Dynamic Programming and Optimal Control

4th Edition, Volume II

by

Dimitri P. Bertsekas

Massachusetts Institute of Technology

APPENDIX B

Regular Policies in Total Cost Dynamic Programming

NEW

July 13, 2016

This is a new appendix for the author's Dynamic Programming and Optimal Control, Vol. II, 4th Edition, Athena Scientific, 2012. It includes new research, and its purpose is to address issues relating to the solutions of Bellman's equation, and the validity of the value iteration (VI) and policy iteration (PI) algorithms in infinite horizon total cost problems, with an emphasis on the undiscounted problems of Chapters 3 and 4.

We adopt an abstract DP viewpoint, similar to the one we used in Sections 1.6, 2.5, and 2.6. As in these sections, we aim to unify the analysis, to highlight the significant structures of the corresponding DP models, and to connect it to the developments of Chapters 3 and 4. In particular, we do not assume a contractive character for the associated DP mappings, requiring them to be just monotone. Abstract DP models are the subject of the author's research monograph [Ber13], which may be consulted for a more extensive analysis, and for proofs of some of the results of the present appendix, which will be given without proof.

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

APPENDIX B: Regular Policies in Total Cost Dynamic Programming

The purpose of this appendix is to address issues relating to the fundamental structure of Bellman's equation, and the validity of the value iteration (VI) and policy iteration (PI) algorithms in infinite horizon total cost problems. We focus on the more complex undiscounted problems of Chapters 3 and 4. In particular, we do not assume a contractive character for the DP mappings T_μ and T, requiring them to be just monotone.

We adopt an abstract DP viewpoint, similar to the one we used in Sections 1.6, 2.5, and 2.6. Similar to these sections, our aim is to unify the analysis, to highlight the significant structures of the corresponding DP models, and to connect it to the developments of Chapters 3 and 4. Abstract DP models are the subject of the author's research monograph [Ber13], which may be consulted for a more extensive analysis, and for proofs of some of the results of the present appendix.

The range of application of contractive models includes discounted problems with bounded cost per stage, and related discounted semi-Markov and zero sum sequential games, as well as SSP problems where all policies are proper (cf. Section 3.3). At the other extreme, we have noncontractive models, such as the positive and negative cost problems of Section 4.1. As a result, Bellman's equation may have multiple solutions, and the VI and PI algorithms may not work.

Between these extremes, we have encountered a number of models that do not have a contractive nature, yet possess enough structure to allow more powerful results. Examples are the SSP models with improper policies of Chapter 3, the deterministic optimal control problems of Section 4.2, the SSP problems of Section 4.4, and the affine monotonic problems

of Section 4.5. These models possess important theoretical characteristics, such as the uniqueness of solution of Bellman's equation within a subset of interest, and the validity of useful forms of VI and PI. An important feature of these models is that some policies (called *regular*) are well-behaved with respect to VI, in the sense that their cost function can be obtained by VI starting from a wide range of initial conditions, while other policies (called *irregular*) are not so well-behaved.

An example of regular policy is a stationary policy μ for which T_μ is a contraction within the set of bounded functions $B(X)$, so that $T_\mu^k \to J_\mu$ for all $J \in B(X)$. In particular, proper policies in SSP models are contractive and regular, while improper policies are not, leading to the characterization of SSP models as *semicontractive*, a term introduced in the monograph [Ber13]. Similarly, stable policies in affine monotonic problems (cf. Section 4.5) and terminating policies in deterministic optimal control (cf. Section 4.2) are regular.†

Our analysis revolves around the optimal cost function over just the regular policies, which we denote by $\hat{J}$. In summary, key insights from this analysis are:

(a) Because the regular policies are well-behaved with respect to VI, $\hat{J}$ *is also well-behaved with respect to VI, and demarcates the location of the fixed points of* T. In particular the limits of VI starting from above $\hat{J}$ as well as all the fixed points of T, lie below $\hat{J}$.

(b) With a judicious choice of the set of regular policies, $\hat{J}$ *can be proved to be the largest solution of Bellman's equation*. Moreover VI converges to $\hat{J}$ starting from above, i.e., from initial conditions $J \geq \hat{J}$, while PI also converges to $\hat{J}$ under favorable circumstances. Note that the optimal cost function over all policies, J^*, does not have such a property: it may be the largest solution of Bellman's equation, as in negative cost problems (cf. Section 4.1), or the smallest solution (among nonnegative functions), as in positive cost problems (cf. Section 4.1), or it may not be a solution at all (cf. the counterexample of Section 4.4).

(c) If the problem structure is such that irregular policies cannot be "better" than regular ones, in the sense that $J^* = \hat{J}$, then J^* *is the largest*

† The intended use of the term "semicontractive" is to characterize models where some (but not all) of the mappings T_μ are contractions with respect to a suitable norm, while the others are not. Typical examples are the SSP problems of Chapter 3 and Section 4.4, and the affine monotonic models of Section 4.5. In the abstract context of this appendix, semicontractive models will be discussed in Sections B.3, B.5, and B.6. The notion of regularity, as developed in Section B.2, goes beyond semicontractiveness since it relates to nonstationary policies as well.

solution of Bellman's equation. Moreover, J^* *can be obtained by VI and PI starting from a wide range of initial conditions*.

(d) Under some special circumstances where irregular policies cannot be optimal, J^* *is the unique solution of Bellman's equation*. Moreover, J^* *can be obtained by VI starting from any real-valued initial conditions, as well as by specially modified forms of PI*. An example are the SSP problems in Chapter 3, under the favorable Assumptions 3.1.1 and 3.1.2.

Our line of development leads to a variety of interesting results, richer in character than the ones we obtained for SSP problems, where the regular policies can be identified with the proper policies. For example our results apply to the infinite-state deterministic and stochastic optimal control problems, as well as to finite-state minimax-type shortest path problems. Moreover, our results can be strengthened in the presence of additional special structure.

In what follows, we first formulate our abstract DP model in Section B.1. Then we develop the main ideas of our approach, first for nonstationary policies (Section B.2), and then for stationary policies (Section B.3). We then apply the results of Sections B.2 and B.3 in a variety of contexts, including monotone increasing and monotone decreasing models (Section B.4), and shortest path-like problems (Section B.5). Special cases of the theory of Section B.5 include the SSP case of Chapter 3 and the affine monotonic problems of Section 4.5 under the infinite cost Assumption 4.5.3. We also discuss in Section B.5, robust shortest path planning problems, a minimax analog of the SSP problem using the analysis of the author's paper [Ber14].

B.1 AN ABSTRACT DYNAMIC PROGRAMMING MODEL

We introduce an abstract DP model, which is similar to the one of Section 1.6, except that it does not possess a contractive structure. Let X and U be two sets, which we loosely refer to as a set of "states" and a set of "controls," respectively. For each $x \in X$, let $U(x) \subset U$ be a nonempty subset of controls that are feasible at state x. We denote by $\mathcal{M}$ the set of all functions $\mu : X \mapsto U$ with $\mu(x) \in U(x)$, for all $x \in X$.

In analogy with DP, we consider policies, which are sequences $\pi = \{\mu_0, \mu_1, \ldots\}$, with $\mu_k \in \mathcal{M}$ for all k. We denote by Π the set of all policies. We refer to a sequence $\{\mu, \mu, \ldots\}$, with $\mu \in \mathcal{M}$, as a *stationary policy*. With slight abuse of terminology, we will also refer to any $\mu \in \mathcal{M}$ as a "policy" and use it in place of $\{\mu, \mu, \ldots\}$, when confusion cannot arise.

We denote by $\Re$ the set of real numbers, by $R(X)$ the set of real-valued functions $J : X \mapsto \Re$, and by $E(X)$ the subset of extended real-valued functions $J : X \mapsto \Re \cup \{-\infty, \infty\}$. We denote by $E^+(X)$ the set of

all nonnegative extended real-valued functions of $x \in X$. Throughout the paper, 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 $x \in X$, and we write $J_k \downarrow J$ if $\{J_k\}$ is monotonically nonincreasing and $J_k \to J$.

We introduce a mapping $H : X \times U \times E(X) \mapsto \Re \cup \{-\infty, \infty\}$, satisfying the following condition.

Assumption B.1.1: (Monotonicity) If $J, J' \in E(X)$ and $J \leq J'$, then

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

We define the mapping T that maps a function $J \in E(X)$ to the function $TJ \in E(X)$, given by

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

Also for each $\mu \in \mathcal{M}$, we define 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 $k = 0, 1, \ldots$,

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

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

which will be used repeatedly in what follows. Here, as in Section 1.6, T^k and T_μ^k denotes the composition of T and T_μ, respectively, with itself k times. More generally, given $\mu_0, \ldots, \mu_k \in \mathcal{M}$, we denote by $T_{\mu_0} \cdots T_{\mu_k}$ the composition of $T_{\mu_0}, \ldots, T_{\mu_k}$, so for all $J \in E(X)$,

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

We now consider cost functions associated with T_μ and T. We introduce a function $\bar{J} \in E(X)$, and we define the infinite horizon cost of a policy as the upper limit of its finite horizon costs with $\bar{J}$ being the cost function at the end of the horizon (note here the similarity with the affine monotonic models of Section 4.5).

Definition B.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

$$J_\pi(x) = \limsup_{k\to\infty} (T_{\mu_0} \cdots T_{\mu_k} \bar{J})(x), \qquad \forall\ x \in X. \tag{B.1}$$

The optimal cost function J^* is defined by

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

A policy $\pi^* \in \Pi$ is said to be optimal if $J_{\pi^*} = J^*$.

Some Examples

The model just described is broadly applicable, and includes as special cases essentially all the total cost infinite horizon DP problems that we have discussed including stochastic and minimax, discounted and undiscounted, semi-Markov, multiplicative, risk-sensitive, etc. As an example, for a deterministic discrete-time optimal control problem involving the system

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

and a cost $g(x_k, u_k)$ for the kth stage (cf. Section 4.2), the mapping H is given by

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

and $\bar{J}$ is the zero function $[\bar{J}(x) \equiv 0]$. It can be seen that the cost function of a policy π, as given by Eq. (B.1), takes the form

$$J_\pi(x) = \limsup_{k\to\infty} (T_{\mu_0} \cdots T_{\mu_k} \bar{J})(x) = \limsup_{k\to\infty} \sum_{t=0}^{k} g\big(x_t, \mu_t(x_t)\big), \tag{B.2}$$

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

For the affine monotonic model of Section 4.5, the mapping H is given by

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

where $g(i, u) \geq 0$ and $A_{ij}(u) \geq 0$ for all i, j, and $u \in U(x)$, and $\bar{J}$ is some nonnegative function; as an example for the multiplicative and exponential cost problems of Section 4.5, we have $\bar{J}(i) \equiv 1$.

For an undiscounted stochastic problem involving a Markov chain with state space $X = \{1, \ldots, n\}$, transition probabilities $p_{xy}(u)$, and expected one-stage cost function g, the mapping H is given by

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

(with the convention $\infty - \infty = \infty$ if J is extended real-valued). The SSP problem arises when one of the states is cost-free and absorbing.

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

$$x_{k+1} = f\big(x_k, \mu_k(x_k), w_k\big), \qquad k = 0, 1, \ldots, \tag{B.3}$$

where w_k is a random disturbance that takes values from a space W. We assume that w_k, $k = 0, 1, \ldots$, are characterized by probability distributions $P(\cdot \mid x_k, u_k)$ that are identical for all k, where $P(w_k \mid x_k, u_k)$ is the probability of occurrence of w_k, when the current state and control are x_k and u_k, respectively. Thus the probability of w_k may depend explicitly on x_k and u_k, but not on values of prior disturbances $w_{k-1}, \ldots, w_0$. We allow infinite state and control spaces, as well as problems with discrete (finite or countable) state space (in which case the underlying system is a Markov chain). However, for technical reasons that relate to measure theoretic issues, we assume that W is a countable set.

