Regular Policies in
Abstract Dynamic Programming

Dimitri P. Bertsekas†

Abstract

We consider an abstract dynamic programming model, and analysis based on regular policies that are well-behaved with respect to value iteration. Regular policies 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 fixed point and value iteration properties, which the optimal cost function over all policies need not have. We accordingly develop a methodology that can deal with long standing analytical and algorithmic issues in undiscounted dynamic programming models, such as stochastic and minimax shortest path, positive cost, negative cost, mixed positive-negative cost, risk-sensitive, and multiplicative cost problems. Among others, we use our approach to obtain new results for convergence of value and policy iteration in deterministic discrete-time optimal control with nonnegative cost per stage.

1. INTRODUCTION

The purpose of this paper is to address issues relating to the value and policy iteration algorithms, and their application in infinite horizon total cost discrete-time optimal control. We do this by expanding the methodology of abstract dynamic programming (DP for short), which aims to unify the analysis of DP models and to highlight their significant structures.

To provide some context for our analysis, let us note two types of established abstract DP 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 cost functions. 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], Ch. 5), for which the abstract mapping is not a contraction of any kind but is instead monotone. Among others, these models cover the important cases of nonpositive and nonnegative cost DP problems of [Bla65] and [Str66], respectively.

In this paper we focus on semicontractive models, which were introduced in the recent monograph [Ber13]. There are several such models with variations in their characteristics, but they all share the property that the abstract mapping corresponding to some policies has a contraction-like property, but the mapping of others does not. For a stationary policy, a central property in this connection is called S-regularity, where S is a set of cost functions. This property, defined formally in Section 3, 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.

† Dimitri Bertsekas is with the Dept. of Electr. Engineering and Comp. Science, and the Laboratory for Information and Decision Systems, M.I.T., Cambridge, Mass., 02139. Many helpful discussions with Huizhen (Janey) Yu on the subject of this paper are gratefully acknowledged.
A prominent case where regularity concepts are central is \(n\)-state stochastic shortest path problems. 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, and some stationary policies called \emph{proper} are guaranteed to terminate starting from every initial state, while others called \emph{improper} are not. This problem has been discussed in many sources, including the books [Pal67], [Der70], [Whi82], [Ber87], [BeT89], [BeT91], [HeL99], and [Ber12], where it is sometimes referred to by earlier names such as “first passage problem” and “transient programming problem.” Here the proper policies involve a (weighted sup-norm) contraction mapping and are regular (see the discussion in Sections 3.1), while the improper ones are not.

The notion of \(S\)-regularity of a stationary policy is patterned after the notion of a proper policy, and was used extensively in [Ber13] and in the subsequent paper [Ber14], as an analytical vehicle to investigate the existence and uniqueness of solution of the abstract form of Bellman’s equation, and the convergence properties of value and policy iteration, in a variety of undiscounted stochastic and minimax problems. Related analysis was given in the context of stochastic shortest path problems in our recent paper [BeY15]. A key idea is that the optimal cost function over \(S\)-regular policies only, call it \(J^*_S\), is the natural limit of value iteration starting from functions \(J\) within \(S\) that satisfy \(J \geq J^*_S\), a property that the optimal cost function over all policies \(J^*\) may lack. This is analytically convenient both in the case where \(J^*_S = J^*\), and also in the case where \(J^*_S \neq J^*\), which arises for example in shortest path problems with zero length cycles (see [BeY15]).

Our purpose in 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 value and policy iteration in semi-contractive models. We show that for important special cases of optimal control problems, including some that are discussed in Section 4, our approach yields substantial improvements over the current state of the art, and highlights the fundamental convergence mechanism of value iteration and, to some extent, policy iteration.

The paper is organized as follows. After formulating our abstract DP model in Section 2, we develop the main ideas of our approach in Section 3, and illustrate them in a variety of contexts, including monotone increasing and monotone decreasing models. In Section 4, we focus on monotone increasing models. In Section 4 we apply our results to some important classes of optimal control problems. In particular, in Section 4.1 we establish the convergence of value and policy iteration under new and easily verifiable conditions in a fundamental class of undiscounted deterministic optimal control problems with nonnegative cost per stage and a terminal set of states. In Section 4.2, we will discuss the convergence of value iteration in undiscounted nonnegative cost stochastic optimal control problems, following the analysis of the paper [YuB13]. In Section 4.3, we consider a discounted version of the problem of Section 4.2.

2. ABSTRACT DYNAMIC PROGRAMMING MODEL

We review the abstract DP model as described in Section 3.1 of [Ber13]. Let \(X\) and \(U\) be two sets, which we loosely refer to as a set of “states” and a set of “controls,” respectively. For each \(x \in X\), let \(U(x) \subset U\) be a nonempty subset of controls that are feasible at state \(x\). We denote by \(\mathcal{M}\) the set of all functions \(\mu : X \to U\) with \(\mu(x) \in U(x)\), for all \(x \in X\).

In analogy with DP, we consider policies, which are sequences \(\pi = \{\mu_0, \mu_1, \ldots\}\), with \(\mu_k \in \mathcal{M}\) for all \(k\). We denote by \(\Pi\) the set of all policies. We refer to a sequence \(\{\mu, \mu, \ldots\}\), with \(\mu \in \mathcal{M}\), as a \emph{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 \( \mathbb{R} \) the set of real numbers, by \( R(X) \) the set of real-valued functions \( J : X \to \mathbb{R} \), and by \( E(X) \) the subset of extended real-valued functions \( J : X \to \mathbb{R} \cup \{-\infty, \infty\} \). We denote by \( E^+(X) \) the set of all nonnegative extended real-valued functions of \( x \in X \). Throughout the paper, when we write \( \lim \), \( \lim \sup \), or \( \lim \inf \) of a sequence of functions we mean it to be pointwise. We also write \( J \downarrow J \) to mean that \( J_k(x) \to J(x) \) for each \( x \in X \), and we write \( J_k \downarrow J \) if \( \{J_k\} \) is monotonically nonincreasing and \( J_k \to J \).

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

**Assumption 2.1: (Monotonicity)** If \( J, J' \in E(X) \) and \( J \leq J' \), then

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

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

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

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

\[
(T_\mu J)(x) = H(x, \mu(x), J), \quad \forall x \in X, J \in E(X).
\]

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

\[
\begin{align*}
J \leq J' & \quad \Rightarrow \quad T^k J \leq T^k J', & \forall \mu \in \mathcal{M}, \\
J \leq TJ & \quad \Rightarrow \quad T^k J \leq T^{k+1} J, & \forall \mu \in \mathcal{M},
\end{align*}
\]

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

\[
(T_{\mu_0} \cdots T_{\mu_k} J)(x) = (T_{\mu_0}(T_{\mu_1} \cdots (T_{\mu_{k-1}}(T_{\mu_k} J) \cdots )))(x), \quad \forall x \in X.
\]

We now consider cost functions associated with \( T_\mu \) and \( T \). We introduce a function \( \bar{J} \in E(X) \), and we define the infinite horizon cost of a policy as the upper limit of its finite horizon costs with \( \bar{J} \) being the cost function at the end of the horizon.

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

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

(2.1)

The optimal cost function \( J^* \) is defined by

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

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 essentially all the interesting total cost infinite horizon DP problems, including stochastic and minimax, discounted and undiscounted, semi-Markov, multiplicative, risk-sensitive, etc (see [Ber13]). As an example, for a deterministic discrete-time optimal control problem involving the system
\begin{equation}
    x_{k+1} = f(x_k, u_k), \quad k = 0, 1, \ldots,
\end{equation}
and a cost \( g(x_k, u_k) \) for the \( k \)th stage, the mapping \( H \) is given by
\begin{equation}
    H(x, u, J) = g(x, u) + J(f(x, u)), \quad x \in X, \ u \in U(x),
\end{equation}
and \( \bar{J} \) is the zero function \([\bar{J}(x) \equiv 0]\). It can be seen that the cost function of a policy \( \pi \), as given by Eq. (2.1), takes the form
\begin{equation}
    J_{\pi}(x) = \limsup_{k \to \infty}(T_{\mu_0} \cdots T_{\mu_k} \bar{J})(x) = \limsup_{k \to \infty} \sum_{t=0}^{k} g(x_t, \mu_t(x_t)),
\end{equation}
since \((T_{\mu_0} \cdots T_{\mu_k} \bar{J})(x)\) is the cost of the first \( k + 1 \) periods using \( \pi \) starting from \( x \), and with terminal cost \( 0 \) (the value of \( \bar{J} \) at the terminal state).

