_{\mu,\delta'} J'$.

(b) A policy that is δ-S-irregular for some $\delta > 0$ is δ'-S-irregular for all $\delta' \in (\delta, \infty)$.

(c) A policy that is δ-S-regular for some $\delta > 0$ is δ'-S-regular for all $\delta' \in (0, \delta)$.

(d) For every S-regular policy μ and sequence $\{\delta_k\}$ with $\delta_k > 0$ for all k, $\delta_k \downarrow 0$, we have $J_{\mu,\delta_k} \downarrow J_\mu$.

(e) Let $\overline{\mu}$ be a policy that satisfies $T_{\overline{\mu}} J_{\mu,\delta} = TJ_{\mu,\delta}$, where $\delta > 0$ and μ is a δ-S-regular policy. Then

$$T_{\overline{\mu},\delta} J_{\mu,\delta} = T_{\overline{\mu}} J_{\mu,\delta} + \delta\, e = TJ_{\mu,\delta} + \delta\, e \le T_\mu J_{\mu,\delta} + \delta\, e = J_{\mu,\delta}.$$

Proof: (a) For $k = 1$, we have $T_{\mu,\delta} J \le T_{\mu,\delta'} J \le T_{\mu,\delta'} J'$. The proof is completed by induction.

(b) Assume, to arrive at a contradiction, that μ is δ-S-irregular for some $\delta > 0$, and is δ'-S-regular for some $\delta' > \delta$. Then, if μ^* is the optimal S-regular policy of Assumption 3.2.3(a), by the δ-S-irregularity of μ and Assumption 3.2.3(b), there exists $x \in X$ such that

$$J_{\mu,\delta'}(x) = \lim_{k\to\infty} (T^k_{\mu,\delta'} J_{\mu^*,\delta})(x) \ge \limsup_{k\to\infty} (T^k_{\mu,\delta} J_{\mu^*,\delta})(x) = \infty,$$

where the first equality holds because μ is δ'-S-regular and $J_{\mu^*,\delta} \in S$, and

the inequality uses part (a). This contradicts the δ'-S-regularity of μ, which implies that $J_{\mu,\delta'}$ belongs to S and is therefore real-valued.

(c) Follows from part (b).

(d) Since μ is an S-regular policy, it is δ-S-regular for all $\delta > 0$ by Assumption 3.2.3(b). The sequence $\{J_{\mu,\delta_k}\}$ is monotonically nonincreasing, belongs to S, and is bounded below by J_μ, so $J_{\mu,\delta_k} \downarrow J^+$ for some $J^+ \geq J_\mu \geq J^*$. Hence, by Assumption 3.2.3(d), we have $J^+ \in S$ and for all $x \in X$,

$$H\big(x,\mu(x),J^+\big) = \lim_{k\to\infty} H\big(x,\mu(x),J_{\mu,\delta_k}\big) = \lim_{k\to\infty} \big(J_{\mu,\delta_k}(x) - \delta_k\big) = J^+(x),$$

where the second equality follows from the definition of J_{μ,δ_k} as the fixed point of T_{μ,δ_k}. Thus J^+ satisfies $T_\mu J^+ = J^+$ and is therefore equal to J_μ, since μ is S-regular. Hence $J_{\mu,\delta_k} \downarrow J_\mu$.

(e) In the desired relation, repeated below for convenience,

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

the inequality is evident, the second equality is an assumption, and the other equalities follow from the definitions of $T_{\overline{\mu},\delta}$ and $T_{\mu,\delta}$, and the fixed point property $J_{\mu,\delta} = T_{\mu,\delta} J_{\mu,\delta}$. **Q.E.D.**

We have the following proposition, whose conclusions are identical to the ones of the earlier Prop. 3.2.3.

Proposition 3.2.4: Under Assumption 3.2.3 the conclusions of Prop. 3.2.3 hold.

Proof: We will show that $J^* = TJ^*$. The proof will then follow from Prop. 3.1.2. Let μ^* be the optimal S-regular policy of Assumption 3.2.3(a), and let $\{\delta_k\}$ be a positive sequence such that $\delta_k \downarrow 0$. Using Assumption 3.2.3(c), we may choose a policy μ^k such that

$$T_{\mu^k} J_{\mu^*,\delta_k} = T J_{\mu^*,\delta_k}.$$

Using Lemma 3.2.5(e) with $\mu = \mu^*$, which applies since μ^* is δ-S-regular for all $\delta > 0$ [cf. Assumption 3.2.3(b)], we have for all $m \geq 1$,

$$T^m_{\mu^k,\delta_k} J_{\mu^*,\delta_k} \leq T_{\mu^k,\delta_k} J_{\mu^*,\delta_k} \leq T J_{\mu^*,\delta_k} + \delta_k\, e \leq J_{\mu^*,\delta_k},$$

