SIAM J. OPTIM.
Vol. 27, No. 3, pp. 1694–1727

REGULAR POLICIES IN ABSTRACT DYNAMIC PROGRAMMING*

DIMITRI P. BERTSEKAS†

Abstract. We consider challenging dynamic programming models where the associated Bellman equation, and the value and policy iteration algorithms commonly exhibit complex and even pathological behavior. Our analysis is based on the new notion of regular policies. These are policies that are well-behaved with respect to value and policy iteration, and are patterned after proper policies, which are central in the theory of stochastic shortest path problems. We show that the optimal cost function over regular policies may have favorable value and policy iteration properties, which the optimal cost function over all policies need not have. We accordingly develop a unifying methodology to address long standing analytical and algorithmic issues in broad classes of undiscounted models, including stochastic and minimax shortest path problems, as well as positive cost, negative cost, risk-sensitive, and multiplicative cost problems.

Key words. abstract dynamic programming, shortest path, value iteration, policy iteration, discrete-time optimal control

AMS subject classifications. 49L20, 90C39, 49J21, 90C40

DOI. 10.1137/16M1090946

1. Introduction.

The purpose of this paper is to address complicating issues that relate to the solutions of Bellman's equation, and the convergence of the value and policy iteration algorithms in total cost infinite horizon dynamic programming (DP for short). We do this in the context of abstract DP, which aims to unify the analysis of DP models and to highlight their fundamental structures.

To describe broadly our analysis, let us note two types of models. The first is the *contractive models*, introduced in [Den67], which involve an abstract DP mapping that is a contraction over the space of bounded functions over the state space. These models apply primarily in discounted infinite horizon problems of various types, with bounded cost per stage. The second is the *noncontractive models*, developed in [Ber75] and [Ber77] (see also [BeS78, Chap. 5]), for which the abstract DP mapping is not a contraction of any kind but is instead monotone. Among others, these models include shortest path problems of various types, as well as the classical nonpositive and nonnegative cost DP problems, introduced in [Bla65] and [Str66], respectively. It is well known that contractive models are analytically and computationally well-behaved, while noncontractive models exhibit significant pathologies, which interfere with their effective solution.

In this paper we focus on *semicontractive models* that were introduced in the recent monograph [Ber13]. These models are characterized by an abstract DP mapping, which for some policies has a contraction-like property, while for others it does not. A central notion in this regard is *S-regularity* of a stationary policy, where S is a set of cost functions. This property, defined formally in section 5, is related to classical notions of asymptotic stability, and it roughly means that value iteration using that policy converges to the same limit, the cost function of the policy, for every starting function in the set S.

*Received by the editors August 24, 2016; accepted for publication (in revised form) May 4, 2017; published electronically August 17, 2017.
http://www.siam.org/journals/siopt/27-3/M109094.html
†Department of Electrical Engineering and Computer Science, and the Laboratory for Information and Decision Systems, M.I.T., Cambridge, MA 02139 (dimitrib@mit.edu).

A prominent case where regularity concepts are central is finite-state problems of finding an optimal stochastic shortest path (SSP for short). These are Markovian decision problems involving a termination state, where one aims to drive the state of a Markov chain to a termination state at minimum expected cost. They have been discussed in many sources, including the books [Pal67, Der70, Whi82, Ber87, BeT89, BeT91, Put94, HeL99], and [Ber12], where they are sometimes referred to by earlier names such as "first passage problems" and "transient programming problems." Here some stationary policies called *proper* are guaranteed to terminate starting from every initial state, while others called *improper* are not. The proper policies involve a (weighted sup-norm) contraction mapping and are S-regular (with S being the set of real-valued functions over the state space), while the improper ones are not.

The notion of S-regularity of a stationary policy is patterned after the notion of a proper policy, but applies more generally in abstract DP. It was used extensively in [Ber13], and in the subsequent papers [Ber15a] and [Ber16] as a unifying analytical vehicle for a variety of total cost stochastic and minimax problems. A key idea is that the optimal cost function over S-regular policies only, call it J_S^*, is the one produced by the standard algorithms, starting from functions $J \in S$ with $J \geq J_S^*$. These are the value and policy iteration algorithms (abbreviated as VI and PI, respectively), as well as algorithms based on linear programming and related methods. By contrast, the optimal cost function over all policies J^* may not be obtainable by these algorithms, and indeed J^* may not be a solution of Bellman's equation; this can happen in particular in SSP problems with zero length cycles (see an example due to [BeY16], which also applies to multiplicative cost problems [Ber16]).

One purpose of this paper is to extend the notion of S-regularity to nonstationary policies, and to demonstrate the use of this extension for establishing convergence of VI and PI. We show that for important special cases of optimal control problems, our approach yields substantial improvements over the current state of the art, and highlights the fundamental convergence mechanism of VI and PI in semicontractive models. A second purpose of the paper is to use the insights of the nonstationary policies extension to refine the stationary regular policies analysis of [Ber13], based on PI-related properties of the set S. The paper focuses on issues of existence and uniqueness of the solution of Bellman's equation, and the convergence properties of the VI and PI algorithms, well beyond the analysis of [Ber13]. A more extensive treatment of the subject of the paper (over 100 pages), which includes elaborations of the analysis, examples, and applications, is given in unpublished internet-posted updated versions of [Ber13, Chapters 3 and 4], which may be found in the author's web site (http://web.mit.edu/dimitrib/www/abstractdp_MIT.html).

The paper is organized as follows. After formulating our abstract DP model in section 2, we develop the main ideas of the regularity approach for nonstationary policies in section 3. In section 4 we illustrate our results by applying them to nonnegative cost stochastic optimal control problems, and we discuss the convergence of VI, following the analysis of the paper [YuB15]. In sections 5–7, we specialize the notion of S-regularity to stationary policies, and we refine and streamline the analysis given in the monograph [Ber13, Chapter 3]. As an example, we establish the convergence of VI and PI under new and easily verifiable conditions in undiscounted deterministic optimal control problems with a terminal set of states. Other applications of the theory of sections 5 and 6 are given in [Ber15a] for robust (i.e., minimax) shortest path planning problems, and in [Ber16] for the class of affine monotonic models, which includes multiplicative and risk-sensitive/exponential cost models.

2. Abstract DP model. We review the abstract DP model that will be used throughout this paper (see [Ber13, section 3.1]). Let X and U be two sets, which we 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 $x \in X$.

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

Assumption 2.1 (monotonicity). If $J, J' \in \mathcal{E}(X)$ 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 $J \in \mathcal{E}(X)$ to the function $TJ \in \mathcal{E}(X)$, given by

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

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

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

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

$$\begin{aligned} &J \leq J' \implies T^k J \leq T^k J', && T_\mu^k J \leq T_\mu^k J', && \forall\ \mu \in \mathcal{M}, \\ &J \leq TJ \implies T^k J \leq T^{k+1} J, && T_\mu^k J \leq T_\mu^{k+1} J, && \forall\ \mu \in \mathcal{M}, \end{aligned}$$

which will be used repeatedly in what follows. Here 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 $J \in \mathcal{E}(X)$,

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

We next consider cost functions associated with T_μ and T. We introduce a function $\bar{J} \in \mathcal{E}(X)$, 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 (limit cannot be used since it may not exist).

DEFINITION 2.1. *Given a function* $\bar{J} \in \mathcal{E}(X)$ *for a policy* $\pi \in \Pi$ *with* $\pi = \{\mu_0, \mu_1, \ldots\}$, *we define the cost function of* π *by*

$$J_\pi(x) = \limsup_{k\to\infty} (T_{\mu_0} \cdots T_{\mu_k} \bar{J})(x) \qquad \forall\ x \in X. \tag{2.1}$$

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* $J_{\pi^*} = J^*$.

The model just described is broadly applicable, and includes as special cases nearly all the interesting types of total cost infinite horizon DP problems, including stochastic and minimax, discounted and undiscounted, semi-Markov, multiplicative, risk sensitive, etc. (see [Ber13]).[1] The following is a stochastic optimal control problem, which we will use in this paper both to obtain new results and also as a vehicle to illustrate our approach.

Example 2.1 (stochastic optimal control—Markovian decision problems). Consider an infinite horizon stochastic optimal control problem involving 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

$$x_{k+1} = f\big(x_k, \mu_k(x_k), w_k\big), \qquad k = 0, 1, \ldots, \tag{2.2}$$

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. A recent analysis that has some common elements with the present paper and addresses measure theoretic issues is given in [YuB15].

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

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

[1] However, our model cannot address those stochastic DP models where measurability issues are an important mathematical concern. In the stochastic optimal control problem of Example 2.1, we bypass these issues by assuming that the disturbance space is countable, which includes the deterministic system case, and the case where the system is stochastic with a countable state space (e.g., a countable state Markovian decision problem). Then, the expected value needed to express the finite horizon cost of a policy [cf. (2.1)] can be written as a summation over a countable index set, and is well-defined for all policies, measurable or not.

subject to the system equation constraint (2.2), where g is the one-stage cost function, and $\alpha \in (0,1]$ is the discount factor. This is a classical problem, which is discussed extensively in various sources, such as the books [BeS78, Whi82, Put94, Ber12]. Under very mild conditions guaranteeing that Fubini's theorem can be applied (see [BeS78, section 2.3.2]), it coincides with the abstract DP problem that corresponds to the mapping

$$H(x,u,J) = E\big\{g(x,u,w) + \alpha J\big(f(x,u,w)\big)\big\}, \tag{2.3}$$

and $\bar{J}(x) \equiv 0$. 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 (the value of $\bar{J}$ at the terminal state).

3. Regular policies, value iteration, and fixed points of T. Generally, in an abstract DP model, one expects to establish that J^* is a fixed point of T. This is known to be true for most DP models under reasonable conditions, and in fact it may be viewed as an indication of exceptional behavior when it does not hold. The fixed point equation $J = TJ$, in the context of standard special cases, is the classical *Bellman equation*, the centerpiece of infinite horizon DP. For some abstract DP models, J^* is the unique fixed point of T within a convenient subset of $\mathcal{E}(X)$; for example, 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. However, in general T may have multiple fixed points within $\mathcal{E}(X)$, including for some popular DP problems, while in exceptional cases, J^* may not be among the fixed points of T (see [BeY16] for a relatively simple SSP example of this type).

A related question is the convergence of VI. This is the algorithm that generates $T^k J$, $k = 0, 1, \ldots$, starting from a function $J \in \mathcal{E}(X)$. 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 from every J; this is certainly true when T has multiple fixed points. One of the purposes of this paper 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 3.1. *For a nonempty set of functions $S \subset \mathcal{E}(X)$, we say that a set $\mathcal{C}$ of policy-state pairs (π, x), with $\pi \in \Pi$ and $x \in X$, is S-regular if*

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

In what follows, when referring to a set $\mathcal{C}$ that is S-regular, we implicitly assume that $\mathcal{C}$ and S are nonempty. A set $\mathcal{C}$ of policy-state pairs (π, x) may be S-regular for many different sets S. The largest such set is

$$S_{\mathcal{C}} = \left\{ J \in \mathcal{E}(X) \;\middle|\; J_\pi(x) = \limsup_{k\to\infty} (T_{\mu_0} \cdots T_{\mu_k} J)(x)\ \forall\ (\pi,x) \in \mathcal{C} \right\},$$

and for any nonempty $S \subset S_{\mathcal{C}}$, we have that $\mathcal{C}$ is S-regular. Moreover, the set $S_{\mathcal{C}}$ is nonempty, since it contains $\bar{J}$. For a given $\mathcal{C}$, consider the function $J^*_{\mathcal{C}} \in \mathcal{E}(X)$, given by

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

Note that $J^*_{\mathcal{C}}(x) \geq J^*(x)$ for all $x \in X$ [for those $x \in X$ for which the set of policies $\{\pi \mid (\pi,x) \in \mathcal{C}\}$ is empty, we have $J^*_{\mathcal{C}}(x) = \infty$]. 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 3.1. *Given a set $S \subset \mathcal{E}(X)$, let $\mathcal{C}$ be an S-regular set.*
(a) *For all $J \in S$, we have*

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

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

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

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

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

By letting $k \to \infty$ and by using the definition of S-regularity, it follows that

$$\begin{aligned}\liminf_{k\to\infty} \left(T^k J\right)(x) &\le \limsup_{k\to\infty} \left(T^k J\right)(x)\\ &\le \limsup_{k\to\infty} (T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x) = J_\pi(x) \qquad \forall\ (\pi, x) \in \mathcal{C},\ J \in S,\end{aligned}$$

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

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

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

Letting $k \to \infty$ and using part (a), we obtain the result. ∎

Part (b) of the proposition shows that given a set $S \subset \mathcal{E}(X)$, a set $\mathcal{C} \subset \Pi \times X$ that is S-regular, and a function $J' \in \mathcal{E}(X)$ with $J' \le TJ' \le J^*_{\mathcal{C}}$, the convergence of VI is characterized by the *valid start region*

$$\left\{J \in \mathcal{E}(X) \mid J' \le J \le \tilde{J} \text{ for some } \tilde{J} \in S\right\},$$

and the *limit region*

$$\left\{J \in \mathcal{E}(X) \mid J' \le J \le J^*_{\mathcal{C}}\right\}.$$

The VI algorithm, starting from the former, ends up asymptotically within the latter; cf. Figure 1. Note that both of these regions depend on $\mathcal{C}$ and J'.

