Value and Policy Iterations in Optimal Control and Adaptive Dynamic Programming

Dimitri P. Bertsekas

***Abstract*— In this paper, we consider discrete-time infinite horizon problems of optimal control to a terminal set of states. These are the problems that are often taken as the starting point for adaptive dynamic programming. Under very general assumptions, we establish the uniqueness of the solution of Bellman's equation, and we provide convergence results for value and policy iterations.**

***Index Terms*— Dynamic programming (DP), optimal control, policy iteration (PI), value iteration (VI).**

I. INTRODUCTION

IN THIS paper, we consider a deterministic discrete-time optimal control problem involving the system

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

where x_k and u_k are the state and the control at stage k, lying in sets X and U, respectively, and f is a function mapping $X \times U$ to X. 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 for the kth stage, denoted by $g(x_k, u_k)$, is assumed nonnegative and may possibly take the value ∞

$$0 \le g(x_k, u_k) \le \infty, \quad x_k \in X, \ u_k \in U(x_k) \tag{2}$$

[values $g(x_k, u_k) = \infty$ may be used to model constraints on x_k, for example]. We are interested in feedback policies of the form $\pi = \{\mu_0, \mu_1, \ldots\}$, where each μ_k is a function mapping every $x \in X$ 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 for convenience, when confusion cannot arise, will be denoted by μ. No restrictions are placed on X and U: for example, they may be finite sets as in the classical shortest path problems involving a graph, or they may be continuous spaces as in the classical problems of control to the origin or some other terminal sets.

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

$$J_\pi(x_0) = \lim_{k\to\infty} \sum_{t=0}^{k} g(x_t, \mu_t(x_t)), \quad x_0 \in X \tag{3}$$

[the limit exists thanks to the nonnegativity assumption (2)]. We view J_π as a function over X that takes values in $[0, \infty]$. 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

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

and a policy π^* is said to be optimal, if it attains the minimum of $J_\pi(x)$ for all $x \in X$, that is

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

In the context of dynamic programming (DP), one hopes to prove that the optimal cost function J^* satisfies Bellman's equation

$$J^*(x) = \inf_{u\in U(x)} \{g(x, u) + J^*(f(x, u))\} \quad \forall\, x \in X \tag{4}$$

and that an optimal stationary policy may be obtained through the minimization in the right side of this equation. Note that Bellman's equation generically has multiple solutions, since adding a positive constant to any solution produces another solution. A classical result, stated in Proposition 4-1) of Section II, is that the optimal cost function J^* is the smallest solution of Bellman's equation. Here, we will focus on deriving conditions under which J^* is the unique solution within a certain restricted class of functions, whose value within a special set of states is fixed at zero.

In this paper, we will also consider finding J^* with the classical algorithms of a value iteration (VI) and a policy iteration (PI). The VI algorithm starts from some nonnegative function $J_0 : X \mapsto [0, \infty]$, and generates a sequence of functions $\{J_k\}$ according to

$$J_{k+1} = \inf_{u\in U(x)} \{g(x, u) + J_k(f(x, u))\}. \tag{5}$$

We will derive conditions under which J_k converges to J^* pointwise.

The PI algorithm starts from a stationary policy μ^0, and generates a sequence of stationary policies $\{\mu^k\}$ through a sequence of policy evaluations to obtain J_{μ^k} from the equation

$$J_{\mu^k}(x) = g(x, \mu^k(x)) + J_{\mu^k}(f(x, \mu^k(x))), \quad x \in X \tag{6}$$

interleaved with policy improvements to obtain μ^{k+1} from J_{μ^k} according to

$$\mu^{k+1}(x) \in \arg\min_{u\in U(x)} \{g(x, u) + J_{\mu^k}(f(x, u))\}, \quad x \in X. \tag{7}$$

Manuscript received May 10, 2015; revised September 25, 2015; accepted November 8, 2015. Date of publication December 22, 2015; date of current version February 15, 2017.

The author is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: dimitrib@mit.edu).

Digital Object Identifier 10.1109/TNNLS.2015.2503980

We implicitly assume, here, that J_{μ^k} satisfies (6), which is true under the cost nonnegativity assumption (2) (see Proposition 4 in Section II). Also for the PI algorithm to be well-defined, the minimum in (7) should be attained for each $x \in X$, which is true under some conditions that guarantee the compactness of the level sets

$$\{u \in U(x) \mid g(x,u) + J_{\mu^k}(f(x,u)) \le \lambda\}, \quad \lambda \in \Re.$$

We will derive conditions under which J_{μ^k} converges to J^* pointwise.