For an undiscounted stochastic problem involving a Markov chain with state space \( X = \{1, \ldots, n\} \), transition probabilities \( p_{xy}(u) \) and expected one-stage cost function \( g \), the mapping \( H \) is given by
\begin{equation}
    H(x, u, J) = g(x, u) + \sum_{y=1}^{n} p_{xy}(u)J(y), \quad x \in X, \ J \in E(X),
\end{equation}
(with the convention \( \infty - \infty = \infty \) if \( J \) is extended real-valued). The stochastic shortest path problem arises when one of the states is cost-free and absorbing, and the cost function provides an incentive to reach that state. The following is a more general undiscounted 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 - Undiscounted 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
\begin{equation}
    x_{k+1} = f(x_k, \mu_k(x_k), w_k), \quad k = 0, 1, \ldots,
\end{equation}
where \( w_k \) is a random disturbance that takes values from a space \( W \). We assume that \( w_k \), \( k = 0, 1, \ldots \), are characterized by probability distributions \( P(\cdot \mid x_k, u_k) \) that are identical for all \( k \), where \( P(w_k \mid x_k, u_k) \) is the probability of occurrence of \( w_k \), when the current state and control are \( x_k \) and \( u_k \), respectively. Thus the probability of \( w_k \) may depend explicitly on \( x_k \) and \( u_k \), but not on values of prior disturbances \( w_{k-1}, \ldots, w_0 \).

We allow infinite state and control spaces, as well as problems with discrete (finite or countable) state space (in which case the underlying system is a Markov chain). However, for technical reasons that relate to measure theoretic issues, we assume that \( W \) is a countable set.

Given an initial state \( x_0 \), we want to find a policy \( \pi = \{\mu_0, \mu_1, \ldots\} \), where \( \mu_k : X \to U \), \( \mu_k(x_k) \in U(x_k) \), for all \( x_k \in X \), \( k = 0, 1, \ldots \), that minimizes
\begin{equation}
    J_{\pi}(x_0) = \limsup_{k \to \infty} E \left\{ \sum_{t=0}^{k} g(x_t, \mu_t(x_t), w_t) \right\},
\end{equation}
subject to the system equation constraint (2.6), where \( g \) is the one-stage cost function. 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

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

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

A discounted version of the problem is defined by the mapping

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

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

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

and will be discussed in Section 4.3.

### 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 of the major abstract 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 \( E(X) \); for example, contractive models where \( T_\mu \) is a contraction mapping for all \( \mu \in M \), with respect to some norm and with a common modulus of contraction. However, in general \( T \) may have multiple fixed points within \( E(X) \), including for some popular DP problems, while in exceptional cases, \( J^* \) may not be among the fixed points of \( T \).

A related question is the convergence of value iteration (VI for short). This is the fixed point algorithm that generates \( T^k J \), \( k = 0, 1, \ldots \), starting from a function \( J \in E(X) \). Generally, for abstract DP models where \( J^* \) is a fixed point of \( T \), VI converges to \( J^* \) starting from within some subset of initial functions \( J \), but not 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 \( S \subset E(X) \), we say that a collection \( \mathcal{C} \) of policy-state pairs \((\pi, x)\), with \( \pi \in \Pi \) and \( x \in X \), is \( S\text{-regular} \) if

\[
J_\pi(x) = \limsup_{k \to \infty} (T_{\mu_0} \cdots T_{\mu_k} J)(x), \quad \forall (\pi, x) \in \mathcal{C}, \ J \in S.
\]
A nonempty set \( C \) of policy-state pairs \((\pi,x)\) may be \( S \)-regular for many different sets \( S \). The largest such set is

\[
S_C = \left\{ J \in E(X) \mid J_\pi(x) = \limsup_{k \to \infty} (T_{\mu_0} \cdots T_{\mu_k} J)(x), \quad \forall \ (\pi,x) \in C \right\},
\]

and for any nonempty \( S \subset S_C \), we have that \( C \) is \( S \)-regular. Moreover, the set \( S_C \) is nonempty, since it contains \( J \). For a given nonempty set \( C \) of policy-state pairs \((\pi,x)\), let us consider the function \( J_C^* \in E(X) \), given by

\[
J_C^*(x) = \inf_{J \in S_C} J_\pi(x), \quad x \in X.
\]

(3.1)

Note that \( J_C^* \geq J^* \) [if for some \( x \in X \), the set of policies \( \{ \pi \mid (\pi,x) \in C \} \) is empty, we have \( J_C^*(x) = \infty \)]. We will try to characterize the sets of fixed points of \( TJ \) such a set is \( J \).

Proposition 3.1: Given a set \( S \subset E(X) \), let \( C \) be a collection of policy-state pairs \((\pi,x)\) that is \( S \)-regular.

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

\[
\liminf_{k \to \infty} T^k J \leq \limsup_{k \to \infty} T^k J \leq J_C^*.
\]

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

\[
J' \leq \liminf_{k \to \infty} T^k J \leq \limsup_{k \to \infty} T^k J \leq J_C^*.
\]

(3.2)

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

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

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

\[
\liminf_{k \to \infty} (T^k J)(x) \leq \limsup_{k \to \infty} (T^k J)(x) \leq \limsup_{k \to \infty} (T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x) = J_\pi(x), \quad \forall \ (\pi,x) \in C, \ J \in S,
\]

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

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

\[
J'(x) \leq (TJ')(x) \leq \cdots \leq (T^k J')(x) \leq (T^k J)(x) \leq (T^k \hat{J})(x).
\]

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

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

\[
\{ J \in E(X) \mid J' \leq J \leq \hat{J} \text{ for some } \hat{J} \in S \},
\]

6
From Prop. 3.1(b), with monotonicity of $T$ follows that VI converges to $J$. $J$ does not address the question whether particular applications to be discussed in Section 4 we will use such a choice. Example in the case where $T$ we obtain shortest path, nonpositive cost, and nonnegative cost problems.

Examples for the most common infinite horizon stochastic optimal control problems: discounted, stochastic [Der70], [Whi82], [Put94], [HeL99], [Ber12], [Ber13] provide extensive analysis, examples, and counterexamples for the most common infinite horizon stochastic optimal control problems: discounted, stochastic shortest path, nonpositive cost, and nonnegative cost problems.

Proposition 3.2: Given a set $S \subset E(X)$, let $C$ be a collection of policy-state pairs $(\pi, x)$ that is $S$-regular, and assume that $J_C^*$ is a fixed point of $T$. Then $J_C^*$ is the only possible fixed point of $T$ within the set of all $J \in E(X)$ such that $J_C^* \leq J \leq \hat{J}$ for some $\hat{J} \in S$. Moreover, $T^k J \to J_C^*$ for all $J \in E(X)$ such that $J_C^* \leq J \leq \hat{J}$ for some $\hat{J} \in S$.

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

$$J_C^* = T^k J_C^* \leq T^k J \leq T^k \hat{J}, \quad k = 0, 1, \ldots.$$  

From Prop. 3.1(b), with $J' = J_C^*$, it follows that $T^k \hat{J} \to J_C^*$, so taking limit in the above relation as $k \to \infty$, we obtain $T^k J \to J_C^*$. Q.E.D.

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

Note that Prop. 3.2 does not say anything about fixed points of $T$ that lie below $J_C^*$. In particular, it does not address the question whether $J^*$ is a fixed point of $T$, or whether VI converges to $J^*$ starting from $J$ 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 specific 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, examples, and counterexamples for the most common infinite horizon stochastic optimal control problems: discounted, stochastic shortest path, nonpositive cost, and nonnegative cost problems.

† For this statement to be meaningful, the set $\{\hat{J} \in E(X) \mid J^* \leq \hat{J}\}$ must be nonempty. Generally, it is possible that this set is empty, even though $S$ is assumed nonempty.
In particular, for discounted problems [the case of the mapping (2.8) 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 stochastic shortest path 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 [BeY15] also considers finite-state stochastic shortest path problems, but without the favorable assumptions of [BeT91] just noted. Instead it was assumed 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 [BeY15] with an example. In the context of the present paper, a useful choice is to take

\[
\mathcal{C} = \{(\mu, x) \mid \mu \text{: proper}\},
\]

in which case \( J_C^* \) is the optimal cost function that can be achieved using proper policies only. It was shown in [BeY15] that \( J_C^* \) is a fixed point of \( T \), so by Prop. 3.2, VI converges to \( J_C^* \) starting from any real-valued \( J \geq J_C^* \). The line of analysis of [BeY15] is related to the proof of Prop. 3.1.

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 [YuB13]. In particular it was shown that \( J^* \) is the unique fixed point of \( T \) within the class of all functions \( J \in E(X) \) that satisfy

\[
0 \leq J \leq cJ^* \quad \text{for some } c > 0.
\] (3.3)

Moreover it was shown that VI converges to \( J^* \) starting from any function satisfying the condition

\[
J^* \leq J \leq cJ^* \quad \text{for some } c > 0,
\]

and under certain standard compactness conditions, starting from any \( J \) that satisfies Eq. (3.3) (we refer to [YuB13] for discussion and references to antecedents of this result). It turns out that one may prove the results just mentioned by using Prop. 3.2, with an appropriate choice of \( \mathcal{C} \). The proof uses the arguments of Appendix E of [YuB13], and will be given in Section 4.2.

