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.

1. INTRODUCTION

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

\[
x_{k+1} = f(x_k, u_k, w_k), \quad k = 0, 1, \ldots
\]

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 \( f : X \times U \times W \to X \) 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.‡ 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 cost per stage, \( g(x, u, w) \), is assumed real-valued and nonnegative:

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


‡ The nature of these difficulties is well-documented; see the monograph by Bertsekas and Shreve [BeS78], and the paper by James and Collins [JaC06], 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 assume that $X$ contains a special cost-free and absorbing state $t$, referred to as the \textit{destination}:

$$f(t, u, w) = t, \quad g(t, u, w) = 0, \quad \forall \ u \in U(t), \ w \in W.$$  \hspace{1cm} (1.3)

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

We are interested in policies of the form \(\pi = \{\mu_0, \mu_1, \ldots\}\), where each \(\mu_k\) is a function mapping \(x \in X\) into the control \(\mu_k(x) \in U(x)\). The set of all policies is denoted by \(\Pi\). Policies of the form \(\pi = \{\mu, \mu, \ldots\}\) are called \textit{stationary}, and will be denoted by \(\mu\), 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 \((x_k, \mu_k(x_k))\), \(k = 0, 1, \ldots\), with cost

$$J_\pi(x_0) = \sum_{k=0}^{\infty} E^\pi_{x_0} \left\{ g(x_k, \mu_k(x_k), w_k) \right\}, \quad x_0 \in X,$$

where \(E^\pi_{x_0} \{\cdot\}\) denotes expectation with respect to the probability measure corresponding to initial state \(x_0\) and policy \(\pi\), and the series converges in view of the nonnegativity of cost per stage \(g\). We view \(J_\pi\) as a function over \(X\), and we refer to it as the cost function of \(\pi\). For a stationary policy \(\mu\), the corresponding cost function is denoted by \(J_\mu\). The optimal cost function is defined as

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

and a policy \(\pi^*\) 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 \textit{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

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

Since \(t\) is cost-free and absorbing, this set contains the cost functions \(J_\pi\) 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

$$J(x) = \inf_{u \in U(x)} E \left\{ g(x, u, w) + J(f(x, u, w)) \right\}, \quad x \in X,$$

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 (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

$$J_{k+1}(x) = \inf_{u \in U(x)} E \left\{ g(x, u, w) + J_k(f(x, u, w)) \right\}, \quad x \in X, \ k = 0, 1, \ldots.$$  \hspace{1cm} (1.5)
However, $\{J_k\}$ may not always converge to $J^*$ because, among other reasons, Bellman’s equation may have multiple solutions within $J$.

In a recent paper [Ber17] 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 [Pal67], [Der70], [Whi82], [BeT89], [Put94], [Ali99], [HeL99], and [Ber12], and the references quoted there). For the case where $g \geq 0$, the paper by Bertsekas and Tsitsiklis [BeT91] 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 $\hat{W} \subset J$ that majorize $\hat{J}$, and that the VI algorithm converges to $\hat{J}$ starting from a function in $\hat{W}$. In particular, all solutions of Bellman’s equation lie in the region bordered by $J^*$ from below and $\hat{J}$ from above. Our line of analysis draws its origin from concepts of regularity introduced by the author in the monograph [Ber13] and the subsequent paper [Ber15].

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 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 [JaC06] uses the properness definition of [Pli78], and extends the analysis of [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; this is the most important difference from the analysis of this paper.

Our analysis is also 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 and that $J^*$ is real-valued, but allows $g$ to take both positive and negative values. Moreover [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 $g$ and cannot be
extended without major modifications.

2. PROPER POLICIES AND THE $\delta$-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 proposition gives 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

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

then $J^* \leq J$.

(b) For all stationary policies $\mu$, $J_\mu$ is a solution of the equation

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

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

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

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

(d) A stationary policy $\mu^*$ is optimal if and only if

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

(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.**
For a given state \( x \in X \), a policy \( \pi \) is said to be proper at \( x \) if
\[
J_\pi(x) < \infty, \quad \sum_{k=0}^{\infty} r_k(\pi, x) < \infty, \tag{2.2}
\]
where \( r_k(\pi, x_0) \) is the probability that \( x_k \neq t \) when using \( \pi \) 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 \( \pi \).

