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, a nonnegative cost per stage, and a termination state. We extend the notion of a proper policy, a policy that terminates within a finite expected number of steps, 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. We show that J^* and $\hat{J}$ are the smallest and largest solutions of Bellman's equation, respectively, within a suitable class of Lyapounov-like functions. If the cost per stage is bounded, these functions are those that are bounded over the effective domain of $\hat{J}$. The standard value iteration algorithm may be attracted to either J^* or $\hat{J}$, depending on the initial condition.

I. Introduction

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

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

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 the complicated mathematical measurability issues in the choice of control.[1] 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, $E\big\{g(x_k, u_k, w_k)\big\}$, is assumed real-valued and nonnnegative:

$$0 \leq E\big\{g(x_k, u_k, w_k)\big\} < \infty, \quad \forall\, x \in X,\ u \in U(x). \tag{2}$$

We assume that X contains a special cost-free and absorbing state t, referred to as the *destination*:

$$f(t, u, w) = t, \qquad g(t, u, w) = 0, \qquad \forall\, u \in U(t),\ w \in W. \tag{3}$$

The essence of the problem is to reach or approach the destination with minimum expected cost.

D. P. Bertsekas is with the Computer Information and Decision Science and Engineering Dept., Arizona State University, Tempe, AZ, and the Laboratory for Information and Decision Systems (LIDS), M.I.T. Email: dimitrib@mit.edu.

[1]The nature of these difficulties is well-documented; see the monograph by Bertsekas and Shreve [1], and the paper by James and Collins [2], which treats stochastic shortest path problems. It may be reasonably conjectured that our analysis can be extended to hold within an appropriate measurability framework, but this undertaking is beyond the scope of the present paper.

We are interested in policies of the form $\pi = \{\mu_0, \mu_1, \dots\}$, 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 $\pi = \{\mu, \mu, \dots\}$ are called *stationary*, and will be denoted by μ, when confusion cannot arise.

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

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

where $E_{x_0}^{\pi}\{\cdot\}$ denotes expectation with respect to the probability measure corresponding to initial state x_0 and policy π, and the series converges in view of the nonnegativity of cost per stage g. 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

$$J(x) = \inf_{u \in U(x)} E\Big\{g(x, u, w) + J\big(f(x, u, w)\big)\Big\}, \qquad x \in X, \tag{4}$$

where the expected value is with respect to the 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 when $J = J^*$ [cf. Prop. 1(d) 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

$$J_{k+1}(x) = \inf_{u \in U(x)} E\Big\{g(x, u, w) + J_k\big(f(x, u, w)\big)\Big\}, \quad x \in X. \tag{5}$$

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 [3] we have addressed the connections between stability and the solutions of Bellman's equation 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 the ones that are guaranteed to reach the termination state within a finite expected number of steps, starting from the states where the optimal cost is finite (a precise definition is given in the next section). Proper policies may be viewed as the analog of stable policies in a deterministic context, and their significance is well known in finite-state SSP problems (see e.g., the books [4], [5], [6], [7], [8], [9], [10], and [11], and the references quoted there). For the case where $g \geq 0$, the paper by Bertsekas and Tsitsiklis [12] provides an analysis that bears similarity with the one of the present paper, but assumes a finite state space and that there exists an optimal policy that is proper. In the infinite-state context of this paper and under weaker assumptions, we show that $\hat{J}$, the optimal cost function over just the proper policies, is the largest solution of Bellman's equation within a set of functions $\widehat{\mathcal{W}} \subset \mathcal{J}$ that majorize $\hat{J}$, and that the VI algorithm converges to $\hat{J}$ starting from a function in $\widehat{\mathcal{W}}$. Our line of analysis draws its origin from concepts of regularity introduced by the author in the monograph [13] and the paper [14].

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 [15], and James and Collins [2], where they are called *transient*. There are a few differences between the frameworks of [15], [2] and this paper, which impact on the results obtained. In particular, the paper [15] uses a similar (but not identical) definition of properness to the one of the present paper, but assumes that all policies are proper, that g is bounded, and that J^* is real-valued. The paper [2] uses the properness definition of [15], and extends the analysis of [11] from finite state space to infinite state space (addressing also measurability issues). Moreover, [2] allows the cost per stage g to take both positive and negative values. However, [2] uses assumptions that guarantee that improper policies cannot be optimal and that $J^* = \hat{J}$, while J^* is real-valued. This is the most important difference from the analysis of this paper.

Our analysis is also related to the one of Bertsekas and Yu [16], where the case $J^* \neq \hat{J}$ was analyzed using perturbation ideas that are similar to the ones of Section 3. The paper [16] assumes that the state space is finite and that J^* is real-valued, but allows g to take negative values. Moreover [16] gives an example showing that J^* may not be a solution of Bellman's equation if improper policies can be optimal, and g can take both positive and negative values. 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 and cannot be extended without major modifications.