The significance of the preceding property depends of course on the choice of $\mathcal{C}$ and S. With an appropriate choice, however, there are important implications regarding the location of the fixed points of T and the convergence of VI from a broad range of starting points. Some of these implications are the following:

(a) $J^*_{\mathcal{C}}$ is an upper bound to every fixed point J' of T that lies below some $\tilde{J} \in S$ (i.e., $J' \le \tilde{J}$).
(b) If $J^*_{\mathcal{C}}$ is a fixed point of T (an important case for our subsequent development), then VI converges to $J^*_{\mathcal{C}}$ starting from any $J \in \mathcal{E}(X)$ such that $J^*_{\mathcal{C}} \le J \le \tilde{J}$ for some $\tilde{J} \in S$. For future reference, we state this result as a proposition.

PROPOSITION 3.2. *Given a set $S \subset \mathcal{E}(X)$, let $\mathcal{C}$ be an S-regular set 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 $J \in \mathcal{E}(X)$ such that $J^*_{\mathcal{C}} \le J \le \tilde{J}$ for some $\tilde{J} \in S$. Moreover, $T^k J \to J^*_{\mathcal{C}}$ for all $J \in \mathcal{E}(X)$ such that $J^*_{\mathcal{C}} \le J \le \tilde{J}$ for some $\tilde{J} \in S$.*

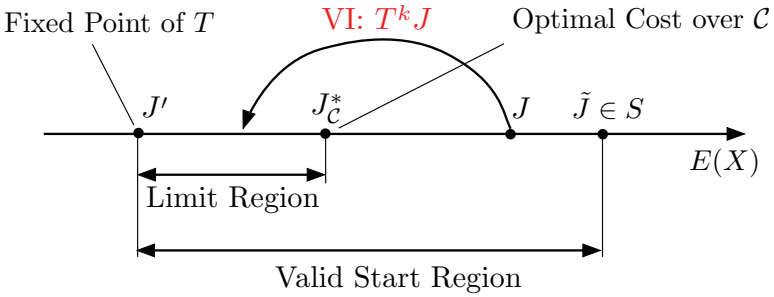

FIG. 1. *Illustration of Proposition* 3.1. *Neither $J^*_{\mathcal{C}}$ nor J^* need to be fixed points of T, but if $\mathcal{C}$ is S-regular, and there exists $\tilde{J} \in S$ with $J^*_{\mathcal{C}} \leq \tilde{J}$, then $J^*_{\mathcal{C}}$ demarcates from above the range of fixed points of T that lie below $\tilde{J}$.*

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 Proposition 3.1(b), with $J' = J^*_{\mathcal{C}}$, it follows that $T^k \tilde{J} \to J^*_{\mathcal{C}}$, so taking the limit in the above relation as $k \to \infty$, we obtain $T^k J \to J^*_{\mathcal{C}}$. □

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 $J \in \mathcal{E}(X)$ such that $J^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$.[2] In the particular applications to be discussed in section 4 we will use such a choice.

The following example illustrates the preceding propositions in the context of a central problem in optimal control. For simplicity we consider a one-dimensional special case, but the example can be generalized to any finite-dimensional linear-quadratic (positive semidefinite) problem (see Example 6.1 in section 6).

Example 3.1 (linear-quadratic optimal control problem). Consider the mapping

$$H(x, u, J) = u^2 + J(\gamma x + u),$$

where x and u are scalars, $\gamma > 1$, and $J : \Re \mapsto \Re$ is a scalar function. This corresponds to the optimal control problem involving the scalar system $x_{k+1} = \gamma x_k + u_k$ and the quadratic cost $\sum_{k=0}^{\infty} u_k^2$. A special feature of this example is that there is no penalty on the state, so the standard observability assumption is not satisfied.

Let S be the set of nonnegative quadratic functions $J(x) = Px^2$ with $P \geq 0$. Let $\mathcal{C}$ be the set of pairs (π, x), where $x \in \Re$ and π is a *linear stable* policy, i.e., a stationary policy $\pi = \{\mu, \mu, \ldots\}$ with μ linear, of the form $\mu(x) = rx$, $r \in \Re$, such that the closed-loop system $x_{k+1} = (\gamma + r)x_k$ is stable, i.e., $|\gamma + r| < 1$. For such a policy, the generated sequence of states is $x_k = (\gamma + r)^k x_0$, and we have for every $J \in \mathcal{C}$ with $J(x) = Px^2$,

$$\left(T^k_\mu J\right)(x_0) = P(x_k)^2 + \sum_{\ell=0}^{k-1} (rx_\ell)^2 = P(\gamma + r)^{2k} x_0^2 + \sum_{\ell=0}^{k-1} r^2 (\gamma + r)^{2\ell} x_0^2.$$

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

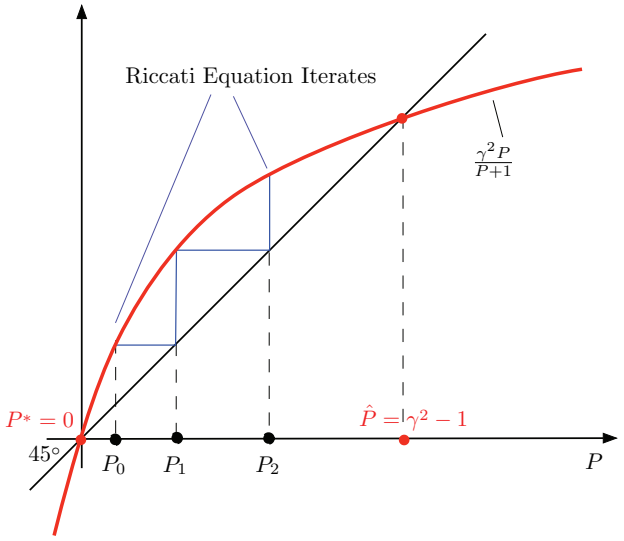

FIG. 2. *Illustration of the Riccati equation for the one-dimensional linear-quadratic problem of Example* 3.1, *where* $q = 0$. *The solutions are* $P^* = 0$ *(corresponds to the optimal cost function* J^**) and* $\hat{P} = \gamma^2 - 1$ *(corresponds to the optimal cost function* $J^*_{\mathcal{C}}$ *that can be achieved with linear stable control laws).*

Since $\lim_{k\to\infty} P(\gamma + r)^{2k} x_0^2 = 0$, it follows that

$$\lim_{k\to\infty} \left(T_\mu^k J\right)(x_0) = J_\mu(x_0) = \sum_{\ell=0}^{\infty} r^2(\gamma + r)^{2\ell} x_0^2 = \frac{r^2}{1 - (\gamma + r)^2} x^2, \tag{3.1}$$

so $\lim_{k\to\infty}(T_\mu^k J)(x_0)$ does not depend on J. Thus $\mathcal{C}$ is S-regular.

Let us consider the Bellman equation $J = TJ$ restricted to quadratic functions of the form $J(x) = Px^2$, $P \geq 0$. It takes the form $Px^2 = \min_{u\in\Re}[u^2 + P(\gamma x + u)^2]$, which after performing the minimization yields

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

This is the well-known algebraic Riccati equation for the problem (see, e.g., [AnM79, Ber17a]). This equation has two nonnegative solutions as shown in Figure 2: $P^* = 0$ and $\hat{P} = \gamma^2 - 1$. The solution $P^* = 0$ corresponds to the optimal cost function, which is $J^*(x) \equiv 0$ with optimal policy $\mu^*(x) \equiv 0$. The other solution corresponds to $\hat{J}(x) = \hat{P}x^2$, which can be verified to be the restricted optimal cost function $J^*_{\mathcal{C}}$. To see this, note that for a linear stable policy $\mu(x) = rx$, the corresponding cost function is quadratic of the form (3.1). By setting to 0 the derivative of the expression $\frac{r^2}{1-(\gamma+r)^2}$ [cf. (3.1)], we can verify that the optimal value of r is $\hat{r} = \frac{1-\gamma^2}{\gamma}$, and that the corresponding cost function is $(\gamma^2 - 1)x^2 = \hat{J}(x)$. Thus the fixed point $\hat{J}$ is equal to the optimal cost function $J^*_{\mathcal{C}}$ that can be achieved with linear stable policies.

Another interesting fact is that the VI method generates the sequence $J_k(x) = P_k x^2$, where $P_{k+1} = \frac{\gamma^2 P_k}{P_k + 1}$, and converges to the second solution $J^*_{\mathcal{C}}$ when started with any $P_0 > 0$ (cf. Figure 2). This is consistent with Propositions 3.1 and 3.2: $J^*_{\mathcal{C}}$ is the largest fixed point of T and is the limit of VI starting from any real-valued J_0 with $J_0 \geq J^*_{\mathcal{C}}$. Note that in this example, J^* is a fixed point of T, but VI does not converge

to J^*, except when started at J^*. It can also be verified that the PI method, when started with a linear stable policy also converges to $J^*_{\mathcal{C}}$, and not to the optimal cost function J^*.

Propostion 3.2 does not say anything about fixed points of T that lie below $J^*_{\mathcal{C}}$. In particular, it does not address the question of whether J^* is a fixed point of T, or whether VI converges to J^* starting from $\bar{J}$ or from below J^*; these are major questions in abstract DP models, which are typically handled by special analytical techniques that are tailored to the particular model's structure and assumptions. Significantly, however, these questions have been already answered in the context of various models, and when available, they can be used to supplement the preceding propositions. For example, the DP books [Pal67, Der70, Whi82, Put94, HeL99, Ber12, Ber13] provide extensive analysis for the most common infinite horizon stochastic optimal control problems: discounted, SSP, nonpositive cost, and nonnegative cost problems.

In particular, for discounted problems [the case of the mapping (2.3) with $\alpha \in (0, 1)$ and g being a bounded function], 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, involving a cost-free termination state, under some favorable conditions (there must exist a proper policy, i.e., a stationary policy that leads to the termination state with probability 1, improper policies must have infinite cost for some states, and some finiteness or compactness conditions on the control space U must be satisfied; see [BeT91, Ber12]).

The paper [BeY16] also considers finite-state SSP problems, but under the weaker assumptions that there exists at least one proper policy, that J^* is real valued, and U satisfies some finiteness or compactness conditions. Under these assumptions, J^* need not be a fixed point of T, as shown in [BeY16] with an example. In the context of the present paper, a useful choice is to take $\mathcal{C} = \{(\mu, x) \mid \mu\colon \text{proper}\}$, in which case $J^*_{\mathcal{C}}$ is the optimal cost function that can be achieved using proper policies only. It was shown in [BeY16] that $J^*_{\mathcal{C}}$ is a fixed point of T, so by Proposition 3.2, VI converges to $J^*_{\mathcal{C}}$ starting from any real valued $J \geq J^*_{\mathcal{C}}$.

For nonpositive and nonnegative cost problems (cf. Example 2.1 with $g \leq 0$ or $g \geq 0$, respectively), J^* is a fixed point of T, but not necessarily unique. However, for nonnegative cost problems, some new results on the existence of fixed points of T and convergence of VI were recently proved in [YuB15]. It turns out that one may prove these results by using Proposition 3.2, with an appropriate choice of $\mathcal{C}$. The proof uses the arguments of [YuB15, Appendix E], and will be given in section 4.1.

A class of DP problems with more complicated structure is the general convergence model discussed in the thesis [Van81] and the survey paper [Fei02]. This is the case of Example 2.1 where the cost per stage g can take both positive and negative values, under some restrictions that guarantee that J_π is defined by (2.1) as a limit. The paper [Yu15] describes the complex issues of convergence of VI for these models, and in an infinite space setting that addresses measurability issues. We note that there are examples of general convergence models where X and U are finite sets, but VI does not converge to J^* starting from $\bar{J}$ (see [Van81, Example 3.2], [Fei02, Example 6.10], and [Yu15, Example 4.1]). The analysis of [Yu15] may also be used to bring to bear Proposition 3.1 on the problem, but this analysis is beyond our scope in this paper.