We denote by \( \tilde{\Pi}_x \) the set of all policies that are proper at \( x \), and we use the notation
\[
C = \{(\pi, x) \mid \pi \in \tilde{\Pi}_x\}. \tag{2.3}
\]
We denote by \( \tilde{J} \) the corresponding restricted optimal cost function,
\[
\tilde{J}(x) = \inf_{(\pi, x) \in C} J_\pi(x) = \inf_{\pi \in \tilde{\Pi}_x} J_\pi(x), \quad x \in X.
\]
Finally we denote by \( \tilde{X} \) the effective domain of \( \tilde{J} \), i.e.,
\[
\tilde{X} = \{x \in X \mid \tilde{J}(x) < \infty\}. \tag{2.4}
\]
Note that \( \tilde{X} \) is the set of all \( x \) such that there \( \tilde{\Pi}_x \) is nonempty, and that \( t \in \tilde{X} \).

The definition of proper policy just given differs from the definition of a transient policy adopted by James and Collins [JaC06]. In particular, the definition of [JaC06] requires that the expected number of steps to reach the destination is uniformly bounded over the initial state \( x \) (see [JaC06], p. 608) and is not tied to a single state \( x \).

For any \( \delta > 0 \), let us consider the \( \delta \)-perturbed optimal control problem. This is the same problem as the original, except that the cost per stage is changed to
\[
g(x, u, w) + \delta, \quad \forall \ x \neq t,
\]
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 \( \delta \)-perturbed problem. The \( \delta \)-perturbed cost function of a policy \( \pi \) is denoted by \( J_{\pi,\delta} \) and is given by
\[
J_{\pi,\delta}(x) = J_\pi(x) + \delta \sum_{k=0}^{\infty} r_k(\pi, x). \tag{2.5}
\]
We denote by \( J^*_\delta \) the optimal cost function of the \( \delta \)-perturbed problem, i.e., \( J^*_\delta(x) = \inf_{\pi \in \Pi} J_{\pi,\delta}(x) \). The following proposition relates the \( \delta \)-perturbed problem with proper policies.

**Proposition 2.2:**

(a) A policy \( \pi \) is proper at a state \( x \in X \) if and only if \( J_{\pi,\delta}(x) < \infty \) for all \( \delta > 0 \).

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

(c) For every \( \epsilon > 0 \), a policy \( \pi_\epsilon \) that is \( \epsilon \)-optimal for the \( \delta \)-perturbed problem is proper at all \( x \in \tilde{X} \).
Proof: (a) Follows from Eq. (2.5) and the definition (2.2) of a proper policy.

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

(c) We have $J_{\pi,\delta}(x) \leq J_\delta^*(x) + \epsilon$ for all $x \in X$. Hence $J_{\pi,\delta}(x) < \infty$ for all $x \in \hat{X}$, implying by part (a) that $\pi_\epsilon$ is proper at all $x \in \hat{X}$. Q.E.D.

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

**Proposition 2.3:** We have $\lim_{\delta \downarrow 0} J_\delta^*(x) = \hat{J}(x)$ for all $x \in X$. Moreover, for any $\epsilon > 0$, a policy $\pi_\epsilon$ that is $\epsilon$-optimal for the $\delta$-perturbed problem is $\epsilon$-optimal within the class of proper policies, i.e.

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

**Proof:** Let $\pi_\epsilon$ be a policy that is $\epsilon$-optimal for the $\delta$-perturbed problem, and is also proper at all $x \in \hat{X}$ [cf. Prop. 2.2(c)]. By using Eq. (2.5), we have for all $\delta > 0$, $\epsilon > 0$, and $\pi \in \hat{\Pi}_x$,

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

where

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

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

$$\hat{J}(x) \leq J_\delta^*(x) \leq J_{\pi}(x) + w_{\pi,\delta}(x), \quad \forall x \in \hat{\Pi}_x.$$ 