II. 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 [17]. For textbook accounts we refer to [1], [8], [11], and for a more abstract development, we refer to the monograph [13]. The following proposition gives the results that we will need.

Proposition 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 for all $x \in X$*

$$J(x) = \inf_{u \in U(x)} E\Big\{ g(x,u,w) + J\big(f(x,u,w)\big) \Big\}, \tag{6}$$

then $J^* \leq J$.

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

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

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 for all $x \in X$, we have*

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

(d) *A stationary policy μ^* is optimal if and only if for all $x \in X$, we have*

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

(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 (5) starting from any J_0 with $0 \leq J_0 \leq J^*$.*

Proof. See [1], Props. 5.2, 5.4, and 5.10, or [11], Props. 4.1.1, 4.1.3, 4.1.5, 4.1.9. □

For a given state $x \in X$, a policy π is said to be *proper at x* if

$$J_\pi(x) < \infty, \qquad \sum_{k=0}^{\infty} r_k(\pi, x) < \infty, \tag{7}$$

where $r_k(\pi, x_0)$ is the probability that $x_k \neq t$ when using π and starting from $x_0 = x$. Note that the sum $\sum_{k=0}^{\infty} r_k(\pi, x)$ is the expected number of steps to reach the destination starting from x and using π.

We denote by $\widehat{\Pi}_x$ the set of all policies that are proper at x, and we use the notation

$$\mathcal{C} = \big\{(\pi, x) \mid \pi \in \hat{\Pi}_x\big\}. \tag{8}$$

We denote by $\hat{J}$ the corresponding restricted optimal cost function,

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

Finally we denote by $\widehat{X}$ the effective domain of $\hat{J}$, i.e.,

$$\widehat{X} = \big\{x \in X \mid \hat{J}(x) < \infty\big\}. \tag{9}$$

Note that $\widehat{X}$ is the set of all x such that $\widehat{\Pi}_x$ is nonempty.

The definition of proper policy just given differs from the definition of a transient policy adopted by James and Collins [2]. In particular, the definition of [2] requires that the expected number of steps to reach the destination is uniformly bounded over the initial state x (see [2], p. 608) and is not tied to a single state x.

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

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

while $g(x,u,w)$ 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

$$J_{\pi,\delta}(x) = J_\pi(x) + \delta \sum_{k=0}^{\infty} r_k(\pi, x). \tag{10}$$

We denote by $\hat{J}_\delta$ the optimal cost function of the δ-perturbed problem, i.e., $\hat{J}_\delta(x) = \inf_{\pi \in \Pi} J_{\pi,\delta}(x)$. The following proposition relates the δ-perturbed problem with proper policies.

Proposition 2. (a) *A policy π is proper at a state $x \in X$ if and only if $J_{\pi,\delta}(x) < \infty$.*

(b) *We have $\hat{J}_\delta(x) < \infty$ for all $\delta > 0$ if and only if $x \in \widehat{X}$.*

(c) *For every $\epsilon > 0$ and $\delta > 0$, there exists a policy π_ϵ that is proper at all $x \in \widehat{X}$ and is ϵ-optimal for the δ-perturbed problem, i.e.,*

$$J_{\pi_\epsilon,\delta}(x) \leq \hat{J}_\delta(x) + \epsilon, \qquad \forall\, x \in X.$$

Proof. (a) Follows from Eq. (10) and the definition (7) of a proper policy.

(b) If $x \in \widehat{X}$ there exists a policy π that is proper at x, and by part (a), $\hat{J}_\delta(x) \leq J_{\pi,\delta}(x) < \infty$ for all $\delta > 0$. Conversely, if $\hat{J}_\delta(x) < \infty$, there exists π such that $J_{\pi,\delta}(x) < \infty$, implying [by part (a)] that $\pi \in \widehat{\Pi}_x$, so that $x \in \widehat{X}$.

(c) By Prop. 1(c), there exists an ϵ-optimal policy π_ϵ for the δ-perturbed problem, so we have $J_{\pi_\epsilon,\delta}(x) \leq \hat{J}_\delta(x) + \epsilon$ for all $x \in X$. Hence $J_{\pi_\epsilon,\delta}(x) < \infty$ for all $x \in \widehat{X}$, implying by part (a) that π_ϵ is proper at all $x \in \widehat{X}$. □

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

Proposition 3. *We have $\lim_{\delta \downarrow 0} \hat{J}_\delta(x) = \hat{J}(x)$ for all $x \in X$.*

