May 2017 (revised Nov. 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, 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. In the favorable case where $J^* = \hat{J}$, strong analytical and algorithmic results are obtained.

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 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 nonnnegative:

$$0 \leq g(x, u, w) < \infty, \quad \forall\ x \in X,\ u \in U(x),\ w \in W. \tag{1.2}$$

† 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.

‡ 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 *destination*:

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

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 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

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

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

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

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 (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

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

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

In two recent papers, [Ber17a] and [Ber17b], we have focused on undiscounted discrete-time deterministic optimal control with nonnegative cost per stage, an infinite number of states, and a termination state. We have addressed there the connections between controllability, stability, and the solutions of Bellman's equation. In this paper we address similar issues in the context of SSP problems, and 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. These problems have been extensively researched (see e.g., the books [Pal67], [Der70], [Whi82], [BeT89], [Put94], [Alt99], [HeL99], and [Ber12], and the references quoted there). For the case where $g \geq 0$, the paper by Bertsekas and Tsitsiklis [13] 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 (which implies that J^* is real-valued and is equal to $\hat{J}$). 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 any function in $\widehat{\mathcal{W}}$. Moreover, we may have $J^* \neq \hat{J}$, while under some boundedness restrictions, we show that all solutions of Bellman's equation lie in the region bordered by J^* from below and $\hat{J}$ from above.

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 infinite-state SSP problems where g takes both positive and negative values may be possible, but the line of analysis of the present paper relies strongly on the nonnegativity of g and cannot be extended without major modifications.

To compare our analysis with the existing literature for infinite-state SSP problems, we note that proper policies have been considered earlier 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 related (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. Our line of analysis is also different, and draws its origin from concepts of regularity introduced by the author in the monograph [Ber13] and the subsequent paper [Ber15].

2. PROPER POLICIES AND THE δ-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

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

then $J^* \leq J$.

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

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

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

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

(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.**

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

$$J_\pi(x) < \infty, \qquad \sum_{k=0}^{\infty} r_k(\pi, x) < \infty, \tag{2.2}$$

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

$$\mathcal{C} = \big\{(\pi, x) \mid \pi \in \widehat{\Pi}_x\big\}. \tag{2.3}$$

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

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

with the convention that the infimum over the empty set is ∞. Finally we denote by $\widehat{X}$ the effective domain of $\hat{J}$, i.e.,

$$\widehat{X} = \big\{x \in X \mid \hat{J}(x) < \infty\big\}. \tag{2.4}$$

Note that $\widehat{X}$ is the set of all x such that $\widehat{\Pi}_x$ is nonempty. Because every policy is proper at the termination state t, which is cost-free and absorbing, we have $\hat{J}(t) = 0$ and $t \in \widehat{X}$.

The preceding definition of proper policy at a state 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 *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

$$J_{\pi,\delta}(x) = J_\pi(x) + \delta \sum_{k=0}^{\infty} r_k(\pi, x). \tag{2.5}$$

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 intuition behind our use of the δ-perturbed problem is that within its context, improper policies are excluded from optimality, since they have infinite cost starting from some states at which there exists a

proper policy that has finite cost. As a consequence of this, we will show that $\hat{J}_\delta$ converges to $\hat{J}$ as $\delta \downarrow 0$ (and not to J^* if $J^* \neq \hat{J}$). In view of the fact that $\hat{J}_\delta$ solves the Bellman equation of the δ-perturbed problem, by using a limiting argument as $\delta \downarrow 0$, it will follow that $\hat{J}$ solves the Bellman equation of the original unperturbed problem. The following proposition relates the δ-perturbed problem and proper policies.

Proposition 2.2:

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

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

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

Proof: (a) Follows from Eq. (2.5) and the defining property (2.2) 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) An ϵ-optimal π_ϵ exists by Prop. 2.1(c). 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 $x \in \widehat{X}$. **Q.E.D.**

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

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

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

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

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

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 $x \in \widehat{X}$. Moreover, by Prop. 2.2(b), $\hat{J}_\delta(x) = \hat{J}(x) = \infty$ for all $x \notin \widehat{X}$, so that $\hat{J}(x) = \lim_{\delta \downarrow 0} \hat{J}_\delta(x)$ for all $x \in X$.