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

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

from which $\hat{J}(x) = \lim_{\delta \downarrow 0} J_\delta^*(x)$ for all $x \in \hat{X}$. By Prop. 2.2(b), we also have $J_\delta^*(x) = \hat{J}(x) = \infty$ for all $x \notin \hat{X}$, so that $\hat{J}(x) = \lim_{\delta \downarrow 0} J_\delta^*(x)$ for all $x \in X$. We also have

$$J_{\pi_\epsilon}(x) \leq J_{\pi,\delta}(x) \leq J_\delta^*(x) + \epsilon \leq J_{\pi}(x) + \delta \sum_{k=0}^{\infty} r_k(\pi, x) + \epsilon, \quad \forall x \in \hat{X}, \pi \in \hat{\Pi}_x.$$
By taking the limit as $\delta \downarrow 0$, we obtain
\[ J_{\pi_\epsilon}(x) \leq J(x) + \epsilon, \quad \forall x \in \hat{X}, \pi \in \hat{\Pi}_x. \]
By taking the infimum over $\pi \in \hat{\Pi}_x$, it follows that $J_{\pi_\epsilon}(x) \leq \hat{J}(x) + \epsilon$ for all $x \in \hat{X}$, which combined with the fact $J_{\pi_\epsilon}(x) = \hat{J}(x) = \infty$ for all $x \notin \hat{X}$, yields the result. Q.E.D.

3. MAIN RESULT

By Prop. 2.1(a), $J^*_\delta$ solves Bellman’s equation for the $\delta$-perturbed problem, while by Prop. 2.3, $\lim_{\delta \downarrow 0} J^*_\delta(x) = \hat{J}(x)$. This suggests that $\hat{J}$ solves the unperturbed Bellman equation, which is the “limit” as $\delta \downarrow 0$ of the $\delta$-perturbed version. Indeed we will show a stronger result, namely that $\hat{J}$ is the unique solution of Bellman’s equation within the set of functions
\[ \hat{W} = \left\{ J \in \mathcal{J} \mid \hat{J} \leq J, E_{x_0}^\pi \{ J(x_k) \} \to 0, \forall (\pi, x_0) \in \mathcal{C} \right\}, \tag{3.1} \]
where
\[ \mathcal{C} = \{ (\pi, x) \mid \pi \in \hat{\Pi}_x \} \]
[cf. Eq. (2.3)], $E_{x_0}^\pi \{ \cdot \}$ denotes expected value with respect to the probability measure corresponding to initial state $x_0$ under policy $\pi$, and $E_{x_0}^\pi \{ J(x_k) \}$ denotes the expected value of the function $J$ along the sequence $\{x_k\}$ generated starting from $x_0$ and using $\pi$. The functions in $\hat{W}$ 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 Lyapunov functions.

Given a policy $\pi = \{ \mu_0, \mu_1, \ldots \}$, we denote by $\pi_k$ the policy
\[ \pi_k = \{ \mu_k, \mu_{k+1}, \ldots \}. \tag{3.2} \]
We first show a preliminary result.

**Proposition 3.1:**

(a) For all pairs $(\pi, x_0) \in \mathcal{C}$ and $k = 0, 1, \ldots$, we have
\[ 0 \leq E_{x_0}^\pi \{ \hat{J}(x_k) \} \leq E_{x_0}^\pi \{ J_{\pi_k}(x_k) \}, \]
where $\pi_k$ is the policy given by Eq. (3.2).

(b) The set $\hat{W}$ of Eq. (3.1) contains $\hat{J}$, as well as all functions $J$ satisfying $0 \leq J \leq h(\hat{J})$ for some function $h : X \mapsto X$ with $h(J) \to 0$ as $J \to 0$. 

7
Proof: (a) For any pair \((\pi, x_0)\) \(\in \mathcal{C}\) and \(\delta > 0\), we have

\[
J_{\pi, \delta}(x_0) = E_{x_0}^\pi \left\{ J_{\pi_k, \delta}(x_k) + k\delta + \sum_{m=0}^{k-1} g(x_m, \mu_m(x_m), w_m) \right\}.
\]

Since \(J_{\pi, \delta}(x_0) < \infty\) [cf. Prop. 2.2(a)], it follows that \(E_{x_0}^\pi \{ J_{\pi_k, \delta}(x_k) \} < \infty\). Hence for all \(x_k\) that can be reached with positive probability using \(\pi\) and starting from \(x_0\), we have \(J_{\pi_k, \delta}(x_k) < \infty\), implying [by Prop. 2.2(a)] that \((\pi_k, x_k) \in \mathcal{C}\) and hence \(\hat{J}(x_k) \leq J_{\pi_k}(x_k)\). By applying \(E_{x_0}^\pi \{ \cdot \}\) to this last inequality, the result follows.