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 \( J_\pi \) is defined by Eq. (2.1) as a limit. The paper [Yu14] 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 Example 3.2 of [Van81], Example 6.10 of [Fei2], and Example 4.1 of [Yu14]). The analysis of [Yu14] may also be used to bring to bear Prop. 3.1 on the problem, but this analysis is beyond our scope in this paper.

**The Case Where \( J_C^* \leq \bar{J} \)**

It is well known that the results for nonnegative cost and nonpositive cost infinite horizon stochastic optimal control problems are markedly different. In particular, 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 $T_\mu J \geq \bar{J}$ and $T_\mu \bar{J} \leq \bar{J}$ for all $\mu \in \mathcal{M}$, respectively (see [Ber13], Ch. 4). In the context of regularity, with $C$ being $S$-regular, it turns out that there are analogous significant differences between the cases $J^*_C \geq \bar{J}$ and $J^*_C \leq \bar{J}$. The favorable aspects of the condition $J^*_C \geq \bar{J}$ will be seen later in the context of PI, where it guarantees the monotonic improvement of the policy iterates (see the subsequent Prop. 3.6). The following proposition establishes some favorable aspects of the condition $J^*_C \leq \bar{J}$ 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 $C$, so $\bar{J}$ can serve as the element $J$ of $S$ with $J^*_C \leq \bar{J}$ in Props. 3.1 and 3.2 (see the proof of the following proposition).

Proposition 3.3: Given a set $S \subset E(X)$, let $C$ be a collection of policy-state pairs $(\pi, x)$ that is $S$-regular, and assume that $J^*_C \leq \bar{J}$. Then:

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

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

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

Proof: (a) If $S$ does not contain $\bar{J}$, we can replace $S$ with $\tilde{S} = S \cup \{\bar{J}\}$, and $C$ will still be $\tilde{S}$-regular. By applying Prop. 3.1(b) with $S$ replaced by $\tilde{S}$ and $\bar{J} = \bar{J}$, the result follows.

(b) Assume without loss of generality that $\bar{J} \in S$ [cf. the proof of part (a)]. By using Prop. 3.2 with $\bar{J} = \bar{J}$, we have $J^*_C = \lim_{k \to \infty} T^k \bar{J}$. This relation yields for any policy $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi$,

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

so by taking the infimum over $\pi \in \Pi$, we obtain $J^*_C \leq J^*$. Since generically we have $J^*_C \geq J^*$, it follows that $J^*_C = J^*$. Finally, from Prop. 3.2, we obtain $T^k \bar{J} \to J^*$ for all $J \in E(X)$ such that $J^* \leq J \leq \bar{J}$ for some $\bar{J} \in S$. Q.E.D.

As a special case of the preceding proposition, we have that if $J^* \leq \bar{J}$ and $J^*$ is a fixed point of $T$, then $J^* = \lim_{k \to \infty} T^k \bar{J}$, and for every other fixed point $J'$ of $T$ we have $J' \leq J^*$ (apply the proposition with $C = \Pi \times X$ and $S = \{\bar{J}\}$, in which case $J^*_C = J^* \leq \bar{J}$). 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, dating to [Bla65], 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 cost $E \{g(x, u, w)\} \geq 0$ for all $u$ except one that leads to a termination state with probability 1 and nonpositive cost. The results of Prop. 3.3, specialized to this case, have been obtained from Prop. 2.4 of the paper [BeY15].
Note that without the assumption $J^*_c \leq \bar{J}$ in the preceding proposition, it is possible that $T^k\bar{J}$ does not converge to $J^*$, even if $J^*_c = J^* = TJ^*$; see Example 4.1 of the present paper, which involves an infinite control space and nonnegative cost per stage, and the deterministic finite-state finite-control shortest path problem in Example 4.1 of [Yu14], which involves both positive and negative costs. The possibility that $T^k\bar{J}$ may not converge to $J^*$ when $\bar{J} \leq J^*$ is also well known in the theory of nonnegative cost infinite horizon stochastic optimal control [Str66].

A Deterministic Shortest Path Example

The following example illustrates the fine points of our analysis under exceptional circumstances. It highlights the intricacies of showing convergence of the VI sequence $T^kJ$ to $J^*$ from a broad range of initial conditions $J$, and the significance of assumptions such as $J^* \geq \bar{J}$ and $J^* \leq \bar{J}$.

![Figure 3.1](image)

A deterministic shortest path problem with a single node 1 and a termination node $t$. At 1 there are two choices; a self-transition, which costs 0, and a transition to $t$, which costs $b$.

Example 3.1

This is a deterministic shortest path problem, involving a termination state $t$ and a single state 1, with a choice between staying at 1 at cost 0, and moving to $t$ at cost $b$; cf. Fig. 3.1. To transcribe the problem into the abstract formalism, let $X = \{1\}$, so $J \in E(X) = \mathbb{R} \cup \{-\infty, \infty\}$. For convenience we will write $J$ in place of $J(1)$. The control space $U$ consists of two choices $u$ and $u'$, where the mapping $H$ has the form

$$H(1, u, J) = J, \quad H(1, u', J) = b, \quad \forall J \in E(X),$$

with $b \in \mathbb{R}$, and $\bar{J} = 0$. Then the Bellman equation $J = TJ$ is written as

$$J = \min \{H(1, u, J), H(1, u', J)\} = \min \{J, b\},$$

which is the familiar equation for the shortest path distance $J$ from 1 to the destination. The mapping $T$ is given by $TJ = \min \{J, b\}$, and the set of its fixed points is $[-\infty, b]$.

There are two stationary policies, $\mu$ and $\mu'$, where $\mu(1) = u$ and $\mu'(1) = u'$. For the policy $\mu$ we have $J_{\mu} = 0$ (this is the policy that cycles around the state 1 in the shortest path problem of Fig. 3.1). Every policy $\pi$ other than $\mu$ has cost $J_\pi = b$. Thus, if $b > 0$, $\mu$ is optimal, and we have $J^* = 0$, while if $b < 0$, every policy other than $\mu$ is optimal, and we have $J^* = b$. We consider the case $b > 0$ (a monotone increasing model with $J^* \geq \bar{J}$) and the case $b < 0$ (a monotone decreasing model with $J^* \leq \bar{J}$):

1. Let $b > 0$ and $C$ be the set of all $(\pi, 1)$ where $\pi$ is other than $\mu$, which is $E(X)$-regular. Then $J^*_c = b > 0 = J^*$, so the valid start region is $[b, \infty]$, while the limit region corresponding to a choice $J' \leq b$ (which
is a fixed point of \(T\) is \([J', b]\). Here VI converges to \(J_c^* = b\) (in a single iteration) starting from the valid start region, but not to \(J^* = 0\).

(2) Let \(b > 0\) and \(C = \Pi \times X\), which is \(\{0\}\)-regular. Then \(J_c^* = J^* = 0\). Here VI converges to \(J^*\) from within the valid start region, but this region consists of just \(J^*\), so this is not a very useful type of result.

(3) Let \(b < 0\) and \(C\) be the set of all \((\pi, 1)\) where \(\pi\) is other than \(\mu\), which is \(E(X)\)-regular. Then \(J_c^* = b = J^*\), so the valid start region is \([b, \infty]\), and VI converges to \(J^*\) from within the region.

(4) Let \(b < 0\) and \(C = \{\mu\} \times X\), which is \(\{0\}\)-regular. Then, \(J_c^* = 0 > b = J^*\), so the valid start region is \(\{0\}\), and VI converges to \(J^*\) from within the region, as well as from \(J \geq b\) with \(J \neq 0\), which do not belong to this region.

(5) Let \(b < 0\) and \(C = \Pi \times X\), which is \(\{0\}\)-regular. Then \(J_c^* = J^* = b\), so the convergence of VI is the same as in case (4).

In the cases where \(b < 0\), it can also be verified that PI can oscillate between the stationary polices \(\mu\) and \(\mu'\). This is consistent with our discussion of PI in Section 3.3. Let us also consider the variant of the problem where \(u\) has cost -1 instead of 0. Then the self-cycle has negative cost and we have \(J_c^* = -1\), so Prop. 3.2 does not apply.

The preceding example shows the importance of choosing properly the set \(C\) in order to obtain meaningful results. Note, however, that in a given problem the interesting choices of \(C\) are usually limited, and that the propositions of this section can guide a favorable choice. One useful approach is to try the set

\[
C = \{ (\pi, x) \mid J_\pi(x) < \infty \},
\]

so that \(J_c^* = J^*\). By the definition of regularity, if \(S\) is any subset of the set

\[
S_C = \left\{ J \in E(X) \mid J_\pi(x) = \limsup_{k \to \infty}(T_{\mu_0} \cdots T_{\mu_k} J)(x), \forall (\pi, x) \in C \right\},
\]

then \(C\) is \(S\)-regular. One may then try to derive a suitable subset of \(S_C\) that admits an interesting characterization. This is the approach followed in Sections 4.1-4.3. A variation of this approach is to focus on

an interesting subset \(\overline{M}\) of stationary policies such that for the set

\[
C = \overline{M} \times X,
\]

we have \(J_c^* = J^*\). This is the approach followed for several applications in [Ber13], [Ber14a], and [Ber14b], including various types of deterministic, stochastic, and minimax shortest path problems, where \(\overline{M}\) is chosen to be the set of proper policies. In the next section we focus on analysis based on stationary policies.
3.1. S-Regular Stationary Policies

We will now specialize the notion of $S$-regularity to stationary policies with the following definition, and obtain results that are useful in a variety of contexts, including policy iteration.