The case where $J^*_{\mathcal{C}} \leq \bar{J}$. It is well known that the results for nonnegative cost and nonpositive cost infinite horizon stochastic optimal control problems are

markedly different. In particular, roughly speaking, 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, Chap. 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 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 Propositions 3.1 and 3.2 (see the proof of the following proposition).

PROPOSITION 3.3. *Given a set $S \subset \mathcal{E}(X)$, let $\mathcal{C}$ be an S-regular set and assume that $J^*_{\mathcal{C}} \leq \bar{J}$. Then*

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

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

(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 $J \in \mathcal{E}(X)$ such that $J^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$.*

Proof. (a) If S does not contain $\bar{J}$, we can replace S with $\bar{S} = S \cup \{\bar{J}\}$, and $\mathcal{C}$ will still be $\bar{S}$-regular. By applying Proposition 3.1(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 Proposition 3.2 with $\tilde{J} = \bar{J}$, we have $J^*_{\mathcal{C}} = \lim_{k\to\infty} T^k \bar{J}$. This relation yields for any policy $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi$,

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

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 Proposition 3.2, we obtain $T^k J \to J^*$ for all $J \in \mathcal{E}(X)$ such that $J^* \leq J \leq \tilde{J}$ for some $\tilde{J} \in S$. ∎

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 (see [Ber13, section 4.3]), where $T_\mu \bar{J} \leq \bar{J}$ for all $\mu \in \mathcal{M}$, in which case it is known that J^* is a fixed point of T under mild conditions. We then obtain a classical result on the convergence of VI for nonpositive cost models. The proposition also applies to a classical type of search problem with both positive and negative costs per stage. This is Example 2.1, where at each $x \in X$ we have $E\{g(x,u,w)\} \geq 0$ for all u except one that leads to a termination state with probability 1 and nonpositive cost. Note that without the assumption $J^*_{\mathcal{C}} \leq \bar{J}$ in the preceding proposition, it is possible that $T^k \bar{J}$ does not converge to J^*, even if $J^*_{\mathcal{C}} = J^* = TJ^*$, as is well known in the theory of nonnegative cost infinite horizon stochastic optimal control.

Generally, it is important to choose properly the set $\mathcal{C}$ in order to obtain meaningful results. Note, however, that in a given problem the interesting choices of $\mathcal{C}$ are usually limited, and that the propositions of this section can guide a favorable choice. One useful approach is to try the set

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

so that $J^*_{\mathcal{C}} = J^*$. By the definition of regularity, if S is any subset of the set

$$S_{\mathcal{C}} = \left\{ J \in \mathcal{E}(X) \;\middle|\; J_\pi(x) = \limsup_{k\to\infty} (T_{\mu_0} \cdots T_{\mu_k} J)(x) \;\forall\; (\pi, x) \in \mathcal{C} \right\},$$

then $\mathcal{C}$ is S-regular. One may then try to derive a suitable subset of $S_{\mathcal{C}}$ that admits an interesting characterization. This is the approach followed in the applications of the next section.

4. Applications in stochastic optimal control. In this section, we will consider the stochastic optimal control problem of Example 2.1, where

$$H(x,u,J) = E\big\{ g(x,u,w) + \alpha J\big(f(x,u,w)\big)\big\}, \tag{4.1}$$

and $\bar{J}(x) \equiv 0$. Here $\alpha \in (0,1]$ is the discount factor and we assume that the expected cost per stage is nonnegative:

$$0 \le E\big\{g(x,u,w)\big\} < \infty \qquad \forall\; x \in X,\; u \in U(x). \tag{4.2}$$

This is a classical problem, also known as the negative DP model [Str66].

We will use some known results for this problem, which we collect in the following proposition (for proofs, see, e.g., [BeS78, Propositions 5.2, 5.4, and 5.10], or [Ber13, Propositions 4.3.3, 4.3.9, and 4.3.14]).

PROPOSITION 4.1. *Consider the stochastic optimal control problem, where H is given by* (4.1)*, g satisfies the nonnegativity condition* (4.2)*, and $\alpha \in (0,1]$. Then*

(a) *$J^* = TJ^*$ and if $J \in \mathcal{E}^+(X)$ satisfies $J \ge TJ$, then $J \ge J^*$;*
(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*

$$U_k(x,\lambda) = \big\{ u \in U(x) \mid H(x,u,T^k\bar{J}) \le \lambda \big\} \tag{4.3}$$

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 $J \in \mathcal{E}^+(X)$ with $J \le J^*$.

Note that there may exist fixed points J' of T with $J' \ge J^*$, 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 $J \ge J^*$ 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 other initial conditions.

We will now establish conditions for the uniqueness of J^* as a fixed point of T, and the convergence of VI and PI. We will consider separately the cases $\alpha = 1$ and $\alpha < 1$. Our analysis will proceed as follows:

(a) Define a set $\mathcal{C}$ such that $J^*_{\mathcal{C}} = J^*$.
(b) Define a set $S \subset \mathcal{E}^+(X)$ such that $J^* \in S$ and $\mathcal{C}$ is S-regular.
(c) Use Proposition 3.2 in conjunction with the fixed point properties of J^* [cf. Proposition 4.1(a)] to show that J^* is the unique fixed point of T within S, and that the VI algorithm converges to J^* starting from J within the set $\{J \in S \mid J \ge J^*\}$.
(d) Use the compactness condition of Proposition 4.1(d), to enlarge the set of functions starting from which VI converges to J^*.

4.1. Nonnegative undiscounted cost stochastic DP. Assume that the problem is undiscounted, i.e., $\alpha = 1$. Consider the set

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

for which we have $J^*_{\mathcal{C}} = J^*$, and assume that $\mathcal{C}$ is nonempty.

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

$$S = \big\{J \in \mathcal{E}^+(X) \mid E^\pi_{x_0}\big\{J(x_k)\big\} \to 0 \ \forall\ (\pi, x_0) \in \mathcal{C}\big\}. \tag{4.4}$$

We will show that $J^* \in S$ and that $\mathcal{C}$ is S-regular. Once this is done, it will follow from Proposition 3.2 and the fixed point property of J^* [cf. Proposition 4.1(a)] that $T^k J \to J^*$ for all $J \in S$ that satisfy $J \geq J^*$. If the sets $U_k(x, \lambda)$ of (4.3) are compact, the convergence of VI starting from below J^* will also be guaranteed. We have the following proposition. The proof uses the line of argument of [YuB15, Appendix E].

PROPOSITION 4.2 (convergence of VI). *Consider the stochastic optimal control problem of this section, assuming $\alpha = 1$ and the cost nonnegativity condition* (4.2). *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* (4.3) *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 $J \in \mathcal{E}(X)$, $(\pi, x_0) \in \mathcal{C}$, and k,

$$(T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x_0) = E^\pi_{x_0}\big\{J(x_k)\big\} + E^\pi_{x_0}\left\{\sum_{t=0}^{k-1} g\big(x_t, \mu_t(x_t), w_t\big)\right\}, \tag{4.5}$$

where μ_t, $t = 0, 1, \ldots$, denote generically the components of π. By the cost nonnegativity condition (4.2), the rightmost term above converges to $J_\pi(x_0)$ as $k \to \infty$, so by taking the 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 (4.4) 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 $J^* \in S$. We have for all $(\pi, x_0) \in \mathcal{C}$

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

and more generally,

$$E^\pi_{x_0}\big\{J_\pi(x_t)\big\} = E^\pi_{x_0}\big\{g\big(x_t, \mu_t(x_t), w_t\big)\big\} + E^\pi_{x_0}\big\{J_\pi(x_{t+1})\big\} \qquad \forall\ t = 0, 1, \ldots, \tag{4.6}$$

where $\{x_t\}$ 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

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

By adding (4.6) for $t = 0, \ldots, k-1$, and canceling the finite terms $E^{\pi}_{x_0}\{J_\pi(x_t)\}$ for $t = 1, \ldots, k-1$,

$$J_\pi(x_0) = E^{\pi}_{x_0}\big\{J_\pi(x_k)\big\} + \sum_{t=0}^{k-1} E^{\pi}_{x_0}\big\{g\big(x_t, \mu_t(x_t), w_t\big)\big\} \qquad \forall\ (\pi, x_0) \in \mathcal{C},\ k = 1, 2, \ldots.$$

The rightmost term above tends to $J_\pi(x_0)$ as $k \to \infty$, so we obtain $E^{\pi}_{x_0}\{J_\pi(x_k)\} \to 0$ for all $(\pi, x_0) \in \mathcal{C}$. Since $0 \le J^* \le J_\pi$ for all π, 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 $J^* \in S$.

From Proposition 3.2 it follows that J^* is the unique fixed point of T within $\{J \in S \mid J \ge J^*\}$. On the other hand, every fixed point $J \in \mathcal{E}^+(X)$ of T satisfies $J \ge J^*$ by Proposition 4.1(a), so J^* is the unique fixed point of T within S. Also from Proposition 3.2 we have that the VI sequence $\{T^k J\}$ converges to J^* starting from any $J \in S$ with $J \ge J^*$. Finally, for any $J \in S$, let us select $\tilde{J} \in S$ with $\tilde{J} \ge J^*$ and $\tilde{J} \ge J$, and note that by the monotonicity of T, we have $T^k \bar{J} \le T^k J \le T^k \tilde{J}$. If we also assume compactness of the sets $U_k(x, \lambda)$ of (4.3), then by Proposition 4.1(d), we have $T^k \bar{J} \to J^*$, which together with the convergence $T^k \tilde{J} \to J^*$ just proved, implies that $T^k J \to J^*$. □

A consequence of the preceding proposition is an interesting condition for VI convergence from above, which was first proved in [YuB15]. In particular, since $J^* \in S$, any J satisfying $J^* \le J \le cJ^*$ for some $c > 0$ belongs to S, so we have the following.

PROPOSITION 4.3 (see [YuB15]). *We have $T^k J \to J^*$ for all $J \in \mathcal{E}(X)$ satisfying $J^* \le J \le cJ^*$ for some $c > 0$.*

The preceding proposition highlights a requirement for the reliable implementation of VI: it is important to know the sets $X_s = \{x \in X \mid J^*(x) = 0\}$ and $X_\infty = \{x \in X \mid J^*(x) = \infty\}$ in order to obtain a suitable initial condition $J \in \mathcal{E}(X)$ satisfying $J^* \le J \le cJ^*$ for some $c > 0$. For finite state and control problems, the set X_s can be computed in polynomial time as shown in the paper [BeY16], which also provides a method for dealing with cases where X_∞ is nonempty, based on adding a high cost artificial control at each state.

Regarding PI, we note that the analysis of section 5.2 will guarantee its convergence for the stochastic problem of this section if somehow it can be shown that J^* is the unique fixed point of T within a subset of $\{J \mid J \ge J^*\}$ that contains the limit J_∞ of PI. This result was given as [YuB15, Corollary 5.2]. Alternatively, there is a mixed VI and PI algorithm proposed in [YuB15], which can be applied under the condition of Proposition 4.3, and applies to a more general problem where w can take an uncountable number of values and measurability issues are an important concern.

Finally, we note that in this section we do not consider any special structure, other than the expected cost nonnegativity condition (4.2). In particular, we do not discuss the implications of the possible existence of a termination state as in finite-state or countable-state SSP problems. The approach of this paper is relevant to the convergence analysis of VI and PI for such problems, and for a corresponding analysis for finite-state problems, we refer to the paper [BeY16].

4.2. Discounted nonnegative cost stochastic DP. We will now consider the case where $\alpha < 1$. The cost function of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ has the form

$$J_\pi(x_0) = \lim_{k\to\infty} E^\pi_{x_0}\left\{\sum_{t=0}^{k-1} \alpha^t g\big(x_t, \mu_t(x_t), w_t\big)\right\},$$

where 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 will assume that X is a normed space.

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

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

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

Let us say that a function $J \in \mathcal{E}^+(X)$ 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

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

We will assume that $J^* \in S$. In practical settings we may be able to guarantee this by finding a stationary policy μ such that the function J_μ is bounded on bounded subsets of X_f. We also assume the following.

Assumption 4.1. In the discounted stochastic optimal control problem of this section, $\mathcal{C}$ 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 Assumption 4.1 is 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. Clearly under this assumption, $J^*_{\mathcal{C}} = J^*$. We have the following proposition.