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

\[
J_\pi(x_0) = E_{x_0}^\pi \left\{ g(x_0, \mu_0(x_0), w_0) \right\} + E_{x_0}^\pi \left\{ J_{\pi_1}(x_1) \right\},
\]

and more generally,

\[
E_{x_0}^\pi \{ J_{\pi_m}(x_m) \} = E_{x_0}^\pi \left\{ g(x_m, \mu_m(x_m), w_m) \right\} + E_{x_0}^\pi \left\{ J_{\pi_{m+1}}(x_{m+1}) \right\}, \quad m = 0, 1, \ldots, \tag{3.3}
\]

where \(\{x_m\}\) is the sequence generated starting from \(x_0\) and using \(\pi\). Using the fact \(J_\pi(x_0) < \infty\), it follows that all the terms in the above relations are finite, and in particular

\[
E_{x_0}^\pi \{ J_{\pi_m}(x_m) \} < \infty, \quad \forall (\pi, x_0) \in \mathcal{C}, \quad m = 0, 1, \ldots.
\]

By adding Eq. (3.3) for \(m = 0, \ldots, k-1\), and canceling the finite terms \(E_{x_0}^\pi \{ J_{\pi_m}(x_m) \}\) for \(m = 1, \ldots, k-1\), we obtain for all \(k = 1, 2, \ldots,\)

\[
J_\pi(x_0) = E_{x_0}^\pi \left\{ J_{\pi_k}(x_k) \right\} + \sum_{m=0}^{k-1} E_{x_0}^\pi \left\{ g(x_m, \mu_m(x_m), w_m) \right\}, \quad \forall (\pi, x_0) \in \mathcal{C}.
\]

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

\[
E_{x_0}^\pi \{ J_{\pi_k}(x_k) \} \to 0, \quad \forall (\pi, x_0) \in \mathcal{C}.
\]

Since by part (a), \(0 \leq E_{x_0}^\pi \{ \hat{J}(x_k) \} \leq E_{x_0}^\pi \{ J_{\pi_k}(x_k) \}\), it follows that

\[
E_{x_0}^\pi \{ \hat{J}(x_k) \} \to 0, \quad \forall (\pi, x_0) \in \mathcal{C},
\]

so that \(\hat{J} \in \hat{\mathcal{W}}\). This also implies that

\[
E_{x_0}^\pi \{ J(x_k) \} \to 0, \quad \forall (\pi, x_0) \in \mathcal{C},
\]

if \(0 \leq J \leq h(\hat{J})\). Q.E.D.
We can now prove our main result.

**Proposition 3.2:**

(a) \( \hat{J} \) is the unique solution of the Bellman Eq. (2.1) within the set \( \hat{W} \) of Eq. (3.1).

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

(c) **(Optimality Condition)** If \( \mu \) is a stationary policy that is proper at all \( x \in \hat{X} \) and

\[
\hat{\mu}(x) \in \arg \min_{u \in U(x)} E \left\{ g(x, u, w) + \hat{J}(f(x, u, w)) \right\}, \quad \forall \ x \in X,
\]

(3.4)

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

**Proof:** (a), (b) By Prop. 3.1(b), \( \hat{J} \in \hat{W} \). We will first show that \( \hat{J} \) is a solution of Bellman’s equation and then show that it is the unique solution within \( \hat{W} \) by showing the convergence of VI [cf. part (b)]. Since \( J^*_\delta \) solves the Bellman equation for the \( \delta \)-perturbed problem, and \( J^*_\delta \geq \hat{J} \) (cf. Prop. 2.3), we have for all \( \delta > 0 \) and \( x \neq t \),

\[
J^*_\delta(x) = \inf_{u \in U(x)} E \left\{ g(x, u, w) + J^*_\delta(f(x, u, w)) \right\}
\]

\[
\geq \inf_{u \in U(x)} E \left\{ g(x, u, w) + J^*_\delta(f(x, u, w)) \right\}
\]