where the first inequality follows from the monotonicity of T_{μ^k,δ_k}. Taking the limit as $m \to \infty$, and using Assumption 3.2.3(b) [cf. Eq. (3.11)], it follows that μ^k is δ^k-S-regular, and we have

$$J_{\mu^*} \leq J_{\mu^k} \leq J_{\mu^k,\delta_k} \leq T J_{\mu^*,\delta_k} + \delta_k\, e \leq J_{\mu^*,\delta_k}, \tag{3.12}$$

where the second inequality follows from Lemma 3.2.5(a). Since $J_{\mu^*,\delta_k} \downarrow J_{\mu^*}$ [cf. Lemma 3.2.5(d)], by taking the limit as $k \to \infty$ in Eq. (3.12), we obtain

$$J_{\mu^*} = \lim_{k\to\infty} TJ_{\mu^*,\delta_k}. \tag{3.13}$$

We thus obtain

$$\begin{aligned}(T_{\mu^*}J_{\mu^*})(x) &= J_{\mu^*}(x)\\ &= \lim_{k\to\infty} \inf_{u\in U(x)} H(x,u,J_{\mu^*,\delta_k})\\ &\le \inf_{u\in U(x)} \lim_{k\to\infty} H(x,u,J_{\mu^*,\delta_k})\\ &= \inf_{u\in U(x)} H(x,u,J_{\mu^*})\\ &= (TJ_{\mu^*})(x)\\ &\le (T_{\mu^*}J_{\mu^*})(x),\end{aligned}$$

where the first equality is the Bellman equation for the S-regular policy μ^*, the second equality is Eq. (3.13), and the third equality follows from Assumption 3.2.3(d) and the fact $J_{\mu^k,\delta_k} \downarrow J_{\mu^*}$. Thus equality holds throughout above and we obtain $J_{\mu^*} = TJ_{\mu^*}$. Since μ^* is optimal, we obtain $J^* = TJ^*$, and the conclusions follow from Prop. 3.1.2. **Q.E.D.**

To see what may happen if there is no optimal S-regular policy, even though there exists a δ-S-regular policy for all $\delta > 0$, consider the example given following the Definition 3.2.1 of a δ-S-regular policy. Here there is a single state, $\bar{J} = 0$, and there is a single policy μ with

$$T_\mu J = TJ = \min\{1, J\},$$

and $J_\mu = J^* = 0$. Then for $S = \Re$, μ is S-irregular, and Assumptions 3.2.3(a) and 3.2.3(b) are violated. As a result, contrary to the assertion of Prop. 3.2.4, the set of fixed points of T is $\{J \mid -\infty < J \le 1\}$ and contains $J < J^* = 0$, while VI starting from every $J \ne 0$ does not converge to J^*.

The following example addresses a class of SSP problems where the perturbation approach applies and yields interesting results.

Example 3.2.4 (Stochastic Shortest Problems with an Optimal Proper Policy)

Consider the finite-spaces SSP problem of Example 1.2.6, and let $S = R(X)$. We assume that there exists an optimal proper policy, and we will show that Assumption 3.2.3 is satisfied, so that Prop. 3.2.4 applies.

Indeed, according to known results for SSP problems discussed in Example 1.2.6 (e.g., [BeT96], Prop. 2.2, [Ber12a], Prop. 3.3.1), a policy μ is proper (stops with probability 1 starting from any $x \in X$) if and only if T_μ is a contraction with respect to a weighted sup-norm. It follows that for a

proper policy μ, $T_{\mu,\delta}$ is a weighted sup-norm contraction for all $\delta \geq 0$, so μ is δ-S-regular and $J_{\mu,\delta} \in R(X)$. Moreover for an improper policy μ we have $J_\mu(x) > -\infty$ for all x (since there exists an optimal policy that is proper and hence its cost function is real-valued). Thus if $\delta > 0$, we obtain

$$\limsup_{k\to\infty} (T^k_{\mu,\delta} J)(x) = \infty, \qquad \forall\ J \in R(X);$$

this is because an additional cost of δ is incurred each time that the policy does not stop. Since for the optimal proper policy μ^*, we have $J_{\mu^*,\delta} \in R(X)$, Assumption 3.2.3(b) holds. In addition the conditions (c) and (d) of Assumption 3.2.3 are clearly satisfied.

Thus if there exists an optimal proper policy, Assumption 3.2.3 holds, and the results of Prop. 3.2.4 apply. In particular, J^* is the unique fixed point of T within the set $\big\{J \in R(X) \mid J \geq J^*\big\}$, and the VI algorithm converges to J^* starting from any function J within this set. These results also apply to the search problem of Example 3.2.1, assuming that there exists an optimal policy that stops with probability 1.

We finally note that similar to the search problem, if we just assume that there exists at least one proper policy, while $J^*(x) > -\infty$ for all $x \in X$, Prop. 3.2.2 applies and shows that $\lim_{\delta\downarrow 0} J^*_\delta$ yields the best that can be achieved when restricted to proper policies only.