In this paper, we will address the preceding questions, for the case where there is a nonempty stopping set $X_s \subset X$, which consists of cost-free and absorbing states in the sense that

$$g(x,u) = 0, \quad x = f(x,u) \quad \forall\, x \in X_s,\ u \in U(x). \tag{8}$$

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

$$X_s = \{x \in X \mid J^*(x) = 0\}. \tag{9}$$

In the applications of primary interest, g is usually taken to be strictly positive outside of X_s to encourage asymptotic convergence of the generated state sequence to X_s, so this assumption is natural and often easily verifiable. Besides X_s, another interesting subset of X is

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

Ordinarily, in practical applications, the states in X_f are those from which one can reach the stopping set X_s, at least asymptotically.

For an initial state x, we say that a policy π terminates starting from x, if the state sequence $\{x_k\}$ generated starting from x and using π reaches X_s in finite time, i.e., satisfies $x_{\bar{k}} \in X_s$ for some index $\bar{k}$. A key assumption, in this paper, is that the optimal cost $J^*(x)$ (if it is finite) can be arbitrarily approximated closely by using policies that terminate from x. In particular, in all the results and discussion of this paper, we make the following assumption (except for Proposition 5, which provides conditions under which the assumption holds).

Assumption 1: The cost nonnegativity condition (2), and the stopping set conditions (8) and (9) hold. Moreover, for every pair (x, ϵ) with $x \in X_f$ and $\epsilon > 0$, there exists a policy π that terminates starting from x and satisfies $J_\pi(x) \le J^*(x) + \epsilon$.

Specific and easily verifiable conditions that imply this assumption will be given in Section IV. A prominent case is when X and U are finite, so the problem becomes a deterministic shortest path problem with nonnegative arc lengths. If all the cycles of the state transition graph have positive length, then all the policies π that do not terminate from a state $x \in X_f$ must satisfy $J_\pi(x) = \infty$, implying that there exists an optimal policy that terminates from all $x \in X_f$. Thus, in this case, Assumption 1 is naturally satisfied.

When X is the n-dimensional Euclidean space $\Re^n$, a primary case of interest for this paper, it may easily happen that the optimal policies are not terminating from some $x \in X_f$, but instead the optimal state trajectories may approach X_s asymptotically. This is true, for example, in the classical linear-quadratic optimal control problem, where $X = \Re^n$, $X_s = \{0\}$, $U = \Re^m$, g is the positive semidefinite quadratic, and f is a linear system of the form $x_{k+1} = Ax_k + Bu_k$, where A and B are the given matrices. However, we will show in Section IV that Assumption 1 is satisfied under some natural and easily verifiable conditions.

Regarding notation, we denote by $\Re$ and $\Re^n$ the real line and the n-dimensional Euclidean space, respectively. We denote by $E^+(X)$ the set of all functions $J : X \mapsto [0,\infty]$, and by $\mathcal{J}$ the set of functions

$$\mathcal{J} = \{J \in E^+(X) \mid J(x) = 0 \quad \forall\, x \in X_s\}. \tag{10}$$

Since X_s consists of cost-free and absorbing states [see (8)], the set $\mathcal{J}$ contains the cost function J_π of all policies π, as well as J^*. In our terminology, all equations, inequalities, and convergence limits involving functions are meant to be pointwise. Our main results are given in the following three propositions.

Proposition 1 (Uniqueness of Solution of Bellman's Equation): Let Assumption 1 hold. The optimal cost function J^* is the unique solution of Bellman's equation (4) within the set of functions $\mathcal{J}$.

There are well-known examples where $g \ge 0$, but Assumption 1 does not hold, and there are additional solutions of Bellman's equation within $\mathcal{J}$. The following is a two-state shortest path example, which is discussed in more detail in [1, Sec. 3.1.2] and [2, Example 1.1].