By taking the limit as \( \delta \downarrow 0 \) and using Prop. 2.3, we obtain

\[
\hat{J}(x) \geq \inf_{u \in U(x)} E \left\{ g(x, u, w) + \hat{J}(f(x, u, w)) \right\}, \quad \forall \ x \in X.
\]

(3.5)

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

\[
E \left\{ g(x, u, w) + \delta_m + J^*_\delta_m(f(x, u, w)) \right\} \geq \inf_{v \in U(x)} E \left\{ g(x, v, w) + \delta_m + J^*_\delta_m(f(x, v, w)) \right\} = J^*_\delta_m(x).
\]

Taking the limit as \( m \to \infty \), and using the monotone convergence theorem (to interchange limit and expectation) and the fact \( \lim_{\delta_m \downarrow 0} J^*_\delta_m = \hat{J} \) (cf. Prop. 2.3), we have

\[
E \left\{ g(x, u, w) + \hat{J}(f(x, u, w)) \right\} \geq \hat{J}(x), \quad \forall \ x \in X, \ u \in U(x),
\]
so that
\[ \inf_{u \in U(x)} E\left\{ g(x, u, w) + \hat{J}\left(f(x, u, w)\right) \right\} \geq \hat{J}(x), \quad \forall x \in X. \]  

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 \( J_0 \in \hat{W} \) [cf. part (b)]. Indeed, for \( x_0 \in \hat{X} \) and any \( \pi = \{\mu_0, \mu_1, \ldots\} \in \hat{\Pi}_{x_0} \), let \( \{x_k\} \) be the generated sequence starting from \( x_0 \). Since from the definition of the VI sequence \( \{J_k\} \) [cf. Eq. (1.5)], we have
\[ J_k(x) \leq E\left\{ g(x, u, w) + J_{k-1}\left(f(x, u, w)\right) \right\}, \quad \forall x \in X, \ u \in U(x), \ k = 1, 2, \ldots, \]
it follows that
\[ J_k(x_0) \leq E_{x_0}^\pi \left\{ J_0(x_0) + \sum_{m=0}^{k-1} g(x_m, \mu_m(x_m), w_m) \right\}. \]

Since \( J_0 \in \hat{W}, \) we have \( E_{x_0}^\pi \{J_0(x_k)\} \to 0 \), so by taking the limit as \( k \to \infty \) in the preceding relation, it follows that \( \limsup_{k \to \infty} J_k(x_0) \leq J_\pi(x_0) \). By taking the infimum over all \( \pi \in \hat{\Pi}_{x_0} \), 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 \( x_0 \in \hat{X} \). We also have \( \hat{J} \leq J_k \) for all \( k \), so that \( \hat{J}(x_0) = J_k(x_0) = \infty \) for all \( x_0 \notin \hat{X} \). This completes the proof of part (b). Finally, since \( \hat{J} \in \hat{W} \) and \( \hat{J} \) is a solution of Bellman’s equation, part (b) implies the uniqueness assertion of part (a).

(c) If \( \mu \) is proper at all \( x \in \hat{X} \) and Eq. (3.4) holds, then
\[ \hat{J}(x) = E\left\{ g(x, \mu(x), w) + \hat{J}\left(f(x, \mu(x), w)\right) \right\}, \quad x \in X. \]

By Prop. 2.1(b), this implies that \( J_\mu \leq \hat{J}, \) so \( \mu \) is optimal over the set of proper policies. Conversely, assume that \( \mu \) is proper at all \( x \in \hat{X} \) and \( J_\mu = \hat{J} \). Then by Prop. 2.1(b), we have
\[ \hat{J}(x) = E\left\{ g(x, \mu(x), w) + \hat{J}\left(f(x, \mu(x), w)\right) \right\}, \quad x \in X, \]
and since [by part (b)] \( \hat{J} \) is a solution of Bellman’s equation,
\[ \hat{J}(x) = \inf_{u \in U(x)} E\left\{ g(x, u, w) + \hat{J}\left(f(x, u, w)\right) \right\}, \quad x \in X. \]
Combining the last two relations, we obtain Eq. (3.4). Q.E.D.