The results for the preceding SSP example cannot be improved, in the sense that uniqueness of the fixed point of T within $R(X)$ cannot be shown. This can be verified using case (a) of the shortest path example of Section 3.1.2. Moreover, as shown by the same example, the PI algorithm may oscillate between an optimal and a nonoptimal policy. This motivates modifications of the PI framework, which we will discuss in Section 3.3.3.

3.3 ALGORITHMS

In this section, we will discuss VI and PI algorithms for finding J^* and an optimal policy under the assumptions of the preceding section. We have already shown that the VI algorithm converges to the optimal cost function J^* for any starting function $J \in S$ in the case of Assumption 3.2.1 (cf. Prop. 3.2.1), and also for any starting function $J \in S$ with $J \geq J^*$ in the case of Assumption 3.2.3 (cf. Prop. 3.2.4). We will discuss asynchronous versions of VI under these two assumptions in Section 3.3.1, and will prove satisfactory convergence properties.

In Section 3.3.2, we will show that there is a valid version of the PI algorithm, which starting from an S-regular μ^0, generates a sequence of S-regular policies $\{\mu^k\}$ such that $J_{\mu^k} \to J^*$. We will briefly discuss this algorithm, and then focus on a modified version of PI that is unaffected by the presence of S-irregular policies. This algorithm is similar to the PI algorithm of Section 2.6.3, and can also be implemented in a distributed asynchronous environment. Finally, we will discuss in Section 3.3.3 a version of PI that is based on the perturbation approach of Section 3.2.2.

3.3.1 Asynchronous Value Iteration

Let us consider the model of Section 2.6.1 for asynchronous distributed computation of the fixed point of a mapping T, and the asynchronous distributed VI method described there. The model involves a partition of X into disjoint nonempty subsets $X_1, \ldots, X_m$, and a corresponding partition of J as $J = (J_1, \ldots, J_m)$, where J_ℓ is the restriction of J on the set X_ℓ.

We consider a network of m processors, each updating asynchronously corresponding components of J. In particular, we assume that J_ℓ is updated only by processor ℓ, and only for times t in a selected subset $\mathcal{R}_\ell$ of iterations. Moreover, as in Section 2.6.1, processor ℓ uses components J_j supplied by other processors $j \neq \ell$ with communication "delays" $t - \tau_{\ell j}(t) \geq 0$:

$$J_\ell^{t+1}(x) = \begin{cases} T\left(J_1^{\tau_{\ell 1}(t)}, \ldots, J_m^{\tau_{\ell m}(t)}\right)(x) & \text{if } t \in \mathcal{R}_\ell,\ x \in X_\ell, \\ J_\ell^t(x) & \text{if } t \notin \mathcal{R}_\ell,\ x \in X_\ell. \end{cases} \tag{3.14}$$

Under the assumptions of Section 3.2, we can prove convergence 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 will consider two types of conditions, corresponding to the Assumptions of Sections 3.2.1 and 3.2.2, respectively. In the first case, we will choose $S = B(X)$, while in the second case we will choose $S = \{J \in B(X) \mid J \geq J^*\}$. The reason for using $B(X)$ instead of $R(X)$ is that it may make it easier for policies to be S-regular, since we allow an infinite state space (cf. the remarks following Assumption 3.2.1).

Consider first the case where Assumption 3.2.1 holds with $S = B(X)$, and assume that the continuous updating and information renewal Assumption 2.6.1. Assume further that we have two functions $\underline{V}, \overline{V} \in S$ such that

$$\underline{V} \leq T\underline{V} \leq T\overline{V} \leq \overline{V}, \tag{3.15}$$

so that, by Prop. 3.2.1, $T^k\underline{V} \leq J^* \leq T^k\overline{V}$ for all k, and

$$T^k\underline{V} \uparrow J^*, \qquad T^k\overline{V} \downarrow J^*.$$

Then we can show asynchronous convergence of the VI algorithm (3.14), 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

$$S(k) = \{J \in S \mid T^k\underline{V} \leq J \leq T^k\overline{V}\}, \qquad k = 0, 1, \ldots.$$

The sets $S(k)$ satisfy $S(k+1) \subset S(k)$ in view of Eq. (3.15) and the monotonicity of T. Using Prop. 3.2.1, 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 $J^0 \in S(0)$.

Consider next the case where Assumption 3.2.3 holds with $S = \{J \in B(X) \mid J \geq J^*\}$. In this case we use the sets $S(k)$ given by

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

where $\overline{V}$ is a function in S with $J^* \leq T\overline{V} \leq \overline{V}$. These sets satisfy the synchronous convergence and box conditions of Prop. 2.6.1, and we can similarly show asynchronous convergence to J^* of the generated sequence $\{J^t\}$ starting from any $J^0 \in S(0)$.