**Definition 3.2:** For a nonempty set of functions $S \subset E(X)$, we say that a stationary policy $\mu$ is $S$-regular if $T^k_\mu J \to J_\mu$ 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 $\mu$ is $S$-regular if the set $C = \{(\mu, x) \mid x \in X\}$ is $S$-regular. As a corollary of Prop. 3.2, we obtain the following proposition, given also as Prop. 3.1.2 in [Ber13]. The proposition is useful when it is known that for a given set $S$, there exists an $S$-regular policy that is optimal over all policies. In some cases this assertion may be natural and easily proved.

**Proposition 3.4:** Given a set $S \subset E(X)$, assume that $J^*$ is a fixed point of $T$, and that there exists an $S$-regular policy that is optimal (i.e., an $S$-regular $\mu^*$ such that $J_{\mu^*} = J^*$). Then:

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

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

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

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

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

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

The following proposition extends the preceding one and provides optimality conditions for a policy $\mu^*$ to be $\mathcal{M}_S$-optimal.

† The definition of $S$-regularity of a stationary policy $\mu$ in [Ber13] is slightly more restrictive, and includes the conditions $J_\mu \in S$ and $J_\mu = T_\mu J_\mu$, in addition to $T^k_\mu J \to J_\mu$ for all $J \in S$. The definition given here allows more refined results (since the conditions $J_\mu \in S$ and $J_\mu = T_\mu J_\mu$ can be included in the statements of various propositions only when needed), and is consistent with the Definition 3.1 for $S$-regularity of a set $C$. 

12
Proposition 3.5: Given a set $S \subset E(X)$, assume that there exists at least one $S$-regular policy and that $J^*_S$ is a fixed point of $T$. Then:

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

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

(c) If $\mu^*$ is $S$-regular, $J^*_S \in S$, and $T_{\mu^*} J^*_S = T J^*_S$, then $\mu^*$ is $M_S$-optimal. Conversely, if $\mu^*$ is $M_S$-optimal and $J_{\mu^*} = T_{\mu^*} J^*_{\mu^*}$, then $T_{\mu^*} J^*_S = T J^*_S$.

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

$$\mathcal{C} = M_S \times X$$

is $S$-regular, and we have

$$J^*_S = J^*_C \geq J^*.$$  \hspace{1cm} (3.4)

The results of parts (a) and (b) follow from Prop. 3.2 with the above definition of $\mathcal{C}$.

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

$$T^2_{\mu^*} J^*_S = T_{\mu^*} (T J^*_S) = T_{\mu^*} J^*_S = T J^*_S = J^*_S,$$

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

$$J^*_S = \lim_{k \rightarrow \infty} T^k_{\mu^*} J^*_S = J^*_{\mu^*},$$

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

$$T J^*_S = J^*_S = J^*_{\mu^*} = T_{\mu^*} J^*_S = T_{\mu^*} J^*_S.$$

Q.E.D.

The preceding proposition and the notion of $S$-regularity of a stationary policy, can be used as the basis for the analysis of $n$-state stochastic shortest path problems (where $n$-regular policies are the proper policies; see the papers [BeT91] and [BeY15]), and also shortest path problems of the minimax type (see [Ber14a], where the definition of a proper policy is appropriately adapted to the minimax context). Further discussion of this topic is beyond the scope of the present paper, but we refer to [Ber13], Section 3.2, and [Ber14a] for analysis.
3.2. Convergence of Policy Iteration

We will now consider policy iteration (PI for short) and its convergence properties. The idea is to generate an improving sequence of policies whose cost functions converge monotonically to some $J^\infty$ that satisfies $J^\infty \geq J^*$ and will be shown to be a fixed point of $T$ under simple conditions. If the preceding analysis can be used to guarantee that $J^*$ is the “largest” fixed point of $T$, it will then follow that $J^\infty = J^*$.

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

$$T_{\mu^{k+1}}J_{\mu^k} = TJ_{\mu^k}. \quad (3.5)$$

This iteration embodies both the policy evaluation step, which computes $J_{\mu^k}$, and the policy improvement step, which computes $\mu^{k+1}$ via the minimization over $U(x)$ for each $x$, which is implicit in Eq. (3.5). We will assume that this minimization can be carried out, so that the algorithm is well-defined. The evaluation of a stationary $\mu$ will ordinarily be done by solving the equation $J_{\mu} = TJ_{\mu}$, which holds for most models of interest, and which we will assume in our analysis (under exceptional circumstances we may have $J_{\mu} \neq TJ_{\mu}$; for an example involving a stochastic shortest path problem with both positive and negative transition costs, see [BeY15]).

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

$$T_{\mu^k}J_k = TJ_k, \quad J_{k+1} = T_{\mu^k}^{m_k}J_k, \quad k = 0, 1, \ldots, \quad (3.6)$$

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

**Proposition 3.6: (Convergence of PI)** Assume that:

1. For all $\mu \in M$, we have $J_\mu = T_{\mu^0}J_\mu$ and there exists $\bar{\mu} \in M$ such that $T_{\bar{\mu}}J_\mu = TJ_\mu$.

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

$$H(x, u, J) \geq \lim_{m \rightarrow \infty} H(x, u, J_m), \quad \forall x \in X, \ u \in U(x). \quad (3.7)$$

Then the PI algorithm (3.5) is well defined and the following hold:

(a) If $J^* \geq \bar{J}$, then the sequence $\{\mu^k\}$ generated by the PI algorithm (3.5) satisfies $J_{\mu^k} \downarrow J^\infty$, where $J^\infty$ is a fixed point of $T$ with $J^\infty \geq J^*$.

(b) If for a set $S \subset E(X)$, the policies $\mu^k$ generated by the PI algorithm (3.5) are $S$-regular and we have $J_{\mu^k} \in S$ for all $k$, then $J_{\mu^k} \downarrow J^*_S$ and $J^*_S$ is a fixed point of $T$. 

14
Proof: (a) Condition (1) of the proposition guarantees that the PI algorithm is well defined. We first show that the condition $J^* \geq \bar{J}$ implies a generic cost improvement property of PI. If $\mu$ is a policy and $\bar{\mu}$ satisfies $T_{\bar{\mu}} J_{\mu} = TJ_{\mu}$, we have

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

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

$$J_{\mu} \geq TJ_{\mu} \geq J_{\bar{\mu}}. \quad (3.8)$$

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

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

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

$$J_{\infty} \geq \lim_{k \to \infty} TJ_{\mu^k} \geq TJ_{\infty}, \quad (3.9)$$

where the second inequality follows from the fact $J_{\mu^k} \geq J_{\infty}$. Using Eq. (3.7), we also have for all $x \in X$ and $u \in U(x)$,

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

By taking the infimum of the left-hand side over $u \in U(x)$, we obtain $TJ_{\infty} \geq J_{\infty}$, which combined with Eq. (3.9), yields $J_{\infty} = TJ_{\infty}$.

(b) We show again a generic cost improvement property of PI. If $\mu$ and $\bar{\mu}$ are S-regular policies, with $J_{\mu} \in S$ and $T_{\bar{\mu}} J_{\mu} = TJ_{\mu}$, we have

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

where the last inequality follows from the monotonicity of $T_{\bar{\mu}}$ and the last equality follows from the assumptions that $\mu$ is S-regular and $J_{\mu} \in S$. It follows similarly to part (a) that $J_{\mu^k} \downarrow \bar{J}$ where $\bar{J} \in S$ is a fixed point of $T$. By applying Prop. 3.1(b) with $J^* = J_{\infty}$, $J = J_{\mu^k}$, and $C$ equal to $M_X$, we have $J_{\infty} \leq \bar{J}$.

Condition (1) of the proposition holds for most DP models of interest, and the same is true for condition (2), which is a technical continuity-type assumption. The condition $J^* \geq \bar{J}$ in part (a) is essential for showing the cost improvement property (3.8) in the preceding proof (if cost improvement can be shown independently, the condition $J^* \geq \bar{J}$ is not needed). The paper [BeY15] provides an example of a two-state deterministic shortest path problem where this condition is violated, and the PI algorithm (3.5) oscillates between an optimal and a suboptimal policy. This example, translated to the terminology of the present paper, is the same as Example 3.1 with $b < 0$, where the oscillation of PI between the two stationary policies, $\mu$ and $\mu'$, can be easily verified by the reader. Note that the condition $J^* \geq \bar{J}$ does not hold for monotone decreasing models where $T_{\mu} \bar{J} \leq \bar{J}$ for all $\mu \in M$ (unless $J^* = \bar{J}$).