Proof. Let us fix $\delta > 0$, and for a given $\epsilon > 0$, let π_ϵ be a policy that is proper at all $x \in \widehat{X}$ and is ϵ-optimal for the δ-perturbed problem [cf. Prop. 2(c)]. By using Eq. (10), we have for all $\epsilon > 0$, $x \in \widehat{X}$, and $\pi \in \widehat{\Pi}_x$,

$$\begin{aligned} \hat{J}(x) - \epsilon &\leq J_{\pi_\epsilon}(x) - \epsilon \\ &\leq J_{\pi_\epsilon,\delta}(x) - \epsilon \\ &\leq \hat{J}_\delta(x) \\ &\leq J_{\pi,\delta}(x) \\ &= J_\pi(x) + w_{\pi,\delta}(x), \qquad \forall\, x \in \widehat{X}, \end{aligned}$$

where

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

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

$$\hat{J}(x) \leq \hat{J}_\delta(x) \leq J_\pi(x) + w_{\pi,\delta}(x), \qquad \forall\, x \in \widehat{\Pi}_x.$$

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

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

from which $\hat{J}(x) = \lim_{\delta \downarrow 0} \hat{J}_\delta(x)$ for all $x \in \widehat{X}$. Since by Prop. 2(b), we also have $\hat{J}_\delta(x) = \hat{J}(x) = \infty$ for all $x \notin \widehat{X}$, the result follows. □

III. Main Results

By Prop. 1(a), $\hat{J}_\delta$ solves Bellman's equation for the δ-perturbed problem, while by Prop. 3, $\lim_{\delta\downarrow 0}\hat{J}_\delta(x)=\hat{J}(x)$. This suggests that $\hat{J}$ solves the unperturbed Bellman equation, which is the "limit" as $\delta\downarrow 0$ of the δ-perturbed version. Indeed, under a certain assumption we will show a stronger result, namely that $\hat{J}$ is the unique solution of Bellman's equation within the set of functions

$$\widehat{\mathcal{W}}=\Big\{J\in\mathcal{J}\mid \hat{J}\le J,\ E^{\pi}_{x_0}\big\{J(x_k)\big\}\to 0,\ \forall\ (\pi,x_0)\in\mathcal{C}\Big\}, \tag{11}$$

where $\mathcal{C}$ is given by Eq. (8), $E^{\pi}_{x_0}\{\cdot\}$ denotes expected value with respect to the probability measure corresponding to initial state x_0 under policy π, and $E^{\pi}_{x_0}\big\{J(x_k)\big\}$ denotes the expected value of the function J along the sequence $\{x_k\}$ generated starting from x_0 and using π. The functions in $\widehat{\mathcal{W}}$ are the ones whose expected value is decreasing to 0 along the trajectories generated by the proper policies, so they may be interpreted as a type of Lyapounov functions.

Given a policy $\pi=\{\mu_0,\mu_1,\ldots\}$, we denote by π_k the policy

$$\pi_k=\{\mu_k,\mu_{k+1},\ldots\}. \tag{12}$$

We first show a preliminary result.

Proposition 4. *The following hold:*

(a) *For all pairs $(\pi,x_0)\in\mathcal{C}$ and $k=0,1,\ldots$, we have*

$$0\le E^{\pi}_{x_0}\big\{\hat{J}(x_k)\big\}\le E^{\pi}_{x_0}\big\{J_{\pi_k}(x_k)\big\},$$

where π_k is the policy given by Eq. (12).

(b) *The set $\widehat{\mathcal{W}}$ of Eq. (11) contains $\hat{J}$, as well as all $J\in\widehat{\mathcal{W}}$ satisfying $\hat{J}\le J\le c\hat{J}$ for some $c\ge 1$.*

Proof. (a) For any pair $(\pi,x_0)\in\mathcal{C}$ and $\delta>0$, we have

$$J_{\pi,\delta}(x_0)=E^{\pi}_{x_0}\Bigg\{J_{\pi_k,\delta}(x_k)+k\delta + \sum_{m=0}^{k-1} g\big(x_m,\mu_m(x_m),w_m\big)\Bigg\}.$$

Since $J_{\pi,\delta}(x_0)<\infty$ [cf. Prop. 2(a)], it follows that $E^{\pi}_{x_0}\big\{J_{\pi_k,\delta}(x_k)\big\}<\infty$. Hence for all x_k that can be reached with positive probability using π and starting from x_0, we have $J_{\pi_k,\delta}(x_k)<\infty$, implying [by Prop. 2(a)] that $(\pi_k,x_k)\in\mathcal{C}$ and hence $\hat{J}(x_k)\le J_{\pi_k}(x_k)$. By applying $E^{\pi}_{x_0}\{\cdot\}$ to this last inequality, the result follows.