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

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

subject to the system equation constraint (B.3), where g is the one-stage cost function. The corresponding mapping of the abstract DP problem is

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

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

A discounted version of the problem is defined by the mapping

$$H(x, u, J) = E\big\{ g(x, u, w) + \alpha J\big(f(x, u, w)\big) \big\}, \tag{B.4}$$

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

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

B.2 REGULAR POLICIES, VALUE ITERATION, AND FIXED POINTS OF T

Generally, in a DP model, one expects to establish that J^* is a solution of Bellman's equation, i.e., it is a fixed point of T. This is known to be true for most of the major DP models under reasonable conditions, and in fact it may be viewed as an indication of exceptional behavior when it does not hold. For some models, J^* is the unique fixed point of T within a convenient subset of $E(X)$, such as the space of bounded functions. An example is contractive models where T_μ is a contraction mapping for all $\mu \in \mathcal{M}$, with respect to some norm and with a common modulus of contraction (cf. Chapters 1 and 2), and SSP problems under the assumptions of Chapter 3. However, in general T may have multiple fixed points within $E(X)$, including for some popular DP problems, while in exceptional cases, J^* may not be among the fixed points of T (as it can happen in SSP problems under the weak conditions of Section 4.4 and the affine monotonic problems of Section 4.5).

A related question is the convergence of VI, which we will view as the fixed point algorithm that generates $T^k J$, $k = 0, 1, \ldots$, starting from a function $J \in E(X)$. Generally, for abstract DP models where J^* is a fixed point of T, VI converges to J^* starting from within some subset of initial functions J, but not necessarily from every J; this is certainly true when T has multiple fixed points. One of the purposes of this appendix is to characterize the set of functions starting from which VI converges to J^*, and the related issue of multiplicity of fixed points, through notions of regularity that we now introduce.

Definition B.2.1: For a nonempty set of functions $S \subset E(X)$, we say that a collection $\mathcal{C}$ of policy-state pairs (π, x), with $\pi \in \Pi$ and $x \in X$, is *S-regular* if

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

Thus for an S-regular collection of pairs (π, x), the value of $J_\pi(x)$ is not affected if the starting function is changed from $\bar{J}$ to any $J \in S$. In particular, if π is a stationary policy μ, VI yields in the limit $J_\mu(x)$ starting from any $J \in S$.

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

$$J^*_{\mathcal{C}}(x) = \inf_{\{\pi \,|\, (\pi, x) \in \mathcal{C}\}} J_\pi(x), \qquad x \in X. \tag{B.5}$$

Note that $J^*_{\mathcal{C}} \geq J^*$ [if for some $x \in X$, the set of policies $\{\pi \mid (\pi, x) \in \mathcal{C}\}$ is empty, we have $J^*_{\mathcal{C}}(x) = \infty$]. We will try to characterize the sets of fixed points of T and limit points of VI in terms of the function $J^*_{\mathcal{C}}$ for an S-regular set $\mathcal{C}$. The following is a key proposition.†

Proposition B.2.1: Given a set $S \subset E(X)$, let $\mathcal{C}$ be a collection of policy-state pairs (π, x) that is S-regular.

(a) For all $J \in S$, we have

$$\liminf_{k\to\infty} T^k J \leq \limsup_{k\to\infty} T^k J \leq J^*_{\mathcal{C}}.$$

(b) For all $J' \in E(X)$ with $J' \leq TJ'$, and all $J \in E(X)$ such that $J' \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$, we have

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

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

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

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

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

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

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

$$J'(x) \leq (TJ')(x) \leq \cdots \leq (T^k J')(x) \leq (T^k J)(x) \leq (T^k \tilde{J})(x).$$

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

Some interesting implications of part (b) of the proposition are that given a set $S \subset E(X)$, and a set $\mathcal{C} \subset \Pi \times X$ that is S-regular:

† In this proposition as well as later, when referring to a collection $\mathcal{C}$ that is S-regular, we implicitly assume that $\mathcal{C}$ and S are nonempty.

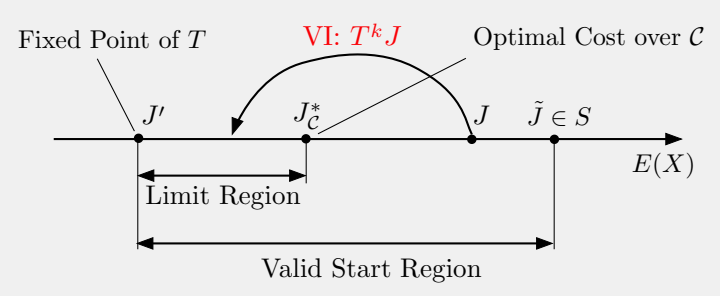

Figure B.2.1. Illustration of Prop. B.2.1. Neither $J^*_{\mathcal{C}}$ nor J^* need to be fixed points of T, but if $\mathcal{C}$ is S-regular, and there exists $\tilde{J} \in S$ with $J^*_{\mathcal{C}} \leq \tilde{J}$, then $J^*_{\mathcal{C}}$ demarcates from above the range of fixed points of T that lie below $\tilde{J}$.

(1) $J^*_{\mathcal{C}}$ is an upper bound to every fixed point J' of T that lies below some $\tilde{J} \in S$ (i.e., $J' \leq \tilde{J}$). Moreover, for such a fixed point J', the convergence of VI is characterized by the *valid start region*

$$\big\{J \in E(X) \mid J^*_{\mathcal{C}} \leq J \leq \tilde{J} \text{ for some } \tilde{J} \in S\big\},$$

and the *limit region*

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

The VI algorithm, starting from the former, ends up asymptotically within the latter; cf. Fig. B.2.1.

(2) If $J^*_{\mathcal{C}}$ is a fixed point of T (a common case in our subsequent analysis), then VI converges to $J^*_{\mathcal{C}}$ starting from any $J \in E(X)$ such that $J^*_{\mathcal{C}} \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$. For future reference, we state this observation as a proposition.

Proposition B.2.2: Given a set $S \subset E(X)$, let $\mathcal{C}$ be a collection of policy-state pairs (π, x) that is S-regular, and assume that $J^*_{\mathcal{C}}$ is a fixed point of T. Then $J^*_{\mathcal{C}}$ is the only possible fixed point of T within the set of all $J \in E(X)$ such that $J^*_{\mathcal{C}} \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$. Moreover, $T^k J \to J^*_{\mathcal{C}}$ for all $J \in E(X)$ such that $J^*_{\mathcal{C}} \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$.

Proof: Let $J \in E(x)$ and $\tilde{J} \in S$ be such that $J^*_{\mathcal{C}} \leq J \leq \tilde{J}$. Using the fixed point property of $J^*_{\mathcal{C}}$ and the monotonicity of T, we have

$$J^*_{\mathcal{C}} = T^k J^*_{\mathcal{C}} \leq T^k J \leq T^k \tilde{J}, \qquad k = 0, 1, \ldots.$$

From Prop. B.2.1(b), with $J' = J^*_{\mathcal{C}}$, it follows that $T^k\tilde{J} \to J^*_{\mathcal{C}}$, so taking limit in the above relation as $k \to \infty$, we obtain $T^k J \to J^*_{\mathcal{C}}$. **Q.E.D.**

The preceding proposition takes special significance when $\mathcal{C}$ is rich enough so that $J^*_{\mathcal{C}} = J^*$, as for example in the case where $\mathcal{C}$ is the set $\Pi \times X$ of all (π, x), or other choices to be discussed later. It then follows that VI converges to J^* starting from any $J \in E(X)$ such that $J^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$.† In the particular applications to be discussed in Section B.4 we will use such a choice.

Note that Prop. B.2.2 does not say anything about fixed points of T that lie below $J^*_{\mathcal{C}}$, and does not give conditions under which $J^*_{\mathcal{C}}$ is a fixed point. In particular, it does not address the question whether J^* is a fixed point of T, or whether VI converges to J^* starting from $\bar{J}$ or from below J^*. Generally, it can happen that both, only one, or none, of the two functions $J^*_{\mathcal{C}}$ and J^* is a fixed point of T! These are major issues in abstract DP models, which we will address in this appendix, under specialized assumptions. Significantly, however, such issues have been already addressed in Chapters 1-4, in the context of various specific models.

In particular, for the discounted problems of Chapters 1 and 2 [the case of the mapping (B.4) with $\alpha \in (0, 1)$ and g: bounded], underlying sup-norm contraction properties guarantee that J^* is the unique fixed point of T within the class of bounded real-valued functions over X, and that VI converges to J^* starting from within that class. This is also true for finite-state SSP problems, under the favorable assumptions of Chapter 3.

For SSP problems under the weak assumptions of Section 4.4, J^* need not be a fixed point of T. In the context of the present appendix, a useful choice is to take

$$\mathcal{C} = \big\{(\mu, x) \mid \mu\colon \text{proper}\big\},$$

in which case $J^*_{\mathcal{C}}$ is the optimal cost function that can be achieved using proper policies only. It was shown in Section 4.4 that $J^*_{\mathcal{C}}$ is a fixed point of T, so by Prop. B.2.2, VI converges to $J^*_{\mathcal{C}}$ starting from any real-valued $J \geq J^*_{\mathcal{C}}$ (cf. Prop. 4.4.2).

For nonpositive and nonnegative cost problems ($g \leq 0$ or $g \geq 0$, respectively, cf. Section 4.1), J^* is a fixed point of T, but not necessarily unique. We will discuss cases of nonnegative cost problems in Section B.4, for appropriate choices of $\mathcal{C}$, we will obtain some interesting results. The following is a nonnegative cost linear-quadratic example, where both J^* and $J^*_{\mathcal{C}}$ are fixed points of T, but $J^* \neq J^*_{\mathcal{C}}$. Moreover VI tends to converge to $J^*_{\mathcal{C}}$ rather than to J^*.

† For this statement to be meaningful, the set $\big\{\tilde{J} \in E(X) \mid J^* \leq \tilde{J}\big\}$ must be nonempty. Generally, it is possible that this set is empty, even though S is assumed nonempty.

Example B.2.1 (Linear-Quadratic Example)

Consider Example 4.2.2, which involves the scalar system $x_{k+1} = \gamma x_k + u_k$, $\gamma > 1$, and the quadratic cost $g(x,u) = u^2$. Here $X = U(x) = \Re$, and Bellman's equation has the form

$$J(x) = \min_{u \in \Re} \big\{u^2 + J(\gamma x + u)\big\}, \qquad x \in \Re.$$

The optimal cost function, $J^*(x) \equiv 0$ is a solution. Let us call *linear stable* a stationary policy $\mu(x) = \beta x$, with β such that the closed-loop system $x_{k+1} = (\gamma + \beta)x$ is stable in the sense that $|\gamma + \beta| < 1$. Let C be the set of pairs

$$\mathcal{C} = \big\{(\mu, x) \mid x \in \Re,\ \mu\text{: linear stable}\big\}.$$

For S being the set of real-valued functions J that satisfy $J(0) = 0$ and are continuous at 0,

$$S = \big\{J \in R(X) \mid J(x_k) \to 0 \text{ if } x_k \to 0\big\},$$

it can be seen that $\mathcal{C}$ is S-regular. Moreover, it can be verified that the policy $\mu(x) = \frac{(1-\gamma^2)x}{\gamma}$ is optimal within the class of linear stable policies, and we have

$$J^*_{\mathcal{C}}(x) = (\gamma^2 - 1)x^2,$$

which is also a fixed point of T, as noted in Example 4.2.2.