PROPOSITION 4.4. *Let Assumption* 4.1 *hold. Then J^* is the unique fixed point of T within S, and we have $T^k J \to J^*$ for all $J \in S$ with $J^* \le J$. If in addition U is a metric space, and the sets $U_k(x, \lambda)$ of* (4.3) *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. Using the notation of section 4.1, we have for all $J \in \mathcal{E}(X)$, $(\pi, x_0) \in \mathcal{C}$, and k,

$$(T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x_0) = \alpha^k E^\pi_{x_0}\big\{J(x_k)\big\} + E^\pi_{x_0}\left\{\sum_{t=0}^{k-1} \alpha^t g\big(x_t, \mu_t(x_t), w_t\big)\right\}$$

[cf. (4.5)]. 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

$\alpha^k E^\pi_{x_0}\{J(x_k)\} \to 0$. Thus for all $(\pi, x_0) \in \mathcal{C}$ and $J \in S$, we have

$$\lim_{k\to\infty} (T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x_0) = \lim_{k\to\infty} E^\pi_{x_0} \left\{ \sum_{t=0}^{k-1} \alpha^t g\big(x_t, \mu_t(x_t), w_t\big) \right\} = J_\pi(x_0),$$

so $\mathcal{C}$ is S-regular. Since $J^*_{\mathcal{C}}$ is equal to J^* which is a fixed point of T [by Proposition 3.1(a)], it follows that $T^k J \to J^*$ for all $J \in S$. Under the compactness assumption on the sets $U_k(x, \lambda)$, the result follows by using Proposition 4.1(d). ∎

5. *S*-regular stationary policies. We will now specialize the notion of S-regularity to stationary policies with the following definition from [Ber13].

DEFINITION 5.1. *For a nonempty set of functions $S \subset \mathcal{E}(X)$, we say that a stationary policy μ is* S-regular *if $J_\mu \in S$, $J_\mu = T_\mu J_\mu$, and $T^k_\mu J \to J_\mu$ for all $J \in S$. A policy that is not S-regular is called* S-irregular.

Comparing this definition with Definition 3.1, 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 $J_\mu = T_\mu J_\mu$. Thus a policy μ is S-regular if the VI algorithm corresponding to μ, $J_{k+1} = T_\mu J_k$, represents a dynamic system that has J_μ as its unique equilibrium within S, and is asymptotically stable in the sense that the iteration converges to J_μ, starting from any $J \in S$.

5.1. Restricted optimization over *S*-regular policies. Given a nonempty set $S \subset \mathcal{E}(X)$, let $\mathcal{M}_S$ be the set of policies that are S-regular, and consider optimization over the S-regular policies only. The corresponding optimal cost function is denoted J^*_S:

$$J^*_S(x) = \inf_{\mu\in\mathcal{M}_S} J_\mu(x) \qquad \forall\ x \in X. \tag{5.1}$$

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

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

A technical point here is that while S is assumed nonempty, it is possible that $\mathcal{M}_S$ is empty. In this case our results will not be useful, but J^*_S is still defined by (5.1) as $J^*_S(x) \equiv \infty$. This is convenient in various proof arguments.

An important question is whether J^*_S is a fixed point of T and can be obtained by the VI algorithm. The following proposition, essentially a specialization of Proposition 3.2, shows that if J^*_S is a fixed point of T, then it can be obtained by VI, when started within the set

$$\mathcal{W}_S = \Big\{ J \in \mathcal{E}(X) \mid J^*_S \le J \le \tilde{J} \text{ for some } \tilde{J} \in S \Big\}, \tag{5.2}$$

which we refer to as the *well-behaved region*. The proposition also provides a necessary and sufficient condition for an S-regular policy μ^* to be $\mathcal{M}_S$-optimal.

PROPOSITION 5.1. *Given a set $S \subset \mathcal{E}(X)$, assume that J^*_S is a fixed point of T. Then*

(a) (uniqueness of fixed point) *J^*_S is the unique fixed point of T within $\mathcal{W}_S$;*
(b) (VI convergence) *we have $T^k J \to J^*_S$ for every $J \in \mathcal{W}_S$;*
(c) (optimality condition) *if μ^* is S-regular, $J^*_S \in S$, and $T_{\mu^*} J^*_S = T J^*_S$, then μ^* is $\mathcal{M}_S$-optimal. Conversely, if μ^* is $\mathcal{M}_S$-optimal, then $T_{\mu^*} J^*_S = T J^*_S$.*

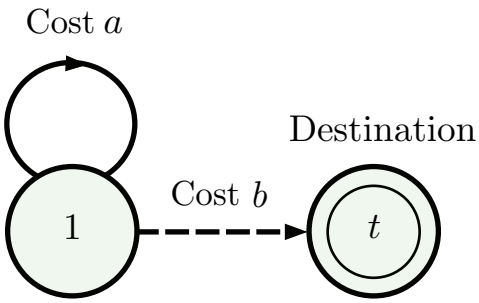

FIG. 3. *A shortest path problem with a single node* 1 *and a termination node* t.

Proof. (a), (b) Follows from Proposition 3.2 with $\mathcal{C} = \{(\mu, x) \mid \mu \in \mathcal{M}_S,\ x \in X\}$, in which case $J^*_{\mathcal{C}} = J^*_S$.

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

The following example illustrates the preceding proposition and demonstrates some of the unusual behaviors that can arise in the context of our model.

Example 5.1. Consider the deterministic shortest path example shown in Figure 3. Here there is a single state 1 in addition to the termination state t. At state 1 there are two choices: a self-transition, which costs a, and a transition to t, which costs b. The mapping H, abbreviating $J(1)$ with just the scalar J, is

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

and the initial function $\bar{J}$ is taken to be 0.

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

$$T_\mu J = b, \quad T_{\mu'} J = a + J, \quad T J = \min\{b,\, a + J\} \qquad \forall\ J \in \Re.$$

For the proper policy μ, the mapping $T_\mu : \Re \mapsto \Re$ is a contraction. For the improper policy μ', the mapping $T_{\mu'} : \Re \mapsto \Re$ is not a contraction, and it has a fixed point within $\Re$ only if $a = 0$, in which case every $J \in \Re$ is a fixed point. Let S be equal to the real line $\Re$ [the set $\mathcal{R}(X)$]. Then a policy is S-regular if and only if it is proper (this is generally true for SSP problems, for $S = \Re^n$). Thus μ is S–regular, while μ' is not.

Let us consider the optimal cost J^*, the fixed points of T within $\Re$, and the behavior of VI and PI for different combinations of values of a and b.

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

(b) If $a = 0$, the set of fixed points of T (within $\Re$) is the interval $(-\infty, b]$. Here the improper policy is optimal if $b \geq 0$, and the proper policy is optimal if $b \leq 0$.

(c) If $a = 0$ and $b > 0$, the proper policy is strictly suboptimal, yet its cost at state 1 (which is b) is a fixed point of T. The optimal cost, $J^* = 0$, lies in the

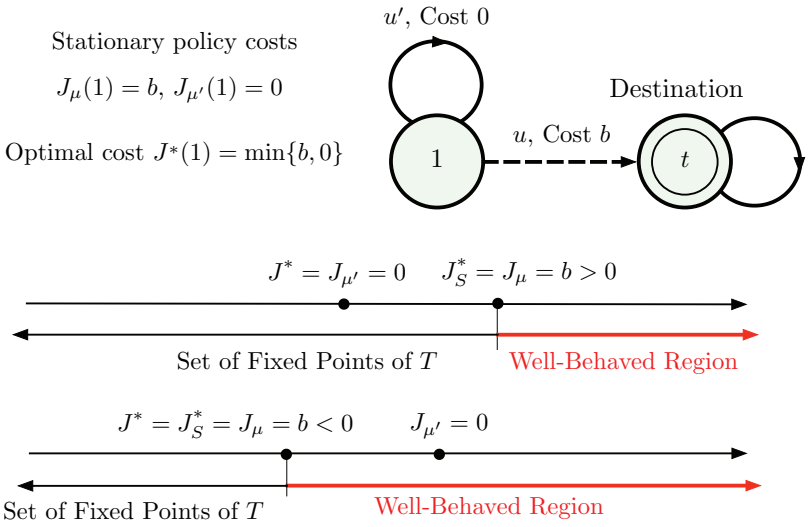

FIG. 4. *The well-behaved region of* (5.2) *for the deterministic shortest path Example* 5.1 *when where there is a zero length cycle* ($a = 0$). *For* $S = \Re$, *the policy* μ *is* S*-regular, while the policy* μ' *is not. The figure illustrates the two cases, where* $b > 0$ *and* $b < 0$.

interior of the set of fixed points of T, which is $(-\infty, b]$. Thus the VI method that generates $\{T^k J\}$ starting with $J \neq J^*$ cannot find J^*. In particular if J is a fixed point of T, VI stops at J, while if J is not a fixed point of T (i.e., $J > b$), VI terminates in two iterations at $b \neq J^*$. Moreover, the standard PI method is unreliable in the sense that starting with the suboptimal proper policy μ, it may stop with that policy because $T_\mu J_\mu = b = \min\{b, J_\mu\} = TJ_\mu$ (the improper/optimal policy μ' also satisfies $T_{\mu'} J_\mu = TJ_\mu$, so a rule for breaking the tie in favor of μ is needed but such a rule may not be obvious in general).

(d) If $a = 0$ and $b < 0$, the improper policy is strictly suboptimal, and we have $J^* = b$. Here it can be seen that the VI sequence $\{T^k J\}$ converges to J^* for all $J \geq b$, but stops at J for all $J < b$, since the set of fixed points of T is $(-\infty, b]$. Moreover, starting with either the proper or the improper policy, PI may oscillate, since $T_\mu J_{\mu'} = TJ_{\mu'}$ and $T_{\mu'} J_\mu = TJ_\mu$, as can be easily verified [the optimal policy μ also satisfies $T_\mu J_\mu = TJ_\mu$ but it is not clear how to break the tie; compare also with case (c) above].

(e) If $a < 0$, the improper policy is optimal and we have $J^* = -\infty$. There are no fixed points of T within $\Re$, but J^* is the unique fixed point of T within the set $[-\infty, \infty]$. Then VI will converge to J^* starting from any $J \in [-\infty, \infty]$, while PI will also converge to the optimal policy starting from either policy.

Let us focus on the case where there is a zero length cycle ($a = 0$). The cost functions J_μ, $J_{\mu'}$, and J^* are fixed points of the corresponding mappings, but the sets of fixed points of $T_{\mu'}$ and T within S are $\Re$ and $(-\infty, b]$, respectively. Figure 4 shows the well-behaved regions $\mathcal{W}_S$ of (5.2) for the two cases $b > 0$ and $b < 0$, and is consistent with the results of Proposition 5.1. In particular, the VI algorithm fails when started outside the well-behaved region, while starting from within the region, it is attracted to J_S^* rather than to J^*.

Note that Proposition 5.1(b) asserts convergence of the VI algorithm to J_S^* only for initial conditions $J \leq \tilde{J}$ for some $\tilde{J} \in S$. For an example where there is a single

policy μ, which is S-regular, but $\{T_\mu^k J\}$ does not converge to J_μ starting from some $J \geq J_\mu$ that lies outside S, consider a mapping $T_\mu : \Re \mapsto \Re$ that has two fixed points: J_μ and another fixed point $J' > J_\mu$. Let $\tilde{J} = (J_\mu + J')/2$ and $S = (-\infty, \tilde{J}]$, and assume that T_μ is a contraction mapping within S (a one-dimensional example of this type, where $S = \Re$, can be easily constructed graphically). Then, $\tilde{J} \in S$, and starting from any $J \in S$, we have $T^k J \to J_\mu$, so that μ is S-regular. However, since J' is a fixed point of T, the sequence $\{T^k J'\}$ stays at J' and does not converge to J_μ. The difficulty here is that $\mathcal{W}_S = [J_\mu, \tilde{J}]$ and $J' \notin \mathcal{W}_S$.

In many contexts where Proposition 5.1 applies, there exists an $\mathcal{M}_S$-optimal policy μ^* such that T_{μ^*} is a contraction with respect to a weighted sup-norm. This is true for example in several types of shortest path problems. In such cases, VI converges to J_S^* linearly, as shown in the following proposition first given in [BeY16] for SSP problems.