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

$$J_\pi(x_0)=E^{\pi}_{x_0}\Big\{g\big(x_0,\mu_0(x_0),w_0\big)\Big\}+E^{\pi}_{x_0}\big\{J_{\pi_1}(x_1)\big\}, \tag{13}$$

and for all $m=1,2,\ldots,$

$$E^{\pi}_{x_0}\big\{J_{\pi_m}(x_m)\big\}=E^{\pi}_{x_0}\Big\{g\big(x_m,\mu_m(x_m),w_m\big)\Big\} + E^{\pi}_{x_0}\big\{J_{\pi_{m+1}}(x_{m+1})\big\}, \tag{14}$$

where $\{x_m\}$ is the sequence generated starting from x_0 and using π. By using repeatedly the expression (14) for $m = 1, \ldots, k-1$, and combining it with Eq. (13), we obtain for all $k = 1, 2, \ldots,$ and $(\pi, x_0) \in \mathcal{C}$,

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

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

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

By part (a), it follows that $E^\pi_{x_0}\big\{\hat{J}(x_k)\big\} \to 0$ for all $(\pi, x_0) \in \mathcal{C}$, so that $\hat{J} \in \widehat{\mathcal{W}}$. This also implies that $E^\pi_{x_0}\big\{J(x_k)\big\} \to 0$ for all $(\pi, x_0) \in \mathcal{C}$, if $\hat{J} \le J \le c\hat{J}$ for some $c \ge 1$. $\square$

We can now prove our main result. We denote by X^* the effective domain of J^*:

$$X^* = \big\{x \in X \mid J^*(x) < \infty\big\}. \tag{15}$$

We also denote by $Q^*_\delta(x, u)$ the optimal Q-factor of a pair (x, u) in the δ-perturbed problem:

$$Q^*_\delta(x, u) = E\Big\{g(x, u, w) + \delta + \hat{J}_\delta\big(f(x, u, w)\big)\Big\}.$$

The following proposition shows that $\hat{J}$ is the unique solution of the Bellman equation within the set $\widehat{\mathcal{W}}$ of Lyapounov functions under a certain assumption relating to the states in X^*. This assumption can often be easily verified in practice. It is satisfied for example if there exists a policy π (necessarily proper at all $x \in X^*$) such that $J_{\pi,\delta}$ is bounded over the set X^*. Later, we will also prove the result under the alternative assumption that the set of disturbances W is finite.

Proposition 5. *Assume that there exists a $\delta > 0$ such that*

$$Q^*_\delta(x, u) < \infty, \qquad \forall\ x \in X^*,\ u \in U(x). \tag{16}$$

Then:

(a) *$\hat{J}$ is the unique solution of the Bellman Eq. (6) within the set $\widehat{\mathcal{W}}$ of Eq. (11).*

(b) *(VI Convergence) If $\{J_k\}$ is the sequence generated by the VI algorithm (5) starting with some $J_0 \in \widehat{\mathcal{W}}$, then $J_k \to \hat{J}$.*

(c) *(Optimality Condition) If μ is a stationary policy that is proper at all $x \in \widehat{X}$, and for all $x \in X$ we have*

$$\mu(x) \in \arg\min_{u \in U(x)} E\Big\{g(x, u, w) + \hat{J}\big(f(x, u, w)\big)\Big\}, \tag{17}$$

then μ is optimal over the set of proper policies, i.e., $J_\mu = \hat{J}$. Conversely, if μ is optimal within the set of proper policies, then it satisfies the preceding condition (17).

Proof. (a), (b) By Prop. 4(b), $\hat{J} \in \widehat{\mathcal{W}}$. We will first show that $\hat{J}$ is a solution of Bellman's equation and then show that it is the unique solution within $\widehat{\mathcal{W}}$ by showing the convergence of VI [cf. part (b)]. Since $\hat{J}_\delta$ solves the Bellman equation for the δ-perturbed problem, and $\hat{J}_\delta \geq \hat{J}$ (cf. Prop. 3), we have for all $\delta > 0$ and $x \neq t$,

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

By taking the limit as $\delta \downarrow 0$ and using Prop. 3, we obtain

$$\hat{J}(x) \geq \inf_{u \in U(x)} E\Big\{g(x,u,w) + \hat{J}\big(f(x,u,w)\big)\Big\}, \qquad \forall\ x \in X. \tag{18}$$

To prove the reverse inequality, we consider two cases:

(1) $x \notin X^*$, i.e., $J^*(x) = \infty$. Then from Bellman's equation, we have

$$\infty = J^*(x) = \inf_{u \in U(x)} E\Big\{g(x,u,w) + J^*\big(f(x,u,w)\big)\Big\}.$$

Since $\hat{J} \geq J^*$ it then follows that

$$\infty = \hat{J}(x) = \inf_{u \in U(x)} E\Big\{g(x,u,w) + \hat{J}\big(f(x,u,w)\big)\Big\},\ \forall\ x \notin X^*. \tag{19}$$

