May 2017 **Report LIDS-P-3507**

Proper Policies in Infinite-State Stochastic Shortest Path Problems

Dimitri P. Bertsekas†

Abstract

We consider stochastic shortest path problems with infinite state and control spaces, and a nonnegative cost per stage. We extend the notion of a proper policy from the context of finite state space to the context of infinite state space. We consider the optimal cost function J^*, and the optimal cost function $\hat{J}$ over just the proper policies. Assuming that there exists at least one proper policy, we show that J^* and $\hat{J}$ are the smallest and largest solutions of Bellman's equation, respectively, within a class of functions with a boundedness property. The standard value iteration algorithm may be attracted to either J^* or $\hat{J}$, depending on the initial condition.

1. INTRODUCTION

In this paper we consider a stochastic discrete-time infinite horizon optimal control problem involving the system

$$x_{k+1} = f(x_k, u_k, w_k), \qquad k = 0, 1, \ldots, \tag{1.1}$$

where x_k and u_k are the state and control at stage k, which belong to sets X and U, w_k is a random disturbance that takes values in a countable set W with given probability distribution $P(w_k \mid x_k, u_k)$, and $f : X \times U \times W \mapsto X$ is a given function. The state and control spaces X and U are arbitrary, but we assume that W is countable to bypass measurability issues in the choice of control. The control u_k must be chosen from a constraint set $U(x_k) \subset U$ that may depend on the current state x_k. The expected cost for the kth stage, $g(x_k, u_k)$, is assumed real-valued and nonnnegative:

$$0 \leq g(x_k, u_k) \leq \infty, \qquad \forall\ x_k \in X,\ u_k \in U(x_k),\ k = 0, 1, \ldots. \tag{1.2}$$

Note that values $g(x, u) = \infty$ may be used to model state constraints in addition to state and control penalties. We assume that X contains a special cost-free and absorbing state t, referred to as the *destination*:

$$f(t, u, w) = t, \qquad g(t, u) = 0, \qquad \forall\ u \in U(t),\ w \in W. \tag{1.3}$$

† 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.

We are interested in policies of the form $\pi = \{\mu_0, \mu_1, \ldots\}$, where each μ_k is a function mapping $x \in X$ into the control $\mu_k(x) \in U(x)$. The set of all policies is denoted by Π. Policies of the form $\pi = \{\mu, \mu, \ldots\}$ are called *stationary*, and will be denoted by μ, when confusion cannot arise.

Given an initial state x_0, a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ when applied to the system (1.1), generates a random sequence of state-control pairs $\big(x_k, \mu_k(x_k)\big)$, $k = 0, 1, \ldots$, with cost

$$J_\pi(x_0) = \sum_{k=0}^{\infty} E_{x_0}^{\pi}\Big\{g\big(x_k, \mu_k(x_k)\big)\Big\}, \qquad x_0 \in X,$$

where $E_{x_0}^{\pi}\{\cdot\}$ denotes expectation with respect to the probability measure corresponding to initial state x_0 and policy π. We view J_π as a function over X, and we refer to it as the cost function of π. For a stationary policy μ, the corresponding cost function is denoted by J_μ. The optimal cost function is defined as

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

and a policy π^* is said to be optimal if $J_{\pi^*}(x) = J^*(x)$ for all $x \in X$. We refer to the problem of finding J^* and an optimal policy as the *stochastic shortest path problem* (SSP problem for short). We denote by $\mathcal{E}^+(X)$ the set of functions $J : X \mapsto [0, \infty]$. All equations, inequalities, limit and minimization operations involving functions from this set are meant to be pointwise. In our analysis, we will use the set of functions

$$\mathcal{J} = \big\{J \in \mathcal{E}^+(X) \mid J(t) = 0\big\}.$$

Since t is cost-free and absorbing, this set contains the cost functions J_π of all $\pi \in \Pi$, as well as J^*.