For this problem, VI starting with any positive definite quadratic initial condition

$$J_0(x) = P_0 x^2, \qquad P_0 > 0,$$

generates the sequence of quadratic functions $J_k(x) = P_k x^2$ according to

$$P_{k+1} = \gamma^2 \frac{P_k}{P_k + 1}, \qquad k = 0, 1, \ldots,$$

(cf. Fig. 4.1.2 in Section 4.1 of Vol. I). It can be seen that $J_k \to J^*_{\mathcal{C}}$ if $P_0 > 0$ and $J_k \to J^*$ if $P_0 = 0$. This is consistent with Props. B.2.1 and B.2.2.

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

We have seen in Section 4.1 that the results for nonnegative cost and nonpositive cost infinite horizon stochastic optimal control problems are markedly different. In particular, PI behaves better when the cost is nonnegative, while VI behaves better if the cost is nonpositive. These differences extend to the so-called *monotone increasing* and *monotone decreasing* abstract DP models, where a principal assumption is that $T_\mu \bar{J} \geq \bar{J}$ and $T_\mu \bar{J} \leq \bar{J}$ for all $\mu \in \mathcal{M}$, respectively (see [Ber13], Ch. 4).

In the context of regularity, with $\mathcal{C}$ being S-regular, it turns out that there are analogous significant differences between the cases $J_C^* \geq \bar{J}$ and $J_C^* \leq \bar{J}$. The favorable aspects of the condition $J_C^* \geq \bar{J}$ will be seen later in the context of PI, where it guarantees the monotonic improvement of the policy iterates (see the subsequent Prop. B.3.4). The following proposition establishes some favorable aspects of the condition $J_{\mathcal{C}}^* \leq \bar{J}$ in the context of VI. These can be attributed to the fact that $\bar{J}$ can always be added to S without affecting the S-regularity of $\mathcal{C}$, so $\bar{J}$ can serve as the element $\tilde{J}$ of S with $J_C^* \leq \tilde{J}$ in Props. B.2.1 and B.2.2 (see the proof of the following proposition).

Proposition B.2.3: Given a set $S \subset E(X)$, let $\mathcal{C}$ be a collection of policy-state pairs (π, x) that is S-regular, and assume that $J_{\mathcal{C}}^* \leq \bar{J}$. Then:

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

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

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

Proof: (a) If S does not contain $\bar{J}$, we can replace S with $\bar{S} = S \cup \{\bar{J}\}$, and $\mathcal{C}$ will still be $\bar{S}$-regular. By applying Prop. B.2.1(b) with S replaced by $\bar{S}$ and $\tilde{J} = \bar{J}$, the result follows.

(b) Assume without loss of generality that $\bar{J} \in S$ [cf. the proof of part (a)]. By using Prop. B.2.2 with $\tilde{J} = \bar{J}$, we have $J_{\mathcal{C}}^* = \lim_{k\to\infty} T^k\bar{J}$. This relation yields for any policy $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi$,

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

so by taking the infimum over $\pi \in \Pi$, we obtain $J_{\mathcal{C}}^* \leq J^*$. Since generically we have $J_{\mathcal{C}}^* \geq J^*$, it follows that $J_{\mathcal{C}}^* = J^*$. Finally, from Prop. B.2.2, we obtain $T^k J \to J^*$ for all $J \in E(X)$ such that $J^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$. **Q.E.D.**

As a special case of the preceding proposition, we have that if $J^* \leq \bar{J}$ and J^* is a fixed point of T, then $J^* = \lim_{k\to\infty} T^k\bar{J}$, and for every other fixed point J' of T we have $J' \leq J^*$ (apply the proposition with $\mathcal{C} = \Pi \times X$ and $S = \{\bar{J}\}$, in which case $J_{\mathcal{C}}^* = J^* \leq \bar{J}$). This special case is relevant, among others, to the monotone decreasing models, where $T_\mu\bar{J} \leq \bar{J}$ for

all $\mu \in \mathcal{M}$. A special case is the convergence of VI for nonpositive cost models [cf. Prop. 4.1.7(b)]. The proposition also applies to a classical type of search problem with both positive and negative costs per stage. This is the SSP problem, where at each $x \in X$ we have cost $E\{g(x,u,w)\} \geq 0$ for all u except one that leads to a termination state with probability 1 and nonpositive cost.

B.3 REGULAR STATIONARY POLICIES

We will now specialize the notion of S-regularity to stationary policies with the following definition, and obtain results that are useful in a variety of contexts, including PI-type of algorithms. We will also address questions of whether the optimal cost function over S-regular policies only is a fixed point of T.

Definition B.3.1: For a nonempty set of functions $S \subset E(X)$, we say that a stationary policy μ is *S-regular* if $J_\mu \in S$, $J_\mu = T_\mu J_\mu$, and $T_\mu^k J \to J_\mu$ for all $J \in S$. A policy that is not S-regular is called *S-irregular*.

Comparing this definition with Definition B.2.1, we see that μ is S-regular if the set $\mathcal{C} = \{(\mu, x) \mid x \in X\}$ is S-regular, and in addition $J_\mu \in S$ and $J_\mu = T_\mu J_\mu$. Given a set $S \subset E(X)$, let $\mathcal{M}_S$ be the set of policies that are S-regular, and let us consider optimization over the S-regular policies only. The corresponding optimal cost function is denoted J_S^*:

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

This notation is consistent with the definition of $J_{\mathcal{C}}^*$ since $J_S^* = J_{\mathcal{C}}^*$ when $\mathcal{C} = \mathcal{M}_S \times X$ and $\mathcal{M}_S$ is nonempty. We say that μ^* is *$\mathcal{M}_S$-optimal* if

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

A key issue is whether J_S^* is a fixed point of T (we will shortly provide conditions that guarantee this). The following proposition shows that if J_S^* is a fixed point of T, it can then be obtained by VI, and provides optimality conditions for a policy μ^* to be $\mathcal{M}_S$-optimal.

Proposition B.3.1: Given a set $S \subset E(X)$, assume that there exists at least one S-regular policy and that J_S^* is a fixed point of T. Then:

(a) J_S^* is the only possible fixed point of T within the set of all $J \in E(X)$ such that $J_S^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$.

(b) We have $T^k J \to J_S^*$ for every $J \in E(X)$ such that $J_S^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$.

(c) If μ^* is S-regular, $J_S^* \in S$, and $T_{\mu^*} J_S^* = T J_S^*$, then μ^* is $\mathcal{M}_S$-optimal. Conversely, if μ^* is $\mathcal{M}_S$-optimal, then $T_{\mu^*} J_S^* = T J_S^*$.

Proof: (a), (b) The definition of S-regularity and J_S^* imply that the nonempty set

$$\mathcal{C} = \mathcal{M}_S \times X$$

is S-regular, and we have

$$J_S^* = J_{\mathcal{C}}^* \geq J^*.$$

The results of parts (a) and (b) follow from Prop. B.2.2 with the above definition of $\mathcal{C}$.

(c) If μ^* is S-regular, in view of the assumptions $T_{\mu^*} J_S^* = T J_S^* = J_S^*$, we have

$$T_{\mu^*}^2 J_S^* = T_{\mu^*}(T J_S^*) = T_{\mu^*} J_S^* = T J_S^* = J_S^*,$$

where the first equality follows by applying T_{μ^*} to the equality $T_{\mu^*} J_S^* = T J_S^*$. Using this argument repeatedly, we have $J_S^* = T_{\mu^*}^k J_S^*$ for all k, so that

$$J_S^* = \lim_{k \to \infty} T_{\mu^*}^k J_S^* = J_{\mu^*},$$

where the last equality follows since μ^* is S-regular and we assume that $J_S^* \in S$. Thus μ^* is $\mathcal{M}_S$-optimal. Conversely, if μ^* is M_S-optimal, we have $J_{\mu^*} = J_S^*$, so that the assumptions imply that

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

Q.E.D.

A weakness of the preceding proposition is the assumption that J_S^* is a fixed point of T. For a specific application, this must be proved with a separate analysis. We will provide three different approaches for a proof, in the following three subsections, respectively.

(a) The first approach is inspired by problems, for which J^* is generically a fixed point of T, in which case if there is a set S such that $J_S^* = J^*$,

Prop. B.3.1 applies and shows that J^* can be obtained by the VI algorithm starting from any $J \geq J^*$. This approach can be used for the positive and negative cost models of Section 4.1, for which we have shown that J^* is a fixed point of T, but it can also be used generically for deterministic and for minimax problems as we will show shortly.

(b) The second approach is based on a perturbation argument similar to the ones used on Sections 4.4 and 4.5 for SSP and affine monotonic problems, respectively. As in these sections, the perturbation approach may be used in the context of problems where in the presence of a perturbation, irregular policies produce infinite cost from some initial state (see the development of Section B.5).

(c) The third approach is based on PI arguments, and in addition to showing that J_S^* is a fixed point of T, it provides valid PI algorithms.

B.3.1 Showing that J_S^* is a Fixed Point of T - The Deterministic and Minimax Cases

We will show that the optimal cost function J^* is a fixed point of T under some assumptions, which among others are satisfied generically in the case of deterministic problems corresponding to the mapping

$$H(x,u,J) = g(x,u) + J\big(f(x,u)\big), \qquad x \in X,\ u \in u(x),\ J \in E(X), \tag{B.6}$$

and in the case of minimax problems corresponding to the mapping

$$H(x,u,J) = \sup_{w \in W(x,u)} \Big[g(x,u,w) + J\big(f(x,u,w)\big)\Big], \qquad x \in X,\ u \in u(x),\ J \in E(X). \tag{B.7}$$

As a first step in this direction, we prove the following proposition.

Proposition B.3.2: Let $\hat{\Pi}$ be a subset of policies such that:

(1) We have

$$(\mu,\pi) \in \hat{\Pi} \qquad \text{if and only if} \qquad \mu \in \mathcal{M},\ \pi \in \hat{\Pi},$$

where for $\mu \in \mathcal{M}$ and $\pi = \{\mu_0, \mu_1, \ldots\}$, we denote by (μ,π) the policy $\{\mu, \mu_0, \mu_1, \ldots\}$.

(2) For every $\pi = \{\mu_0, \mu_1, \ldots\} \in \hat{\Pi}$, we have

$$J_\pi = T_{\mu_0} J_{\pi_1},$$

where π_1 is the policy $\pi_1 = \{\mu_1, \mu_2, \ldots\}$.

(3) We have

$$\inf_{\mu \in \mathcal{M},\, \pi \in \hat{\pi}} T_\mu J_\pi = \inf_{\mu \in \mathcal{M}} T_\mu \hat{J},$$

where the function $\hat{J}$ is given by

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

Then $\hat{J}$ is a fixed point of T. In particular, if $\hat{\Pi} = \Pi$, then J^* is a fixed point of T.

Proof: For every $x \in X$, we have

$$\hat{J}(x) = \inf_{\pi \in \hat{\Pi}} J_\pi(x) = \inf_{\mu \in \mathcal{M},\, \pi \in \hat{\Pi}} (T_\mu J_\pi)(x) = \inf_{\mu \in \mathcal{M}} (T_\mu \hat{J})(x) = (T\hat{J})(x),$$

where the second equality holds by conditions (1) and (2), and the third equality holds by condition (3). **Q.E.D.**

The assumptions of the preceding proposition can be shown to hold when $\hat{\Pi} = \Pi$ in the case of the deterministic mapping (B.6) and the minimax mapping (B.7) with $\hat{\Pi}$ being the set of all policies Π.† As a result J^*, which is equal to $\hat{J}$ when $\hat{\Pi} = \Pi$, is a fixed point of T. Moreover, if we choose a set S such that J_S^* can be shown to be equal to J^*, then Prop. B.3.1 applies and shows that J^* is the unique fixed point of T with the set $\left\{J \in E(X) \mid J_S^* \leq J \leq \tilde{J}\right\}$ for some $\tilde{J} \in S$. In addition the VI sequence $\{T^k J\}$ converges to J^* starting from every J within that set. This idea underlies the analysis of the deterministic problem of Section 4.2, where J^* is known to be a fixed point of T because the cost per stage is nonnegative and the analysis of Section 4.1 applies.