PROPOSITION 5.2 (convergence rate of VI). *Let S be equal to $\mathcal{B}(X)$, the space of all functions over X that are bounded with respect to a weighted sup-norm $\|\cdot\|_v$ corresponding to a positive function $v : X \mapsto \Re$. Assume that J_S^* is a fixed point of T, and that there exists an $\mathcal{M}_S$-optimal policy μ^* such that T_{μ^*} is a contraction with respect to $\|\cdot\|_v$, with corresponding modulus of contraction β. Then*

$$\big\|TJ - J_S^*\|_v \leq \beta \|J - J_S^*\|_v \qquad \forall\ J \geq J_S^*, \tag{5.3}$$

and we have

$$\|J - J_S^*\|_v \leq \frac{1}{1-\beta} \sup_{x \in X} \frac{J(x) - (TJ)(x)}{v(x)} \qquad \forall\ J \geq J_S^*. \tag{5.4}$$

Proof. By using the $\mathcal{M}_S$-optimality of μ^* and Proposition 5.1(c), we have $J_S^* = T_{\mu^*} J_S^* = T J_S^*$, so that

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

for all $x \in X$ and $J \geq J_S^*$. By taking the supremum of the left-hand side over $x \in X$, and by using the fact that the inequality $J \geq J_S^*$ implies that $TJ \geq TJ_S^* = J_S^*$, we obtain (5.3).

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

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

By taking the supremum of both sides over x, we obtain (5.4). ∎

A critical assumption of Propositions 5.1 and 5.2 is that J_S^* is a fixed point of T. For a specific application, this must be proved with a separate analysis after a suitable set S is chosen. There are several approaches that guide the choice of S and facilitate the analysis. One approach applies to problems where J^* is a fixed point of T; this is true generically in wide classes of problems, including deterministic and minimax

models (we give a proof for the deterministic case later, in section 6). Then for every set S such that $J_S^* = J^*$, Proposition 5.1 applies and shows that J^* can be obtained by VI starting from any $J \in \mathcal{W}_S$. Other important models where J^* is guaranteed to be a fixed point of T are the monotone increasing and monotone decreasing models of [Ber13, section 4.3], a fact known since [Ber77]. In what follows we will use the PI algorithm as the basis for a new and different line of analysis to show that J_S^* is a fixed point of T.

5.2. Policy iteration-based analysis of Bellman's equation. The approach of this section is applicable under assumptions that guarantee that there is a sequence $\{\mu^k\}$ of S-regular policies that can be generated by the PI algorithm, which generates a sequence of policies $\{\mu^k\}$ according to

$$T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k}, \qquad k = 0, 1, \ldots, \tag{5.5}$$

starting from an initial policy μ^0. To be able to carry out the policy improvement step, which computes $\mu^{k+1}(x)$ as a minimum over $u \in U(x)$ of $H(x, u, J_{\mu^k})$ for each $x \in X$ [cf. (5.5)], there should be enough assumptions to guarantee that this minimum is attained for every x. One such assumption is that $U(x)$ is a finite set for each $x \in X$. A more general assumption, involving a form of compactness of the constraint set, is given in the next section (see Lemma 6.1).

The significance of all μ^k being S-regular lies in that *the corresponding cost function sequence $\{J_{\mu^k}\}$ lies within the well-behaved region of equation* (5.2)*, and is monotonically nonincreasing*. We have the following proposition.

PROPOSITION 5.3 (policy improvement under S-regularity). *Given a set $S \subset \mathcal{E}(X)$, assume that $\{\mu^k\}$ is a sequence generated by the PI algorithm* (5.5) *that consists of S-regular policies. Then $J_{\mu^k} \geq J_{\mu^{k+1}}$ for all k.*

Proof. Using the S-regularity of μ^k, we have

$$J_{\mu^k} = T_{\mu^k} J_{\mu^k} \geq T J_{\mu^k} = T_{\mu^{k+1}} J_{\mu^k}. \tag{5.6}$$

By repeatedly applying $T_{\mu^{k+1}}$ to both sides, we obtain

$$J_{\mu^k} \geq \lim_{m \to \infty} T_{\mu^{k+1}}^m J_{\mu^k} = J_{\mu^{k+1}},$$

where the equation on the right holds since μ^{k+1} is S-regular and $J_{\mu^k} \in S$ (since μ^k is S-regular). ∎

The preceding proposition shows that for a sequence of S-regular policies $\{\mu^k\}$ that is generated by PI, the cost function sequence $\{J_{\mu^k}\}$ converges pointwise to a limit J_∞. Under mild conditions, we will show that J_∞ is a fixed point of T and is equal to J_S^*, thus bringing to bear Proposition 5.1. Let us first formalize the property that the PI algorithm can generate a sequence of S-regular policies.

DEFINITION 5.2 (weak PI property). *A set $S \subset \mathcal{E}(X)$ has the* weak PI property *if there exists a sequence $\{\mu^k\}$ that satisfies* (5.5) *and consists of S-regular policies.*

The following proposition shows that J_S^* is a fixed point of T, assuming the weak PI property and a mild continuity-type condition.

PROPOSITION 5.4 (weak PI property theorem). *Given a set $S \subset \mathcal{E}(X)$, assume that*

(1) *S has the weak PI property;*

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

$$H(x,u,J) = \lim_{m\to\infty} H(x,u,J_m) \qquad \forall\ x \in X,\ u \in U(x). \tag{5.7}$$

Then

(a) J_S^* *is a fixed point of* T *and the conclusions of Proposition* 5.1 *hold;*

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

Proof. (a) Let $\{\mu^k\}$ be a sequence of S-regular policies generated by the PI algorithm (there exists such a sequence by the weak PI property). Then by Proposition 5.3, the sequence $\{J_{\mu^k}\}$ is monotonically nonincreasing and must converge to some $J_\infty \geq J_S^*$. We will show that J_∞ is a fixed point of T and then invoke Proposition 3.2.

Indeed, we have

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

[cf. (5.6)], so by letting $k \to \infty$, we obtain $J_\infty \geq TJ_\infty$. To prove the reverse inequality, we first note that from the definition of the PI iteration and the nonincreasing property $J_{\mu^k} \geq J_{\mu^{k+1}}$, we have

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

By using (5.7) together with the preceding relation, we obtain for all $x \in X$ and $u \in U(x)$,

$$H(x,u,J_\infty) = \lim_{k\to\infty} H(x,u,J_{\mu^k}) \geq \lim_{k\to\infty} (TJ_{\mu^k})(x) \geq \lim_{k\to\infty} J_{\mu^{k+1}} = J_\infty(x).$$

By taking the infimum of the left-hand side over $u \in U(x)$, it follows that $TJ_\infty \geq J_\infty$. Thus $J_\infty = TJ_\infty$. Finally, by applying Proposition 3.2 with

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

we have $J_\infty = J_{\mathcal{C}}^* = J_S^*$.

(b) The limit of $\{J_{\mu^k}\}$ was shown to be equal to J_S^* in the preceding proof. Moreover, the finiteness of $\mathcal{M}_S$ and the policy improvement property of Proposition 5.3 imply that some $\mu^{\bar{k}}$ is $\mathcal{M}_S$-optimal. ∎

Note that the preceding proposition shows that under the weak PI property, PI converges to J_S^*. However, this does not imply convergence to J^*. We next introduce a stronger type of PI property, which we will use to obtain stronger results.

DEFINITION 5.3 (strong PI property). *A set* $S \subset \mathcal{E}(X)$ *has the* strong PI property *if*

(a) *there exists at least one* S*-regular policy;*

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

The strong PI property clearly implies the weak PI property. On the other hand, the strong PI property may be harder to verify in a given setting. The following proposition provides conditions guaranteeing the strong PI property. The key implication of these conditions is that they preclude optimality of an S-irregular policy [see condition (4) of the proposition]. Condition (3) of the proposition is implied by finiteness of the constraint set or by a more general compactness assumption that will be given in the next section.

PROPOSITION 5.5 (verifying the strong PI property). *Given a set $S \subset \mathcal{E}(X)$, assume that*

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

$$\limsup_{k\to\infty} (T_\mu^k J)(x) = \infty. \tag{5.8}$$

Then

(a) *if a policy μ satisfies $T_\mu J \le J$ for some function $J \in S$, then μ is S-regular;*
(b) *S has the strong PI property.*

Proof. (a) By the monotonicity of T_μ, we have $\limsup_{k\to\infty} T_\mu^k J \le J$, and since by condition (1), $J(x) < \infty$ for all x, it follows from (5.8) that μ is S-regular.

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

$$T_{\bar\mu} J_\mu = TJ_\mu \le T_\mu J_\mu = J_\mu,$$

so $\bar\mu$ is S-regular by part (a). □

By using the strong PI property and assuming also that $J_S^* \in S$, we will now show that J_S^* is the unique fixed point of T within S. This result will be the starting point for the analysis of section 6.

PROPOSITION 5.6 (strong PI property theorem). *Let S satisfy the conditions of Proposition* 5.5.

(a) (uniqueness of fixed point) *If T has a fixed point within S, then this fixed point is equal to J_S^*.*
(b) (fixed point property and optimality condition) *If $J_S^* \in S$, then J_S^* is the unique fixed point of T within S. Moreover, every policy μ that satisfies $T_\mu J_S^* = TJ_S^*$ is $\mathcal{M}_S$-optimal and there exists at least one such policy.*
(c) (PI convergence) *If for each sequence $\{J_m\} \subset S$ with $J_m \downarrow J$ for some $J \in \mathcal{E}(X)$, we have*

$$H(x,u,J) = \lim_{m\to\infty} H(x,u,J_m) \qquad \forall\ x \in X,\ u \in U(x), \tag{5.9}$$

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

Proof. (a) Let $J' \in S$ be a fixed point of T. By applying Proposition 3.2 with $\mathcal{C} = \{(\mu, x) \mid \mu \in \mathcal{M}_S,\ x \in X\}$, we have $J' \le J_{\mathcal{C}}^* = J_S^*$. For the reverse inequality, let μ' be such that $J' = TJ' = T_{\mu'} J'$ [cf. condition (3) of Proposition 5.5]. Then by Proposition 5.5(a), it follows that μ' is S-regular, and since $J' \in S$, by the definition of S-regularity, we have $J' = J_{\mu'} \ge J_S^*$, showing that $J' = J_S^*$.

(b) For every $\mu \in \mathcal{M}_S$ we have $J_\mu \ge J_S^*$, so that $J_\mu = T_\mu J_\mu \ge T_\mu J_S^* \ge TJ_S^*$. Taking the infimum over all $\mu \in \mathcal{M}_S$, we obtain $J_S^* \ge TJ_S^*$. Let μ be a policy such that $TJ_S^* = T_\mu J_S^*$ [there exists one by condition (3) of Proposition 5.5, since we assume that $J_S^* \in S$]. The preceding two relations yield $J_S^* \ge T_\mu J_S^*$, so by Proposition 5.5(a), μ is S-regular. Therefore, we have

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

where the second equality holds by S-regularity of μ and $J_S^* \in S$ by assumption. Hence equality holds throughout in the above relation, proving that J_S^* is a fixed point of T and that μ is $\mathcal{M}_S$-optimal.

(c) Since the strong PI property [which holds by Proposition 5.5(b)] implies the weak PI property, the result follows from Proposition 5.4. □

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

Note that for PI to be valid, as per Proposition 5.6(c), an initial S-regular policy must be available. Chapter 3 of [Ber13] describes a combined VI and PI algorithm, which does not require an initial S-regular policy, and can tolerate the generation of S-irregular policies. Let us also consider two additional algorithmic approaches for computing J_S^*, not given in [Ber13], which can be justified based on the preceding analysis.

A mathematical programming solution method. We will show that J_S^* is an upper bound to all functions $J \in S$ that satisfy $J \leq TJ$, and we will exploit this fact to obtain an algorithm to compute J_S^*. We have the following proposition.

PROPOSITION 5.7. *Given a set $S \subset \mathcal{E}(X)$ for all functions $J \in S$ satisfying $J \leq TJ$, we have $J \leq J_S^*$.*

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

Assuming that J_S^* is a fixed point of T, we can use the preceding proposition to compute J_S^* by maximizing an appropriate monotonically increasing function of J subject to the constraints $J \in S$ and $J \leq TJ$.[3] This approach is well known in finite-state finite-control Markovian decision problems, where it is usually referred to as the *linear programming solution method*, because in this case the resulting optimization problem is a linear program (see, e.g., the books [Kal83, Put94, Ber12]).

For a more general finite-state case, suppose that $X = \{1, \ldots, n\}$ and $S = \Re^n$. Then Proposition 5.7 shows that $J_S^* = (J_S^*(1), \ldots, J_S^*(n))$ is the unique solution of the following optimization problem:

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

where $\beta_1, \ldots, \beta_n$ are any positive scalars. If H is linear in J and each $U(i)$ is a finite set, this is a linear program, which can be solved with standard methods.