It is well known that when $g \geq 0$, J^* satisfies the Bellman equation given by

$$J(x) = \inf_{u \in U(x)} \Big\{g(x,u) + E\big\{J\big(f(x,u,w)\big)\big\}\Big\}, \qquad x \in X, \tag{1.4}$$

where the expected value is with respect to the distribution distribution $P(w \mid x, u)$. Moreover, an optimal stationary policy (if it exists) may be obtained through the minimization in the right side of this equation (cf. Prop. 2.1 in the next section). One hopes to obtain J^* in the limit by means of value iteration (VI for short), which starting from some function $J_0 \in \mathcal{J}$, generates a sequence $\{J_k\} \subset \mathcal{J}$ according to

$$J_{k+1} = \inf_{u \in U(x)} \big\{g(x,u) + E\big\{J_k\big(f(x,u,w)\big)\big\}\big\}, \qquad x \in X,\ k = 0, 1, \ldots. \tag{1.5}$$

However, $\{J_k\}$ may not always converge to J^* because, among other reasons, Bellman's equation may have multiple solutions within $\mathcal{J}$.

In a recent paper [Ber17] we have addressed the connections between stability and optimal control in the context of undiscounted discrete-time deterministic optimal control with a termination state. In this paper we address similar issues in the context of SSP problems but we focus attention on proper policies,

which are guaranteed to reach the termination state with probability one from the states where the optimal cost is finite (a precise definition is given in the next section). The significance of proper policies is well known in finite-state SSP problems (see e.g., the books [Pal67], [Der70], [Whi82], [BeT89], [Put94], [Alt99], [HeL99], and [Ber12], and the references quoted there). In the inifinite-state context of this paper and under suitable assumptions, we show that $\hat{J}$, the optimal cost function over just the proper policies, is the largest solution of Bellman's equation, and that the VI algorithm converges to $\hat{J}$ starting from within a set of functions $\widehat{\mathcal{W}} \subset \mathcal{J}$ that majorize $\hat{J}$. We also consider the favorable special case where $J^* = \hat{J}$.

To compare our analysis with the existing literature, we note that proper policies for infinite-state SSP problems have been considered earlier, notably in the works of Pliska [Pli78], and James and Collins [JaC06], where they are called *transient*. There are a few differences between the frameworks of [Pli78], [JaC06] and this paper, which impact on the results obtained. In particular, the paper [Pli78] uses the same definition of properness as we do, but assumes that all policies are proper, g is assumed bounded, and J^* is real-valued. The paper [JaC06] uses the properness definition of [Pli78], and extends the analysis of Bertsekas and Tsitsiklis [BeT91] from finite state space to infinite state space (addressing also measurability issues). Moreover, [JaC06] allows the cost per stage g to take both positive and negative values. However, [JaC06] uses assumptions that guarantee that improper policies cannot be optimal and that $J^* = \hat{J}$, while J^* is real-valued. Our analysis is most closely related to the one of Bertsekas and Yu [BeY16], where the case $J^* \neq \hat{J}$ was analyzed using perturbation ideas that are similar to the ones of Section 3. The paper [BeY16] assumes that the state space is finite, but allows g to take both positive and negative value. Also [BeY16] gives an example showing that J^* may not be a solution of Bellman's equation if improper policies can be optimal. The extension of our results to SSP problems where g takes both positive and negative values may be possible, but our line of analysis relies strongly on the nonnegativity of g.

2. PROPER POLICIES AND THE PERTURBED PROBLEM

In this section, we will lay the groundwork for our analysis and introduce the notion of a proper policy. To this end, we will use some classical results for stochastic optimal control with nonnegative cost per stage, which stem from the original work of Strauch [Str66]. For textbook accounts we refer to [BeS78], [Put94], [Ber12], and for a more abstract development, we refer to the monograph [Ber13]. The following two propositions give the results that we will need.

Proposition 2.1: The following hold:

(a) J^* is a solution of Bellman's equation and if $J \in \mathcal{E}^+(X)$ is another solution, i.e., J satisfies

$$J(x) = \inf_{u \in U(x)} \Big\{ g(x,u) + E\big\{ J\big(f(x,u,w)\big)\big\}\Big\}, \qquad \forall\ x \in X, \tag{2.1}$$

then $J^* \leq J$.