Figure 3.1 Illustration of the solutions of Bellman’s equation. All solutions either lie between \( J^* \) and \( \hat{J} \), or they lie outside the set \( \hat{W} \). The VI algorithm converges to \( \hat{J} \) starting from any \( J_0 \in \hat{W} \).
We illustrate Prop. 3.2 in Fig. 3.1. 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. 3.2, it follows that \( J^* \) is the unique solution of Bellman’s equation within \( \hat{W} \), and the VI algorithm converges to \( J^* \) starting from any \( J_0 \in \hat{W} \). 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 \( J_0 \in J \) with \( E_{x_0}^\pi \{ J_0(x_k) \} \to 0 \), for all \( (\pi, x_0) \in C \).

4. 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

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

We know from Props. 2.1(a) and 3.2(a) that \( J^* \) and \( \hat{J} \) are solutions, and that all other solutions \( J \) must satisfy either \( J^* \leq J \leq \hat{J} \) or \( J \notin \hat{W} \).

In the special case of a deterministic problem (one where the disturbance \( w_k \) takes a single value), it was shown in the paper [Ber17] 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 \( J \geq \hat{J} \), as shown by the following example. Of course, by Prop. 3.2(a), these solutions must lie outside \( \hat{W} \).

Example 4.1

Let \( X = \mathbb{R} \), \( t = 0 \), and assume that there is only one control at each state, and a single policy \( \pi \). 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 \( x_k/\alpha \) with probability \( \alpha \) and to the termination state \( t = 0 \) with probability \( 1 - \alpha \). Here, the only admissible policy is proper, and we have

\[ J^* (x) = \hat{J}(x) = 0, \quad \forall x \in X. \]

Bellman’s equation has the form

\[ J(x) = (1 - \alpha) J(0) + \alpha J \left( \frac{x}{\alpha} \right), \quad x \in X, \]
and has an infinite number of solutions within $J$ 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. 3.2(a), none of these solutions belongs to $\hat{W}$, since $x_k$ is either equal to $x_0/\alpha_k$ (with probability $\alpha_k$) or equal to 0 (with probability $1 - \alpha_k$). For example, in the case of $J(x) = \gamma |x|$, we have

$$E_{x_0}^\pi \{ J(x_k) \} = \alpha^k \gamma \frac{x_0}{\alpha^k} = \gamma |x_0|, \quad \forall \; k \geq 0,$$

so $J(x_k)$ does not converge to 0. Moreover, none of these solutions seems to be significant in some discernible way.

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 [Wil71] and Kucera [Kuc72], [Kuc73], 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.

5. THE CASE OF BOUNDED COST PER STAGE

We will now consider the case where the cost per stage $g$ is bounded over $X \times U \times W$, i.e.,

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

We will show that the function $\hat{J}$ is the maximal solution of Bellman’s equation if we restrict ourselves to functions that are bounded within the effective domain $\hat{X}$ of $\hat{J}$ [cf. Eq. (2.4)].

We denote by $B$ the set of functions

$$B = \left\{ J \in \mathcal{J} \mid \sup_{x \in \hat{X}} J(x) < \infty \right\}.$$

From the definition (2.2) of properness, we have

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

while for all $J \in B$, we have

$$E_{x_0}^\pi \{ J(x_k) \} \leq \left( \sup_{x \in \hat{X}} J(x) \right) \cdot r_k(\pi,x_0) \to 0, \quad \forall \; \pi \in \hat{\Pi}_{x_0}.$$
It follows that the set 
\[ \hat{W}_b = \{ J \in J \mid \hat{J} \leq J, J \in B \} \]
is contained in \( \hat{W} \), while the function \( \hat{J} \) belongs to \( \hat{W}_b \).

By applying Prop. 3.2, we see that \( \hat{J} \) is the maximal solution of Bellman’s equation within the set \( \hat{W}_b \).
It follows that \( \hat{J} \) is the maximal solution of Bellman’s equation within \( B \). Moreover, if \( J^* = \hat{J} \), then \( J^* \) is the unique solution of Bellman’s equation within \( B \). This result is consistent with Example 4.1, where \( J^* \) and \( \hat{J} \) are equal and bounded, and all the additional solutions of Bellman’s equation are unbounded.

6. 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 within this set. When compared with their deterministic counterparts of the paper [Ber17], 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, unbounded as well as extended real-valued, whereas this need not be true for stochastic problems.

7. REFERENCES