(2) $x \in X^*$. Then we let $\{\delta_m\}$ be a sequence with $\delta_m \downarrow 0$. We have for all m, $x \neq t$, and $u \in U(x)$,

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

We now take limit as $m \to \infty$ in the preceding relation. The condition (16) implies that for all m sufficiently large the left side is finite for each $u \in U(x)$, so we can apply the monotone convergence theorem to interchange limit as $m \to \infty$ and expectation.[2] Since $\lim_{\delta_m \downarrow 0} \hat{J}_{\delta_m} = \hat{J}$ (cf. Prop. 3), we obtain

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

[2] We are using here the following version of the monotone convergence theorem: Let $\{h_m\}$ be a sequence of monotonically nonincreasing functions $h_m : \{1,2,\ldots\} \mapsto \Re$, let $\{p_1, p_2, \ldots\}$ be a probability distribution, and assume that for some function $\bar{h} : \{1,2,\ldots\} \mapsto \Re$ such that $h_m(i) \leq \bar{h}(i)$ for all m and i, we have $\sum_{i=1}^\infty p_i \bar{h}(i) < \infty$. Then

$$\lim_{m \to \infty} \sum_{i=1}^\infty p_i h_m(i) = \sum_{i=1}^\infty p_i \lim_{m \to \infty} h_m(i).$$

We give the proof, which is simple in the discrete distribution case considered here: Let h be the pointwise limit of $\{h_m\}$, i.e., $h(i) = \lim_{m \to \infty} h_m(i)$ for all i. Since $\{h_m\}$ is nonincreasing, we have

$$\sum_{i=1}^\infty p_i h_m(i) \geq \sum_{i=1}^\infty p_i h(i), \qquad \forall\ m = 0, 1, \ldots,$$

so that

$$\lim_{m \to \infty} \sum_{i=1}^\infty p_i h_m(i) \geq \sum_{i=1}^\infty p_i h(i).$$

so that

$$\inf_{u\in U(x)} E\Big\{g(x,u,w)+\hat J\big(f(x,u,w)\big)\Big\}\ge \hat J(x),\qquad \forall\ x\in X^*. \tag{20}$$

Thus by combining Eqs. (18), (19), and (20), we see that

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

and 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 $x_0\in\widehat X$ and any $\pi=\{\mu_0,\mu_1,\ldots\}\in\widehat\Pi_{x_0}$, let $\{x_k\}$ be the generated sequence starting from x_0. Since from the definition of the VI sequence $\{J_k\}$ [cf. Eq. (5)], we have for all $x\in X$, $u\in U(x)$, $k=1,2,\ldots$,

$$J_k(x)\le E\Big\{g(x,u,w)+J_{k-1}\big(f(x,u,w)\big)\Big\},$$

it follows that

$$J_k(x_0)\le E^\pi_{x_0}\left\{J_0(x_k)+\sum_{m=0}^{k-1} g\big(x_m,\mu_m(x_m),w_m\big)\right\}.$$

Since $J_0\in\widehat{\mathcal{W}}$, we have $E^\pi_{x_0}\big\{J_0(x_k)\big\}\to 0$, so by taking the limit as $k\to\infty$ in the preceding relation, it follows that $\limsup_{k\to\infty} J_k(x_0)\le J_\pi(x_0)$. By taking the infimum over all $\pi\in\widehat\Pi_{x_0}$, we obtain $\limsup_{k\to\infty} J_k(x_0)\le\hat J(x_0)$. Conversely, since $\hat J\le J_0$ and $\hat J$ is a solution of Bellman's equation (as shown earlier), it follows by induction that $\hat J\le J_k$ for all k. Thus $\hat J(x_0)\le\liminf_{k\to\infty} J_k(x_0)$, implying that $J_k(x_0)\to\hat J(x_0)$ for all $x_0\in\widehat X$. We also have $\hat J\le J_k$ for all k, so that $\hat J(x_0)=J_k(x_0)=\infty$ for all $x_0\notin\widehat X$. 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).

(c) If μ is proper at all $x\in\widehat X$ and Eq. (17) holds, then

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

By Prop. 1(b), this implies that $J_\mu\le\hat J$, so μ is optimal over the set of proper policies. Conversely, assume that μ is proper at all $x\in\widehat X$ and $J_\mu=\hat J$. Then by Prop. 1(b), we have

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

Conversely, since $\bar h(i)-h_m(i)\ge 0$ for all m, we have for every $N\ge 1$

$$\sum_{i=1}^\infty p_i\big(\bar h(i)-h_m(i)\big)\ge\sum_{i=1}^N p_i\big(\bar h(i)-h_m(i)\big),$$

and hence

$$\lim_{m\to\infty}\sum_{i=1}^\infty p_i\big(\bar h(i)-h_m(i)\big)\ge\lim_{m\to\infty}\sum_{i=1}^N p_i\big(\bar h(i)-h_m(i)\big),$$

so that

$$\sum_{i=1}^\infty p_i\bar h(i)-\lim_{m\to\infty}\sum_{i=1}^\infty p_i h_m(i)\ge\sum_{i=1}^N p_i\bar h(i)-\sum_{i=1}^N p_i h(i).$$