† This is evident in the case of the deterministic mapping (B.6), and it is also true for the case of the minimax mapping (B.7) because the operation of maximization over w commutes with lim sup. The assumptions of the proposition also hold for other choices of $\hat{\Pi}$. For example, when $\hat{\Pi}$ is the set of all *eventually stationary* policies, i.e., policies of the form $\{\mu_0, \ldots, \mu_k, \mu, \mu, \ldots\}$, where $\mu_0, \ldots, \mu_k, \mu \in \mathcal{M}$ and k is some positive integer.

We note, however, that for stochastic optimal control problems such as the SSP problem of Section 4.4, condition (2) of the preceding proposition need not be satisfied (because the expected value operation need not commute with lim sup), and for this reason it is possible that J^* is not a fixed point T, as illustrated by the example given in Section 4.4. We also note that the preceding proposition cannot be used with $\hat{\Pi}$ equal to a set of all stationary policies, because condition (1) would be violated in this case.

B.3.2 Showing that J_S^* is a Fixed Point of T - A Perturbation Approach

We will now discuss a perturbation approach for showing that J_S^* is a fixed point of T. This approach was used in the cases of the SSP problem of Section 4.4.1 [cf. Prop. 4.4.2(a)], and the affine monotonic problem of Section 4.5 [cf. Prop. 4.5.6(a)]. We will generalize these analyses and show that J_S^* is a fixed point of T if the problem obtained by adding a positive perturbation to H is well-behaved with respect to S-regular policies. The idea, illustrated in Section 4.4.1 for SSP problems, 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.†

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

$$(T_{\mu,\delta}J)(x) = H\big(x, \mu(x), J\big) + \delta, \qquad x \in X,$$

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

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

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

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

We refer to the problem associated with the mappings $T_{\mu,\delta}$ as the δ-*perturbed problem*. Note that by the monotonicity of H, we have $T_{\mu,\delta}J \geq T_\mu J$ for all $\delta > 0$, $\mu \in \mathcal{M}$, and $J \in S$, and hence also $J_{\pi,\delta} \geq J_\pi$ for all $\pi \in \Pi$, and $J_\delta^* \geq J^*$.

The following proposition shows that if the δ-perturbed problem is "well-behaved" with respect to the S-regular policies, then its cost function

† Here, we consider adding to H a constant perturbation $\delta > 0$. A more general approach, which may be useful in some contexts, is to add an (x,u)-dependent perturbation $\delta(x,u) \geq 0$.

J_δ^* can be used to approximate the optimal cost function J_S^* over the S-regular policies only, and moreover J_S^* is a fixed point of T.

Proposition B.3.3: Given a set $S \subset E(X)$, assume that:

(1) For every $\delta > 0$, we have $J_\delta^* = T_\delta J_\delta^*$, and there exists an S-regular policy μ_δ^* that is optimal for the δ-perturbed problem, i.e., $J_{\mu_\delta^*,\delta} = J_\delta^*$.

(2) For every S-regular policy μ, we have

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

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

Consider J_S^*, the optimal cost function over the S-regular policies only,

$$J_S^* = \inf_{\mu:\ S\text{-regular}} J_\mu.$$

(a) We have $\lim_{\delta \downarrow 0} J_\delta^* = J_S^*$.

(b) Assume in addition that H has the property that for every sequence $\{J_m\} \subset S$ with $J_m \downarrow J$, we have

$$\lim_{m \to \infty} H(x, u, J_m) \geq H(x, u, J), \qquad \forall\ x \in X,\ u \in U(x). \quad \text{(B.8)}$$

Then J_S^* is a fixed point of T and the conclusions of Prop. B.3.1 hold.

Proof: (a) For all $\delta > 0$, by using conditions (1) and (2), we have for all S-regular μ,

$$J_S^* \leq J_{\mu_\delta^*} \leq J_{\mu_\delta^*,\delta} = J_\delta^* \leq J_{\mu,\delta} \leq J_\mu + w_\mu(\delta).$$

By taking the limit as $\delta \downarrow 0$ and then the infimum over all S-regular μ, it follows that

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

(b) From condition (1), for all $\delta > 0$, we have

$$J_\delta^* = T_\delta J_\delta^* \geq T J_\delta^* = T J_{\mu_\delta^*,\delta} \geq T J_S^*,$$

and by taking the limit as $\delta \downarrow 0$ and using part (a), we obtain $J_S^* \geq T J_S^*$.

To prove the reverse inequality, let $\{\delta_m\}$ be a sequence with $\delta_m \downarrow 0$. Using condition (1), we have $T_{\delta_m} J^*_{\delta_m} = J^*_{\delta_m}$, so that for all m,

$$H(x, u, J^*_{\delta_m}) + \delta_m \geq (T_{\delta_m} J^*_{\delta_m})(x) = J^*_{\delta_m}(x), \qquad \forall\ x \in X,\ u \in U(x).$$

Taking the limit as $m \to \infty$, and using Eq. (B.8) and the fact $J^*_{\delta_m} \downarrow J^*_S$ [cf. part (a)], we have

$$H(x, u, J^*_S) \geq J^*_S(x), \qquad \forall\ x \in X,\ u \in U(x),$$

so that $TJ^*_S \geq J^*_S$. Thus J^*_S is a fixed point of T, and the assumptions of Prop. B.3.1 are satisfied. **Q.E.D.**

The preceding proposition does not require the 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, given in Section 3.2 of [Ber13]. A very similar example is the deterministic shortest path Example 4.4.1 of Chapter 4. Another example is given by the SSP problem of Example 4.4.2, where in addition J^* is not a fixed point of T.

Example B.3.1

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. B.3.1). We also have

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

consistent with Prop. B.3.3.

The perturbation line of analysis of Prop. B.3.3 has been already used in the context of the SSP problem of Section 4.4 (cf. Prop. 4.4.1), and the affine monotonic problem of Section 4.5 (cf. Prop. 4.5.5). In particular, we showed there that the optimal cost function over the S-regular policies only, J^*_S (or $\tilde{J}$ in the notation of Sections 4.4 and 4.5), is a fixed point of T, and the conclusions of Prop. B.3.1 hold (cf. Props. 4.4.2 and 4.5.6).

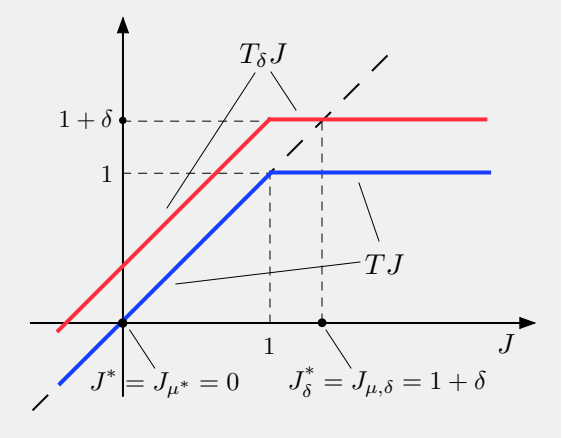

Figure B.3.1: The mapping T and its perturbed version T_δ in Example B.3.1. Here, the assumptions of Prop. B.3.3 hold, and we have $\lim_{\delta \downarrow 0} J^*_\delta = J^*_S$. However, J^* is also a fixed point of T and is not equal to J^*_S.

B.3.3 Policy Iteration and its Convergence

We will now consider the PI algorithm and its convergence properties. The idea is to generate an improving sequence of policies whose cost functions J_{μ^k} converge monotonically to some J_∞ that satisfies $J_\infty \geq J^*$ and will be shown to be a fixed point of T under simple conditions. If for some set $S \subset E(X)$, the generated policies μ^k are S-regular and their cost functions J_{μ^k} belong to S, then J_∞ is equal to J^*_S, since by Prop. B.2.2, J^*_S is the "largest" fixed point of T over the set of J such that $J^*_S \leq J \leq J_{\mu^k}$. Moreover, if we have $J^*_S = J^*$, then the PI sequence $\{J_{\mu^k}\}$ converges to J^*. This line of analysis was used for example in Section 4.2 (cf. Prop. 4.2.3).

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

$$T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k}. \tag{B.9}$$

This iteration embodies both the policy evaluation step, which computes J_{μ^k}, and the policy improvement step, which computes μ^{k+1} via the minimization over $U(x)$ for each x, which is implicit in Eq. (B.9). We will assume that this minimization can be carried out, so that the algorithm is well-defined. The evaluation of a stationary μ will ordinarily be done by solving the equation $J_\mu = T_\mu J_\mu$, which holds for most models of interest, and which we will assume in our analysis (under exceptional circumstances we may have $J_\mu \neq T_\mu J_\mu$, as shown in Section 4.4 for SSP problems under weak conditions).

We have the following proposition, the proof of which is patterned after the proofs of Props. 4.4.2 and 4.4.3 of [Ber13] that relate to PI algorithms for monotone increasing abstract DP models.

Proposition B.3.4: (Convergence of PI) Assume that:

(1) For all $\mu \in \mathcal{M}$, we have $J_\mu = T_\mu J_\mu$ and there exists $\bar{\mu} \in \mathcal{M}$ such that $T_{\bar{\mu}} J_\mu = T J_\mu$.

(2) For each sequence $\{J_m\} \subset E(X)$ with $J_m \downarrow J$ for some $J \in E(X)$, we have

$$H(x,u,J) \geq \lim_{m\to\infty} H(x,u,J_m), \quad \forall\ x \in X,\ u \in U(x). \tag{B.10}$$

Then the PI algorithm (B.9) is well defined and the following hold:

(a) If $J^* \geq \bar{J}$, then a sequence $\{\mu^k\}$ generated by the PI algorithm (B.9) satisfies $J_{\mu^k} \downarrow J_\infty$, where J_∞ is a fixed point of T with $J_\infty \geq J^*$. Moreover, if for a set $S \subset E(X)$ and some $\bar{k} \geq 0$, $\mu^{\bar{k}}$ is S-regular, then $J_{\mu^k} \downarrow J_S^*$ and J_S^* is a fixed point of T.

(b) If for a set $S \subset E(X)$ and some $\bar{k} \geq 0$, all the policies μ^k, $k \geq \bar{k}$, generated by the PI algorithm (B.9) are S-regular, then $J_{\mu^k} \downarrow J_S^*$ and J_S^* is a fixed point of T.

Proof: (a) Condition (1) of the proposition guarantees that the PI algorithm is well defined. We first show that the condition $J^* \geq \bar{J}$ implies a generic cost improvement property of PI. If μ is a policy and $\bar{\mu}$ satisfies $T_{\bar{\mu}} J_\mu = T J_\mu$, we have

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

from which, by repeatedly applying $T_{\bar{\mu}}$ to both sides, we obtain $J_\mu \geq \lim_{k\to\infty} T_{\bar{\mu}}^k J_\mu$. Since $J_\mu \geq J^* \geq \bar{J}$ and by definition $J_{\bar{\mu}} = \lim_{k\to\infty} T_{\bar{\mu}}^k \bar{J}$, it follows that

$$J_\mu \geq T J_\mu \geq J_{\bar{\mu}}. \tag{B.11}$$

Using this relation with $\mu = \mu^k$ and $\bar{\mu} = \mu^{k+1}$, we have

$$J_{\mu^k} \geq T J_{\mu^k} \geq J_{\mu^{k+1}}, \qquad k = 0, 1, \ldots,$$

so that $J_{\mu^k} \downarrow J_\infty$ for some $J_\infty \geq J^*$. By taking the limit as $k \to \infty$,

$$J_\infty \geq \lim_{k\to\infty} T J_{\mu^k} \geq T J_\infty, \tag{B.12}$$

where the second inequality follows from the fact $J_{\mu^k} \geq J_\infty$. Using Eq. (B.10), we also have for all $x \in X$ and $u \in U(x)$,

$$H(x,u,J_\infty) \geq \lim_{k\to\infty} H(x,u,J_{\mu^k}) \geq \lim_{k\to\infty} (TJ_{\mu^k})(x) = J_\infty(x).$$

By taking the infimum of the left-hand side over $u \in U(x)$, we obtain $TJ_\infty \geq J_\infty$, which combined with Eq. (B.12), yields $J_\infty = TJ_\infty$. Moreover, by the definition of S-regularity, $J_{\mu^{\bar{k}}} \in S$, so by Prop. B.2.2 with $\mathcal{C}$ equal to $\mathcal{M}_S \times X$, J_S^* (which is equal to $J_{\mathcal{C}}^*$) is the only possible fixed point of T within the set of all $J \in E(X)$ such that $J_S^* \leq J \leq J_{\mu^{\bar{k}}}$. This set includes J_∞ (since $J_S^* \leq J_{\mu^k} \leq J_{\mu^{\bar{k}}}$ for all $k \geq \bar{k}$). Hence $J_\infty = J_S^*$.