3.3.2 Asynchronous Policy Iteration

In this section, we focus on PI methods, under Assumption 3.2.1 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 3.2.1(b)]. At the typical iteration k, we have an S-regular policy μ^k, and we compute a policy μ^{k+1} such that $T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k}$ (this is possible by Lemma 3.2.1). Then μ^{k+1} is S-regular, by Lemma 3.2.2, and we have

$$J_{\mu^k} = T_{\mu^k} J_{\mu^k} \geq T J_{\mu^k} = T_{\mu^{k+1}} J_{\mu^k} \geq \lim_{m \to \infty} T^k_{\mu^{k+1}} J_{\mu^k} = J_{\mu^{k+1}}. \qquad (3.16)$$

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 $J_{\mu^k} \downarrow J^*$ (see Exercise 3.6).

Unfortunately, when there are S-irregular policies, the preceding PI algorithm is somewhat limited, because an initial S-regular policy may not be known, and also because 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. We will need a few assumptions that are in addition to the ones of Section 3.2.1. For analytical simplicity, we include in these assumptions finiteness of the state and control spaces.

Assumption 3.3.1: The set S is equal to $R(X)$, and Assumption 3.2.1 holds with this choice of S. Furthermore, the following hold:

(a) $H(x, u, J)$ is real-valued for all $J \in S$, $x \in X$, and $u \in U(x)$.

(b) X and U are finite sets.

(c) For all scalars $r > 0$ and functions $J \in S$, we have

$$H(x, u, J + r\,e) \leq H(x, u, J) + r\,e, \qquad \forall\ x \in X,\ u \in U(x), \qquad (3.17)$$

where e is the unit function.

In view of the requirement that $S = R(X)$ with X being a finite set, part (c) of the preceding assumption is a nonexpansiveness condition for $H(x, u, \cdot)$, which also implies continuity of $H(x, u, \cdot)$.

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 function with a component $V(x)$ for each x.
- Q is a function with a component $Q(x, u)$ for each pair (x, u) with $u \in U(x)$.

The mapping produces a pair

$$\big(MF_\mu(V, Q), F_\mu(V, Q)\big),$$

where

- $F_\mu(V, Q)$ is a function with a component $F_\mu(V, Q)(x, u)$ for each (x, u), defined by
$$F_\mu(V, Q)(x, u) = H\big(x, u, \min\{V, Q_\mu\}\big), \tag{3.18}$$
where for any Q and μ, we denote by Q_μ the function of x defined by
$$Q_\mu(x) = Q\big(x, \mu(x)\big), \qquad x \in X,$$
and for any two functions V_1 and V_2, we denote by $\min\{V_1, V_2\}$ the function of x given by
$$\min\{V_1, V_2\}(x) = \min\big\{V_1(x), V_2(x)\big\}, \qquad x \in X.$$

- $MF_\mu(V, Q)$ is a function with a component $\big(MF_\mu(V, Q)\big)(x)$ for each x, where M is the operator of pointwise minimization over u:
$$(MQ)(x) = \min_{u \in U(x)} Q(x, u),$$
so that
$$\big(MF_\mu(V, Q)\big)(x) = \min_{u \in U(x)} F_\mu(V, Q)(x, u).$$

We consider an algorithm that is similar to the asynchronous PI algorithm given in Section 2.6.3 for contractive models. It applies asynchronously the mapping $MF_\mu(V, Q)$ for local policy improvement and update of V and μ, and the mapping $F_\mu(V, Q)$ for local policy evaluation and update of Q. The algorithm involves a partition of the state space into sets $X_1, \ldots, X_m$, and assignment of each subset X_ℓ to a processor $\ell \in \{1, \ldots, m\}$. For each ℓ, there are two infinite disjoint subsets of times

$\mathcal{R}_\ell, \overline{\mathcal{R}}_\ell \subset \{0, 1, \ldots\}$, corresponding to policy improvement and policy evaluation iterations, respectively. At time t, each processor ℓ operates on $V^t(x)$, $Q^t(x,u)$, and $\mu^t(x)$, only for x in its "local" state space X_ℓ. In particular, at each time t, each processor ℓ does one of the following:

(a) *Local policy improvement*: If $t \in \mathcal{R}_\ell$, processor ℓ sets for all $x \in X_\ell$,

$$V^{t+1}(x) = \min_{u \in U(x)} H\big(x, u, \min\{V^t, Q^t_{\mu^t}\}\big) = \big(MF_{\mu^t}(V^t, Q^t)\big)(x), \tag{3.19}$$

sets $\mu^{t+1}(x)$ to a u that attains the minimum, and leaves Q unchanged, i.e., $Q^{t+1}(x,u) = Q^t(x,u)$ for all $x \in X_\ell$ and $u \in U(x)$.