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

[3] For the mathematical programming approach to apply, it is sufficient that $J_S^* \leq TJ_S^*$. However, we generally have $J_S^* \geq TJ_S^*$ (this follows by writing for all $\mu \in \mathcal{M}_S$, $J_\mu = T_\mu J_\mu \geq TJ_\mu \geq TJ_S^*$, and taking the infimum over all $\mu \in \mathcal{M}_S$), so the condition $J_S^* \leq TJ_S^*$ is equivalent to J_S^* being a fixed point of T.

$\{J_k, \mu^k\}$ according to

$$T_{\mu^k} J_k = TJ_k, \qquad J_{k+1} = T_{\mu^k}^{m_k} J_k, \qquad k = 0, 1, \ldots, \tag{5.10}$$

where m_k is a positive integer for each k.

The following proposition shows that optimistic PI converges under mild assumptions to a fixed point of T, independently of any S-regularity framework. However, when such a framework is introduced, and the sequence generated by optimistic PI generates a sequence of S-regular policies, then the algorithm converges to J_S^*, which is in turn a fixed point of T, similar to the PI convergence result under the weak PI property; cf. Proposition 5.4(b). Thus the proposition serves both an analytical purpose (as a tool for establishing that J_S^* is a fixed point of T), and a computational purpose [establishing the validity of the optimistic PI algorithm (5.10) as a means for computing J_S^*].

PROPOSITION 5.8 (convergence of optimistic PI). *Let $J_0 \in \mathcal{E}(X)$ be a function such that $J_0 \geq TJ_0$, and assume that*

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

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

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

Then the optimistic PI algorithm (5.10) is well-defined and the following hold:

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

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

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

$$J_0 \geq TJ_0 = T_{\mu^0} J_0 \geq T_{\mu^0}^{m_0} J_0 = J_1 \geq T_{\mu^0}^{m_0+1} J_0 = T_{\mu^0} J_1 \geq TJ_1 = T_{\mu^1} J_1 \geq \cdots \geq J_2, \tag{5.11}$$

and continuing similarly, we obtain $J_k \geq TJ_k \geq J_{k+1}$ for all $k = 0, 1, \ldots$. 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 (5.5) in Proposition 5.4.

(b) In the case where all the policies μ^k are S-regular and $\{J_k\} \subset S$, from (5.11), we have $J_{k+1} \geq J_{\mu^k}$ for all k, so it follows that

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

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

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

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

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

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

Note that the fixed point J_∞ in Proposition 5.8(a) need not be equal to J_S^* or J^*. As an illustration, consider the shortest path Example 5.1 with $S = \Re$, and $a = 0$, $b > 0$. Then if $0 < J_0 < b$, it can be seen that $J_k = J_0$ for all k, so $J^* = 0 < J_\infty$ and $J_\infty < J_S^* = b$.

6. Infinite and finite cost cases for irregular policies. The results of the preceding section do not assert that J^* is a fixed point of T or that $J^* = J_S^*$. In this section we address this issue with some additional assumptions. A critical part of the analysis is based on the strong PI property theorem of Proposition 5.6.

6.1. The case where all irregular policies have infinite cost. We will first assume that all S-irregular policies have infinite cost for some initial state [cf. (5.8)]. The following assumption and proposition were given in [Ber13, section 3.2], but the line of proof given here is considerably streamlined thanks to the use of the strong PI property analysis of the preceding section, which was developed after [Ber13] was published.

Assumption 6.1. We have a subset $S \subset \mathcal{R}(X)$ satisfying the following:

(a) S contains $\bar{J}$, and has the property that if J_1, J_2 are two functions in S, then S contains all functions J with $J_1 \leq J \leq J_2$.

(b) The function $J_S^* = \inf_{\mu \in \mathcal{M}_S} J_\mu$ belongs to S.

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

$$\limsup_{k \to \infty} (T_\mu^k J)(x) = \infty. \tag{6.1}$$

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

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

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 $J \in S$,

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

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

The conditions (b) and (c) of the preceding assumption were introduced in Propositions 5.5 and 5.6. New conditions are (a), (d), (e), and (f). In the case where S is the set of real-valued functions $\mathcal{R}(X)$ and $\bar{J} \in \mathcal{R}(X)$, condition (a) is automatically satisfied, while condition (e) is typically verified easily. The verification of condition (f) may be nontrivial in some cases. We postpone the discussion of this issue for later (see Proposition 6.2).

The main result of this section is the following proposition.

PROPOSITION 6.1. *Let Assumption* 6.1 *hold. Then*

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

(b) *we have $T^k J \to J^*$ for all $J \in S$;*

(c) *a policy μ is optimal if and only if $T_\mu J^* = TJ^*$. Moreover, there exists an optimal S-regular policy;*

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

(e) *if in addition for each sequence* $\{J_m\} \subset S$ *with* $J_m \downarrow J$ *for some* $J \in S$, *we have*

$$H(x,u,J) = \lim_{m\to\infty} H(x,u,J_m) \qquad \forall\ x \in X,\ u \in U(x), \tag{6.2}$$

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

The proof of Proposition 6.1 will be developed through a sequence of lemmas. We first state without proof a result given in [Ber13, Lemma 3.2.1]. It guarantees that starting from an S-regular policy, the PI algorithm is well-defined. Similar results are well known in DP theory.

LEMMA 6.1. *Let Assumption* 6.1(d) *hold. For every* $J \in S$, *there exists a policy* μ *such that* $T_\mu J = TJ$.

Next we restate, for easy reference, some of the results of the preceding section in the next two lemmas.

LEMMA 6.2. *Let Assumption* 6.1(c) *hold. A policy* μ *that satisfies* $T_\mu J \leq J$ *for some* $J \in S$ *is* S*-regular.*

Proof. This is Proposition 5.5(b). ∎

LEMMA 6.3. *Let Assumptions* 6.1(b), (c), (d) *hold. Then*

(a) *the function* J_S^* *of Assumption* 6.1(b) *is the unique fixed point of* T *within* S;

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

Proof. This is Proposition 5.6, parts (a) and (b) [Assumption 6.1(d) guarantees that for every $J \in S$, there exists a policy μ such that $T_\mu J = TJ$ (cf. Lemma 6.1)]. ∎

Let us also prove the following technical lemma that relies on the continuity Assumption 6.1(e).

LEMMA 6.4. *Let Assumptions* 6.1(d), (e) *hold. Then if* $J \in S$, $\{T^k J\} \subset S$, *and* $T^k J \uparrow J_\infty$ *for some* $J_\infty \in S$, *we have* $J_\infty = J_S^*$.

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

$$U_k(x) = \Big\{u \in U(x) \mid H\left(x,u,T^k J\right) \leq J_\infty(x)\Big\}, \qquad k = 0,1,\ldots, \tag{6.3}$$

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

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

(such a point exists by Lemma 6.1). Then $u_k \in U_k(x)$.

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

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

Therefore from the definition (6.3), we have $\{u_i\}_{i=k}^\infty \subset U_k(x)$. Since $U_k(x)$ is compact, all the limit points of $\{u_i\}_{i=k}^\infty$ belong to $U_k(x)$ and at least one limit point exists. Hence

the same is true for the limit points of the whole sequence $\{u_i\}$. Thus if $\tilde{u}$ is a limit point of $\{u_i\}$, we have

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

By (6.3), this implies that

$$H\big(x, \tilde{u}, T^k J\big) \leq J_\infty(x), \qquad k = 0, 1, \ldots.$$

Taking the limit as $k \to \infty$ and using Assumption 6.1(e), we obtain

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

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

We are now ready to show Proposition 6.1 by using the additional parts (a) and (f) of Assumption 6.1.

Proof of Proposition 6.1. (a), (b) We will first prove that $T^k J \to J_S^*$ for all $J \in S$, and we will use this to prove that $J_S^* = J^*$ and that there exists an optimal S-regular policy. Thus parts (a) and (b), together with the existence of an optimal S-regular policy, will be shown simultaneously.

We fix $J \in S$, and choose $J' \in S$ such that $J' \leq J$ and $J' \leq TJ'$ [cf. Assumption 6.1(f)]. By the monotonicity of T, we have $T^k J' \uparrow J_\infty$ for some $J_\infty \in \mathcal{E}(X)$. Let μ be an S-regular policy such that $J_\mu = J_S^*$ [cf. Lemma 6.3(b)]. Then we have, using again the monotonicity of T,

$$J_\infty = \lim_{k\to\infty} T^k J' \leq \limsup_{k\to\infty} T^k J \leq \lim_{k\to\infty} T_\mu^k J = J_\mu = J_S^*. \tag{6.4}$$

Since J' and J_S^* belong to S, and $J' \leq T^k J' \leq J_\infty \leq J_S^*$, Assumption 6.1(a) implies that $\{T^k J'\} \subset S$, and $J_\infty \in S$. From Lemma 6.4, it then follows that $J_\infty = J_S^*$. Thus equality holds throughout in (6.4), proving that $\lim_{k\to\infty} T^k J = J_S^*$.

There remains to show that $J_S^* = J^*$ and that there exists an optimal S-regular policy. To this end, we note that by the monotonicity Assumption 2.1, for any policy $\pi = \{\mu_0, \mu_1, \ldots\}$, we have

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

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

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

where the equality follows since $T^k J \to J_S^*$ for all $J \in S$ (as shown earlier), and $\bar{J} \in S$ [cf. Assumption 6.1(a)]. Thus for all $\pi \in \Pi$, $J_\pi \geq J_S^* = J_\mu$, implying that the policy μ that is optimal within the class of S-regular policies is optimal over all policies, and that $J_S^* = J^*$.

(c) If μ is optimal, then $J_\mu = J^* \in S$, so by Assumption 6.1(c), μ is S-regular and therefore $T_\mu J_\mu = J_\mu$. Hence, $T_\mu J^* = T_\mu J_\mu = J_\mu = J^* = TJ^*$. Conversely, if $J^* = TJ^* = T_\mu J^*$, μ is S-regular (cf. Lemma 6.2), so $J^* = \lim_{k\to\infty} T_\mu^k J^* = J_\mu$. Therefore, μ is optimal.

(d) If $J \in S$ and $J \leq TJ$, by repeatedly applying T to both sides and using the monotonicity of T, we obtain $J \leq T^k J$ for all k. Taking the limit as $k \to \infty$ and

using the fact $T^k J \to J^*$ [cf. part (b)], we obtain $J \leq J^*$. The proof that $J \geq TJ$ implies $J \geq J^*$ is similar.

(e) As in the proof of Proposition 5.4(b), the sequence $\{J_{\mu^k}\}$ converges monotonically to a fixed point of T, call it J_∞. Since J_∞ lies between $J_{\mu^0} \in S$ and $J_S^* \in S$, it must belong to S, by Assumption 6.1(a). Since the only fixed point of T within S is J^* [cf. part (a)], it follows that $J_\infty = J^*$. □

Finally let us give a proposition, which provides an approach to verify part (f) of Assumption 6.1. The proposition will be used later in this section (cf. the proof of Proposition 6.4).

PROPOSITION 6.2. *Let S be equal to $\mathcal{R}_b(X)$, the subset of $\mathcal{R}(X)$ that consists of functions J that are bounded below, i.e., for some $b \in \Re$, satisfy $J(x) \geq b$ for all $x \in X$. Let parts* (b), (c), *and* (d) *of Assumption* 6.1 *hold, and assume further that for all scalars $r > 0$, we have*

$$TJ_S^* - re \leq T(J_S^* - re), \tag{6.5}$$

where e is the unit function, $e(x) \equiv 1$. Then part (f) *of Assumption* 6.1 *also holds.*

Proof. Let $J \in S$, and let $r > 0$ be a scalar such that $J_S^* - re \leq J$ [such a scalar exists since $J_S^* \in R_b(x)$ by Assumption 6.1(b)]. Define $J' = J_S^* - re$, and note that by Lemma 6.3, J_S^* is a fixed point of T. By using (6.5), we have

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

thus proving part (f) of Assumption 6.1. □

Several examples of applications of Proposition 6.1 are given in recent papers of the author, such as [Ber15a] that considers the minimax-type of shortest problems, and [Ber16] that considers SSP problems with multiplicative or exponential cost functions (see also [DeR79, Pat01, Ber13, CaR14]). The paper [Ber15b] considers an infinite-space optimal control problem with nonnegative cost per stage, where the objective is to steer a deterministic system towards a set of termination states. We consider a similar but more general version of this problem, where we remove the assumption of nonnegativity for the cost per stage (the paper [Ber15b] considers also a related minimax problem, as well as PI-related methodology that we do not address here).