Example 1 (Counterexample for Uniqueness of Solution of Bellman's Equation): Let $X = \{0,1\}$, where 0 is the unique cost-free and absorbing states, $X_s = \{0\}$, and assume that at state 1, we can stay at 1 at no cost, or move to 0 at cost 1. Here $J^*(0) = J^*(1) = 0$, so (9) is violated. Bellman's equation is

$$J^*(0) = J^*(0), \quad J^*(1) = \min\{J^*(1),\ 1 + J^*(0)\}$$

while

$$\mathcal{J} = \{J \mid J(0) = 0,\ J(1) \ge 0\}.$$

It can be seen that the set of solutions of Bellman's equation within $\mathcal{J}$, namely $\{J \mid J(0) = 0,\ 0 \le J(1) \le 1\}$, is infinite.

Proposition 2 (Convergence of VI): Let Assumption 1 hold.

1) The VI sequence $\{J_k\}$ generated by (5) converges pointwise to J^* starting from any function $J_0 \in \mathcal{J}$ with $J_0 \ge J^*$.
2) Assume further that U is a metric space, and the sets $U_k(x,\lambda)$ given by

$$U_k(x,\lambda) = \{u \in U(x) \mid g(x,u) + J_k(f(x,u)) \le \lambda\}$$

are compact for all $x \in X$, $\lambda \in \Re$, and k, where $\{J_k\}$ is the VI sequence $\{J_k\}$ generated by (5) starting from $J_0 \equiv 0$. Then, the VI sequence $\{J_k\}$ generated by (5) converges pointwise to J^* starting from any function $J_0 \in \mathcal{J}$.

The compactness assumption of Proposition 2-2) is satisfied, if $U(x)$ is finite for all $x \in X$. Other easily verifiable assumptions implying this compactness assumption will be given later. Note that when there are solutions to Bellman's equation within $\mathcal{J}$, in addition to J^*, VI will not converge to J^* starting from any of these solutions. However, it is also possible that Bellman's equation has J^* as its unique solution within $\mathcal{J}$, and yet VI does not converge to J^* starting from the zero function because the compactness assumption of Proposition 2-2) is violated. There are several examples of this type in the literature, and the following example, an adaptation in [1, Example 4.3.3], is a deterministic problem for which Assumption 1 is satisfied.

Example 2 (Counterexample for Convergence of VI): Let $X = [0, \infty) \cup \{s\}$, with s being the cost-free and absorbing states, and let $U = (0, \infty) \cup \{\bar{u}\}$, where $\bar{u}$ is a special stopping control, which moves the system from states $x \geq 0$ to state s at unit cost. The system has the form

$$x_{k+1} = \begin{cases} x_k + u_k & \text{if } x_k \geq 0 \text{ and } u_k \neq \bar{u} \\ s & \text{if } x_k \geq 0 \text{ and } u_k = \bar{u} \\ s & \text{if } x_k = s \text{ and } u_k \in U. \end{cases}$$

The cost per stage has the form

$$g(x_k, u_k) = \begin{cases} x_k & \text{if } x_k \geq 0 \text{ and } u_k \neq \bar{u} \\ 1 & \text{if } x_k \geq 0 \text{ and } u_k = \bar{u} \\ 0 & \text{if } x_k = s \text{ and } u_k \in U. \end{cases}$$

Let also $X_s = \{s\}$. Then, it can be verified that

$$J^*(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x = s \end{cases}$$

and that an optimal policy is to use the stopping control $\bar{u}$ at every state (since using any other controls at states $x \geq 0$, leads to unbounded accumulation of positive cost). Thus, it can be seen that Assumption 1 is satisfied. On the other hand, the VI algorithm is

$$J_{k+1}(x) = \min\left\{1 + J_k(s),\ \inf_{u>0}\{x + J_k(x+u)\}\right\}$$

for $x \geq 0$ and $J_{k+1}(s) = J_k(s)$, and it can be verified by induction that starting from $J_0 \equiv 0$, the sequence $\{J_k\}$ is given for all k by

$$J_k(x) = \begin{cases} \min\{1,\ kx\} & \text{if } x \geq 0 \\ 0 & \text{if } x = s. \end{cases}$$

Thus, $J_k(0) = 0$ for all k, while $J^*(0) = 1$, so the VI algorithm fails to converge for the state $x = 0$. The difficulty, here, is that the compactness assumption of Proposition 2-2) is violated.

Proposition 3 (Convergence of PI): Let Assumption 1 hold. A sequence $\{J_{\mu^k}\}$ generated by the PI algorithm (6) and (7) satisfies $J_{\mu^k}(x) \downarrow J^*(x)$ for all $x \in X$.