(b) *Local policy evaluation*: If $t \in \overline{\mathcal{R}}_\ell$, processor ℓ sets for all $x \in X_\ell$ and $u \in U(x)$,

$$Q^{t+1}(x,u) = H\big(x, u, \min\{V^t, Q^t_{\mu^t}\}\big) = F_{\mu^t}(V^t, Q^t)(x,u), \tag{3.20}$$

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 3.3.1, we will show convergence of the algorithm to (J^*, Q^*), where Q^* is defined by

$$Q^*(x,u) = H(x,u,J^*), \qquad x \in X,\ u \in U(x). \tag{3.21}$$

To this end, we first show that Q^* is the unique fixed point of the mapping F defined by s

$$(FQ)(x,u) = H\big(x,u,MQ\big), \qquad x \in X,\ u \in U(x).$$

Indeed, under our assumption, Prop. 3.2.1 applies, so J^* is the unique fixed point of T, and we have $MQ^* = TJ^* = J^*$. Thus, from the definition (3.21), Q^* is a fixed point of F. To show uniqueness of the fixed point of F, note that if $\overline{Q}$ is a fixed point of F, then $\overline{Q}(x,u) = H(x,u,M\overline{Q})$ for all $x \in X$, $u \in U(x)$, and by minimization over $u \in U(x)$, we have $M\overline{Q} = T(M\overline{Q})$. Hence $M\overline{Q}$ is equal to the unique fixed point J^* of T, so that the equation $\overline{Q} = F\overline{Q}$ yields $\overline{Q}(x,u) = H(x,u,M\overline{Q}) = H(x,u,J^*)$, for all (x,u). From the definition (3.21) of Q^*, it then follows that $\overline{Q} = Q^*$.

We introduce the μ-dependent mapping

$$L_\mu(V,Q) = \big(MQ, F_\mu(V,Q)\big), \tag{3.22}$$

where $F_\mu(V,Q)$ is given by Eq. (3.18). Note that the policy evaluation part of the algorithm [cf. Eq. (3.20)] amounts to applying the second component

of L_μ, while the policy improvement part of the algorithm [cf. Eq. (3.19)] 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 3.3.1: Let Assumption 3.3.1 hold. Then for all $\mu \in \mathcal{M}$, the mapping L_μ of Eq. (3.22) is monotone and has (J^*, Q^*) as its unique fixed point.

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 $Q^* = FQ^*$, as shown earlier. Then, using also the definition of F_μ, we have

$$J^* = MQ^*, \qquad Q^* = FQ^* = F_\mu(J^*, Q^*),$$

which shows that (J^*, Q^*) is a fixed point of L_μ. To show uniqueness, let $(\overline{V}, \overline{Q})$ be a fixed point of L_μ, i.e., $\overline{V} = M\overline{Q}$ and $\overline{Q} = F_\mu(\overline{V}, \overline{Q})$. Then

$$\overline{Q} = F_\mu(\overline{V}, \overline{Q}) = F\overline{Q},$$

where the last equality follows from $\overline{V} = M\overline{Q}$. Thus $\overline{Q}$ is a fixed point of F, and since Q^* is the unique fixed point of F, as shown earlier, we have $\overline{Q} = Q^*$. It follows that $\overline{V} = MQ^* = J^*$, so (J^*, Q^*) is the unique fixed point of L_μ. **Q.E.D.**

The uniform fixed point property of L_μ just shown is, however, insufficient for the convergence proof of the asynchronous algorithm, in the absence of a contraction property. For this reason, we introduce two mappings $\underline{L}$ and $\overline{L}$ that are associated with the mappings L_μ and satisfy

$$\underline{L}(V, Q) \leq L_\mu(V, Q) \leq \overline{L}(V, Q), \qquad \forall\ \mu \in \mathcal{M}. \tag{3.23}$$

These are the mappings defined by

$$\underline{L}(V, Q) = \Big(MQ, \min_{\mu \in \mathcal{M}} F_\mu(V, Q) \Big), \quad \overline{L}(V, Q) = \Big(MQ, \max_{\mu \in \mathcal{M}} F_\mu(V, Q) \Big), \tag{3.24}$$

where the min and max over μ are attained in view of the finiteness of $\mathcal{M}$ [cf. Assumption 3.3.1(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. (3.24), uniformly for all V and (x, u), namely a policy $\bar{\mu}$ for which

$$Q\big(x, \bar{\mu}(x)\big) = \max_{u \in U(x)} Q(x, u), \qquad \forall\ x \in X,$$

[cf. Eq. (3.18)]. Similarly, there exists $\underline{\mu}$ that attains the minimum in Eq. (3.24), uniformly for all V and (x, u). Thus for any given (V, Q), we have

$$\underline{L}(V, Q) = L_{\underline{\mu}}(V, Q), \qquad \overline{L}(V, Q) = L_{\bar{\mu}}(V, Q), \tag{3.25}$$

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 3.3.2: Let Assumption 3.3.1 hold. Then the mappings $\underline{L}$ and $\overline{L}$ of Eq. (3.24) are monotone, and have (J^*, Q^*) as their unique fixed point.