6.2. Application to deterministic continuous-state problems. Let us consider a deterministic optimal control problem with the system equation

$$x_{k+1} = f(x_k, u_k), \qquad k = 0, 1, \ldots, \tag{6.6}$$

where x_k and u_k are the state and control at stage k, lying in sets X and U, respectively, and f is a function mapping $X \times U$ to X. The control u_k must be chosen from a constraint set $U(x_k)$. The cost per stage is denoted $g(x, u)$ (note that g can take both positive and negative values). No restrictions are placed on X and U: for example, they may be finite sets as in deterministic shortest path problems, or they may be continuous spaces as in classical problems of control to the origin or some other terminal set.

The cost function of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ starting at an initial state x_0 is

$$J_\pi(x_0) = \limsup_{N \to \infty} \sum_{k=0}^{N-1} g\left(x_k, \mu_k(x_k)\right), \qquad x_0 \in X, \tag{6.7}$$

where $(x_k, \mu_k(x_k))$, $k = 0, 1, \ldots$, are the state-control pairs using π. We assume that there is a nonempty stopping set $X_0 \subset X$, consisting of cost-free and absorbing states in the sense that

$$g(x,u) = 0, \qquad x = f(x,u) \qquad \forall\ x \in X_0,\ u \in U(x). \tag{6.8}$$

Clearly, for $x \in X_0$, we have $J^*(x) = 0$, as well as $J_\pi(x) = 0$ for all policies $\pi \in \Pi$. Besides X_0, another interesting subset of X is

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

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

To formulate a corresponding abstract DP problem, we introduce the mapping $T_\mu : \mathcal{R}(X) \mapsto \mathcal{R}(X)$ by

$$(T_\mu J)(x) = g\big(x, \mu(x)\big) + J\big(f(x, \mu(x))\big), \qquad x \in X, \tag{6.9}$$

and the mapping $T : \mathcal{E}(X) \mapsto \mathcal{E}(X)$ given by

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

The initial function $\bar{J}$ is the zero function $[\bar{J}(x) \equiv 0]$. An important fact is that *because the problem is deterministic, J^* is a fixed point of T.*[4]

We say that a policy μ is *terminating* if the state sequence $\{x_k\}$ generated starting from any $x \in X_f$ and using μ reaches X_0 in finite time, i.e., satisfies $x_{\bar{k}} \in X_0$ for some index $\bar{k}$. The set of terminating policies is denoted by $\mathcal{T}$. Our key assumption is that for $x \in X_f$, the optimal cost $J^*(x)$ can be approximated arbitrarily closely by using terminating policies. In particular, we assume the following.

Assumption 6.2 (near-optimal termination). For every pair (x, ϵ) with $x \in X_f$ and $\epsilon > 0$, there exists a terminating policy μ that satisfies $J_\mu(x) \le J^*(x) + \epsilon$.

This assumption implies in particular that the optimal cost function over terminating policies,

$$\hat{J}(x) = \inf_{\mu \in \mathcal{T}} J_\mu(x), \qquad x \in X,$$

is equal to J^*. Moreover since J^* is a fixed point of T (because we are dealing with a deterministic problem), it follows that $\hat{J}$ is a fixed point of T, which brings to bear Proposition 5.1.

There are easily verifiable conditions that imply Assumption 6.2, some of which are discussed in [Ber15b], where it is assumed in addition that $g \ge 0$. A prominent case is when X and U are finite, so the problem becomes a deterministic shortest path problem. If all cycles of the state transition graph have positive length, all policies π that do not terminate from a state $x \in X_f$ must satisfy $J_\pi(x) = \infty$, implying that there exists an optimal policy that terminates from all $x \in X_f$. Thus, in this case Assumption 6.2 is naturally satisfied. Another interesting case arises when $g(x,u) = 0$ for all (x,u) except if $x \notin X_0$ and $f(x,u) \in X_0$, in which case we have $g(x,u) < 0$, i.e., there is no cost incurred except for a negative cost (positive

[4]For any policy $\pi = \{\mu_0, \mu_1, \ldots\}$, using the definition of J_π, we have for all x, $J_\pi(x) = g\big(x, \mu_0(x)\big) + J_{\pi_1}\big(f(x, \mu_0(x))\big)$, where $\pi_1 = \{\mu_1, \mu_2, \ldots\}$. By taking the infimum of the left-hand side over π and the infimum of the right-hand side over π_1 and then μ_0, we obtain $J^* = TJ^*$.

reward) upon termination. Then, assuming that X_0 can be reached from all states, Assumption 6.2 is satisfied. This is also an example of a deterministic problem where zero length cycles are common.

When X is the n-dimensional Euclidean space $\Re^n$, a primary case of interest in control system design contexts, it may easily happen that the optimal policies are not terminating from some $x \in X_f$. Instead the optimal state trajectories may approach X_0 asymptotically. This is true for example in the classical linear-quadratic optimal control problem, where under some natural controllability and observability conditions, the optimal closed-loop system is linear and stable, so the state will typically never reach the termination set $X_0 = \{0\}$ in finite time, although it will approach it asymptotically (see, e.g., [Ber17a, section 3.1]). However, the Assumption 6.2 is satisfied (see [Ber15b]).

Let us denote by S the set of functions

(6.10)
$$S = \big\{J \in \mathcal{E}(X) \mid \ J(x) = 0 \ \forall\ x \in X_0,\ J(x) \in \Re \ \forall\ x \in X_f,\ J(x) > -\infty \ \forall\ x \in X\big\}.$$

Since X_0 consists of cost-free and absorbing states [cf. (6.8)], and $J^*(x) > -\infty$ for all $x \in X$ (by Assumption 6.2), the set S contains the cost function J_μ of all policies μ, as well as J^*. Moreover it can be seen that every terminating policy is S-regular, i.e., $\mathcal{T} \subset \mathcal{M}_S$, which implies that $J_S^* = \hat{J} = J^*$. The reason is that the terminal cost is zero after termination for any terminal cost function $J \in S$, i.e., $(T_\mu^k J)(x) = (T_\mu^k \bar{J})(x) = J_\mu(x)$ for $\mu \in \mathcal{T}$, $x \in X_f$, and k sufficiently large.

The following proposition is a consequence of Proposition 5.1, the deterministic character of the problem (which guarantees that J^* is a fixed point of T), and Assumption 6.2 (which guarantees that $J_S^* = \hat{J} = J^*$).

Proposition 6.3. *Let Assumption* 6.2 *hold. Then*

(a) *J^* is the only fixed point of T within the set of all $J \in S$ such that $J \geq J^*$;*
(b) *we have $T^k J \to J^*$ for every $J \in S$ such that $J \geq J^*$;*
(c) *if μ^* is terminating and $T_{\mu^*} J^* = TJ^*$, then μ^* is optimal. Conversely, if μ^* is terminating and is optimal, then $T_{\mu^*} J^* = TJ^*$.*

For an example of what may happen in the absence of Assumption 6.2, consider the deterministic shortest path Example 5.1 with $a = 0$, $b > 0$, and $S = \Re$. Here we have $0 = J^* < \hat{J} = b$, while the set of fixed points of T is the interval $(-\infty, b]$.

We will now consider additional assumptions, which guarantee the stronger conclusions of Proposition 6.1. We first replace the set S of (6.10) with the following subset of functions that are bounded below:

$$\hat{S} = \big\{J \in \mathcal{E}(X) \mid \ J(x) = 0 \ \forall\ x \in X_0,\ J(x) \in \Re \ \forall\ x \in X_f,\ J \text{ is uniformly bounded below by a scalar}\big\}.$$

We have the following proposition.

Proposition 6.4. *Let Assumption* 6.2 *hold, and assume further that*

(1) *$J_{\hat{S}}^* \in \hat{S}$;*
(2) *for each $\hat{S}$-irregular policy μ and each $J \in \hat{S}$, there is at least one state $x \in X$ such that $\limsup_{k\to\infty} (T_\mu^k J)(x) = \infty$;*
(3) *the control set U is a metric space, and the set*

$$\{u \in U(x) \mid g(x,u) + J(f(x,u)) \leq \lambda\}$$

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

Then

(a) *the optimal cost function J^* is the unique fixed point of T within the set $\hat{S}$;*
(b) *we have $T^k J \to J^*$ for all $J \in \hat{S}$;*
(c) *a policy μ is optimal if and only if $T_\mu J^* = TJ^*$. Moreover, there exists an optimal $\hat{S}$-regular policy;*
(d) *for any $J \in \hat{S}$, if $J \le TJ$ we have $J \le J^*$, and if $J \ge TJ$ we have $\hat{J} \ge J^*$;*
(e) *a sequence $\{\mu^k\}$ generated by the PI algorithm starting from an $\hat{S}$-regular policy μ^0 satisfies $J_{\mu^k} \downarrow J^*$.*

Proof. The proof consists of showing that all parts of Assumption 6.1 are satisfied with $\hat{S}$ used in place of S, so Proposition 6.1 applies. Indeed, parts (a) and (e) of this assumption are trivially satisfied, while parts (b)–(d) are the conditions (1)–(3) of the proposition. Then Lemma 6.3 is used to assert that $J^*_{\hat{S}}$ is a fixed point of T. Moreover, Assumption 6.1(f) is shown using the line of proof of Proposition 6.2. In particular, for any $J \in S$, we let $r > 0$ be a scalar such that $J^*_S - re \le J$ [such a scalar exists since $J^*_S \in \hat{S}$ by condition (1)]. Defining $J' = J^* - re$, where $r > 0$ is sufficiently large so that $J' \le J$, we have

$$J' = J^*_S - re = TJ^*_S - re \le T(J^*_S - re) = TJ',$$

so Assumption 6.1(f) holds. Finally the additional assumption needed to apply Proposition 6.1(e) is clearly satisfied in this deterministic problem. ∎

6.3. The case of irregular policies with finite cost. In this section, we consider problems where some S-irregular policies may have finite cost for all states, so Proposition 6.1 cannot be used. We address this issue by introducing a perturbation that allows us to use Proposition 6.1 for the perturbed cost problem, and take the limit as the perturbation vanishes. The idea is that with a perturbation, the cost functions of S-irregular policies may increase disproportionately relative to the cost functions of the S-regular policies, thereby making the problem more amenable to analysis.

In particular, given $p : X \mapsto [0, \infty)$, a nonnegative "perturbation function" of x, for each $\delta \ge 0$ and policy μ, we consider the mappings $T_{\mu,\delta}$ and T_δ given by

$$(T_{\mu,\delta}J)(x) = H\big(x, \mu(x), J\big) + \delta p(x), \quad x \in X, \qquad T_\delta J = \inf_{\mu \in \mathcal{M}} T_{\mu,\delta} J. \tag{6.11}$$

The cost functions of policies $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi$ and $\mu \in \mathcal{M}$, and optimal cost function J^*_δ are

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

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

The following proposition shows that if the δ-perturbed problem is "well-behaved" with respect to a subset of S-regular policies, then its cost function J^*_δ can be used to approximate the optimal cost function over this subset of policies only, and moreover J^*_S is a fixed point of T.

PROPOSITION 6.5. *Given a set $S \subset \mathcal{E}(X)$ and a subset $\widehat{\mathcal{M}}$ of S-regular policies, assume that for every $\delta > 0$,*

(1) *the Bellman equation $J^*_\delta = T_\delta J^*_\delta$ holds for the δ-perturbed problem;*
(2) *for every $\epsilon > 0$, there exists a policy $\mu_\epsilon \in \widehat{\mathcal{M}}$ that is ϵ-optimal for the δ-perturbed problem, i.e., $J_{\mu_\epsilon,\delta} \le J^*_\delta + \epsilon\, e$, where e is the unit function $e(x) \equiv 1$;*

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

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

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

Consider $\hat{J}$, *the optimal cost function over the policies in* $\widehat{\mathcal{M}}$ *only:* $\hat{J} = \inf_{\mu \in \widehat{\mathcal{M}}} J_\mu$.

(a) *We have* $\lim_{\delta \downarrow 0} J^*_\delta = \hat{J}$.