Part (b) of the proposition shows that PI restricted to S-regular policies will converge to $J_S^*$ but not necessarily to $J^*$. Indeed in Example 3.1 with $b > 0$, let $S = [b, \infty)$. Then the only S-regular stationary policy is the nonoptimal policy $\mu'$ that terminates at state 1 $[\mu'(1) = u']$, and we have $J_S^* = b > 0 = J^*$.
Starting with $\mu' \in \mathcal{M}$, PI generates $J_{\mu'} = b$, and since $b = J_{\mu'} = T_{\mu'} J_{\mu'} = T J_{\mu'}$, it terminates with $\mu'$. Thus even when $J^* \geq J$, convergence of PI to $J^*$ is problematic, unless $J^*$ is the only possible fixed point of $T$ within \( \{ J \in E^+(X) \mid J \geq J^* \} \).

For the case of the optimistic version of PI (3.6), it turns out that the conditions for convergence are less restrictive. There is no need for the condition $J^* \geq \bar{J}$ or the $S$-regularity of the generated policies, as shown in the following proposition. This is due to the fact that optimistic PI embodies the characteristics of VI, which has favorable properties when $J^* \leq \bar{J}$ (see the discussion in connection with Prop. 3.3).

**Proposition 3.7: (Convergence of Optimistic PI)** Assume that:

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

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

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

Then the optimistic PI algorithm (3.6) is well defined and, under the condition $J_0 \geq TJ_0$, the following hold:

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

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

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

\[
J_0 \geq TJ_0 = T_{\mu^0} J_0 \geq T_{\mu^0}^{m_0} J_0 = J_1 \geq T_{\mu^0}^{m_0+1} J_0 = T_{\mu^0} J_1 \geq TJ_1 = T_{\mu^1} J_1 \geq \cdots \geq J_2,
\]

and continuing similarly, we obtain

\[
J_k \geq TJ_k \geq J_{k+1}, \quad k = 0, 1, \ldots
\]

Thus $J_k \downarrow J_\infty$ for some $J_\infty$. The proof that $J_\infty$ is a fixed point of $T$ is the same as in the case of the PI algorithm (3.5).

(b) In the case where all the policies $\mu^k$ are $S$-regular and $\{ J_k \} \subset S$, from Eq. (3.11), we have $J_{k+1} \geq J_{\mu^k}$ for all $k$, so by using the $S$-regularity of $\mu^k$, we have

\[
J_\infty = \lim_{k \to \infty} J_k \geq \lim \inf_{k \to \infty} J_{\mu^k} \geq J^*_S.
\]

Using the fixed point property of $J_\infty$ proved in part (a), and applying Prop. 3.1(b) with $J' = J_\infty$, $\hat{J} = J_k \geq J^*_S$ and $C$ equal to the set $\mathcal{M}_S \times X$, we have $J_\infty \leq J^*_S$, which combined with the preceding relation yields $J_\infty = J^*_S$. Q.E.D.
The preceding two propositions can be used to ascertain convergence to $J^*$ of the PI algorithms (3.5) and (3.6) (i.e., $J_\infty = J^*$) if $J^*$ is known to be the only possible fixed point of $T$ within a subset of $E(X)$ to which $J_\infty$ can also be shown to belong. For example this is true under the assumptions of Prop. 3.2 or the assumptions of Prop. 3.4, assuming also that $J_\infty \leq \tilde{J}$ for some $\tilde{J} \in S$. We will see an example of such use of the proposition in Section 4.1, where we will show convergence of the PI algorithms (3.5) and (3.6), in the sense that $J_{\mu_k} \downarrow J^*$ and $J_{k} \downarrow J^*$, respectively, for deterministic optimal control problems with nonnegative cost per stage. Generally, the sequence $\{\mu^k\}$ of generated policies need not converge to some policy, and even if it converges, the limit policy need not be optimal.

Another example of application of the PI convergence results of this section is the classical discounted linear-quadratic problems discussed in Section 4.4.3 of [Ber13]. There, with $S$ being the set of positive definite quadratic functions plus a nonnegative constant, and with $M$ being the set of linear stable controllers, it was shown that all $\mu \in M$ are $S$-regular, that the conditions of Props. 3.6(b) and 3.7(b) hold, and that the PI algorithms (3.5) and (3.6) converge to $J^*$. The convergence of the PI algorithm (3.5) for linear-quadratic problems is well-known and dates to the paper [Kle68].

4. MONOTONE INCREASING MODELS

An important type of abstract DP model is one where $\tilde{J} \leq T_\mu \tilde{J}$ for all $\mu \in M$. In this model, the finite horizon costs $T_{\mu_0} \cdots T_{\mu_k} \tilde{J}$ of any policy $\pi = \{\mu_0, \mu_1, \ldots\}$ monotonically increase to $J_\pi$. Consequently this model is known as monotone increasing, and among others, it can be used to represent problems where nonnegative costs accumulate additively over time. An important example is the classical nonnegative cost stochastic optimal control problem [Str66] (cf. Example 2.1, with $g \geq 0$). Note that if the optimal cost $J^*(x)$ at a state $x$ is to be finite, the accumulation of nonnegative costs must be diminishing starting from $x$. In the absence of discounting, this must be accomplished through the presence of cost-free states, which in optimal control problems are typically desirable states that we aim to reach, perhaps asymptotically, from the remaining states. The applications of this section are of this type.

For the monotone increasing model, $J^*$ is known to be a fixed point of $T$ under certain relatively mild assumptions. However, VI may not converge to $J^*$ starting from below $J^*$ (e.g., starting from $\tilde{J}$), and also starting from above $J^*$. In this section we will address the question of convergence of VI from above $J^*$ by using the regularity ideas of the preceding section. The starting point for the analysis is the following assumption, introduced in [Ber75], [Ber77] (see also [BeS78], Ch. 5, and [Ber13], Section 4.3).

**Assumption I: (Monotone Increase)**

(a) We have

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

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

$$\lim_{m \to \infty} H(x, u, J_m) = H(x, u, J), \quad \forall x \in X, u \in U(x).$$
There exists a scalar $\alpha \in (0, \infty)$ such that for all scalars $r \in (0, \infty)$ and functions $J \in E(X)$ with $J \leq J$, we have

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

We summarize the results that are relevant to our development in the following proposition (see [BeS78], Props. 5.2, 5.4, and 5.10, or [Ber13], Props. 4.3.3, 4.3.9, and 4.3.14).

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

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

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

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

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

$$U_k(x, \lambda) = \{u \in U(x) \mid H(x, u, T_k \bar{J}) \leq \lambda\} \quad (4.1)$$

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

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

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

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

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

(c) Use Prop. 3.2 in conjunction with the fixed point properties of $J^*$ [cf. Prop. 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 $\{J \in S \mid J \geq J^*\}$.

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

Some statements regarding the validity of PI, using Props. 3.6 and 3.7, will also be made.
4.1. Deterministic Optimal Control to a Terminal Set of States

Let us consider the undiscounted deterministic optimal control problem of Eqs. (2.2) and (2.3), where

\[ H(x, u, J) = g(x, u) + J(f(x, u)) , \]

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

\[ x_{k+1} = f(x_k, u_k) . \]

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

\[ 0 \leq g(x, u) < \infty, \quad \forall x \in X, \ u \in U(x) . \]

As in Eq. (2.4), the cost function \( \bar{J}_\pi \) of a policy \( \pi \) is the upper limit of the finite horizon cost functions

\[ T_{\mu_0} \cdots T_{\mu_k} \bar{J} \]

of the policy, with \( \bar{J}(x) \equiv 0 \). Viewing the problem within the abstract DP framework with \( H \)

given by Eq. (4.2), clearly Assumption I holds, so Prop. 4.1 applies.

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

\[ g(x, u) = 0, \quad x = f(x, u), \quad \forall x \in X_s, \ u \in U(x) . \]

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

\[ X_s = \{ x \in X \mid J^*(x) = 0 \} . \]

In the applications of primary interest, \( g \) is taken to be strictly positive outside of \( X_s \) to encourage asymptotic convergence of the generated state sequence to \( X_s \), so the assumption is natural. Besides \( X_s \), two other interesting subsets of \( X \) are

\[ X_f = \{ x \in X \mid J^*(x) < \infty \}, \quad X_\infty = \{ x \in X \mid J^*(x) = \infty \} . \]

The problem just described is natural in the context of optimal control, and is a commonly used model for regulation of the state of a dynamic system to the origin or more generally to a terminal set of states \( X_s \). It is usually taken as the starting point for analyses relating to the adaptive DP methodology: for a selective list of references that is biased towards recent works, see the papers [Hey14a], [Hey14b], [JiJ14], [LiW13], [VVL13], [WWL14], the survey papers in the edited volumes [SBP04] and [LeL13], and the special issue [LLL08]. Some of these works relate to continuous-time problems, and in their treatment of algorithmic convergence, may include unnecessarily restrictive assumptions such as \( X \) and \( U \) being Euclidean spaces, continuity and other conditions on \( g \), special structure of the system (4.3), etc. The problem of this section also arises in many other related control contexts (see e.g., [Ran06], [RDK09]).