(b) By using the assumption of S-regularity of μ^k, we show again a generic cost improvement property of PI. If μ and $\bar{\mu}$ are S-regular policies, and $T_{\bar{\mu}}J_\mu = TJ_\mu$, we have

$$J_\mu = T_\mu J_\mu \geq TJ_\mu = T_{\bar{\mu}}J_\mu \geq \lim_{k\to\infty} T_{\bar{\mu}}^k J_\mu = J_{\bar{\mu}},$$

where the last inequality follows from the monotonicity of $T_{\bar{\mu}}$ and the last equality follows from the assumption that μ and $\bar{\mu}$ are S-regular. It follows similar to part (a) that $J_{\mu^k} \downarrow J_\infty$ where J_∞ is a fixed point of T. The proof from this point is identical to the one of part (a). **Q.E.D.**

The proposition shows that *PI restricted to S-regular policies will converge to J_S^* but not necessarily to J^**. Indeed this can be so, as we have seen in the deterministic shortest path Example 4.1.3 with $b > 0$ and $S = [b, \infty)$.

Condition (1) of the proposition holds for most DP models of interest, and the same is true for condition (2), which is a technical continuity-type assumption. The condition $J^* \geq \bar{J}$ in part (a) is essential for showing the cost improvement property (B.11) in the preceding proof (if cost improvement can be shown independently, the condition $J^* \geq \bar{J}$ is not needed). In Example 4.1.3, we have seen an instance of a two-state deterministic shortest path problem where this condition is violated, and the PI algorithm (B.9) oscillates between an optimal and a suboptimal policy. Note that the condition $J^* \geq \bar{J}$ does not hold for monotone decreasing models where $T_\mu \bar{J} \leq \bar{J}$ for all $\mu \in \mathcal{M}$ (unless $J^* = \bar{J}$).

Optimistic PI

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

$$T_{\mu^k}J_k = TJ_k, \qquad J_{k+1} = T_{\mu^k}^{m_k} J_k, \qquad k = 0, 1, \ldots, \tag{B.13}$$

where m_k is a positive integer for each k. For this algorithm, it turns out that the conditions for convergence are less restrictive. There is no need for the condition $J^* \geq \bar{J}$ or the S-regularity of the generated policies, as shown in the following proposition. This is due to the fact that optimistic PI embodies the characteristics of VI, which has favorable properties when $J^* \leq \bar{J}$ (see the discussion in connection with Prop. B.2.3).

Proposition B.3.5: (Convergence of Optimistic PI) Assume that:

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

(2) For each sequence $\{J_m\} \subset E(X)$ with $J_m \downarrow J$ for some $J \in E(X)$, we have

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

Then the optimistic PI algorithm (B.13) is well defined, and under the condition $J_0 \geq TJ_0$, the following hold:

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

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

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

$$\begin{aligned} J_0 \geq TJ_0 = T_{\mu^0} J_0 \geq T_{\mu^0}^{m_0} J_0 \\ = J_1 \geq T_{\mu^0}^{m_0+1} J_0 = T_{\mu^0} J_1 \geq TJ_1 = T_{\mu^1} J_1 \geq \cdots \geq J_2, \end{aligned} \tag{B.14}$$

and continuing similarly, we obtain

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

Thus $J_k \downarrow J_\infty$ for some J_∞. The proof that J_∞ is a fixed point of T is the same as in the case of the PI algorithm (B.9).

(b) In the case where all the policies μ^k are S-regular and $\{J_k\} \subset S$, from Eq. (B.14), we have $J_{k+1} \geq J_{\mu^k}$ for all k, so we have

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

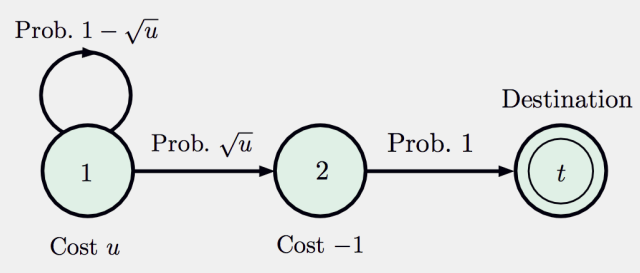

Figure B.3.2. An SSP problem with two states 1, 2, and a termination state t. Here, for $S = \Re^2$ the optimal cost function J^*_S over the S-regular policies (i.e., the proper policies) is equal to J^*, but there is no optimal policy (proper or not). Any sequence of proper policies $\{\mu^k\}$ with $\mu^k(1) \to 0$ is asymptotically optimal in the sense that $J_{\mu^k} \to J^*$, and yet $\{\mu^k\}$ converges to the strictly suboptimal improper policy for which $u = 0$ at state 1.

Using the fixed point property of J_∞ proved in part (a), and applying Prop. B.2.1(b) with

$$J' = J_\infty, \qquad \tilde{J} = J_k \geq J^*_S, \qquad \mathcal{C} = \mathcal{M}_S \times X,$$

we have $J_\infty \leq J^*_S$, which combined with the preceding relation yields $J_\infty = J^*_S$. **Q.E.D.**

The preceding two propositions can be used to ascertain convergence to J^* of the PI algorithms (B.9) and (B.13) (i.e., $J_\infty = J^*$) if J^* is known to be the only possible fixed point of T within a subset of $E(X)$ to which J_∞ can also be shown to belong. For example this is true under the assumptions of Prop. B.2.2, assuming also that $J_\infty \leq \tilde{J}$ for some $\tilde{J} \in S$. We have seen examples of such use of the proposition in Section 4.2, where we showed convergence of the PI algorithms (B.9) and (B.13), in the sense that $J_{\mu^k} \downarrow J^*$ and $J_k \downarrow J^*$, respectively, for positive cost deterministic optimal control problems.

Generally, the sequence $\{\mu^k\}$ of generated policies by PI-like algorithms need not converge to some policy, and even if it converges, the limit policy need not be optimal. This is illustrated with the following example from [BeY16], involving an SSP problem and a sequence of proper policies $\{\mu^k\}$ that satisfy $\lim_{k\to\infty} J_{\mu^k} \to J^*$, and yet $\{\mu^k\}$ converges to an improper policy that is strictly suboptimal.

Example B.3.2 (Policy Convergence - A Counterexample)

Consider an SSP problem with two states 1, 2, in addition to the termination state t; cf. Fig. B.3.2. At state 1 we must choose $u \in [0, 1]$, with expected cost equal to u. Then, we transition to state 2 with probability $\sqrt{u}$, and we self-transition to state 1 with probability $1 - \sqrt{u}$. From state 2 we transition

to t with cost -1. Thus we have

$$H(1,u,J) = u + \big(1-\sqrt{u}\big)J(1) + \sqrt{u}\,J(2), \qquad \forall\ J \in \Re^2,\ u \in [0,1],$$

$$H(2,u,J) = -1, \qquad \forall\ J \in \Re^2,\ u \in U(2).$$

Here for $S = \Re^2$, the optimal cost function J_S^* over the S-regular policies (i.e., the proper policies) is equal to J^*. There is a unique improper policy μ: it chooses $u = 0$ at state 1, and has cost $J_\mu(1) = 1$. Every policy μ with $\mu(1) \in (0,1]$ is proper, and J_μ can be obtained by solving the equation $J_\mu = T_\mu J_\mu$. We have $J_\mu(2) = -1$, so that

$$J_\mu(1) = \mu(1) + \Big(1 - \sqrt{\mu(1)}\Big)J_\mu(1) - \sqrt{\mu(1)},$$

and we obtain

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

Thus, $J^*(1) = -1$. Consider a sequence of proper policies $\{\mu^k\}$ with $\mu^k(1) \to 0$. Any such sequence satisfies $J_{\mu^k} \to J^*$, yet it converges to the strictly suboptimal improper policy.

Finally, let us note the possibility of combining PI with our earlier perturbation approach, to obtain a PI algorithm where the policy evaluation is performed on the perturbed problem. We developed such an algorithm for SSP problems in Section 4.4.2. This algorithm can be generalized nearly verbatim to the context of this appendix; see also [Ber13], Section 3.3.3.

B.4 MONOTONE INCREASING MODELS

An important type of abstract DP model is one where $\bar{J} \leq T_\mu \bar{J}$ for all $\mu \in \mathcal{M}$. In this model, the finite horizon costs $T_{\mu_0} \cdots T_{\mu_k} \bar{J}$ of any policy $\pi = \{\mu_0, \mu_1, \ldots\}$ monotonically increase to J_π. Consequently this model is known as monotone increasing, and among others, it can be used to represent problems where nonnegative costs accumulate additively over time. A major example is the nonnegative cost stochastic optimal control problem of Section 4.1. Note that if the optimal cost $J^*(x)$ at a state x is to be finite, the accumulation of nonnegative costs must be diminishing starting from x. In the absence of discounting, this must be accomplished through the presence of cost-free states, which in optimal control problems are typically desirable states that we aim to reach, perhaps asymptotically, from the remaining states. The applications of this section are of this type.

For the monotone increasing model, J^* is known to be the smallest fixed point of T within the class of functions $J \geq \bar{J}$, under certain relatively mild assumptions. An example is the positive cost model of Section 4.1 [cf. Prop. 4.1.3(a)]. However, VI may not converge to J^* starting from below J^* (e.g., starting from $\bar{J}$), and also starting from above J^*. In this section

we will address the question of convergence of VI from above J^* by using the regularity ideas of the preceding section. The starting point for the analysis is the following assumption, introduced in [Ber75], [Ber77] (see also [BeS78], Ch. 5, and [Ber13], Section 4.3).

Assumption I: (Monotone Increase)

(a) We have

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

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

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

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

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

We summarize the results that are relevant to our development in the following proposition (see [BeS78], Props. 5.2, 5.4, and 5.10, or [Ber13], Props. 4.3.3, 4.3.9, and 4.3.14). Actually for the examples of this section, we will only need the special cases of the various parts of the proposition that were proved in Section 4.1, in the context of stochastic optimal control.

Proposition B.4.1: Let Assumption I hold. Then:

(a) $J^* = TJ^*$, and if $J \in E(X)$ satisfies $J \geq TJ$, then $J \geq J^*$.

(b) For all $\mu \in \mathcal{M}$ we have $J_\mu = T_\mu J_\mu$.

(c) $\mu^* \in \mathcal{M}$ is optimal if and only if $T_{\mu^*} J^* = TJ^*$.

(d) If U is a metric space and the sets

$$U_k(x,\lambda) = \left\{u \in U(x) \,\middle|\, H(x,u,T^k\bar{J}) \leq \lambda\right\} \tag{B.15}$$

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

Note that under Assumption I there may exist fixed points J' of T with $J^* \leq J'$, while VI or PI may not converge to J^* starting from above J^*. However, convergence of VI to J^* from above, if it occurs, is often much faster than convergence from below, so starting points $J \geq J^*$ may be desirable. One well-known such case is deterministic finite-state shortest path problems where major algorithms, such as the Bellman-Ford method or other label correcting methods have polynomial complexity, when started from J above J^*, but only pseudopolynomial complexity when started from $J = 0$.