We also have

$$J_{\pi_\epsilon}(x) \leq J_{\pi_\epsilon,\delta}(x) \leq \hat{J}_\delta(x) + \epsilon \leq J_\pi(x) + \delta \sum_{k=0}^{\infty} r(\pi, x) + \epsilon, \qquad \forall\ x \in \widehat{X},\ \pi \in \widehat{\Pi}_x.$$

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

$$J_{\pi_\epsilon}(x) \leq J_\pi(x) + \epsilon, \qquad \forall\ x \in \widehat{X},\ \pi \in \widehat{\Pi}_x.$$

By taking the infimum over $\pi \in \widehat{\Pi}_x$, it follows that $J_{\pi_\epsilon}(x) \leq \hat{J}(x) + \epsilon$ for all $x \in \widehat{X}$, which combined with the fact $J_{\pi_\epsilon}(x) = \hat{J}(x) = \infty$ for all $x \notin \widehat{X}$, yields the result. **Q.E.D.**

3. MAIN RESULT

By Prop. 2.1(a), $\hat{J}_\delta$ solves Bellman's equation for the δ-perturbed problem, while by Prop. 2.3, $\lim_{\delta \downarrow 0} \hat{J}_\delta(x) = \hat{J}(x)$. This suggests that $\hat{J}$ solves the unperturbed Bellman equation, which is the "limit" as $\delta \downarrow 0$ of the δ-perturbed version. Indeed we will show a stronger result, namely that $\hat{J}$ is the unique solution of Bellman's equation within the set of functions

$$\widehat{\mathcal{W}} = \Big\{ J \in \mathcal{J} \mid \hat{J} \leq J,\ E^\pi_{x_0}\big\{J(x_k)\big\} \to 0,\ \forall\ (\pi, x_0) \in \mathcal{C} \Big\}, \tag{3.1}$$

where

$$\mathcal{C} = \big\{(\pi, x) \mid \pi \in \widehat{\Pi}_x\big\}$$

[cf. Eq. (2.3)], $E^\pi_{x_0}\{\cdot\}$ 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. The set $\widehat{\mathcal{W}}$ is also related to abstract DP concepts of regularity, which are central in the monograph [Ber13] and the paper [Ber15].

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

$$\pi_k = \{\mu_k, \mu_{k+1}, \ldots\}. \tag{3.2}$$

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

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

where π_k is the policy given by Eq. (3.2).

(b) The set $\widehat{\mathcal{W}}$ of Eq. (3.1) contains $\hat{J}$, as well as all functions $J \in S$ satisfying $\hat{J} \leq J \leq c\hat{J}$ for some $c \geq 1$.

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

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

Since $J_{\pi,\delta}(x_0) < \infty$ [cf. Prop. 2.2(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. 2.2(a)] that $(\pi_k, x_k) \in \mathcal{C}$ and hence $\hat{J}(x_k) \leq J_{\pi_k}(x_k)$. By applying $E^{\pi}_{x_0}\{\cdot\}$ to this last inequality, the result follows.

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

$$J_{\pi}(x_0) = E^{\pi}_{x_0}\Big\{g\big(x_0, \mu_0(x_0), w_0\big)\Big\} + E^{\pi}_{x_0}\big\{J_{\pi_1}(x_1)\big\}, \tag{3.3}$$

and

$$E^{\pi}_{x_0}\big\{J_{\pi_m}(x_m)\big\} = E^{\pi}_{x_0}\Big\{g\big(x_m, \mu_m(x_m), w_m\big)\Big\} + E^{\pi}_{x_0}\big\{J_{\pi_{m+1}}(x_{m+1})\big\}, \qquad m = 1, 2, \ldots, \tag{3.4}$$

where $\{x_m\}$ is the sequence generated starting from x_0 and using π. By using repeatedly the expression (3.4) for $m = 1, \ldots, k-1$, and combining it with Eq. (3.3), we obtain for all $k = 1, 2, \ldots$,

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

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

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

so that $\hat{J} \in \widehat{\mathcal{W}}$. This also implies that

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

if $\hat{J} \leq J \leq c\hat{J}$ for some $c \geq 1$. **Q.E.D.**

We can now prove our main result.