Proof: Monotonicity is clear 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. 3.3.1), and there exists $\underline{\mu}$ such that $\underline{L}(J^*, Q^*) = L_{\underline{\mu}}(J^*, Q^*)$ [cf. Eq. (3.25)], 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. (3.25), we have

$$(V, Q) = \underline{L}(V, Q) = L_{\underline{\mu}}(V, Q)$$

for some $\underline{\mu} \in \mathcal{M}$. Since by Prop. 3.3.1, (J^*, Q^*) is the only fixed point of $L_{\underline{\mu}}$, it follows that $(V, Q) = (J^*, Q^*)$, so (J^*, Q^*) is the only fixed point of $\underline{L}$. Similarly, we show that (J^*, Q^*) is the unique fixed point of $\overline{L}$. **Q.E.D.**

We are now ready to construct a sequence of sets needed to apply Prop. 2.6.1 and prove convergence. For a scalar $c \geq 0$, we denote

$$J_c^- = J^* - c\,e, \qquad Q_c^- = Q^* - c\,e_Q,$$

$$J_c^+ = J^* + c\,e, \qquad Q_c^+ = Q^* + c\,e_Q,$$

with e and e_Q are the unit functions in the spaces of J and Q, respectively.

Proposition 3.3.3: Let Assumption 3.3.1 hold. Then for all $c > 0$,

$$\underline{L}^k(J_c^-, Q_c^-) \uparrow (J^*, Q^*), \qquad \overline{L}^k(J_c^+, Q_c^+) \downarrow (J^*, Q^*), \tag{3.26}$$

where $\underline{L}^k$ (or $\overline{L}^k$) denotes the k-fold composition of $\underline{L}$ (or $\overline{L}$, respectively).

Proof: For any μ, using the assumption (3.17), we have for all (x,u),

$$\begin{aligned} F_\mu(J_c^+,Q_c^+)(x,u) &= H\big(x,u,\min\{J_c^+,Q_c^+\}\big) \\ &= H\big(x,u,\min\{J^*,Q^*\}+c\,e\big) \\ &\le H\big(x,u,\min\{J^*,Q^*\}\big)+c \\ &= Q^*(x,u)+c \\ &= Q_c^+(x,u), \end{aligned}$$

and similarly

$$Q_c^-(x,u) \le F_\mu(J_c^-,Q_c^-)(x,u).$$

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. 3.3.1),

$$(J_c^-,Q_c^-) \le L_\mu(J_c^-,Q_c^-) \le (J^*,Q^*) \le L_\mu(J_c^+,Q_c^+) \le (J_c^+,Q_c^+).$$

Using this relation and Eqs. (3.23) and (3.25), we obtain

$$(J_c^-,Q_c^-) \le \underline{L}(J_c^-,Q_c^-) \le (J^*,Q^*) \le \overline{L}(J_c^+,Q_c^+) \le (J_c^+,Q_c^+). \qquad (3.27)$$

Denote for $k = 0,1,\ldots,$

$$(\overline{V}_k,\overline{Q}_k) = \overline{L}^k(J_c^+,Q_c^+), \qquad (\underline{V}_k,\underline{Q}_k) = \underline{L}^k(J_c^-,Q_c^-).$$

From the monotonicity of $\overline{L}$ and $\underline{L}$ and Eq. (3.27), we have that $(\overline{V}_k,\overline{Q}_k)$ converges monotonically from above to some $(\overline{V},\overline{Q}) \ge (J^*,Q^*)$, while $(\underline{V}_k,\underline{Q}_k)$ converges monotonically from below to some $(\underline{V},\underline{Q}) \le (J^*,Q^*)$. 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 property of $\overline{L}$, implied by Eq. (3.17) and the finiteness of the control constraint set, 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 (3.19)-(3.20), consider the sets

$$S(k) = \big\{(V,Q) \mid \underline{L}^k(J_c^-,Q_c^-) \le (V,Q) \le \overline{L}^k(J_c^+,Q_c^+)\big\}, \qquad k = 0,1,\ldots,$$

whose intersection is (J^*,Q^*) [cf. Eq. (3.26)]. By Prop. 3.3.3 and Eq. (3.23), 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 (3.19)-(3.20) 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.

3.3.3 Policy Iteration With Perturbations

Let us now consider PI for problems where there exists an optimal S-regular policy, but S-irregular policies may have real-valued cost functions, and the perturbation approach of Section 3.2.2 applies. We will develop an algorithm that generates a sequence of policies $\{\mu^k\}$ such that $J_{\mu^k} \to J^*$, under the following assumption, which among others, is satisfied in the SSP problem of Example 3.2.4.

Assumption 3.3.2: Assumption 3.2.3 holds, and in addition:

(a) For every $\delta > 0$ and δ-S-regular policy μ, there exists a policy $\overline{\mu}$ such that $T_{\overline{\mu}} J_{\mu,\delta} = T J_{\mu,\delta}$.