We will now use the results of the preceding section to establish conditions regarding the uniqueness of J^* as a fixed point of T, and the convergence of VI and PI for various optimal control problems. In all these problems, our analysis will proceed as follows:

(a) Define a collection $\mathcal{C}$ such that $J^*_{\mathcal{C}} = J^*$.

(b) Define a set $S \subset E^+(X)$ such that $J^* \in S$ and $\mathcal{C}$ is S-regular.

(c) Use Prop. B.2.2 (which shows that $J^*_{\mathcal{C}}$ is the largest fixed point of T within S) in conjunction with Prop. B.4.1(a) (which shows that J^* is the smallest fixed point of T within S) to show that J^* is the unique fixed point of T within S. Use also Prop. B.2.2 to show that the VI algorithm converges to J^* starting from J within the set $\{J \in S \mid J \geq J^*\}$.

(d) Use the compactness condition of Prop. B.4.1(d), to enlarge the set of functions starting from which VI converges to J^*.

Some statements regarding the validity of PI, using Props. B.3.4 and B.3.5, will also be made.

B.4.1 Deterministic Optimal Control

Let us consider the undiscounted deterministic optimal control problem of Section 4.2, where

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

with g being the one-stage cost function and f being the function defining the associated discrete-time system

$$x_{k+1} = f(x_k, u_k).$$

We allow X and U to be arbitrary sets, and we consider the case where

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

As in Eq. (B.2), the cost function J_π of a policy π is the upper limit of the finite horizon cost functions $T_{\mu_0} \cdots T_{\mu_k} \bar{J}$ of the policy, with $\bar{J}(x) \equiv 0$.

We assume that there is a nonempty set $X_0 \subset X$, which is cost-free and absorbing in the sense

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

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

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

Two other interesting subsets of X are

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

Following Section 4.2, given a state x, we say that a policy π *terminates from* x if the sequence $\{x_k\}$, which is generated starting from x and using π, reaches X_0 in the sense that $x_{\bar{k}} \in X_0$ for some index $\bar{k}$. We assumed that for every $x \in X_f$ and $\epsilon > 0$, there exists a policy π that terminates from x and satisfies $J_\pi(x) \le J^*(x) + \epsilon$.

We now introduce the set

$$\mathcal{C} = \big\{(\pi, x) \mid x \in X_f,\ \pi \text{ terminates from } x\big\},$$

and we note that under our preceding assumption, $\mathcal{C}$ is nonempty and $J^*_{\mathcal{C}} = J^*$. The reason is that for $x \in X_f$, we have

$$J^*_{\mathcal{C}}(x) = \inf_{\{\pi \mid (\pi,x) \in \mathcal{C}\}} J_\pi(x) = J^*(x),$$

while for $x \in X_\infty$ we also have $J^*_{\mathcal{C}}(x) = J^*(x) = \infty$ by the definition of $J^*_{\mathcal{C}}$ [cf. Eq. (B.5)], since for such x, the set of policies $\big\{\pi \mid (\pi, x) \in \mathcal{C}\big\}$ is empty.

We next consider the set

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

Clearly $J^* \in S$ and we also claim that $\mathcal{C}$ is S-regular. Indeed for π that terminates from x we have

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

since the choice of J within S does not affect $(T_{\mu_0} \cdots T_{\mu_k} J)(x)$ for k larger than the termination time, when the state enters X_0. Thus, since $J^*_{\mathcal{C}} = J^*$ and J^* is a fixed point of T [cf. Prop. 4.1.1 or Prop. B.4.1(a)], the theory of Section B.3 applies, and the results of that section yield the results of Section 4.2 (in fact the proofs of various results in Section 4.2 are specializations of corresponding proofs of Section B.3).

B.4.2 Positive Cost Stochastic DP

Let us consider the undiscounted stochastic optimal control problem of Section 4.1, involving the mapping

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

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

$$0 \leq g(x,u,w), \qquad \forall\ x \in X,\ u \in U(x),\ w \in W. \tag{B.16}$$

We consider the abstract DP model with H as above, and with $\bar{J}(x) \equiv 0$. We will apply the analysis of Section B.2 with

$$\mathcal{C} = \big\{(\pi, x) \mid J_\pi(x) < \infty\big\}, \tag{B.17}$$

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

Let us denote by $E^\pi_{x_0}\{\cdot\}$ the expected value with respect to the probability distribution induced by $\pi \in \Pi$ under initial state x_0, and consider the set

$$S = \big\{J \in E^+(X) \mid E^\pi_{x_0}\big\{J(x_k)\big\} \to 0,\ \forall\ (\pi, x_0) \in \mathcal{C}\big\}. \tag{B.18}$$

We will show that $J^* \in S$ and that $\mathcal{C}$ is S-regular. Once this is done, it will follow from Prop. B.2.2 and the fixed point property of J^* (cf. Prop. 4.1.1) that $T^k J \to J^*$ for all $J \in S$ that satisfy $J \geq J^*$. If the sets $U_k(x,\lambda)$ of Eq. (B.15) are compact, the convergence of VI starting from below J^* will also be guaranteed. We have the following proposition.

Proposition B.4.2: (Convergence of VI) Consider the stochastic optimal control problem of this section, assuming Eq. (B.16). Then J^* is the unique fixed point of T within S, and we have $T^k J \to J^*$ for all $J \geq J^*$ with $J \in S$. If in addition U is a metric space, and the sets $U_k(x,\lambda)$ of Eq. (B.15) are compact for all $x \in X$, $\lambda \in \Re$, and k, we have $T^k J \to J^*$ for all $J \in S$, and an optimal stationary policy is guaranteed to exist.

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

$$(T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x_0) = E^\pi_{x_0}\big\{J(x_k)\big\} + E^\pi_{x_0}\left\{\sum_{m=0}^{k-1} g\big(x_m, \mu_m(x_m), w_m\big)\right\}, \tag{B.19}$$

where μ_m, $m = 0, 1, \ldots$, denote generically the components of π. The rightmost term above converges to $J_\pi(x_0)$ as $k \to \infty$, so by taking upper limit, we obtain

$$\limsup_{k\to\infty}(T_{\mu_0}\cdots T_{\mu_{k-1}}J)(x_0) = \limsup_{k\to\infty} E^\pi_{x_0}\big\{J(x_k)\big\} + J_\pi(x_0).$$

Thus in view of the definition of S, we see that for all $(\pi, x_0) \in \mathcal{C}$ and $J \in S$, we have

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

so $\mathcal{C}$ is S-regular.

We next show that $J^* \in S$. Given a policy $\pi = \{\mu_0, \mu_1, \ldots\}$, we denote by π_k the policy

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

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

$$J_\pi(x_0) = E^\pi_{x_0}\big\{g\big(x_0, \mu_0(x_0), w_0\big)\big\} + E^\pi_{x_0}\big\{J_{\pi_1}(x_1)\big\},$$

and more generally,

$$E^\pi_{x_0}\big\{J_{\pi_m}(x_m)\big\} = E^\pi_{x_0}\big\{g\big(x_m, \mu_m(x_m), w_m\big)\big\} + E^\pi_{x_0}\big\{J_{\pi_{m+1}}(x_{m+1})\big\}, \tag{B.20}$$

for all $m = 0, 1, \ldots$, where $\{x_m\}$ is the sequence generated starting from x_0 and using π. Using the defining property $J_\pi(x_0) < \infty$ of $\mathcal{C}$, it follows that all the terms in the above relations are finite, and in particular

$$E^\pi_{x_0}\big\{J_{\pi_m}(x_m)\big\} < \infty, \qquad \forall\ (\pi, x_0) \in \mathcal{C},\ m = 0, 1, \ldots.$$

By adding Eq. (B.20) for $m = 0, \ldots, k-1$, and canceling the finite terms $E^\pi_{x_0}\big\{J_{\pi_m}(x_m)\big\}$ for $m = 1, \ldots, k-1$, we obtain for all $k = 1, 2, \ldots$,

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

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

$$E^\pi_{x_0}\big\{J_{\pi_k}(x_k)\big\} \to 0, \qquad \forall\ (\pi, x_0) \in \mathcal{C}.$$

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

$$E^\pi_{x_0}\big\{J^*(x_k)\big\} \to 0, \qquad \forall\ x_0 \text{ with } J^*(x_0) < \infty.$$

Thus $J^* \in S$, while by Prop. 4.1.1, J^* (which is equal to $J^*_{\mathcal{C}}$) is a fixed point of T. Hence, by Prop. B.2.2, J^* is the unique fixed point of T within the set $\{J \in S \mid J \geq J^*\}$. Similarly, by Prop. B.2.2, we have $T^k J \to J^*$ for all $J \in S$. The last conclusion follows from Prop. 4.1.8. **Q.E.D.**

A consequence of the preceding proposition is the following condition for VI convergence from above, first proved in [YuB13], which was noted in Section 4.1.3.

Proposition B.4.3: If a function $J \in E(X)$ satisfies

$$J^* \leq J \leq cJ^* \qquad \text{for some } c > 0, \tag{B.21}$$

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

Proof: Since $J^* \in S$ as shown in Prop. B.4.2, any J satisfying Eq. (B.21), also belongs to the set S of Eq. (B.18), and the result follows from Prop. B.4.2. **Q.E.D.**

Let us finally specialize Prop. B.4.2 to the case of a deterministic problem involving the system $x_{k+1} = f(x_k, u_k)$, the (nonnegative) cost per stage $g(x, u)$, and a set of cost-free and absorbing states X_0 (cf. Section 4.2). We assume that X is a metric space, and that for every policy π and sequence $\{x_k\}$ generated by using π we have

$$J_\pi(x_0) < \infty \qquad \Rightarrow \qquad \text{dist}(x_k, X_0) \to 0, \tag{B.22}$$

where $\text{dist}(x, X_0)$ denotes the distance from a state x to the set X_0. For example, this condition is satisfied if

$$g\big(x_k, \mu_k(x_k)\big) \to 0 \qquad \Rightarrow \qquad \text{dist}(x_k, X_0) \to 0,$$

or more specifically if for some $p > 0$,

$$g(x, u) \geq \text{dist}(x, X_0)^p, \qquad \forall\ x \in X,\ u \in U(x).$$

Let

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

[cf. Eq. (B.17)], and

$$S = \big\{J \in E^+(X) \mid J(x_k) \to 0 \text{ if } \text{dist}(x_k, X_0) \to 0\big\}.$$

Since in view of Eq. (B.22), S is equal to the set (B.18), it follows that $J^* \in S$ and that $\mathcal{C}$ is S-regular, the conclusions of Prop. B.4.2 follow. One may compare these results with the ones of Section 4.2. The two sets of results are similar: in Section 4.2 we did not assume that X is a metric space, while here we have assumed that X is a metric space in order to use the assumption (B.22), which is expressed in terms of the distance $\text{dist}(x, X_0)$.

B.4.3 Discounted Positive Cost Stochastic DP

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

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

where $\alpha \in (0,1)$ is the discount factor, and as earlier $E^\pi_{x_0}\{\cdot\}$ denotes expected value with respect to the probability measure induced by $\pi \in \Pi$ under initial state x_0. We can view this problem within the abstract DP framework by defining the mapping H as

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

[cf. Eq. (B.4)], and $\bar{J}(x) \equiv 0$. We continue to assume that the one-stage cost is nonnegative,

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

We also assume that X is a normed space with norm denoted $\|\cdot\|$. Note that because of the discount factor, the existence of a terminal set of states is not essential for the optimal costs to be finite.