It is implicitly assumed in the preceding proposition that the PI algorithm is well-defined in the sense that the minimization in the policy improvement operation (7) can be carried out for every $x \in X$. Easily verifiable conditions that guarantee this also guarantee the compactness condition of Proposition 2-2), and will be noted following Proposition 4 in Section II. Moreover, in Section IV, we will prove a similar convergence result for a variant of the PI algorithm where the policy evaluation is approximately carried out through a finite number of VIs.

Example 3 (Counterexample for Convergence of PI): For a simple example where the PI sequence J_{μ^k} does not converge to J^*, if Assumption 1 is violated, consider the two-state shortest path Example 1. Let μ be the suboptimal policy that moves from state 1 to state 0. Then, $J_\mu(0) = 0$ and $J_\mu(1) = 1$, and it can be seen that μ satisfies the policy improvement equation

$$\mu(1) \in \arg\min\{1 + J_\mu(0),\ J_\mu(1)\}.$$

Thus, PI may stop with the suboptimal policy μ.

The results of the preceding three propositions are new at the level of generality given here. For example, there has been no proposal of a valid PI algorithm in the classical literature on nonnegative cost infinite horizon Markovian decision problems (exceptions are special cases, such as linear-quadratic problems [3]). The ideas of this paper stem from a more general analysis regarding the convergence of VI, which was recently presented in the author's research monograph on abstract DP [1], and various extensions given in [2] and [4]. Bertsekas and Yu [5], [6] deal with issues that relate in part to the intricacies of the convergence of VI and PI in undiscounted infinite horizon DP.

This paper is organized as follows. In Section II, we provide the background and references, which place in context our results and methods of analysis in relation to the literature. In Section III, we give the proofs of Propositions 1–3. In Section IV, we discuss the special cases and easily verifiable conditions that imply our assumptions, and we provide the extensions of our analysis.

II. Background

The issues discussed in this paper have received attention since the 1960s, originally in [7], who considered the case $g \leq 0$, and the work in [8], who considered the case $g \geq 0$. For textbook accounts, we refer to [9]–[11], and for a more abstract development, we refer to the monograph [1]. These works showed that the cases where $g \leq 0$ (which corresponds to the maximization of nonnegative rewards) and $g \geq 0$ (which is most relevant to the control problems of this paper) are quite different in structure. In particular, while VI converges to J^* starting for $J_0 \equiv 0$ when $g \leq 0$, this is not so when $g \geq 0$; a certain compactness condition is needed to guarantee this [see Example 2 and part 4) of the following proposition]. Moreover when $g \geq 0$, Bellman's equation may have solutions $\hat{J} \neq J^*$ with $\hat{J} \geq J^*$ (see Example 1), and VI will not converge to J^* starting from such $\hat{J}$. In addition, it is known that, in general, PI need not converge to J^* and may instead stop with a suboptimal policy (see Example 3).

The following proposition gives the standard results when $g \geq 0$ (see [9, Propositions 5.2, 5.4, and 5.10], [11, Propositions 4.1.1, 4.1.3, 4.1.5, and 4.1.9], or [1, Propositions 4.3.3, 4.3.9, and 4.3.14]). These results hold for stochastic infinite horizon DP problems with nonnegative cost per stage, and do not consider the favorable structure of deterministic problems or the presence of the stopping set X_s.

Proposition 4: Let the nonnegativity condition (2) hold.

1) J^* satisfies Bellman's equation (4), and if $\hat{J} \in E^+(X)$ is another solution, i.e., $\hat{J}$ satisfies
$$\hat{J}(x) = \inf_{u \in U(x)} \{g(x,u) + \hat{J}(f(x,u))\} \quad \forall\, x \in X \tag{11}$$
then $J^* \leq \hat{J}$.

2) For all stationary policies μ, we have
$$J_\mu(x) = g(x, \mu(x)) + J_\mu(f(x, \mu(x))) \quad \forall\, x \in X. \tag{12}$$

3) A stationary policy μ^* is optimal, if and only if
$$\mu^*(x) \in \arg\min_{u \in U(x)} \{g(x,u) + J^*(f(x,u))\} \quad \forall\, x \in X. \tag{13}$$

4) If U is a metric space and the sets
$$kU_k(x, \lambda) = \{u \in U(x) \mid g(x,u) + J_k(f(x,u)) \leq \lambda\} \tag{14}$$
are compact for all $x \in X$, $\lambda \in \Re$, and k, where $\{J_k\}$ is the sequence generated by VI [see (5)] starting from $J_0 \equiv 0$; then, there exists at least one optimal stationary policy, and we have $J_k \to J^*$.