(b) For every $\delta > 0$ and δ-S-irregular policy μ, and for every $J \in S$, there exists at least one state $x \in X$ such that

$$\limsup_{k\to\infty} (T^k_{\mu,\delta} J)(x) = \infty.$$

We generate a sequence $\{\mu^k\}$ with a perturbed version of PI as follows. Let $\{\delta_k\}$ be a positive sequence with $\delta_k \downarrow 0$, and let μ^0 be any δ_0-S-regular policy. At the typical iteration k, we have a δ_k-S-regular policy μ^k, and we generate μ^{k+1} according to

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

Such μ^{k+1} exists by Assumption 3.3.2(a), and we claim that μ^{k+1} is δ_{k+1}-S-regular. To see this, note that from Lemma 3.2.5(e), we have

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

so that

$$T^m_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k} \le T_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k} = T J_{\mu^k,\delta_k} + \delta_k\, e \le J_{\mu^k,\delta_k}, \qquad \forall\ m \ge 1. \tag{3.29}$$

Since $J_{\mu^k,\delta_k} \in R(X)$, from Assumption 3.3.2(b) it follows that μ^{k+1} is δ_k-S-regular, and hence also δ_{k+1}-S-regular, by Lemma 3.2.5(c). Thus the sequence $\{\mu^k\}$ generated by the perturbed PI algorithm (3.28) is well-defined and consists of δ_k-S-regular policies. We have the following proposition.

Proposition 3.3.4: Let Assumption 3.3.2 hold. Then the sequence $\{J_{\mu^k}\}$ generated by the algorithm (3.28) satisfies $J_{\mu^k} \to J^*$.

Proof: Using Eq. (3.29), we have

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

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

$$J^* \leq J^+ = \lim_{k\to\infty} TJ_{\mu^k,\delta_k}. \tag{3.30}$$

We also have

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

where the equality follows from Assumption 3.2.3(d). It follows that equality holds throughout above, so that

$$\lim_{k\to\infty} TJ_{\mu^k,\delta_k} = TJ^+. \tag{3.31}$$

Combining Eqs. (3.30) and (3.31), we obtain $J^* \leq J^+ = TJ^+$. Since by Prop. 3.2.4, J^* is the unique fixed point of T within $\{J \in S \mid J \geq J^*\}$, it follows that $J^+ = J^*$. Thus $J_{\mu^k,\delta_k} \downarrow J^*$, and since $J_{\mu^k,\delta_k} \geq J_{\mu^k} \geq J^*$, we have $J_{\mu^k} \to J^*$. **Q.E.D.**

Note that when X and U are finite sets, as in the SSP problem of Example 3.2.4, Prop. 3.3.4 implies that the generated policies μ^k will be optimal for all k sufficiently large. The reason is that in this case, 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 $J_\mu(x) \geq J^*(x) + \epsilon$.

In the absence of finiteness of X and U, Prop. 3.3.4 guarantees the monotonic pointwise convergence of $\{J_{\mu^k,\delta_k}\}$ to J^* (see the preceding proof) and the (possibly nonmonotonic) pointwise convergence of $\{J_{\mu^k}\}$ to J^*. This convergence behavior should be contrasted with the behavior of PI without perturbations, which may lead to oscillation between two nonoptimal policies, as noted earlier.

3.4 NOTES, SOURCES, AND EXERCISES

The semicontractive model framework of this chapter and the material of Section 3.1 are new. 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.

The line of proof of Prop. 3.1.1 dates back to an analysis of SSP problems with finite state and control spaces, given in the author's [Ber87], Section 6.4, which assumes existence of an optimal proper policy and nonnegativity of the one-stage cost. Proposition 3.2.1 is patterned after a similar result in Bertsekas and Tsitsiklis [BeT91] for SSP problems with finite state space and compact control constraint sets. The proof given there contains an intricate part (Lemma 3 of [BeT91]) to show a lower bound on the cost functions of proper policies, which is assumed here in part (b) of the semicontraction Assumption 3.2.1.

The perturbation analysis of Section 3.2.2, including the PI algorithm of Section 3.3.3, are new and are based on unpublished collaboration of the author with H. Yu. The results for SSP problems using this analysis (cf. Prop. 3.2.4) strengthen the results of [Ber87] (Section 6.4) and [BeT91] (Prop. 3), in that the one-stage cost need not be assumed nonnegative. We have given two different perturbation approaches in Section 3.2.2. The first approach places assumptions on the optimal cost function J_δ^* of the δ-perturbed problem (cf. Prop. 3.2.2 and Assumption 3.2.2), while the second places assumptions on policies (cf. Assumption 3.2.3) and separates them into δ-S-regular and δ-S-irregular. The first approach is simpler analytically, and at least in part, does not require existence of an S-regular policy (cf. Prop. 3.2.2). The second approach allows the development of a perturbed PI algorithm and the corresponding analysis of Section 3.3.3 (under the extra conditions of Assumption 3.3.2).