We note that \( T \) typically has an infinite number of fixed points within \( E^+(X) \), the set of all nonnegative extended real-valued functions of \( x \in X \). In particular, every \( J \in E^+(X) \) of the form

\[ J = J^* + ce , \]
where \( c \geq 0 \) and \( e \) is the unit function \( [e(x) \equiv 1] \) is a fixed point of \( T \). We will derive conditions under which \( J^* \) is the unique fixed point of \( T \) within the subset of functions in \( E^+(X) \) that take the value 0 within the terminal set \( X_s \), and also show that VI converges to \( J^* \) starting from any \( J \) within that subset. Our analysis involves a notion of termination, which we now introduce.

Given a state \( x \), we say that a policy \( \pi \) terminates from \( x \) if the sequence \( \{x_k\} \), which is generated starting from \( x \) and using \( \pi \), reaches \( X_s \) in the sense that \( x_k \in X_s \) for some index \( k \). We will use the following assumption, and we will later derive easily verifiable conditions under which it holds.

**Assumption 4.1:** In the deterministic optimal control problem of this section, for every \( x \in X_f \) and \( \epsilon > 0 \), there exists a policy \( \pi \) that terminates from \( x \) and satisfies \( J^* \leq J^*(x) + \epsilon \).

We introduce the set
\[
C = \{ (\pi, x) \mid x \in X_f, \pi \text{ terminates from } x \},
\]
and we note that under Assumption 4.1, \( C \) is nonempty and \( J^*_C = J^* \). The reason is that for \( x \in X_f \), Assumption 4.1 implies that \( J^*_C(x) = \inf_{(\pi, x) \in C} J_\pi(x) = J^*(x) \), while for \( x \in X_\infty \) we also have \( J^*_C(x) = J^*(x) = \infty \) by the definition of \( J^*_C \) [cf. Eq. (3.1)], since for such \( x \), the set of policies \( \{ \pi \mid (\pi, x) \in C \} \) is empty.

We next consider the set
\[
S = \{ J \in E^+(X) \mid J(x) = 0, \forall x \in X_s \},
\]
Clearly \( J^* \in S \) and we also claim that \( C \) is S-regular. Indeed for \( \pi \) that terminates from \( x \) we have
\[
\limsup_{k \to \infty} (T_{\mu_0} \cdots T_{\mu_k} J)(x) = \limsup_{k \to \infty} (T_{\mu_0} \cdots T_{\mu_k} J)(x) = J_\pi(x), \quad \forall J \in S,
\]
since for such a policy \( \pi \), the choice of \( J \) within \( S \) does not affect \( (T_{\mu_0} \cdots T_{\mu_k} J)(x) \) for \( k \) larger than the termination time, when the state enters \( X_s \).

We can now prove our main result.

**Proposition 4.2:** (Uniqueness of Solution of Bellman’s Equation and Convergence of VI)

Let Assumption 4.1 hold. Then \( J^* \) is the unique fixed point of \( T \) within the set \( S \) of Eq. (4.5), and we have \( T^k J \to J^* \) for all \( J \in S \) with \( J \geq J^* \). If in addition \( U \) is a metric space, and the sets \( U_k(x, \lambda) \) of Eq. (4.1) are compact for all \( x \in X, \lambda \in \mathbb{R} \), 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 \( J^* \in S \), while the discussion preceding the proposition has shown that \( J^*_C = J^* \) and \( C \) is \( S \)-regular. From Prop. 3.2 it follows that \( J^* \) is the unique fixed point of \( T \) within \( \{ J \in S \mid J \geq J^* \} \). On
the other hand, every fixed point \( J \in E^+(X) \) of \( T \) satisfies \( J \geq J^* \) by the general results that hold under Assumption I [cf. Prop. 4.1(a)], so \( J^* \) is the unique fixed point of \( T \) within \( S \). Also from Prop. 3.2 we have that the VI sequence \( \{T^kJ\} \) converges to \( J^* \) starting from any \( J \in S \) with \( J \geq J^* \).

Finally, for any \( J \in S \), let us select \( \hat{J} \in S \) with \( \hat{J} \geq J \) and \( \hat{J} \geq J^* \), and note that by the monotonicity of \( T \), we have

\[
T^kJ \leq T^k\hat{J} \leq T^k\hat{J}.
\]

If we also assume compactness of the sets \( U_k(x, \lambda) \) of Eq. (4.1), then by Prop. 4.1(d), we have \( T^kJ \rightarrow J^* \), which together with the convergence \( T^k\hat{J} \rightarrow J^* \) just proved, implies that \( T^kJ \rightarrow J^* \).

\[
\text{Q.E.D.}
\]

The following example, is based on the ideas of the related Example 4.3.3 of [Ber13]. It illustrates how VI can fail to converge to \( J^* \) starting from some \( J \) with \( 0 \leq J \leq J^* \) because the compactness assumption of the proposition does not hold. While there are other known examples under Assumption I where this phenomenon occurs, in this particular example Assumption 4.1 is satisfied, thus illustrating the fact that even in deterministic optimal control, convergence of VI to \( J^* \) from above and from below involve fundamentally different mechanisms when \( g \geq 0 \).

**Example 4.1**

Let \( X = \mathbb{R} \) and assume that all states \( x < 0 \) are cost-free and absorbing. We assume that at states \( x \geq 0 \), there are two types of control choice:

(a) A control \( u > 0 \), in which case a cost \( x \) is incurred and the next state is \( x + u \).

(b) A special “termination” control, in which case a cost of 1 is incurred and the next state becomes some negative number (say -1 for concreteness).

It can be seen that any policy that never uses the “termination” control incurs an infinite cost for any starting \( x \geq 0 \). It follows that the stationary policy that terminates from every \( x \geq 0 \) is optimal and we have

\[
J^*(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}
\]

Therefore \( X_s = \{ x \in \mathbb{R} \mid x < 0 \} \).

Turning now to VI, the mapping \( T \) has the form

\[
(TJ)(x) = \begin{cases} \min \{1 + J(-1), \inf_{u>0} \{x + J(x+u)\} \} & \text{if } x \geq 0, \\ J(x) & \text{if } x < 0. \end{cases}
\]

It can be seen that \( J^* \) is the unique fixed point of \( T \) within the set \( S = \{ J \in E^+(X) \mid J(x) = 0, \forall x \in X_s \} \), and VI converges to \( J^* \) in a single iteration starting from any \( J \geq J^* \) within this set, consistent with Prop. 4.2. On the other hand, by induction we can verify that starting from \( \bar{J} = 0 \), the VI iterates are

\[
(T^k\bar{J})(x) = \begin{cases} \min\{1, kx\} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases}
\]

so that \( \lim_{k \to \infty} (T^k\bar{J})(0) = 0 < 1 = J^*(1) \) and VI does not converge to \( J^* \) starting from \( \bar{J} \). Here the compactness assumption of Prop. 4.2 is violated.

A case where Assumption 4.1 holds is when for every \( x \in X_f \) there exists an optimal policy that terminates from \( x \), as in the preceding example. For some types of problems this property can be easily
verified. A prominent case is when \( X \) and \( U \) are finite, so the problem becomes a deterministic shortest path problem with nonnegative arc lengths. If every cycle of the state transition graph has positive length, all policies \( \pi \) 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 \( x \), assuming a terminating policy exists from every \( x \).

When \( X \) is infinite, e.g., when \( X \) is the \( n \)-dimensional Euclidean space \( \mathbb{R}^n \), it may easily happen that the optimal policies are not terminating from some or all \( x \in X_f \). This is true for example in the classical linear-quadratic optimal control problem, where \( X = \mathbb{R}^n \), \( U = \mathbb{R}^m \), \( g \) is positive semidefinite quadratic, and \( f \) represents a linear system of the form \( x_{k+1} = Ax_k + Bu_k \), where \( A \) and \( B \) are given matrices. To address this situation we introduce a notion of asymptotic termination, which is related to classical notions of asymptotic stability. We assume that \( X \) is a normed space with norm denoted \( \| \cdot \| \), and we say that \( \pi \) asymptotically terminates from \( x \) if the sequence \( \{ x_k \} \) generated starting from \( x \) and using \( \pi \) converges to \( X_s \) in the sense that

\[
\lim_{k \to \infty} \text{dist}(x_k, X_s) = 0,
\]

where for all \( x \in X \), \( \text{dist}(x, X_s) \) denotes the minimum distance from \( x \) to \( X_s \),

\[
\text{dist}(x, X_s) = \inf_{y \in X_s} \| x - y \|, \quad x \in X.
\]

The following proposition provides readily verifiable conditions that guarantee Assumption 4.1.