(b) For all stationary policies μ, J_μ is a solution of the equation

$$J(x) = g\big(x,\mu(x)\big) + E\big\{ J\big(f\big(x,\mu(x),w\big)\big)\big\}, \qquad \forall\ x \in X,$$

and if $J \in \mathcal{E}^+(X)$ is another solution, then $J_\mu \leq J$.

(c) For every $\epsilon > 0$ there exists an ϵ-optimal policy, i.e., a policy π_ϵ such that

$$J_{\pi_\epsilon}(x) \leq J^*(x) + \epsilon, \qquad \forall\ x \in X.$$

(d) A stationary policy μ^* is optimal if and only if

$$\mu^*(x) \in \arg\min_{u \in U(x)} \Big\{ g(x,u) + E\big\{ J^*\big(f\big(x,u,w\big)\big)\big\}\Big\}, \qquad \forall\ x \in X.$$

(e) If $U(x)$ is finite for all $x \in X$, then $J_k \to J^*$, where $\{J_k\}$ is the sequence generated by the VI algorithm (1.5) starting from any J_0 with $0 \leq J_0 \leq J^*$.

Proof: See [BeS78], Props. 5.2, 5.4, and 5.10, or [Ber12], Props. 4.1.1, 4.1.3, 4.1.5, 4.1.9. **Q.E.D.**

Proposition 2.2: Let $\pi = \{\mu_0, \mu_1, \ldots\}$ be a policy, and for a given initial state $x_0 \in X$, let $\{x_k\}$ be the sequence of states generated by starting from x_0 and using π.

(a) If $J_\pi(x_0) < \infty$, then $E^\pi_{x_0}\big\{J_{\pi_k}(x_k)\big\} \downarrow 0$, where π_k is the policy $\{\mu_k, \mu_{k+1}, \ldots\}$.

(b) If π is stationary of the form $\{\mu, \mu, \ldots\}$ and $J_\mu(x_0) < \infty$, then $E^\mu_{x_0}\big\{J_\mu(x_k)\big\} \downarrow 0$.

Proof: (a) We have by definition

$$E^\pi_{x_0}\big\{J_{\pi_m}(x_m)\big\} = E^\pi_{x_0}\big\{g\big(x_m,\mu_m(x_m)\big) + J_{\pi_{m+1}}(x_{m+1})\big\}, \qquad m = 0,1,\ldots.$$

By repeatedly applying this relation, we obtain

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

Since g is nonnegative, $\sum_{m=0}^{k-1} E_{x_0}^{\pi}\{g(x_m,\mu_m(x_m))\}$ is monotonically nondecreasing, and it follows that $E_{x_0}^{\pi}\{J_{\pi_k}(x_k)\}$ is monotonically nonincreasing and real valued since $J_\pi(x_0)<\infty$. Moreover, by taking the limit as $k\to\infty$, we obtain $\lim_{k\to\infty} E_{x_0}^{\pi}\{J_{\pi_k}(x_k)\}=0$.

(b) This is a special case of part (a), with $\pi=\{\mu,\mu,\ldots\}$. **Q.E.D.**

Our analysis will focus primarily on the values of J_π within the set

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

since $J_\pi(x)$ is infinite for x outside this set. We denote by $\mathcal{B}$ the set of functions in $\mathcal{J}$, which are bounded on X_f,

$$\mathcal{B}=\left\{J\in\mathcal{J}\;\middle|\;\sup_{x\in X_f} J(x)<\infty\right\}.$$

A policy π is said to be *proper* if

$$J_\pi\in\mathcal{B},\qquad \sup_{x_0\in X_f}\sum_{k=0}^{\infty} r_k(\pi,x_0)<\infty, \tag{2.2}$$

where $r_k(\pi,x_0)$ is the probability that $x_k\neq t$ when using π and starting from x_0. Note that the second condition in Eq. (2.2) is equivalent to

$$\sum_{k=0}^{\infty} r_k(\pi,\cdot)\in\mathcal{B},$$

and states that the expected number of steps to termination using π is uniformly bounded over X_f. The set of all proper policies is denoted by $\widehat{\Pi}$ and the corresponding restricted optimal cost function is denoted by $\hat{J}$:

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

The condition $J_\pi\in\mathcal{B}$ in the definition of a proper policy is unnecessary if the cost per stage g is bounded over $X\times U$, since

$$J_\pi(x_0)\leq \sup_{(x,u)\in X\times U} g(x,u)\cdot\sum_{k=0}^{\infty} r_k(\pi,x_0).$$

For any $\delta>0$, let us consider the *δ-perturbed optimal control problem*. This is the same problem as the original, except that the cost per stage is changed to

$$g(x,u)+\delta,\qquad \forall\ x\neq t,$$

while $g(x,u)$ is left unchanged at 0 when $x=t$. Thus t is still cost-free as well as absorbing in the δ-perturbed problem. The δ-perturbed cost function of a policy π is denoted by $J_{\pi,\delta}$ and is given by

$$J_{\pi,\delta}(x)=J_\pi(x)+\delta\sum_{k=0}^{\infty} r_k(\pi,x). \tag{2.3}$$

We denote by J_δ^*, the optimal cost function of the δ-perturbed problem, i.e., $J_\delta^*(x) = \inf_{\pi\in\Pi} J_{\pi,\delta}(x)$.

Since the cost function of the δ-perturbed problem is nonnegative, Prop. 2.1 applies and shows that J_δ^* is the smallest solution of the corresponding Bellman equation. Our first objective will be to show that as $\delta \downarrow 0$, the δ-perturbed optimal cost function J_δ^* converges to $\hat{J}$, the restricted optimal cost function of the original unperturbed problem over just the set of proper policies $\widehat{\Pi}$. We will then use a limiting as $\delta \downarrow 0$ to show that $\hat{J}$ is a solution of the Bellman equation for the original problem. This is the subject of the next section.

3. MAIN RESULTS

The following proposition shows, among others, that within the set of states X_f, $J_{\pi,\delta}(x)$ differs from $J_\pi(x)$ by $O(\delta)$ if π is proper.

Proposition 3.1:

(a) A policy π is proper if and only if $J_{\pi,\delta} \in \mathcal{B}$.

(b) If there exists at least one proper policy, then $J_\delta^* \in \mathcal{B}$ for all $\delta > 0$. Moreover, for every $\epsilon > 0$, there exists a proper policy π_ϵ that is ϵ-optimal, i.e.,

$$J_{\pi_\epsilon,\delta}(x) \leq J_\delta^*(x) + \epsilon, \qquad \forall\ x \in X.$$

Proof: (a) Follows from Eq. (2.3) and the definition of a proper policy.

(b) If π is a proper policy, we have $J_\delta^* \leq J_{\pi,\delta}$ and by part (a), $J_\delta^* \in \mathcal{B}$. By Prop. 2.1(c), there exists an ϵ-optimal policy π_ϵ for the δ-perturbed problem, so we have $J_{\pi_\epsilon,\delta}(x) \leq J_\delta^*(x) + \epsilon$ for all $x \in X$. Hence $J_{\pi_\epsilon,\delta} \in \mathcal{B}$, implying by part (a) that π_ϵ is proper. **Q.E.D.**

The next proposition shows that the cost function J_δ^* of the δ-perturbed problem can be used to approximate $\hat{J}$.

Proposition 3.2: If there exists at least one proper policy, we have $\lim_{\delta\downarrow 0} J_\delta^*(x) = \hat{J}(x)$ for all $x \in X$.