By taking the limit as $N\to\infty$ and using the fact that $\sum_{i=1}^\infty p_i\bar h(i)$ is finite (so we can cancel it from both sides of the inequality), we obtain

$$\lim_{m\to\infty}\sum_{i=1}^\infty p_i h_m(i)\le\sum_{i=1}^\infty p_i h(i),$$

thus completing the proof.

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

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

Combining the last two relations, we obtain Eq. (17). $\square$

Let us also state our main result under an alternative assumption, which makes the connection with our earlier deterministic results of the paper [3], where an assumption such as Eq. (16) is not needed.

Proposition 6. *Assume that the disturbance set W is finite. Then the conclusions of Prop. 5 hold.*

Proof. The monotone convergence argument for the proof of Eq. (20) goes through using the finiteness of W in place of the assumption (16). $\square$

We note that some additional assumption, like Eq. (16) or the finiteness of W, is necessary to prove our results for SSP problems. In this respect, we note that the original version of the proposition, which appeared in the IEEE Trans. on Aut. Control, was flawed in that it was valid only for the case where W is finite. This was pointed out to us by Yi Zhang (private communication), who constructed the following example.

Example 1. *Let $X = \{t, 0, 1, 2, \ldots\}$, where t is the termination state, and let $g(x,u,w) \equiv 0$, so that $J^*(x) \equiv 0$. There is only one control at each state, and hence only one policy. The transitions are as follows:*

From each state $x = 2, 3, \ldots$ we move deterministically to state $x-1$, from state 1 we move deterministically to state t, and from state 0 we move to state $x = 1, 2, \ldots$, with probability p_x such that $\sum_{x=1}^{\infty} x p_x = \infty$ [so at state 0, the assumption (16) and the finiteness of W are violated].

Here the unique policy is proper at all $x = 1, 2, \ldots$, and we have $\hat{J}(x) = J^*(x) = 0$. However, the policy is not proper at $x = 0$, since the expected number of transitions from $x = 0$ to termination is $\sum_{x=1}^{\infty} x p_x = \infty$. As a result the set $\widehat{\Pi}_0$ is empty and we have $\hat{J}(0) = \infty$. Thus $\hat{J}$ does not satisfy the Bellman equation for $x = 0$, since

$$\infty = \hat{J}(0) \neq E\Big\{g(0,u,w) + \hat{J}\big(f(0,u,w)\big)\Big\} = \sum_{x=1}^{\infty} p_x \hat{J}(x) = 0.$$

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, Prop. 5 or Prop. 6 (under the corresponding assumptions) imply that J^* is the unique solution of Bellman's equation within $\widehat{\mathcal{W}}$, and the VI algorithm converges to J^* starting from any $J_0 \in \widehat{\mathcal{W}}$. Under additional conditions, such as finiteness of $U(x)$ for all $x \in X$ [cf. Prop. 1(e)], the VI algorithm converges to J^* starting from any $J_0 \in \mathcal{J}$ with $E_{x_0}^{\pi}\big\{J(x_k)\big\} \to 0$, for all $(\pi, x_0) \in \mathcal{C}$.

IV. The Multiplicity of Solutions of Bellman's Equation

Let us now discuss the issue of multiplicity of solutions of Bellman's equation within the set of functions

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

We know from Prop. 1(a) and Prop. 5(a) (or Prop. 6) that J^* and $\hat{J}$ are solutions, and that all other solutions J must satisfy either $J^* \leq J \leq \hat{J}$ or $J \notin \widehat{\mathcal{W}}$.

In the special case of a deterministic problem (one where the disturbance w_k takes a single value), it was shown in the paper [3] that $\hat{J}$ is the largest solution of Bellman's equation within $\mathcal{J}$, so all solutions $J \in \mathcal{J}$ satisfy $J^* \leq J \leq \hat{J}$. Moreover, it was shown through examples that there can be any number of solutions that lie between J^* and $\hat{J}$: a finite number, an infinite number, or none at all.

In stochastic problems, however, the situation is strikingly different. There can be an infinite number of solutions $J \in \mathcal{J}$ such that $J \neq \hat{J}$ and $J \geq \hat{J}$, even when the set W is finite, as illustrated by the following example. Of course, by Prop. 5(a) or Prop. 6, under the corresponding assumptions, these solutions must lie outside $\widehat{\mathcal{W}}$.