The compactness assumption of part 4) was originally given in [12] and [13] (a related condition was independently given in [14]). It has been used in several other works and related contexts, such as [1, Proposition 3.2.1], [5], and [15]. In particular, the condition of part 4) holds when $U(x)$ is a finite set for all $x \in X$. The condition of part 4) also holds when $X = \Re^n$, and for each $x \in X$, the set
$$\{u \in U(x) \mid g(x,u) \leq \lambda\}$$
is a compact subset of $\Re^m$, for all $\lambda \in \Re$, and g and f are continuous in u. The proof consists of showing by induction that the VI iterates J_k have compact level sets, and hence are lower semicontinuous.

Let us also note a recent result in [6], where it was shown that J^* is the unique solution of Bellman's equation within the class of all functions $J \in E^+(X)$ that satisfy
$$0 \leq J \leq cJ^* \quad \text{for some } c > 0 \tag{15}$$
(we refer to [6] for discussion and references to antecedents of this result). Moreover, it was shown that VI converges to J^* starting from any function satisfying the condition
$$J^* \leq J \leq cJ^* \quad \text{for some } c > 0$$
and under the compactness conditions of Proposition 4-4), starting from any J that satisfies (15). The same paper and a related paper [5] extensively discuss the PI algorithms for stochastic nonnegative cost problems.

For deterministic problems, there has been substantial research in the adaptive DP literature, regarding the validity of Bellman's equation and the uniqueness of its solution, as well as the attendant questions of convergence of VI and PI. In particular, the infinite horizon deterministic optimal control for both discrete-time and continuous-time systems has been considered since the early days of DP in the works of Bellman. For continuous-time problems, the questions discussed in this paper involve substantial technical difficulties, since the analog of the (discrete-time) Bellman equation (4) is the steady-state form of the (continuous-time) Hamilton–Jacobi–Bellman equation, a nonlinear partial differential equation, the solution and analysis of which is, in general, very complicated. A formidable difficulty is the potential lack of differentiability of the optimal cost function, even for simple problems, such as the time-optimal control of the second-order linear systems to the origin.

The analog of VI for continuous-time systems essentially involves the time integration of the Hamilton–Jacobi–Bellman equation, and its analysis must deal with difficult issues of stability and convergence to a steady-state solution. Nonetheless, there have been the proposals of the continuous-time PI algorithms in the early papers [3], [16]–[18], and the thesis [19], as well as more recently in several works (see the book [20], the survey [21], and the references quoted there). These works also address the possibility of value function approximation, similar to other approximation-oriented methodologies, such as neuro-DP [22] and reinforcement learning [23], which primarily consider discrete-time systems. For example, among the restrictions of the PI method is that it must be started with a stabilizing controller [3], which considered linear-quadratic continuous-time problems, and showed convergence to the optimal policy of the PI algorithm, assuming that an initial stabilizing linear controller is used. By contrast, no such restriction is needed in the PI methodology of this paper; questions of stability are only indirectly addressed through the finiteness of the values $J^*(x)$ and Assumption 1.

For discrete-time systems, there has been much research, both for VI and PI algorithms. For a selective list of recent references, which themselves contain extensive lists of other references, see the book [20], the papers [25]–[29], the survey papers in the edited volumes [30] and [31], and the special issue [24]. Some of these works relate to continuous-time problems as well, and in their treatment of algorithmic convergence, typically assume that X and U are the Euclidean spaces, as well as continuity and other conditions on g, a special structure of the system, and so on. Moreover, some of these works are motivated by the problems of adaptive control of systems with unknown parameters, using simulation-based methods, such as Q-learning, as first proposed in [32]. It is beyond our scope to provide a detailed survey of the state of the art of the VI and PI methodologies in the context of adaptive DP. However, it should be clear that the works in this field involve more restrictive assumptions than our corresponding results of Propositions 1–3. Of course, these works also address the questions that we do not, such as issues of stability of the obtained controllers, the use of approximations, and so on. Thus, the results of this paper may be viewed as new in that they rely on very general assumptions, yet do not address some important practical issues. The line of analysis of this paper, which is based on general results of Markovian decision problem theory and abstract forms of DP, is also different from the lines of analysis of works in adaptive DP, which make heavy use of the deterministic character of the problem and control theoretic methods, such as Lyapunov stability.