**Proposition 4.3:** Let the following two conditions hold.

1. For every \( x \in X_f \), any policy \( \pi \) with \( J_\pi(x) < \infty \) asymptotically terminates from \( x \).

2. For every \( \epsilon > 0 \), there exists a \( \delta_\epsilon > 0 \) such that for each \( x \in X_f \) with

\[
\text{dist}(x, X_s) \leq \delta_\epsilon,
\]

there is a policy \( \pi \) that terminates from \( x \) and satisfies \( J_\pi(x) \leq \epsilon \).

Then Assumption 4.1 also holds, and the conclusions of Prop. 4.2 apply.

**Proof:** Fix \( x \in X_f \) and \( \epsilon > 0 \). Condition (1) guarantees that for any fixed \( x \in X_f \) and \( \epsilon > 0 \), there exists a policy \( \pi \) that asymptotically terminates from \( x \), and satisfies

\[
J_\pi(x) \leq J^*(x) + \epsilon/2.
\]

Starting from \( x \), this policy will generate a sequence \( \{ x_k \} \) such that for some index \( \bar{k} \) we have

\[
\lim_{k \to \infty} \text{dist}(x_{\bar{k}}, X_s) \leq \delta_\epsilon/2,
\]

so by condition (2), there exists a policy \( \bar{\pi} \) that terminates from \( x_{\bar{k}} \) and is such that \( J_{\bar{\pi}}(x_{\bar{k}}) \leq \epsilon/2 \). Consider the policy \( \pi' \) that follows \( \pi \) up to index \( \bar{k} \) and follows \( \bar{\pi} \) afterwards. This policy terminates from \( x \) and satisfies

\[
J_{\pi'}(x) = J_{\pi, \bar{k}}(x) + J_{\pi}(x_{\bar{k}}) \leq J_{\pi}(x) + J_{\bar{\pi}}(x_{\bar{k}}) \leq J^*(x) + \epsilon.
\]
where $J_{\pi,k}(x)$ is the cost incurred by $\pi$ starting from $x$ up to reaching $x_k$. \textbf{Q.E.D.}

Cost functions for which condition (1) of the preceding proposition holds are those involving a cost per stage that is strictly positive outside of $X_s$. More precisely, condition (1) holds if for each $\delta > 0$ there exists $\epsilon > 0$ such that

$$\inf_{u \in U(x)} g(x,u) \geq \epsilon, \quad \forall x \in X \text{ such that dist}(x,X_s) \geq \delta.$$  

Then for any $x$ and policy $\pi$ that does not asymptotically terminate from $x$, we will have $J_\pi(x) = \infty$. From an applications point of view, the condition is natural and consistent with the aim of steering the state towards the terminal set $X_s$ with finite cost.

Condition (2) is a “controllability” condition implying that the state can be steered into $X_s$ with arbitrarily small cost from a starting state that is sufficiently close to $X_s$. As an example, condition (2) is satisfied when $X_s = \{0\}$ and the following hold:

(a) $X = \mathbb{R}^n$, $U = \mathbb{R}^m$, and there is an open sphere $R$ centered at the origin such that $U(x)$ contains $R$ for all $x \in X$.

(b) $f$ represents a controllable linear system of the form

$$x_{k+1} = Ax_k + Bu_k,$$

where $A$ and $B$ are given matrices.

(c) $g$ satisfies

$$0 \leq g(x,u) \leq \beta(\|x\|^{p} + \|u\|^{p}), \quad \forall (x,u) \in V,$$

where $V$ is some open sphere centered at the origin, $\beta, p$ are some positive scalars, and $\| \cdot \|$ is the standard Euclidean norm.

There are straightforward extensions of the preceding conditions to a nonlinear system. Note that even for a controllable system, it is possible that there exist states from which the terminal set cannot be reached, because $U(x)$ may imply constraints on the magnitude of the control vector. Still the preceding analysis allows for this case.

Finally let us consider the PI algorithm (3.5) and its optimistic version (3.6). We assume that these algorithms are well defined, that the standard version is initialized with a policy $\mu^0$ satisfying $J_{\mu,0}(x) = 0$ for all $x \in X_s$, and that the optimistic version is initialized with a function $J_0 \in E^+(X)$ satisfying $J_0(x) = 0$ for all $x \in X_s$ and $J_0 \geq TJ_0$ [and hence also $J_0 \geq J^*$ by Prop. 4.1(a)].

**Proposition 4.4: (Convergence of PI)** Let Assumption 4.1 hold. Then for the PI algorithm (3.5) we have $J_{\mu,k} \downarrow J^*$, while for its optimistic version (3.6) we have $J_k \downarrow J^*$.

**Proof:** The proof of Prop. 4.2 shows that $J^*_C = J^*$, where

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

23
and that $C$ is $S$-regular. We have $J^* \geq \bar{J}$ by the nonnegativity of $g$, while by Prop. 4.1(b), we have $J_{\mu} = T_{\mu}J_{\mu}$ for all $\mu \in \mathcal{M}$. Thus from Props. 3.6 and 3.7 it follows that $J_{\infty}$, the limit of each of the PI algorithms, is a fixed point of $T$.

For the PI algorithm (3.5) we have that $J_{\infty}$ belongs to $S$, since we have assumed that for all $x \in X_s$, $J_{\mu}0(x) = 0$ and hence $J_{\mu}k(x) = 0$, so that $J_{\mu}k \in S$. For the optimistic version (3.6) we also have that $J_{\infty}$ belongs to $S$, since $J_0(x) = 0$ for all $x \in X_s$, and hence also $J_k(x) = 0$ for all $x \in X_s$. Moreover, $J_k \geq J^*$ as can be easily verified by induction using the assumption $J_0 \geq T_{J_0}$, and the facts $J^* = TJ^*$ and $J_0 \geq J^*$ [cf. Prop. 4.1(a)]. Thus $J_{\infty}$ is a fixed point of $T$, belongs to $S$, and satisfies $J_{\infty} \geq J^*$. Since by Prop. 4.2, $J^*$ is the unique fixed point of $T$ within $S$, it follows that $J_{\infty} = J^*$. For the PI algorithm (3.5), this implies that $J_k \downarrow J^*$, while for the optimistic version (3.6), this implies that $J_k \downarrow J^*$. Q.E.D.

We finally note that approximate versions of the PI algorithm, involving cost function approximation and sometimes simulation, are broadly advocated in the literature of adaptive DP, in the spirit of approximate DP or reinforcement learning [see e.g., the books [BeT96], [SuB98], [Pow11]]. The analysis of such algorithms is an interesting subject for further research, possibly using the ideas of this section.

4.2. Nonnegative Cost Stochastic DP

Let us consider the undiscounted stochastic optimal control problem of Example 2.1, involving the mapping

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

(4.6)

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

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

(4.7)

We consider the abstract DP model with $H$ as above, and with $\bar{J}(x) \equiv 0$. Clearly Assumption I holds. Our purpose in this section is to show the connection of our approach with the convergence results for VI that were given in the paper [YuB13].

Consider the set

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

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

Let us denote by $E_0^\pi(\cdot)$ 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

$$S = \{ J \in E^+(X) \mid E_0^\pi(J(x_k)) \to 0, \ \forall (\pi, x_0) \in C \}.$$

(4.8)

We will show that $J^* \in S$ and that $C$ is $S$-regular. Once this is done, it will follow from Prop. 3.2 and the fixed point property of $J^*$ [cf. Prop. 4.1(a)] that $T_J J \to J^*$ for all $J \in S$ that satisfy $J \geq J^*$. If the sets $U_k(x, \lambda)$ of Eq. (4.1) are compact, the convergence of VI starting from below $J^*$ will also be guaranteed. We have the following proposition, which may be viewed as a stochastic analog of Prop. 4.2. The proof uses the argument of Appendix E of [YuB13].
Proposition 4.5: (Convergence of VI) Consider the stochastic optimal control problem of this section, assuming Eq. (4.7). Then \( J^* \) is the unique fixed point of \( T \) within \( S \), and we have \( T^k J \to J^* \) for all \( J \geq J^* \) with \( J \in S \). If in addition \( U \) is a metric space, and the sets \( U_k(x, \lambda) \) of Eq. (4.1) are compact for all \( x \in X, \lambda \in \mathbb{R} \), and \( k \), we have \( T^k J \to J^* \) for all \( J \in S \), and an optimal stationary policy is guaranteed to exist.

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

\[
(T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x_0) = E_{x_0}^\pi \{ J(x_k) \} + E_{x_0}^\pi \left\{ \sum_{t=0}^{k-1} g(x_t, \mu_t(x_t), w_t) \right\},
\]

where \( \mu_t, t = 0, 1, \ldots \), denote generically the components of \( \pi \). The rightmost term above converges to \( J^*_\pi(x_0) \) as \( k \to \infty \), so by taking upper limit, we obtain

\[
\limsup_{k \to \infty}(T_{\mu_0} \cdots T_{\mu_{k-1}} J)(x_0) = \limsup_{k \to \infty} E_{x_0}^\pi \{ J(x_k) \} + J^*_\pi(x_0).
\]

Thus in view of the definition of \( S \), we see that for all \( (\pi, x_0) \in C \) 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 \( C \) is \( S \)-regular.