Proof: Let π_ϵ be a proper ϵ-optimal policy for the δ-perturbed problem [cf. Prop. 3.1(b)]. By using Eq. (2.3), we have for all $\delta > 0$, $\epsilon > 0$, and $\pi \in \widehat{\Pi}$,

$$\hat{J}(x) - \epsilon \leq J_{\pi_\epsilon}(x) - \epsilon \leq J_{\pi_\epsilon,\delta}(x) - \epsilon \leq J^*_\delta(x) \leq J_{\pi,\delta}(x) = J_\pi(x) + w_{\pi,\delta}(x), \qquad \forall\ x \in X_f,$$

where

$$w_{\pi,\delta}(x) = \delta \sum_{k=0}^{\infty} r_k(\pi, x), \qquad x \in X.$$

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

$$\hat{J}(x) \leq J^*_\delta(x) \leq J_\pi(x) + w_{\pi,\delta}(x), \qquad \forall\ x \in X_f.$$

We have $\lim_{\delta \downarrow 0} w_{\pi,\delta}(x) = 0$ for all $x \in X_f$ and $\pi \in \widehat{\Pi}$, so by taking the limit as $\delta \downarrow 0$ and then the infimum over all $\pi \in \widehat{\Pi}$,

$$\hat{J}(x) \leq \lim_{\delta \downarrow 0} J^*_\delta(x) \leq \inf_{\pi \in \widehat{\Pi}} J_\pi(x) = \hat{J}(x), \qquad \forall\ x \in X_f,$$

from which $\hat{J}(x) = \lim_{\delta \downarrow 0} J^*_\delta(x)$ for all $x \in X_f$. Since we also have $J^*_\delta(x) = \hat{J}(x) = \infty$ for all $x \notin X_f$, the result follows. **Q.E.D.**

The next proposition sets the stage for our main result.

Proposition 3.3: Assume that there exists at least one proper policy $\pi \in \widehat{\Pi}$. For all $\delta > 0$, J^*_δ is the unique solution within $\mathcal{B}$ of Bellman's equation for the δ-perturbed problem,

$$J(t) = 0, \qquad J(x) = \inf_{u \in U(x)} \Big\{ g(x,u) + \delta + E\big\{ J\big(f(x,u,w)\big)\big\}\Big\}, \qquad x \neq t. \tag{3.1}$$

Proof: We have that J^*_δ is a solution of Bellman's equation (3.1) by Prop. 2.1(a). Moreover, by Prop. 3.1(b) $J^*_\delta \in \mathcal{B}$. To show that J^*_δ is the unique solution within $\mathcal{B}$, let $\tilde{J} \in \mathcal{B}$ be another solution, so that using also Prop. 2.1(a), we have

$$J^*_\delta(x) \leq \tilde{J}(x) \leq g(x,u) + \delta + E\big\{ \tilde{J}\big(f(x,u,w)\big)\big\}, \qquad \forall\ x \in X,\ u \in U(x). \tag{3.2}$$

For a given $\epsilon > 0$, let $\pi_\epsilon = \{\mu_0, \mu_1, \ldots\}$ be a proper ϵ-optimal policy [which exists by Prop. 3.1(b)]. By repeatedly applying the preceding relation, we have for any $x_0 \in X_f$

$$J^*_\delta(x_0) \leq \tilde{J}(x_0) \leq E^{\pi_\epsilon}_{x_0}\left\{ \tilde{J}(x_k) + \delta \sum_{m=0}^{k-1} r_m(\pi_\epsilon, x_0) + \sum_{m=0}^{k-1} g\big(x_m, \mu_m(x_m)\big)\right\}, \qquad \forall\ k \geq 1,$$

where $\{x_k\}$ is the sequence generated starting from x_0 and using π_ϵ. Since π_ϵ is proper, $J_{\pi_\epsilon}(x_0) < \infty$, so by Prop. 2.2(a), $x_k \in X_f$ with probability one for all k, and therefore $E_{x_0}^{\pi_\epsilon}\{\tilde{J}(x_k)\} \to 0$ (in view of $\tilde{J} \in \mathcal{B}$). It follows that

$$\lim_{k\to\infty} E_{x_0}^{\pi}\left\{\tilde{J}(x_k) + \delta \sum_{m=0}^{k-1} r_m(\pi_\epsilon, x_0) + \sum_{m=0}^{k-1} g\big(x_m, \mu_m(x_m)\big)\right\} = J_{\pi_\epsilon,\delta}(x_0) \leq J_\delta^*(x_0) + \epsilon.$$