The asynchronous PI algorithm of Section 3.3.2 is essentially the same as one of the optimistic PI algorithms of Yu and Bertsekas [YuB11a] for the SSP problem of Example 1.2.6. This paper also analyzed asynchronous stochastic iterative versions of the algorithms, and proved convergence results that parallel those for classical Q-learning for SSP, given in Tsitsiklis [Tsi94] and Yu and Bertsekas [YuB11b]. We follow the line of analysis of that paper. A related paper, which deals with a slightly different asynchronous PI algorithm in an abstract setting and without a contraction structure, is Bertsekas and Yu [BeY10b].

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 system with minimum expected cost. Generalized forms of SSP problems, which involve an infinite (uncountable) number of states, in addition to the termination state, are analyzed by Pliska [Pli78], Hernandez-Lerma et al. [HCP99], and James and Collins [JaC06]. The latter paper allows improper

policies, assumes that J^* is bounded below, and generalizes the results of [BeT91] to infinite (Borel) state spaces, using a similar line of proof.

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] proves results that are similar to the ones for SSP problems with perfect state information. These results can also be derived from the 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 3.2.1(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 (Blackmailer's Dilemma)

Consider an SSP problem where there is only one state $x = 1$, in addition to the termination state 0. At state 1, we can choose a control u with $0 < u \leq 1$, while incurring a cost $-u$; we then move to state 0 with probability u^2, and stay in state 1 with probability $1 - u^2$. We may regard u as a demand made by a blackmailer, and state 1 as the situation where the victim complies. State 0 is the situation where the victim (permanently) refuses to yield to the blackmailer's demand. The problem then can be seen as one whereby the blackmailer tries to maximize his total gain by balancing his desire for increased demands with keeping his victim compliant. In terms of abstract DP we have

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

(a) Verify that T_μ is a sup-norm contraction for each μ. In addition, show that $J_\mu(1) = -\frac{1}{\mu(1)}$, so that $J^*(1) = -\infty$, that there is no optimal policy, and that T has no fixed points within $\Re$. Which parts of Assumption 3.2.1 with $S = \Re$ are violated?

(b) Consider 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 $1 - u$. Here we have

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

Verify that $J^*(1) = -1$, that there is no optimal policy, and that T has multiple fixed points within $\Re$. Which parts of Assumption 3.2.1 with $S = \Re$ are violated?

(c) Repeat part (b) for the case where at state 1, we may also choose $u = 0$ at a cost c. Show that the policy $\overline{\mu}$ that chooses $\overline{\mu}(1) = 0$ is $\Re$-irregular. What are the optimal policies and the fixed points of T in the three cases where $c > 0$, $c = 0$, and $c < 0$. Which parts of Assumption 3.2.1 with $S = \Re$ are violated in each of these three cases?

3.2 (Equivalent Semicontractive Conditions)

Let S be a given subset of $E(X)$. Show that the assumptions of Prop. 3.1.1 hold if and only if $J^* \in S$, $TJ^* \leq J^*$, and there exists an S-regular policy μ such that $T_\mu J^* = TJ^*$.

3.3

Consider the three-node shortest path example of Section 3.1.2. Try to apply Prop. 3.1.1 with $S = [-\infty, \infty) \times [-\infty, \infty)$. What conclusions can you obtain for various values of a and b, and how do they compare with those for $S = \Re^2$?

3.4 (Changing $\bar{J}$)

Let the assumptions of Prop. 3.1.1 hold, and let J^* be the optimal cost function. Suppose that $\bar{J}$ is changed to some function $J \in S$.

(a) Show that following the change, J^* continues to be the optimal cost function over just the S-regular policies.

(b) Consider the three-node shortest path problem of Section 3.1.2 for the case where $a = 0$, $b < 0$. Change $\bar{J}$ from $\bar{J} = 0$ to $\bar{J} = r\,e$ where $r \in \Re$. Verify the result of part (a) for this example. For which values of r is the $\Re^2$-irregular policy optimal?

3.5 (Alternative Semicontractive Conditions)

The purpose of this exercise and the next one is to provide conditions that imply the results of Prop. 3.1.1. Let S be a given subset of $E(X)$. Assume that:

(1) There exists an optimal S-regular policy.

(2) For every S-irregular policy $\overline{\mu}$, we have $T_{\overline{\mu}} J^* \geq J^*$.

Then the assumptions and the conclusions of Prop. 3.1.1 hold.

3.6 (Convergence of PI)

Let Assumption 3.2.1 hold, and let $\{\mu^k\}$ be the sequence generated by the PI algorithm described at the start of Section 3.3.2 [cf. Eq. (3.16)]. Let also $J_\infty = \lim_{k\to\infty} J_{\mu^k}$, and assume that $H(x, u, J_{\mu^k}) \to H(x, u, J_\infty)$ for all $x \in X$ and $u \in U(x)$. Show that $J_\infty = J^*$.