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

$$\lim_{m \to \infty} H(x,u,J_m) = H(x,u,J) \qquad \forall\ x \in X,\ u \in U(x). \tag{6.12}$$

Then J^*_S *is a fixed point of* T, *we have* $J^*_S = \hat{J}$, *and the conclusions of Proposition* 5.1 *hold.*

Proof. (a) By using conditions (2) and (3), we have

$$\hat{J} - \epsilon\, e \leq J_{\mu_\epsilon} - \epsilon\, e \leq J_{\mu_\epsilon,\delta} - \epsilon\, e \leq J^*_\delta \leq J_{\mu,\delta} \leq J_\mu + w_{\mu,\delta} \qquad \forall\ \delta > 0,\ \mu \in \widehat{\mathcal{M}}.$$

By taking the limit as $\epsilon \downarrow 0$, we obtain

$$\hat{J} \leq J^*_\delta \leq J_\mu + w_{\mu,\delta} \qquad \forall\ \delta > 0,\ \mu \in \widehat{\mathcal{M}}.$$

By taking the limit as $\delta \downarrow 0$ and then the infimum over all $\mu \in \widehat{\mathcal{M}}$, it follows that

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

(b) From condition (1) and the fact $J^*_\delta \geq \hat{J}$ shown in part (a), we have for all $\delta > 0$,

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

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

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

Taking the limit as $m \to \infty$, and using (6.12) and the fact $J^*_{\delta_m} \downarrow \hat{J}$ [cf. part (a)], we have

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

so that $T\hat{J} \geq \hat{J}$. Thus $\hat{J}$ is a fixed point of T, and also satisfies $\hat{J} \leq J^*_{\delta_0} \leq J_{\mu_{\delta_0}} \in S$. By Proposition 3.2, we have that $J^*_S = \hat{J}$. It follows that the assumptions of Proposition 5.1 are satisfied. ∎

The preceding proposition applies even if $\lim_{\delta \downarrow 0} J^*_\delta(x) > J^*(x)$ for some $x \in X$. This is illustrated by the deterministic shortest path Example 5.1, for the zero-cycle case where $a = 0$ and $b > 0$. Then for $S = \Re$, we have $J^*_S = b > 0 = J^*$, while the proposition applies because its assumptions are satisfied with $p(x) \equiv 1$. Consistently with the conclusions of the proposition, we have $J^*_\delta = b + \delta$, so $J^*_S = \lim_{\delta \downarrow 0} J^*_\delta$ and J^*_S is a fixed point of T. We refer to [Ber13] and [BeY16] for a more detailed discussion of the approach of this section, applications, examples, and counterexamples, and also for a PI algorithm to find J^*_S, which is based on perturbations. The paper [Ber17b] explores the connections of the perturbation approach of this section with classical notions of feedback control stability. The following example shows how the perturbation approach provides an analysis of linear-quadratic problems, which is consistent with the behavior illustrated in Example 3.1.

Example 6.1 (linear-quadratic optimal control problem). Consider the classical linear-quadratic problem, which involves the deterministic linear system

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

where $x_k \in \Re^n$, $u_k \in \Re^m$ for all k, and A and B are given matrices. The cost function of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ has the form

$$J_\pi(x_0) = \lim_{N\to\infty} \sum_{k=0}^{N-1} \left(x_k' Q x_k + \mu_k(x_k)' R \mu_k(x_k)\right),$$

where x' denotes the transpose of a column vector x, Q is a positive semidefinite symmetric $n \times n$ matrix, and R is a positive definite symmetric $m \times m$ matrix.

The theory of this problem is well known and is discussed in various forms in many sources, including the textbooks [AnM79] and [Ber17a, section 3.1]. The solution revolves around stationary policies μ that are *linear*, in the sense that $\mu(x) = Lx$, where L is some $n \times m$ matrix, and *stable*, in the sense that the matrix $A + BL$ has eigenvalues that are strictly within the unit circle. Thus for a linear stable policy, the closed loop system $x_{k+1} = (A+BL)x_k$ is stable. We assume that there exists at least one linear stable policy.

The solution also revolves around the algebraic matrix Riccati equation

$$P = A' \left(P - PB(B'PB + R)^{-1}B'P\right) A + Q,$$

where the unknown is P, a symmetric $n \times n$ matrix. It is well known that if Q is positive definite, then the Riccati equation has a unique solution P^* within the class of positive semidefinite symmetric matrices, and that the optimal cost function has the form $J^*(x) = x'P^*x$. Moreover, there is a unique optimal policy, and this policy is linear stable (the existence of an optimal linear stable policy can be extended to the case where Q is instead positive semidefinite, but satisfies a certain "detectability" condition; see the textbooks cited earlier).

However, in the general case where Q is positive semidefinite without further assumptions (e.g., $Q = 0$), Example 3.1 shows that the optimal policy need not be stable, and that the optimal cost function over just the linear stable policies may be different than J^*. We address this situation with the aid of the perturbation-based analysis of this section.

The problem can be converted to our abstract format with the identifications $X = \Re^n$, $U(x) \equiv \Re^m$, $\bar{J}(x) \equiv 0$, and

$$H(x, u, J) = x'Qx + u'Ru + J(Ax + Bu).$$

Let S be the set of functions of the form $J(x) = x'Px$, where P is a positive semidefinite symmetric matrix, let $\widehat{\mathcal{M}}$ be the set of linear stable policies, and note that similarly to Example 3.1, every linear stable policy is S-regular. This is due to the fact that for every function $J(x) = x'Px$ and linear stable policy $\mu(x) = Lx$, $(T_\mu^k J)(x_0)$ and $(T_\mu^k \bar{J})(x_0)$ differ by the term $x_0'(A + BL)^{k\prime} P (A + BL)^k x_0$, which vanishes in the limit.

Consider the perturbation function $p(x) = \|x\|^2$. Then for $\delta > 0$, the mapping $T_{\mu,\delta}$ of (6.11) has the form

$$(T_{\mu,\delta}J)(x) = x'(Q + \delta I)x + \mu(x)'R\mu(x) + J\left(Ax + B\mu(x)\right),$$

where I is the identity, and corresponds to the linear-quadratic problem where Q is replaced by the positive definite matrix $Q + \delta I$. This problem admits a quadratic positive definite optimal cost $J^*_\delta(x) = x' P^*_\delta x$, and an optimal linear stable policy. Moreover, the conditions of Proposition 6.5 are satisfied. It follows that J^*_S is equal to the optimal cost over just the linear stable policies $J^*_{\widehat{\mathcal{M}}}$, and is obtained as $\lim_{\delta \to 0} J^*_\delta$, which also implies that $J^*_{\widehat{\mathcal{M}}} = x' \hat{P} x$ where $\hat{P} = \lim_{\delta \to 0} P^*_\delta$.

7. Concluding remarks. We have provided an analysis of challenging abstract DP models based on the notion of regularity. In particular, we have extended this notion to nonstationary policies, and we have highlighted its connection to an earlier development for stationary policies. We have also streamlined and strengthened the corresponding analysis based on PI-related ideas. The main approach is to start from an interesting set of policy-state pairs satisfying a regularity property, and then characterize the region of convergence of VI. We have shown that this approach can lead to new results in the context of a variety of optimal control problems. In addition to the applications described in this paper, our approach has been applied to minimax and exponential cost shortest path problems [Ber15a, Ber16]. Our approach may also be applied to other types of problems that involve a termination state and fit the abstract DP framework of this paper, including SSP game problems [PaB99, Yu11]. These and other related applications are interesting subjects for further research.

Our analysis in this paper focuses on exact forms of DP. However, there are approximation frameworks (such as aggregation and others) that preserve the essential monotonicity property of the DP mapping. For such an approximation setting our analysis applies, but this direction has not been investigated so far, except for the data-perturbed context of section 6.3, which has been analyzed in detail in the paper [BeY16] for the case of an SSP problem.

Acknowledgment. Many helpful discussions with Huizhen (Janey) Yu on the subject of this paper are gratefully acknowledged.

REFERENCES

[AnM79] B.D.O. Anderson and J.B. Moore, *Optimal Filtering*, Prentice Hall, Englewood Cliffs, NJ, 1979.

[BeS78] D.P. Bertsekas and S.E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, NY, 1978, http://web.mit.edu/dimitrib/www/home.html.

[BeT89] D.P. Bertsekas and J.N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[BeT91] D.P. Bertsekas and J.N. Tsitsiklis, *An analysis of stochastic shortest path problems*, Math. Oper. Res., 16 (1991), pp. 580–595.

[BeY16] D.P. Bertsekas and H. Yu, *Stochastic Shortest Path Problems under Weak Conditions*, Laboratory for Information and Decision Systems Report LIDS-2909, MIT, Cambridge, MA, 2016.

[Ber75] D.P. Bertsekas, *Monotone mappings in dynamic programming*, in Proceedings of the 1975 IEEE Conference on Decision and Control, Houston, TX, IEEE, New York, 1975, pp. 20–25.

[Ber77] D.P. Bertsekas, *Monotone mappings with application in dynamic programming*, SIAM J. Control Optim., 15 (1977), pp. 438–464.

[Ber87] D.P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[Ber12] D.P. Bertsekas, *Dynamic Programming and Optimal Control, Vol. II: Approximate Dynamic Programming*, 4th ed., Athena Scientific, Belmont, MA, 2012.

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

[Ber15a] D.P. Bertsekas, *Robust shortest path planning and semicontractive dynamic programming*, Naval Res. Logist., to appear.

[Ber15b] D.P. Bertsekas, *Value and policy iteration in deterministic optimal control and adaptive dynamic programming*, IEEE Trans. Neural Netw. Learn. Syst., to appear.

[Ber16] D.P. Bertsekas, *Affine Monotonic and Risk-Sensitive Models in Dynamic Programming*, Laboratory for Information and Decision Systems Report LIDS-3204, Technical report, MIT, Cambridge, MA, 2016.

[Ber17a] D.P. Bertsekas, *Dynamic Programming and Optimal Control*, Vol. I, 3rd ed., Athena Scientific, Belmont, MA, 2017.

[Ber17b] D.P. Bertsekas, *Stable Optimal Control and Semicontractive Dynamic Programming*, Technical report LIDS-P-3506, MIT, Cambridge, MA, 2017.

[Bla65] D. Blackwell, *Positive dynamic programming*, Proceedings of the Fifth Berkeley Symposium on Mathimatical Statistics and Probability, University of California Press, Berkeley, CA, 1965, pp. 415–418.

[CaR14] O. Çavuş and A. Ruszczyński, *Risk-averse control of undiscounted transient Markov models*, SIAM J. Control Optim., 52 (2014), pp. 3935–3966.

[DeR79] E.V. Denardo and U.G. Rothblum, *Optimal stopping, exponential utility, and linear programming*, Math. Program., 16 (1979), pp. 228–244.

[Den67] E.V. Denardo, *Contraction mappings in the theory underlying dynamic programming*, SIAM Rev., 9 (1967), pp. 165–177.

[Der70] C. Derman, *Finite State Markovian Decision Processes*, Academic Press, N.Y, 1970.

[Fei02] E.A. Feinberg, *Total reward criteria*, in Handbook of Markov Decision Processes, E.A. Feinberg and A. Shwartz, eds., Springer, NY, 2002.

[HeL99] O. Hernandez-Lerma and J.B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer, NY, 1999.

[Kal83] L.C.M. Kallenberg, *Linear Programming and Finite Markov Control Problems*, Mathematical Centre Tracks 148, Amsterdam, 1983.

[PaB99] S.D. Patek and D.P. Bertsekas, *Stochastic shortest path games*, SIAM J. Control Optim., 37 (1999), pp. 804–824.

[Pal67] R. Pallu de la Barriere, *Optimal Control Theory*, Saunders, Philadelphia, 1967.

[Pat01] S.D. Patek, *On terminating Markov decision processes with a risk averse objective function*, Automatica J. IFAC, 37 (2001), pp. 1379–1386.

[Put94] M.L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, New York, 1994.

[Str66] R. Strauch, *Negative dynamic programming*, Ann. Math. Statist., 37 (1966), pp. 871–890.

[Van81] J. Van der Wal, *Stochastic Dynamic Programming*, Thesis, Mathematisch Centrum, Amsterdam, 1981.

[Whi82] P. Whittle, *Optimization Over Time*, Vol. 1, Wiley, New York, 1982.

[YuB15] H. Yu and D.P. Bertsekas, *A mixed value and policy iteration method for stochastic control with universally measurable policies*, Math. Oper. Res., 40 (2015), pp. 926–968.

[Yu11] H. Yu, *Stochastic Shortest Path Games and Q-Learning*, Technical report, Laboratory for Information and Decision Systems Report LIDS-P-2875, MIT, Cambridge, MA, 2011.

[Yu15] H. Yu, *On convergence of value iteration for a class of total cost Markov decision processes*, SIAM J. Control Optim., 53 (2015), pp. 1982–2016.