By combining the preceding two relations, we obtain,

$$J_\delta^*(x_0) \leq \tilde{J}(x_0) \leq J_{\pi_\epsilon,\delta}(x_0) \leq J_\delta^*(x_0) + \epsilon, \qquad \forall\ x_0 \in X_f.$$

By letting $\epsilon \to 0$, it follows that $J_\delta^*(x_0) = \tilde{J}(x_0)$ for all $x_0 \in X_f$. Also for $x_0 \notin X_f$, we have $J^*(x_0) = J_\delta^*(x_0) = \tilde{J}(x_0) = \infty$ [since $J^* \leq J_\delta^* \leq \tilde{J}$, cf. Eq. (3.2)]. Thus $J_\delta^* = \tilde{J}$, proving that J_δ^* is the unique solution of the Bellman Eq. (3.1) within $\mathcal{B}$. **Q.E.D.**

Using the preceding propositions, we will now show our main result: $\hat{J}$ is the unique solution of Bellman's equation within the set of functions

$$\widehat{\mathcal{W}} = \{J \in \mathcal{B} \mid \hat{J} \leq J\}, \tag{3.3}$$

and the VI algorithm yields $\hat{J}$ in the limit for any initial $J_0 \in \widehat{\mathcal{W}}$.

Proposition 3.4: Assume that there exists at least one proper policy. Then:

(a) $\hat{J}$ is the unique solution of the Bellman Eq. (2.1) within the set $\widehat{\mathcal{W}}$ of Eq. (3.3).

(b) (*VI Convergence*) If $\{J_k\}$ is the sequence generated by the VI algorithm (1.5) starting with some $J_0 \in \widehat{\mathcal{W}}$, then $J_k \to \hat{J}$.

(c) (*Optimality Condition*) If $\hat{\mu}$ is a proper stationary policy and

$$\hat{\mu}(x) \in \arg\min_{u\in U(x)} \Big\{g(x,u) + E\big\{\hat{J}\big(f(x,u,w)\big)\big\}\Big\}, \qquad \forall\ x \in X, \tag{3.4}$$

then $\hat{\mu}$ is optimal over the set of proper policies. Conversely, if $\hat{\mu}$ is optimal within the set of proper policies, then it satisfies the preceding condition (3.4).

Proof: (a), (b) We note that $\hat{J} \in \widehat{\mathcal{W}}$, since by Props. 3.1(a) and 3.2, we have $J_\delta^* \in \mathcal{B}$ and $\hat{J} \leq J_\delta^*$. We first show that $\hat{J}$ is a solution of Bellman's equation and then show that it is the unique solution within $\mathcal{B}$ by showing the convergence of VI [cf. part (b)].

From Prop. 3.3 and the fact $J_\delta^* \geq \hat{J}$ shown in Prop. 3.2, we have for all $\delta > 0$ and $x \neq t$,

$$\begin{aligned} J_\delta^*(x) &= \inf_{u \in U(x)} \Big\{ g(x,u) + \delta + E\big\{ J_\delta^*\big(f(x,u,w)\big)\big\}\Big\} \\ &\geq \inf_{u \in U(x)} \Big\{ g(x,u) + E\big\{ J_\delta^*\big(f(x,u,w)\big)\big\}\Big\} \\ &\geq \inf_{u \in U(x)} \Big\{ g(x,u) + E\big\{ \hat{J}\big(f(x,u,w)\big)\big\}\Big\}. \end{aligned}$$

By taking the limit as $\delta \downarrow 0$ and using Prop. 3.2, we obtain

$$\hat{J}(x) \geq \inf_{u \in U(x)} \Big\{ g(x,u) + E\big\{ \hat{J}\big(f(x,u,w)\big)\big\}\Big\}, \qquad \forall\ x \in X. \tag{3.5}$$

For the reverse inequality, let $\{\delta_m\}$ be a sequence with $\delta_m \downarrow 0$. From Prop. 3.3, we have for all m, $x \neq t$, and $u \in U(x)$,

$$g(x,u) + \delta_m + E\big\{ J_{\delta_m}^*\big(f(x,u,w)\big)\big\} \geq \inf_{v \in U(x)} \Big\{ g(x,v) + \delta_m + E\big\{ J_{\delta_m}^*\big(f(x,v,w)\big)\big\}\Big\} = J_{\delta_m}^*(x).$$