Proposition 3.2: The following hold for the restricted optimal cost function $\hat{J}$.

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

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

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

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 (3.5).

Proof: (a), (b) By Prop. 3.1(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. 2.3), 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. 2.3, we obtain

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

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

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

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

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

so that

$$\inf_{u\in U(x)} E\Big\{g(x,u,w) + \hat{J}\big(f(x,u,w)\big)\Big\} \geq \hat{J}(x), \qquad \forall\ x \in X. \tag{3.7}$$

By combining Eqs. (3.6) and (3.7), 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 $x_0 \in \widehat{X}$ 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. (1.5)], we have

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

it follows that

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

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 $\limsup_{k\to\infty} J_k(x_0) \leq J_\pi(x_0)$. By taking the infimum over all $\pi \in \widehat{\Pi}_{x_0}$, we obtain $\limsup_{k\to\infty} J_k(x_0) \leq \hat{J}(x_0)$. 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 $x_0 \in \widehat{X}$. We also have $\hat{J} \leq J_k$ for all k, so that $\hat{J}(x_0) = J_k(x_0) = \infty$ for all $x_0 \notin \widehat{X}$. 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. (3.5) 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. 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 at all $x \in \widehat{X}$ and $J_\mu = \hat{J}$. Then by Prop. 2.1(b), we have

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

while [by part (a)] $\hat{J}$ 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. (3.5). **Q.E.D.**

A complementary result is given in the following proposition, which was first proved in the paper by Yu and Bertsekas [YuB15], Theorem 5.1, in a broader context where measurability issues are fully addressed. The proposition was also proved later in the paper [Ber15] within the simpler context of the present paper.

Proposition 3.3: The optimal cost function J^* is the unique solution of Bellman's equation within the set of functions

$$\big\{J \in \mathcal{E}^+(X) \mid 0 \leq J \leq cJ^* \text{ for some } c > 0\big\}.$$

Moreover, if $\{J_k\}$ is the sequence generated by the VI algorithm (1.5) starting with some J_0 in the set

$$\mathcal{W}^* = \big\{J \mid J^* \leq J \leq cJ^* \text{ for some } c > 0\big\},$$

then $J_k \to J^*$.

We illustrate Props. 3.2 and 3.3 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 from every state 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 $\widehat{\mathcal{W}}$, 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. 2.1(e)], VI converges to J^* starting from any $J_0 \in \mathcal{J}$ with $E^\pi_{x_0}\big\{J_0(x_k)\big\} \to 0$, for all $(\pi, x_0) \in \mathcal{C}$; see [Ber15], and also [Ber17b], which focuses on the deterministic case where w_k can take only one value.

Proposition 3.2 does not say anything about the existence of a proper policy that is optimal within the class of proper policies. For a simple example where $J^* = \hat{J}$ but the only optimal policy is improper,

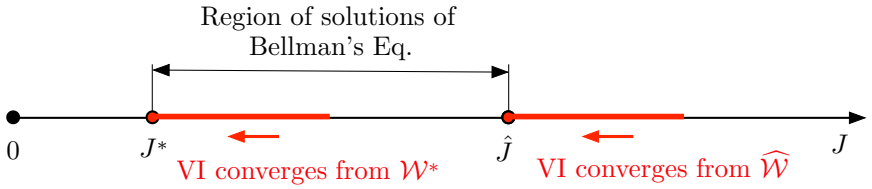

Figure 3.1 Illustration of the solutions of Bellman's equation (cf. Props. 3.2 and 3.3). All solutions either lie between J^* and $\hat{J}$, or they lie outside the set $\widehat{\mathcal{W}}$. The VI algorithm converges to $\hat{J}$ starting from any $J_0 \in \widehat{\mathcal{W}}$, and converges to J^* starting from any $J_0 \in \mathcal{W}^*$.

consider a deterministic shortest path problem with a single node 1 plus the destination t. At node 1 we may choose $u \in [0,1]$ with cost u, and move to t if $u \neq 0$ and stay at 1 if $u = 0$. Note that here we have $J^*(1) = \hat{J}(1) = 0$, and the infimum over $u \in [0,1]$ is attained in Bellman's equation, which has the form $J^*(1) = \min\big\{\inf_{u\in(0,1]} u,\, J^*(1)\big\}$.