Example 2. *Let $X = \Re$, $t = 0$, and assume that there is only one control at each state. The disturbance w_k takes two values: 1 and 0 with probabilities $\alpha \in (0,1)$ and $1-\alpha$, respectively. The system equation is*

$$x_{k+1} = \frac{w_k x_k}{\alpha},$$

and there is no cost at each state and stage [$g(x,u,w) \equiv 0$]. Thus from state x_k we move to x_k/α with probability α and to the termination state $t = 0$ with probability $1-\alpha$. Here, the only admissible policy is proper, and we have

$$J^*(x) = \hat{J}(x) = 0, \qquad \forall\, x \in X.$$

Bellman's equation has the form

$$J(x) = (1-\alpha)J(0) + \alpha J\left(\frac{x}{a}\right), \qquad x \in X,$$

and has an infinite number of solutions within $\mathcal{J}$ in addition to J^* and $\hat{J}$: any positively homogeneous function, such as, for example, $J(x) = \gamma|x|$, $\gamma > 0$, is a solution. Consistently with Prop. 5(a), none of these solutions belongs to $\widehat{\mathcal{W}}$, since x_k is either equal to x_0/α^k (with probability α^k) or equal to 0 (with probability $1-\alpha^k$), and, for example, $E\big\{\gamma|x_k|\big\} = \gamma|x_0|$ for all k.

Let us also note that in the case of linear-quadratic problems, the number of solutions of the Riccati equation has been the subject of considerable investigation, starting with the papers by Willems [18] and Kucera [19], [20], which were followed up by several other papers. These works adopt various assumptions relating to controllability and observability. Because of these assumptions and also because solutions of the Riccati equation give rise to solutions of the Bellman equation, but not reversely, it appears that the full characterization of the set of solutions of the Bellman equation remains an interesting open research question at present.

V. The Case of Bounded Cost per Stage

Let us consider the special case where the cost per stage g is bounded over $X \times U \times W$, i.e.,

$$\sup_{(x,u,w) \in X \times U \times W} g(x,u,w) < \infty. \tag{21}$$

We will show that $\hat{J}$ is the largest solution of Bellman's equation within the class of functions that are bounded over the effective domain $\widehat{X}$ of $\hat{J}$ [cf. Eq. (15)].

We say that a policy π is *uniformly proper* if there is a uniform bound on the expected number of steps to reach the destination from states $x \in \widehat{X}$ using π:

$$\sup_{x \in \widehat{X}} \sum_{k=0}^{\infty} r_k(\pi, x) < \infty.$$

Since we have for all $\pi \in \widehat{\Pi}_{x_0}$,

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

it follows that the cost function J_π of a uniformly proper π belongs to the set $\mathcal{B}$, defined by

$$\mathcal{B} = \left\{ J \in \mathcal{J} \;\middle|\; \sup_{x \in \widehat{X}} J(x) < \infty \right\}. \tag{22}$$

When $\widehat{X} = X$, the notion of a uniformly proper policy coincides with the notion of a transient policy used in [2] and [15], which itself descends from earlier works. However, our definition is somewhat more general, since it also applies to the case where $\widehat{X}$ is a strict subset of X.

Let us denote by $\widehat{\mathcal{W}}_b$ the set of functions

$$\widehat{\mathcal{W}}_b = \{ J \in \mathcal{B} \mid \hat{J} \leq J \}.$$

The following proposition provides conditions for $\hat{J}$ to be the largest fixed solution of the Bellman equation within $\mathcal{B}$. Its assumptions include the existence of a uniformly proper policy, which implies that $\hat{J}$ belongs to $\mathcal{B}$.

Proposition 7. *Assume that the cost per stage g is bounded over $X \times U \times W$ [cf. Eq. (21)], and that there exists a uniformly proper policy. Assume further that Eq. (16) holds or that the set W is finite. Then:*

(a) *$\hat{J}$ is the unique solution of the Bellman Eq. (6) within the set $\widehat{\mathcal{W}}_b$. Moreover, if $\hat{J} = J^*$, then J^* is the unique solution of Bellman's equation within $\mathcal{B}$.*

(b) *If $\{J_k\}$ is the sequence generated by the VI algorithm (5) starting with some $J_0 \in \mathcal{B}$ with $J_0 \geq \hat{J}$, then $J_k \to \hat{J}$.*

Proof. Since, as noted earlier, the cost function of a uniformly proper policy belongs to $\mathcal{B}$, it follows that $\hat{J}$ also belongs to $\mathcal{B}$. On the other hand, for all $J \in \mathcal{B}$, we have

$$E^{\pi}_{x_0}\big\{ J(x_k) \big\} \leq \left(\sup_{x \in \widehat{X}} J(x) \right) \cdot r_k(\pi, x_0) \to 0, \qquad \forall\, \pi \in \widehat{\Pi}_{x_0}.$$

It follows that the set $\widehat{\mathcal{W}}_b$ is contained in $\widehat{\mathcal{W}}$, while the function $\hat{J}$ belongs to $\widehat{\mathcal{W}}_b$. Since by Prop. 5(a) (or Prop. 6, depending on the assumptions), $\hat{J}$ is the unique solution of Bellman's equation within $\widehat{\mathcal{W}}$, it follows that $\hat{J}$ is the unique solution of Bellman's equation within $\widehat{\mathcal{W}}_b$.