We next show that \( J^* \in S \). We have for all \( (\pi, x_0) \in C \)

\[
J^*_\pi(x_0) = E_{x_0}^\pi \{ g(x_0, \mu_0(x_0), w_0) \} + E_{x_0}^\pi \{ J^*_\pi(x_1) \},
\]

and more generally,

\[
E_{x_0}^\pi \{ J^*_\pi(x_t) \} = E_{x_0}^\pi \{ g(x_t, \mu_t(x_t), w_t) \} + E_{x_0}^\pi \{ J^*_\pi(x_{t+1}) \}, \quad \forall \ t = 0, 1, \ldots,
\]

where \( \{ x_t \} \) is the sequence generated starting from \( x_0 \) and using \( \pi \). Using the defining property \( J^*_\pi(x_0) < \infty \) of \( C \), it follows that all the terms in the above relations are finite, and in particular

\[
E_{x_0}^\pi \{ J^*_\pi(x_t) \} < \infty, \quad \forall \ (\pi, x_0) \in C, \ t = 0, 1, \ldots.
\]

By adding Eq. (4.10) for \( t = 0, \ldots, k - 1 \), and canceling the finite terms \( E_{x_0}^\pi \{ J^*_\pi(x_t) \} \) for \( t = 1, \ldots, k - 1 \), we obtain

\[
J^*_\pi(x_0) = E_{x_0}^\pi \{ J^*_\pi(x_k) \} + \sum_{t=0}^{k-1} E_{x_0}^\pi \{ g(x_t, \mu_t(x_t), w_t) \}, \quad \forall \ (\pi, x_0) \in C, \ k = 1, 2, \ldots.
\]

The rightmost term above tends to \( J^*_\pi(x_0) \) as \( k \to \infty \), so we obtain

\[
E_{x_0}^\pi \{ J^*_\pi(x_k) \} \to 0, \quad \forall \ (\pi, x_0) \in C.
\]
Since $0 \leq J^* \leq J_\pi$ for all $\pi$, it follows that
\[ E_{x_0}^\pi \{ J^*(x_k) \} \to 0, \quad \forall \ x_0 \text{ with } J^*(x_0) < \infty. \] (4.11)
Thus $J^* \in S$. From this point the proof uses Prop. 3.2, and follows the corresponding argument of the proof of Prop. 4.2. Q.E.D.

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

\[ J^* \in S \]  

**Proposition 4.6:** If a function $J \in E(X)$ satisfies
\[ J^* \leq J \leq cJ^* \quad \text{for some } c > 0, \] (4.12)
we have $T^kJ \to J^*$.

**Proof:** Since $J^* \in S$ as shown in Prop. 4.5, any $J$ satisfying Eq. (4.12), also belongs to $S$, and the result follows from Prop. 4.5. Q.E.D.

The condition (4.12) 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. For finite-state problems, the set $X_s$ can be computed in polynomial time as shown in the paper [BeY15], which also provides a method for dealing with cases where the set $X_\infty$ is nonempty.

Regarding PI, we note that Prop. 3.6 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 $\{ J \mid J \geq J^* \}$ that contains the limit $J_\infty$ of PI. This result was given as Corollary 5.2 in [YuB13]. Alternatively, there is a mixed VI and PI algorithm proposed in [YuB13], which can be applied under the condition (4.12), 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 (4.7). In particular, we do not discuss the implications of the possible existence of a termination state as in finite-state or countable-state stochastic shortest path problems. As noted in Section 3, 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 paper [BeY15]. Moreover, one may consider stochastic analogs of the approach of Section 4.1. For example assume that we know the set
\[ X_s = \{ x \mid J^*(x) = 0 \}, \]
which we view as a termination set. Then, instead of the set $S$ of Eq. (4.8), we may consider the set
\[ S = \{ J \in E(X) \mid J(x) = 0, \forall x \in X_s \}, \]
and instead of the set $C = \{ (\pi, x_0) \mid J_\pi(x_0) < \infty \}$, we may consider the smaller set
\[ \hat{C} = \{ (\pi, x_0) \mid J_\pi(x_0) < \infty, \ E_{x_0}^\pi \{ J(x_k) \} \to 0, \forall J \in S \}. \]
Then based on the proof of Prop. 4.5, we have that $\hat{C}$ is $S$-regular. Suppose now that we can show that $J^* = J^*$, based on some stochastic version of Assumption 4.1. Then similar to Prop. 3.2, we can use Prop. 4.2, we can use Prop. 4.2 to show that $J^*$ is the unique fixed point of $T$ within $S$, and that we have $T^k J \rightarrow J^*$ for all $J \in S$ with $J \geq J^*$. This is a stronger result than Prop. 4.5 since the set $S$ above is larger and more convenient than the set $S$ of Eq. (4.8).

4.3. Discounted Nonnegative Cost Stochastic DP

We will now consider a discounted version of the stochastic optimal control problem of the preceding section. The cost function of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ has the form

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

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

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

cf. Eq. (2.8), and $\bar{J}(x) \equiv 0$. We continue to assume that the one-stage cost is nonnegative,

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

so the monotone increase Assumption I holds.

We also assume that $X$ is a normed space with norm denoted $\| \cdot \|$. We will extend our earlier analysis by combining the ideas of Sections 4.1 and 4.2. Note that because of the discount factor, the existence of a terminal set of states is not essential for the optimal costs to be finite.

Similar to Section 4.1, we introduce the set

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

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

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

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

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

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

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

The following assumption parallels Assumption 4.1.

27
**Assumption 4.2:** In the 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 $\pi$ that is stable from $x$ and satisfies $J^*(x) \leq J^*(x) + \epsilon$.

Note that under this assumption, we have $J_C^* = J^*$, similar to Section 4.1. We have the following proposition.

**Proposition 4.7:** Let Assumption 4.2 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 $J^* \leq J$. If in addition $U$ is a metric space, and the sets $U_k(x, \lambda)$ of Eq. (4.1) are compact for all $x \in X$, $\lambda \in \mathbb{R}$, 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 4.2, we have for all $J \in E(X)$, $(\pi, x_0) \in \mathcal{C}$, and $k$,

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

[cf. Eq. (4.9)]. 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 $\alpha^k E_{x_0} \left\{ J(x_k) \right\} \to 0$. Thus for all $(\pi, x_0) \in \mathcal{C}$ and $J \in S$, we have

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

so $\mathcal{C}$ is $S$-regular. Since $J_C^*$ is equal to $J^*$ which is a fixed point of $T$ [by Prop. 3.1(c)], 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 Prop. 4.1(d). Q.E.D.

Let us finally note that Assumption 4.2 is natural in control contexts where the objective is to keep the state from becoming unbounded, under the influence of random disturbances represented by $w_k$. In such contexts one expects that optimal or near optimal policies should produce bounded state sequences starting from states with finite optimal cost.

5. **CONCLUDING REMARKS**

We have extended the notion of a regular policy in abstract DP, and we have highlighted its connection to several earlier analyses. In particular, we have developed an approach for establishing convergence of VI by using regularity ideas. Starting from an interesting collection of policy-state pairs, a regularity property is
established that characterizes the region of convergence of VI. This approach can lead to new results, as we have shown in the context of optimal control problems with nonnegative cost per stage.

Our approach may also be applied to other problems that involve a termination state and fit the abstract DP framework of this paper. An example is minimax problems of the shortest path type, where similar to Section 4.1, we have \( g \geq 0 \), and there is a termination set \( X_s = \{ x \in X \mid J^*(x) = 0 \} \). Here the problem is defined a mapping of the form

\[
H(x, u, J) = \sup_{w \in W} \left\{ g(x, u, w) + J(f(x, u, w)) \right\},
\]

[cf. Eq. (2.7)], with \( W \) being some set representing the options of an antagonistic opponent, and with \( J^*(x) = 0 \). The finite-state version of this problem has been discussed in \[Ber14a\] for the case where \( g \) can take both positive and negative values. To extend the analysis of Section 4.1, we need to adapt the definition of termination. In particular, given a state \( x \), in the minimax context we say that a policy \( \pi \) terminates from \( x \) if there exists an index \( \bar{k} \) [which depends on \((\pi, x)\)] such that the sequence \( \{x_k\} \), which is generated starting from \( x \) and using \( \pi \), satisfies \( x_{\bar{k}} \in X_s \) for all sequences \( \{w_0, \ldots, w_{\bar{k}-1}\} \) with \( w_t \in W \) for all \( t = 0, \ldots, \bar{k}-1 \). Then the analysis of Section 4.1 can be readily extended to this problem, with the Props. 4.2-4.4 and their proofs holding essentially as stated.

Other classes of interesting problems that may admit a similar treatment are stochastic shortest path game problems \[PaB99\], \[Yu11\], and risk sensitive shortest path-type problems \[DeR79\], \[Pat01\], \[Ber13\], \[CaR14\]. These and other related applications are interesting subjects for further research.

6. REFERENCES