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

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

We know from Props. 2.1(a) and 3.2(a) that J^* and $\hat{J}$ are solutions, and that all solutions J of Bellman's equation 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 [Ber17b] 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 $\widehat{\mathcal{W}}$.

Example 4.1

Let $X = \Re$, $t = 0$, and assume that there is only one control at each state, and hence a single policy π. 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 state x_k/α with probability α and to the termination state $t = 0$ with probability $1 - \alpha$. Here, the only admissible policy is stationary and proper at all $x \in X$, and we have

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

Bellman's equation has the form

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

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. 3.2(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 $1 - \alpha^k$). For example, in the case of $J(x) = \gamma|x|$, we have

$$E^{\pi}_{x_0}\big\{J(x_k)\big\} = \alpha^k \gamma \left|\frac{x_0}{\alpha^k}\right| = \gamma|x_0|, \qquad \forall\, k \geq 0,$$

so $J(x_k)$ does not converge to 0 unless $x_0 = 0$. Moreover, none of these additional 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. The author's paper [Ber15] shows that $\hat{J}$, which is the optimal cost function over all linear stable policies, corresponds to the largest solution of the Riccati equation. 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, even in linear-quadratic problems. We will next elaborate on the preceding observations and refine our analysis regarding multiplicity of solutions of Bellman's equation for problems where the cost per stage is bounded.

5. THE CASE OF BOUNDED COST PER STAGE

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

$$\sup_{(x,u,w) \in X \times U \times W} g(x, u, w) < \infty. \tag{5.1}$$

This includes the case where g is real-valued and the state space X is finite. 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. (2.4)].

We say that a policy π is *uniformly proper* if

$$\sup_{x\in\widehat{X}} \sum_{k=0}^{\infty} r_k(\pi, x) < \infty. \tag{5.2}$$

Since we have

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

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

$$\mathcal{B} = \left\{ J \in \mathcal{J} \;\middle|\; \sup_{x\in\widehat{X}} J(x) < \infty \right\}. \tag{5.3}$$

Note that when $\widehat{X} = X$, the notion of a uniformly proper policy coincides with the notion of a transient policy used in [Pli78] and [JaC06], 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

$$\widehat{\mathcal{W}}_b = \{J \in \mathcal{B} \mid \hat{J} \leq J\}, \tag{5.4}$$

and by X^* the effective domain of J^*,

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

The following proposition provides conditions for $\hat{J}$ to be the unique fixed point of T within $\widehat{\mathcal{W}}_b$. Its assumptions include the existence of a uniformly proper policy, which implies that $\hat{J}$ belongs to $\widehat{\mathcal{W}}_b$. The proposition also uses the earlier Prop. 3.3 in order to provide conditions for $J^* = \hat{J}$, in which case J^* is the unique fixed point of T within $\mathcal{B}$.

Proposition 5.1: Assume that the cost per stage g is nonnegative and bounded over $X \times U \times W$ [cf. Eq. (5.1)], and that there exists a uniformly proper policy. Then:

(a) $\hat{J}$ is the unique solution of the Bellman Eq. (2.1) within the set $\widehat{\mathcal{W}}_b$ of Eq. (5.4). Moreover, if $\hat{J} = J^*$, then J^* is the unique solution of Bellman's equation within the set $\mathcal{B}$ of Eq. (5.3).

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

(c) Assume in addition that X is finite, that $J^*(x) > 0$ for all $x \neq t$, and that $X^* = \widehat{X}$. Then $\hat{J} = J^*$. Moreover, J^* is the unique solution of Bellman's equation within the set $\mathcal{B}$ and the sequence $\{J_k\}$ generated by the VI algorithm converges to J^* starting from any $J_0 \in \widehat{\mathcal{W}}_b$.