We introduce the set

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

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

$$\mathcal{C} = \big\{(\pi, x) \mid x \in X_f,\ \pi \text{ is stable from } x\big\},$$

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

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

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

We assume that $\mathcal{C}$ is nonempty, $J^* \in S$, and for every $x \in X_f$ and $\epsilon > 0$, there exists a policy π that is stable from x and satisfies $J_\pi(x) \le J^*(x) + \epsilon$. Note that under this assumption, we have $J^*_{\mathcal{C}} = J^*$, similar to Section 4.2. We have the following proposition.

Proposition B.4.4: Under the preceding assumptions, J^* is the unique fixed point of T within S, and we have $T^k J \to J^*$ for all $J \in S$ with $J^* \leq J$. If in addition U is a metric space, and the sets $U_k(x,\lambda)$ of Eq. (B.15) are compact for all $x \in X$, $\lambda \in \Re$, and k, we have $T^k J \to J^*$ for all $J \in S$, and an optimal stationary policy is guaranteed to exist.

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

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

[cf. Eq. (B.19)]. The fact $(\pi, x_0) \in \mathcal{C}$ implies that there is a bounded subset of X_f such that $\{x_k\}$ belongs to that subset with probability 1, so if $J \in S$ it follows that $\alpha^k E^{\pi}_{x_0}\big\{J(x_k)\big\} \to 0$. Thus for all $(\pi, x_0) \in \mathcal{C}$ and $J \in S$,

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

so $\mathcal{C}$ is S-regular. Since $J^*_{\mathcal{C}}$ is equal to J^* which is a fixed point of T (by Prop. 4.1.1), it follows from Prop. B.2.2 that $T^k J \to J^*$ for all $J \in S$. The last conclusion follows from Prop. 4.1.8. **Q.E.D.**

Let us finally note that our assumptions are natural in control contexts where the objective is to keep the state from becoming unbounded, under the influence of random disturbances represented by w_k. In such contexts one expects that optimal or near optimal policies should produce bounded state sequences starting from states with finite optimal cost.

B.5 PROBLEMS WITH INFINITE COST IRREGULAR POLICIES

We will now consider the fixed point properties of J^*, and the convergence of VI for an abstract DP model which is neither monotone increasing nor monotone decreasing, but instead uses the assumption that follows (given as Assumption 3.2.1 in [Ber13]). Key features of this assumption are a condition implying that S-irregular policies cannot be optimal [condition (c) below], and a compactness condition on the level sets of the function $H(x,\cdot,J)$ [condition (d) below]. The assumption is modeled after the SSP conditions of Chapter 3, with $S = \Re^n$ and proper policies playing the role of $\Re^n$-regular policies. The following line of analysis applies, among others, to the SSP problems of Chapter 3, as well to the affine monotonic problems

of Section 4.5 under Assumption 4.5.3, with stable policies playing the role of $\Re_+^n$-regular policies (cf. Prop. 4.5.3).

Assumption B.5.1: 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 J_S^* given by

$$J_S^*(x) = \inf_{\mu:\, S\text{-regular}} J_\mu(x), \qquad x \in X, \tag{B.23}$$

belongs to S.

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

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

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

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

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 $J' \leq TJ'$.

Part (c) of the preceding assumption implies that for each S-irregular μ, there is at least one state such that $J_\mu(x) = \infty$. Since by part (b), $J^* \leq J_S^* \in S$, part (c) implies that an S-irregular policy cannot be optimal. Parts (e) and (f) are technical conditions that are needed for the subsequent analysis. The compactness part (d) plays a key role for asserting the existence of an optimal S-regular policy, as well as for various proof arguments. It implies that *for every $J \in S$, the infimum in the equation*

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

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 B.5.1(c) can be verified in a few interesting cases:

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

(2) Cases where for each x, $U(x)$ is compact, and H satisfies some continuity conditions guaranteeing that the set $\{u \in U \mid H(x,u,J) \leq \lambda\}$ is closed for all $x \in X$ and $J \in S$.

The following proposition, first given as Prop. 3.2.1 in [Ber13], is the main result of this section. Its proof uses the line of argument of its specialized versions, Prop. 3.2.2 (for SSP), and Prop. 4.5.3 (for affine monotonic problems), but is considerably longer and will not be given; we refer to [Ber13].

Proposition B.5.1: Let Assumption B.5.1 hold. Then:

(a) The optimal cost function J^* is the unique fixed point of T within 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 $J \geq J^*$.

Let us also give another proposition, which is useful in situations where only some of the conditions of Assumption B.5.1 are satisfied. For a proof, see [Ber13], Lemma 3.2.4.

Proposition B.5.2: Let Assumption B.5.1(b),(c),(d) hold. Then:

(a) The function J_S^* of Eq. (B.23) is the unique fixed point of T within S.

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

Note that when the number of states is finite, $X = \{1, \ldots, n\}$, Prop. B.5.1(c) shows that J^* is the unique solution of the optimization problem of maximizing $\sum_{i=1}^n \beta_i J(i)$ over the set $\{J \mid J \leq TJ\}$, where $\beta_1, \ldots, \beta_n$

are any positive scalars. Special cases of this problem, including linear programming formulations, were encountered in Sections 2.5, 3.5, 4.1, 4.4, and 4.5.

B.5.1 An Application to Robust Shortest Path Planning

We noted that the analysis of this section applies to the SSP problems of Chapter 3, as well as to the affine monotonic problems of Section 4.5. We will now discuss how it applies to minimax shortest path-type problems, following the author's paper [Ber14], to which we refer for further discussion.

To formally describe the problem, we consider a graph with a finite set of nodes $X \cup \{t\}$ and a finite set of directed arcs $\mathcal{A} \subset \big\{(x,y) \mid x, y \in X \cup \{t\}\big\}$, where t is a special node called the *destination*. At each node $x \in X$ we may choose a control u from a nonempty set $U(x)$, which is a subset of a finite set U. Then a successor node y is selected by an antagonistic opponent from a nonempty set $Y(x,u) \subset X \cup \{t\}$ and a cost $g(x,u,y)$ is incurred. The destination node t is absorbing and cost-free, in the sense that the only outgoing arc from t is (t,t), and we have $Y(t,u) = \{t\}$ and $g(t,u,t) = 0$ for all $u \in U(t)$.

As earlier, we denote the set of all policies by Π, and the finite set of all stationary policies by $\mathcal{M}$. Also, we denote the set of functions $J : X \mapsto [-\infty, \infty]$ by $E(X)$, and the set of functions $J : X \mapsto (-\infty, \infty)$ by $R(X)$. Note that since X is finite, $R(X)$ can be viewed as a finite-dimensional Euclidean space. We introduce the mapping $H : X \times U \times E(X) \mapsto [-\infty, \infty]$ given by

$$H(x,u,J) = \max_{y \in Y(x,u)} \big[g(x,u,y) + \tilde{J}(y)\big], \qquad x \in X, \tag{B.24}$$

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

$$\tilde{J}(y) = \begin{cases} J(y) & \text{if } y \in X, \\ 0 & \text{if } y = t. \end{cases} \tag{B.25}$$

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

$$(TJ)(x) = \min_{u \in U(x)} H(x,u,J), \qquad x \in X, \tag{B.26}$$

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

$$(T_\mu J)(x) = H\big(x, \mu(x), J\big), \qquad x \in X. \tag{B.27}$$

Letting $\bar{J}$ be the zero function,

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

the cost function of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ is given by the earlier Definition B.1.1, i.e.,

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

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

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

$$p = \big\{(x_0, x_1), (x_1, x_2), \ldots\big\}, \tag{B.28}$$

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

$$L_\mu(p) = \limsup_{m\to\infty} \sum_{k=0}^{m} g\big(x_k, \mu(x_k), x_{k+1}\big). \tag{B.29}$$

Using Eqs. (B.24)-(B.27), we see that for any $\mu \in \mathcal{M}$ and $x \in X$, $(T_\mu^k \bar{J})(x)$ is the result of the k-stage DP algorithm that computes $\sup_{p\in P(x,\mu)} L_p^k(\mu)$, the length of the longest path under μ that starts at x and consists of k arcs, so that

$$(T_\mu^k \bar{J})(x) = \sup_{p\in P(x,\mu)} L_p^k(\mu), \qquad x \in X,$$

For completeness, we also define the length of a portion

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

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

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

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

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

$$p = \big\{(x_0, x_1), (x_1, x_2), \ldots, (x_m, t), (t, t), \ldots\big\}, \tag{B.30}$$

where m is a positive integer, and $x_0, \ldots, x_m$ are distinct nondestination nodes. Since $g(t, u, t) = 0$ for all $u \in U(t)$, the length of a terminating path p of the form (B.30), corresponding to μ, is given by

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

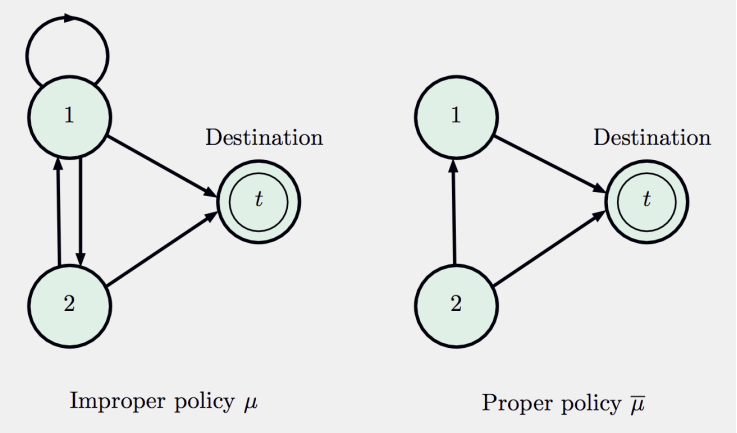

Figure B.5.1. A robust shortest path problem with $X = \{1, 2\}$, two controls at node 1, and one control at node 2. There are two policies, μ and $\overline{\mu}$, corresponding to the two controls at node 1. The figure shows the subgraphs of arcs $\mathcal{A}_\mu$ and $\mathcal{A}_{\overline{\mu}}$.

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

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

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

We will view $\mathcal{A}_\mu$ as a subgraph of the original graph. Note that $\mathcal{A}_\mu$ is defined by the set of paths $\cup_{x \in X} P(x, \mu)$, in the sense that it contains this set of paths and no other paths. We say that $\mathcal{A}_\mu$ is *destination-connected* if for each $x \in X$ there exists a terminating path in $P(x, \mu)$. We say that μ is *proper* if the subgraph of arcs $\mathcal{A}_\mu$ is acyclic (i.e., contains no cycles). Thus μ is proper if and only if all the paths in $\cup_{x \in X} P(x, \mu)$ are simple and hence terminating (equivalently μ is proper if and only if $\mathcal{A}_\mu$ is destination-connected and has no cycles). The term "proper" is consistent with the one used in Chapter 3 for SSP problems, where it indicates a policy under which the destination is reached with probability 1. If μ is not proper, it is called *improper*, in which case the subgraph of arcs $\mathcal{A}_\mu$ must contain a cycle; see the examples of Fig. B.5.1.

Clearly if μ is proper, we have $J_\mu \in R(X)$ and $J_\mu = T_\mu J_\mu$. The following proposition clarifies the properties of J_μ when μ is improper.

Proposition B.5.3: Let μ be an improper policy.

(a) If all cycles in the subgraph of arcs $\mathcal{A}_\mu$ have nonpositive length, $J_\mu(x) < \infty$ for all $x \in X$.

(b) If all cycles in the subgraph of arcs $\mathcal{A}_\mu$ have nonnegative length, $J_\mu(x) > -\infty$ for all $x \in X$.

(c) If all cycles in the subgraph of arcs $\mathcal{A}_\mu$ have zero length, J_μ is real-valued.

(d) If there is a positive length cycle in the subgraph of arcs $\mathcal{A}_\mu$, we have $J_\mu(x) = \infty$ for at least one node $x \in X$. More generally, for each $J \in R(X)$, we have $\limsup_{k\to\infty}(T_\mu^k J)(x) = \infty$ for at least one $x \in X$.