The proof of part (b) and that $\hat{J}$ is the unique solution of the Bellman Eq. (6) within the set $\widehat{\mathcal{W}}_b$ follow as in the proof of Prop. 5.

Assume now that $\hat{J} = J^*$. Then from the preceding proof, J^* is the unique solution of Bellman's equation within the set $\widehat{\mathcal{W}}_b = \{ J \in \mathcal{B} \mid J^* \leq J \}$. If there were another solution J' within $\mathcal{B}$, then by Prop. 1(a), we would have

$J^* \leq J'$ so that $J' \in \widehat{\mathcal{W}}_b$. This shows that $J' = J^*$, so J^* is the unique solution of Bellman's equation within $\mathcal{B}$. $\square$

The uniqueness of solution of Bellman's equation within $\mathcal{B}$ when $\hat{J} = J^*$ [cf. part (a) of the preceding proposition] is consistent with Example 2. In that example, J^* and $\hat{J}$ are equal and bounded, and all the additional solutions of Bellman's equation are unbounded.

Note that without the assumption of existence of a uniformly proper π, $\hat{J}$ and J^* need not belong to $\mathcal{B}$. As an example, let X be the set of nonnegative integers, let $t = 0$, and let there be a single policy that moves the system deterministically from a state $x \geq 1$ to the state $x - 1$ at cost $g(x, x-1) = 1$. Then

$$\hat{J}(x) = J^*(x) = x, \qquad \forall\, x \in X,$$

so $\hat{J}$ and J^* do not belong to $\mathcal{B}$, even though g is bounded. Here the unique policy is proper at all x, but is not uniformly proper.

VI. Concluding Remarks

We have considered nonnegative cost SSP problems, which involve arbitrary state and control spaces, and a Bellman equation with possibly multiple solutions. Within this context, we have generalized the notion of a proper policy and we have discussed the restricted optimization over just the proper policies. The restricted optimal cost function $\hat{J}$ is a solution of Bellman's equation, and if the cost per stage is bounded, $\hat{J}$ is the maximal solution within the set of nonnegative functions that are bounded within their effective domain. By contrast, J^* is the minimal solution. When compared with their deterministic counterparts of the paper [3], the results of the present paper highlight an interesting difference: in deterministic problems $\hat{J}$ is the maximal solution of Bellman's equation within all functions in $\mathcal{J}$ (unbounded as well as extended real-valued), whereas this need not be true for stochastic problems.

VII. References

[1] 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.

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

[3] Bertsekas, D. P., 2018. "Stable Optimal Control and Semicontractive Dynamic Programming," SIAM J. on Control and Optimization, Vol. 56, pp. 231-252.

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

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

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

[7] Bertsekas, D. P., and Tsitsiklis, J. N., 1989. Parallel and Distributed Computation: Numerical Methods, Prentice-Hall, Englewood Cliffs, N. Y. (republished by Athena Scientific, Belmont, MA, 1996); may be downloaded from http://web.mit.edu/dimitrib/www/home.html.

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

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

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

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

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

[13] Bertsekas, D. P., 2018. Abstract Dynamic Programming, 2nd Edition, Athena Scientific, Belmont, MA.

[14] Bertsekas, D. P., 2015. "Regular Policies in Abstract Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-P-3173, MIT, May 2015; arXiv preprint arXiv:1609.03115; SIAM J. on Control and Optimization, Vol. 27, 2017, pp. 1694-1727.

[15] Pliska, S. R., 1978. "On the Transient Case for Markov Decision Chains with General State Spaces," in Dynamic Programming and its Applications, by M. L. Puterman (ed.), Academic Press, N. Y.

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

[17] Strauch, R., 1966. "Negative Dynamic Programming," Ann. Math. Statist., Vol. 37, pp. 871-890.

[18] Willems, J., 1971. "Least Squares Stationary Optimal Control and the Algebraic Riccati Equation," IEEE Trans. on Automatic Control, Vol. 16, pp. 621-634.

[19] Kucera, V., 1972. "The Discrete Riccati Equation of Optimal Control," Kybernetika, Vol. 8, pp. 430-447.

[20] Kucera, V., 1973. "A Review of the Matrix Riccati Equation," Kybernetika, Vol. 9, pp. 42-61.