Still there is a connection between our line of analysis and Lyapunov stability. In particular, if π^* is an optimal controller, i.e., $J_{\pi^*} = J^*$, then for every $x_0 \in X_f$, the state sequence $\{x_k\}$

generated using π^* and starting from x_0 remains within X_f and satisfies $J^*(x_k) \downarrow 0$. This can be seen by writing

$$J^*(x_0) = \sum_{t=0}^{k-1} g(x_t, \mu_t^*(x_t)) + J^*(x_k), \quad k = 1, 2, \ldots$$

and using the facts $g \geq 0$ and $J^*(x_0) < \infty$. Thus, an optimal controller, restricted to the subset X_f, may be viewed as a Lyapunov-stable controller where the Lyapunov function is J^*.

On the other hand, existence of a stable controller does not necessarily imply that J^* is real-valued. In particular, it may not be true that if the generated sequence $\{x_k\}$ by an optimal controller starting from some x_0 converges to X_s, then we have $J^*(x_0) < \infty$. The reason is that the cost per stage g may not decrease fast enough as we approach X_s. As an example, let

$$X = \{0\} \cup \{1/m \mid m\text{: is a positive integer}\}$$

with $X_s = \{0\}$, and assume that there is a unique controller, which moves from $1/m$ to $1/(m+1)$ with incurred cost $1/m$. Then, we have $J^*(x) = \infty$ for all $x \neq 0$, despite the fact that the controller is stable in the sense that it generates a sequence $\{x_k\}$ converging to 0 starting from every $x_0 \neq 0$.

III. Proofs of the Main Results

Let us denote for all $x \in X$

$$\Pi_{T,x} = \{\pi \in \Pi \mid \pi \text{ terminates from } x\}$$

and note the following key implication of Assumption 1:

$$J^*(x) = \inf_{\pi \in \Pi_{T,x}} J_\pi(x) \quad \forall\, x \in X_f. \tag{16}$$

In the subsequent arguments, the significance of policies that terminate starting from some initial state x_0 is that the corresponding generated sequences $\{x_k\}$ satisfy $J(x_k) = 0$ for all $J \in \mathcal{J}$ and k sufficiently large.

Proof of Proposition 1: Let $\hat{J} \in \mathcal{J}$ be a solution of the Bellman equation (11), so that

$$\hat{J}(x) \leq g(x,u) + \hat{J}(f(x,u)) \quad \forall\, x \in X,\ u \in U(x) \tag{17}$$

while by Proposition 4-1), $J^* \leq \hat{J}$. For any $x_0 \in X_f$ and policy $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi_{T,x_0}$, we have by using repeatedly (17)

$$J^*(x_0) \leq \hat{J}(x_0) \leq \hat{J}(x_k) + \sum_{t=0}^{k-1} g(x_t, \mu_t(x_t)), \quad k = 1, 2, \ldots$$

where $\{x_k\}$ is the state sequence generated starting from x_0 and using π. Also, since $\pi \in \Pi_{T,x_0}$ and hence $x_k \in X_s$ and $\hat{J}(x_k) = 0$ for all sufficiently large k, we have

$$\begin{aligned} \limsup_{k\to\infty} &\left\{ \hat{J}(x_k) + \sum_{t=0}^{k-1} g(x_t, \mu_t(x_t)) \right\} \\ &= \lim_{k\to\infty} \left\{ \sum_{t=0}^{k-1} g(x_t, \mu_t(x_t)) \right\} \\ &= J_\pi(x_0). \end{aligned}$$

By combining the last two relations, we obtain

$$J^*(x_0) \leq \hat{J}(x_0) \leq J_\pi(x_0) \quad \forall\, x_0 \in X_f,\ \pi \in \Pi_{T,x_0}.$$

Taking the infimum over $\pi \in \Pi_{T,x_0}$ and using (16), it follows that $J^*(x_0) = \hat{J}(x_0)$ for all $x_0 \in X_f$. Also for $x_0 \notin X_f$, we have $J^*(x_0) = \hat{J}(x_0) = \infty$ [since $J^* \leq \hat{J}$ by Proposition 4-1)], so we obtain $J^* = \hat{J}$. □