Taking the limit as $m \to \infty$, and using the fact $\lim_{\delta_m \downarrow 0} J_{\delta_m}^* = \hat{J}$ (cf. Prop. 3.2), we have

$$g(x,u) + E\big\{ \hat{J}\big(f(x,u,w)\big)\big\} \geq \hat{J}(x), \qquad \forall\ x \in X,\ u \in U(x),$$

so that

$$\inf_{u \in U(x)} \Big\{ g(x,u) + E\big\{ \hat{J}\big(f(x,u,w)\big)\big\}\Big\} \geq \hat{J}(x), \qquad \forall\ x \in X. \tag{3.6}$$

By combining Eqs. (3.5) and (3.6), we see that $\hat{J}$ is a solution of Bellman's equation.

We will next show that $J_k \to \hat{J}$ starting from every initial $J_0 \in \widehat{\mathcal{W}}$ [cf. part (b)]. Indeed, for $x_0 \in X_f$ and any $\pi \in \widehat{\Pi}$, let $\{x_k\}$ be the generated sequence starting from x_0. Since from the definition of the VI sequence $\{J_k\}$, we have

$$J_k(x) \leq g(x,u) + E\big\{ J_{k-1}\big(f(x,u,w)\big)\big\}, \qquad \forall\ x \in X,\ u \in U(x),\ k = 1, 2, \ldots,$$

it follows that

$$J_k(x_0) \leq E_{x_0}^\pi \left\{ J_0(x_k) + \sum_{m=0}^{k-1} g\big(x_m, \mu_m(x_m)\big) \right\}.$$

We have $E_{x_0}^\pi\big\{J_0(x_k)\big\} \to 0$ since π is proper, $x_k \in X_f$ with probability one, and $J_0 \in \mathcal{B}$, so by taking the limit as $k \to \infty$, it follows that $\limsup_{k\to\infty} J_k(x_0) \leq J_\pi(x_0)$. By taking the infimum over all $\pi \in \widehat{\Pi}$, we obtain $\limsup_{k\to\infty} J_k(x_0) \leq \hat{J}(x_0)$. Conversely, since $\hat{J} \leq J_0$ and $\hat{J}$ is a solution of Bellman's equation (as shown earlier), it follows by induction that $\hat{J} \leq J_k$ for all k. Thus $\hat{J}(x_0) \leq \liminf_{k\to\infty} J_k(x_0)$, implying that $J_k(x_0) \to \hat{J}(x_0)$ for all $x_0 \in X_f$. We also have $J^* \leq \hat{J} \leq J_k$ for all k, so that $\hat{J}(x_0) = J_k(x_0) = \infty$ for all $x_0 \notin X_f$. This completes the proof of part (b). Finally, since $\hat{J} \in \widehat{\mathcal{W}}$ and $\hat{J}$ is a solution of Bellman's equation., part (b) implies the uniqueness assertion of part (a).

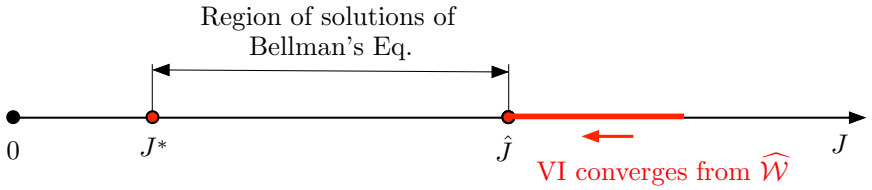

Figure 3.1 Illustration of the solutions of Bellman's equation. The smallest and the largest solutions within $\mathcal{B}$ are J^* and $\hat{J}$, respectively. The VI algorithm converges to $\hat{J}$ starting from any $J_0 \in \mathcal{B}$ with $J_0 \geq \hat{J}$.