Proof: (a) 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

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

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. 3.2(a), $\hat{J}$ 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 $\widehat{\mathcal{W}}_b$.

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. 2.1(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 $\mathcal{B}$.

(b) Follows from Prop. 3.2(b), since $\widehat{\mathcal{W}}_b \subset \widehat{\mathcal{W}}$, as shown in the proof of part (a).

(c) We have by assumption $0 < J^*(x) \leq \hat{J}(x)$ for all $x \neq t$, while $\hat{J}(x) < \infty$ for all $x \in X^*$ since $X^* = \widehat{X}$. In view of the finiteness of X, we can find a sufficiently large c such that $\hat{J} \leq cJ^*$, so by Prop. 3.3, it follows that $\hat{J} = J^*$. From parts (a) and (b), this implies that J^* is the unique solution within $\mathcal{B}$, and that VI converges to J^* starting from within $\widehat{\mathcal{W}}_b$. **Q.E.D.**

The uniqueness of solution of Bellman's equation within $\mathcal{B}$ in the case where $\hat{J} = J^*$ [cf. part (a) of the preceding proposition] is consistent with Example 4.1. 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 1. Then $\hat{J}(x) = J^*(x) = x$ for all $x \in X$, so $\hat{J}$ and J^* do not belong to $\mathcal{B}$.

In a given practical application, we may be interested in computing either J^* or $\hat{J}$. If the cost per stage is bounded, we may compute $\hat{J}$ with the VI algorithm, assuming that an initial function in the set $\widehat{\mathcal{W}}_b$ of Eq. (5.4) can be found. The computation of J^* is also possible by using the VI algorithm and starting from the zero initial condition; cf. Prop. 2.1(d). An alternative possibility for the case of a finite spaces SSP is to approximate the problem with a sequence of α_k-discounted problems where the discount factors α_k tend to 1. This approach, developed in some detail in Exercise 5.28 of the book [Ber17c], has the advantage that the discounted problems can be solved more reliably and with a broader variety of methods than the original undiscounted SSP.

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 [Ber17b], 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. The structural characteristic responsible for this may be related to the fact that stochastic problems include as special cases discounted cost problems, as illustrated by Example 4.1. By contrast, a discounted cost problem cannot be viewed as a special case of a deterministic undiscounted cost problem with a termination state.

7. 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.

[BeT89] 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.

[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, 4th Edition, Athena Scientific, Belmont, MA.

[Ber13] Bertsekas, D. P., 2013. Abstract Dynamic Programming, 1st Edition, Athena Scientific, Belmont, MA.

[Ber15] 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; to appear in SIAM J. on Control and Optimization.

[Ber17a] Bertsekas, D. P., 2017. "Value and Policy Iteration in Deterministic Optimal Control and Adaptive Dynamic Programming," IEEE Transactions on Neural Networks and Learning Systems, Vol. 28, pp. 500-509.

[Ber17b] Bertsekas, D. P., 2017. "Stable Optimal Control and Semicontractive Dynamic Programming," Report LIDS-P-3506, MIT; to appear in SIAM J. on Control and Optimization.

[Ber17c] Bertsekas, D. P., 2017. Dynamic Programming and Optimal Control, Vol. I, 4th Edition, Athena Scientific, Belmont, MA.

[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.

[Kuc72] Kucera, V., 1972. "The Discrete Riccati Equation of Optimal Control," Kybernetika, Vol. 8, pp. 430-447.

[Kuc73] Kucera, V., 1973. "A Review of the Matrix Riccati Equation," Kybernetika, Vol. 9, pp. 42-61.

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

[Pli78] 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.

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

[Str66] Strauch, R., 1966. "Negative Dynamic Programming," Ann. Math. Statist., Vol. 37, pp. 871-890.

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

[Wil71] Willems, J., 1971. "Least Squares Stationary Optimal Control and the Algebraic Riccati Equation," IEEE Trans. on Automatic Control, Vol. 16, pp. 621-634.

[YuB15] Yu, H., and Bertsekas, D. P., 2015. "A Mixed Value and Policy Iteration Method for Stochastic Control with Universally Measurable Policies," Math. of Operations Research, Vol. 40, pp. 926-968.