Proof: Any path with a finite number of arcs, can be decomposed into a simple path, and a finite number of cycles (see e.g., the path decomposition theorem of [Ber98], Prop. 1.1, and Exercise 1.4). Since there is only a finite number of simple paths under μ, their length is bounded above and below. Thus in part (a) the length of all paths with a finite number of arcs is bounded above, and in part (b) it is bounded below, implying that $J_\mu(x) < \infty$ for all $x \in X$ or $J_\mu(x) > -\infty$ for all $x \in X$, respectively. Part (c) follows by combining parts (a) and (b).

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

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

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

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

We will apply the regularity ideas of this section with $S = R(X)$. We recall that μ is $R(X)$-regular if $T^k J \to J_\mu$ for all $J \in R(X)$ (cf. Definition B.3.1). A key fact in our analysis is the following characterization of the notion of $R(X)$-regularity and its connection to the notion of properness.

It shows that proper policies are $R(X)$-regular, but the set of $R(X)$-regular policies may also contain some improper policies, which are characterized in terms of the sign of the lengths of their associated cycles.

Proposition B.5.4: The following are equivalent for a policy μ:

(i) μ is $R(X)$-regular.

(ii) The subgraph of arcs $\mathcal{A}_\mu$ is destination-connected and all its cycles have negative length.

(iii) μ is either proper or else it is improper, all the cycles of the subgraph of arcs $\mathcal{A}_\mu$ have negative length, and $J_\mu \in R(X)$.

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

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

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

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

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

To show that (ii) implies (iii), we assume that μ is improper and show that $J_\mu \in R(X)$. By (ii) $\mathcal{A}_\mu$ is destination-connected, so the set $P(x,\mu)$ contains a simple path for all $x \in X$. Moreover, since by (ii) the cycles of $\mathcal{A}_\mu$ have negative length, each path in $P(x,\mu)$ that is not simple has smaller length than some simple path in $P(x,\mu)$. This implies that $J_\mu(x)$ is equal to the largest path length among simple paths in $P(x,\mu)$, so $J_\mu(x)$ is a real number for all $x \in X$.

To show that (iii) implies (i), we note that if μ is proper, it is $R(X)$-regular, so we focus on the case where μ is improper. Then by (iii), $J_\mu \in$

$R(X)$, so to show $R(X)$-regularity of μ, we must show that $(T_\mu^k J)(x) \to J_\mu(x)$ for all $x \in X$ and $J \in R(X)$, and that $J_\mu = T_\mu J_\mu$. Indeed, from the definition of T_μ, we have

$$(T_\mu^k J)(x) = \sup_{p \in P(x,\mu)} \big[L_\mu^k(p) + J(x_p^k)\big], \tag{B.31}$$

where x_p^k is the node reached after k arcs along the path p, and $J(t)$ is defined to be equal to 0. Thus as $k \to \infty$, for every path p that contains an infinite number of cycles (each necessarily having negative length), the sequence $L_p^k(\mu) + J(x_p^k)$ approaches $-\infty$. It follows that for sufficiently large k, the supremum in Eq. (B.31) is attained by one of the simple paths in $P(x,\mu)$, so $x_p^k = t$ and $J(x_p^k) = 0$. Thus the limit of $(T_\mu^k J)(x)$ does not depend on J, and is equal to the limit of $(T_\mu^k \bar{J})(x)$, i.e., $J_\mu(x)$. To show that $J_\mu = T_\mu J_\mu$, we note that by the preceding argument, $J_\mu(x)$ is the length of the longest path among paths that start at x and terminate at t. Moreover, we have

$$(T_\mu J_\mu)(x) = \max_{y \in Y(x,\mu(x))} \big[g(x,\mu(x),y) + J_\mu(y)\big],$$

where we denote $J_\mu(t) = 0$. Thus $(T_\mu J_\mu)(x)$ is also the length of the longest path among paths that start at x and terminate at t, and hence it is equal to $J_\mu(x)$. **Q.E.D.**

We illustrate the preceding proposition with a two-node example involving an improper policy with a cycle that may have positive, zero, or negative length.

Example B.5.1:

Let $X = \{1\}$, and consider the policy μ where at state 1, the antagonistic opponent may force either staying at 1 or terminating, i.e., $Y\big(1,\mu(1)\big) = \{1,t\}$. Then μ is improper since its subgraph of arcs $\mathcal{A}_\mu$ contains the self-cycle $(1,1)$; cf. Fig. B.5.2. Let

$$g\big(1,\mu(1),1\big) = a, \qquad g\big(1,\mu(1),t\big) = 0.$$

Then,

$$(T_\mu J_\mu)(1) = \max\big[0,\, a + J_\mu(1)\big],$$

and

$$J_\mu(1) = \begin{cases} \infty & \text{if } a > 0, \\ 0 & \text{if } a \le 0. \end{cases}$$

Consistently with Prop. B.5.4, the following hold:

(a) For $a > 0$, the cycle $(1,1)$ has positive length, and μ is $R(X)$-irregular because $J_\mu(1) = \infty$.

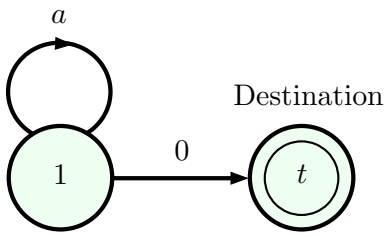

Figure B.5.2. The subgraph of arcs $\mathcal{A}_\mu$ corresponding to an improper policy μ, for the case of a single node 1 and a destination node t. The arcs lengths are shown in the figure.

(b) For $a = 0$, the cycle $(1, 1)$ has zero length, and μ is $R(X)$-irregular because for a function $J \in R(X)$ with $J(1) > 0$,

$$\limsup_{k\to\infty}(T_\mu^k J)(x) = J(1) > 0 = J_\mu(1).$$

(c) For $\alpha < 0$, the cycle $(1, 1)$ has negative length, and μ is $R(X)$-regular because $J_\mu(1) = 0$, and we have $J_\mu \in R(X)$, $J_\mu(1) = \max[0, a + J_\mu(1)] = (T_\mu J_\mu)(1)$, and for all $J \in R(X)$,

$$\limsup_{k\to\infty}(T_\mu^k J)(1) = 0 = J_\mu(1).$$

We now introduce assumptions that will allow the use of Prop. B.5.1 in order to prove our main results.

Assumption B.5.2:

(a) There exists at least one $R(X)$-regular policy.

(b) For every $R(X)$-irregular policy μ, some cycle in the subgraph of arcs $\mathcal{A}_\mu$ has positive length.

Assumption B.5.2 is implied by the weaker conditions given in the following proposition. These conditions may be more easily verifiable in some contexts.

Proposition B.5.5: Assumption B.5.2 holds if anyone of the following two conditions is satisfied.

(1) There exists at least one proper policy, and for every improper policy μ, all cycles in the subgraph of arcs $\mathcal{A}_\mu$ have positive length.

(2) Every policy μ is either proper or else it is improper and its subgraph of arcs $\mathcal{A}_\mu$ is destination-connected with all cycles having negative length, and $J_\mu \in R(X)$.

Proof: Under condition (1), by Prop. B.5.4, a policy is $R(X)$-regular, if and only if it is proper. Moreover, since each $R(X)$-irregular and hence improper policy μ has cycles with positive length, it follows that for all $J \in R(X)$, we have $\limsup_{k\to\infty}(T_\mu^k J)(x) = \infty$ for some $x \in X$. The proof under condition (2) is similar, using Prop. B.5.4. **Q.E.D.**

We now show our main result regarding the minimax shortest path problem.

Proposition B.5.6: Let Assumption B.5.2 hold. Then:

(a) The optimal cost function J^* is the unique fixed point of T within $R(X)$.

(b) A policy μ^* is optimal if and only if $T_{\mu^*}J^* = TJ^*$. Moreover, there exists an optimal proper policy.

(c) We have $T^k J \to J^*$ for all $J \in R(X)$.

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

Proof: We verify the parts (a)-(f) of Assumption B.5.1 with $S = R(X)$. The result then will follow from Prop. B.5.1. To this end we argue as follows:

(1) Part (a) is satisfied since $S = R(X)$.

(2) Part (b) is satisfied since by Assumption B.5.2(a), there exists at least one $R(X)$-regular policy. Moreover, for each $R(X)$-regular policy μ, we have $J_\mu \in R(X)$. Since the number of all policies is finite, it follows that $J_S^* \in R(X)$.

(3) To show that part (c) is satisfied, note that since by Prop. B.5.4 every $R(X)$-irregular policy μ must be improper, so by Assumption B.5.2(b), the subgraph of arcs $\mathcal{A}_\mu$ contains a cycle of positive length. By Prop. B.5.3(d), this implies that for each $J \in R(X)$, we have $\limsup_{k\to\infty}(T_\mu^k J)(x) = \infty$ for at least one $x \in X$.

(4) Part (d) is satisfied since $U(x)$ is a finite set.

(5) Part (e) is satisfied since X is finite and T_μ is a continuous function mapping the finite-dimensional space $R(X)$ into itself.

(6) To show that part (f) is satisfied, we note that by applying Prop. B.5.2 with $S = R(X)$, we have that J_S^* is the unique fixed point of T within $R(X)$. It follows that for each $J \in R(X)$, there exists a sufficiently large scalar $r > 0$ such that the function J' given by

$$J' = J_S^* - re, \qquad \forall\ x \in X,$$

where e is the unit function, $e(x) \equiv 1$, satisfies $J' \leq J$. Moreover, we have

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

where the inequality holds in view of Eqs. (B.24) and (B.26), and the fact $r > 0$.

Thus all parts of Assumption B.5.1 are satisfied, and Prop. B.5.1 applies, with $S = R(X)$. Since under Assumption B.5.2, improper policies are $R(X)$-irregular [cf. Prop. B.5.3(d)] and so cannot be optimal, the conclusions of Prop. B.5.1 are precisely the results we want to prove. **Q.E.D.**

For further analysis and algorithms for the robust shortest path planning problem, we refer to the paper [Ber14]. In particular, this paper applies the perturbation approach of Prop. B.3.3 to the case where it may be easier to guarantee nonnegativity rather than positivity of the lengths of cycles corresponding to improper policies, which is required by Assumption B.5.2. The paper also provides a Dijkstra-like algorithm for problems with nonnegative arc lengths.

Additional References

The following are new references for this appendix, which are not already included in the reference list of the printed book.

[BeY16] Bertsekas, D. P., and Yu, H., 2016. "Stochastic Shortest Path Problems Under Weak Conditions," Lab. for Information and Decision Systems Report LIDS-2909, August 2013, revised March 2015 and January 2016; to appear in Math. of OR.

[Ber98] Bertsekas, D. P., 1998. Network Optimization: Continuous and Discrete Models, Athena Scientific, Belmont, MA.

[Ber13] Bertsekas, D. P., 2013. Abstract Dynamic Programming, Athena Scientific, Belmont, MA.

[Ber14] Bertsekas, D. P., 2014. "Robust Shortest Path Planning and Semicontractive Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-P-2915, MIT; revised Jan. 2015.

[Ber15a] Bertsekas, D. P., 2015. "Regular Policies in Abstract Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-P-3173, MIT, May 2015.

[Ber15b] Bertsekas, D. P., 2015. "Value and Policy Iteration in Deterministic Optimal Control and Adaptive Dynamic Programming," to appear in IEEE Transactions on Neural Networks and Learning Systems; arXiv preprint arXiv:1507.01026.

[YuB13] Yu, H., and Bertsekas, D. P., 2013. "A Mixed Value and Policy Iteration Method for Stochastic Control with Universally Measurable Policies," Lab. for Information and Decision Systems Report LIDS-P-2905, MIT, July 2013; arXiv preprint arXiv:1308.3814, to appear in Math. of OR.