(c) If μ is proper and Eq. (3.4) holds, then

$$\hat{J}(x) = g\big(x, \mu(x)\big) + E\big\{\hat{J}\big(f(x, \mu(x), w)\big)\big\}, \qquad x \in X.$$

By Prop. 2.1(b), this implies that $J_\mu \leq \hat{J}$, so μ is optimal over the set of proper policies. Conversely, assume that μ is proper and $J_\mu = \hat{J}$. Then by Prop. 2.1(b), we have

$$\hat{J}(x) = g\big(x, \mu(x)\big) + E\big\{\hat{J}\big(f(x, \mu(x), w)\big)\big\}, \qquad x \in X,$$

and since [by part (a)] $\hat{J}$ is a solution of Bellman's equation.,

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

Combining the last two relations, we obtain Eq. (3.4). **Q.E.D.**

We illustrate Prop. 3.4 in Fig. 3.1. Examples given in [Ber17] show that between J^* and $\hat{J}$ there can be any number of solutions of Bellman's equation: a finite number, an infinite number, or none at all.

Suppose now that the set of proper policies is sufficient in the sense that it can achieve the same optimal cost as the set of all policies, i.e., $\hat{J} = J^*$. Then, from Prop. 3.4, it follows that J^* is the unique solution of Bellman's equation within $\mathcal{B}$, and the VI algorithm converges to J^* starting from any $J_0 \in \mathcal{B}$ with $J_0 \geq J^*$. Under additional conditions, such as finiteness of $U(x)$ for all $x \in X$ [cf. Prop. 2.1(e)], VI converges to J^* starting from any $J_0 \in \mathcal{B}$.

4. CONCLUDING REMARKS

We have considered SSP problems, which involve arbitrary state and control spaces, and a Bellman's equation with possibly multiple solutions. Within this context, we have considered the restricted optimization problem over the proper policies only. The weakness of our main result is that it assumes existence of a proper policy, which by definition has a bounded cost function. Thus the proposition may apply naturally to problems with bounded cost per stage (e.g., problems with state space that is bounded with respect to some metric), but

may not apply naturally to problems with unbounded cost per stage, such as the classical linear-quadratic optimal control models. If $X_f = X$, it is possible to check the existence of a proper policy by introducing an additional stopping action with high cost c at every $x \neq t$. The policy that uses the stopping action at all states is proper for the resulting modified problem. It follows that there exists a proper policy, if and only if for c high enough, the stopping action is nowhere optimal for the modified problem.

5. REFERENCES

[Alt99] Altman, E., 1999. Constrained Markov Decision Processes, CRC Press, Boca Raton, FL.

[BeS78] Bertsekas, D. P., and Shreve, S. E., 1978. Stochastic Optimal Control: The Discrete Time Case, Academic Press, N. Y. (republished by Athena Scientific, Belmont, MA, 1996); may be downloaded from http://web.mit.edu/dimitrib/www/home.html.

[BeT91] Bertsekas, D. P., and Tsitsiklis, J. N., 1991. "An Analysis of Stochastic Shortest Path Problems," Math. of OR, Vol. 16, pp. 580-595.

[BeY16] Bertsekas, D. P., and Yu, H., 2016. "Stochastic Shortest Path Problems Under Weak Conditions," Lab. for Information and Decision Systems Report LIDS-2909.

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

[Ber17] Bertsekas, D. P., 2017. "Stable Optimal Control and Semicontractive Dynamic Programming," Report LIDS-P-3506, MIT, May 2017.

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

[JaC06] James, H. W., and Collins, E. J., 2006. "An Analysis of Transient Markov Decision Processes," J. Appl. Prob., Vol. 43, pp. 603-621.

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

[Pal67] Pallu de la Barriere, R., 1967. Optimal Control Theory, Saunders, Phila; Dover, N. Y., 1980.

[Put94] Puterman, M. L., 1994. Markov Decision Processes: Discrete Stochastic Dynamic Programming, J. Wiley, N. Y.

[Whi82] Whittle, P., 1982. Optimization Over Time, Wiley, N. Y., Vol. 1, 1982, Vol. 2, 1983.