Proof of Proposition 2:

1) Suppose that $J_0 \in \mathcal{J}$ and $J_0 \geq J^*$. Starting with J_0, let us apply the VI operation to both sides of the inequality $J_0 \geq J^*$. Since J^* is a solution of Bellman's equation and VI has a monotonicity property that maintains the direction of functional inequalities, we see that $J_1 \geq J^*$. Continuing similarly, we obtain $J_k \geq J^*$ for all k. Moreover, we clearly have $J_k(x) = 0$ for all $x \in X_s$, so $J_k \in \mathcal{J}$ for all k. We now argue that, since J_k is produced by k steps of VI starting from J_0, it is the optimal cost function of the k-stage version of the problem with terminal cost function J_0. Therefore, we have for every $x_0 \in X$ and policy $\pi = \{\mu_0, \mu_1, \ldots\}$

$$J^*(x_0) \leq J_k(x_0) \leq J_0(x_k) + \sum_{t=0}^{k-1} g(x_t, \mu_t(x_t)) \qquad k = 1, 2, \ldots$$

where $\{x_t\}$ is the state sequence generated starting from x_0 and using π. If $x_0 \in X_f$ and $\pi \in \Pi_{T,x_0}$, we have $x_k \in X_s$ and $J_0(x_k) = 0$ for all sufficiently large k, so that

$$\begin{aligned} \limsup_{k\to\infty} &\left\{ J_0(x_k) + \sum_{t=0}^{k-1} g(x_t, \mu_t(x_t)) \right\} \\ &= \lim_{k\to\infty} \left\{ \sum_{t=0}^{k-1} g(x_t, \mu_t(x_t)) \right\} \\ &= J_\pi(x_0). \end{aligned}$$

By combining the last two relations, we obtain

$$J^*(x_0) \leq \liminf_{k\to\infty} J_k(x_0) \leq \limsup_{k\to\infty} J_k(x_0) \leq J_\pi(x_0)$$

for all $x_0 \in X_f$ and $\pi \in \Pi_{T,x_0}$. Taking the infimum over $\pi \in \Pi_{T,x_0}$ and using (16), it follows that $\lim_{k\to\infty} J_k(x_0) = J^*(x_0)$ for all $x_0 \in X_f$. Since for $x_0 \notin X_f$, we have $J^*(x_0) = J_k(x_0) = \infty$, we obtain $J_k \to J^*$.

2) Let $\{J_k\}$ be the VI sequence generated starting from some function $J \in \mathcal{J}$. By the monotonicity of the VI operation, $\{J_k\}$ lies between the sequence of VI iterates starting from the zero function [which converges to J^* from below by Proposition 4-4)], and the sequence of VI iterates starting from $J_0 = \max\{J, J^*\}$ [which converges to J^* from above by part 1)]. □

Proof of Proposition 3: If μ is a stationary policy and $\bar{\mu}$ satisfies the policy improvement equation

$$\bar{\mu}(x) \in \arg\min_{u \in U(x)} \{g(x,u) + J_\mu(f(x,u))\}, \quad x \in X$$

[see (7)], we have for all $x \in X$

$$\begin{aligned} J_\mu(x) &= g(x, \mu(x)) + J_\mu(f(x, \mu(x))) \\ &\geq \min_{u \in U(x)} \{g(x,u) + J_\mu(f(x,u))\} \\ &= g(x, \bar{\mu}(x)) + J_\mu(f(x, \bar{\mu}(x))) \end{aligned} \tag{18}$$

where the first equality follows from Proposition 4-2) and the second equality follows from the definition of $\bar{\mu}$. Let us fix x and let $\{x_k\}$ be the sequence generated starting from x and using μ. By repeatedly applying (18), we see that the sequence $\{\tilde{J}_k(x)\}$ defined by

$$\tilde{J}_0(x) = J_\mu(x)$$
$$\tilde{J}_1(x) = J_\mu(x_1) + g(x, \bar{\mu}(x))$$

and more generally

$$\tilde{J}_k(x) = J_\mu(x_k) + \sum_{t=0}^{k-1} g(x_t, \bar{\mu}(x_t)), \quad k = 1, 2, \ldots$$

is monotonically nonincreasing. Thus, also using (18), we have

$$\begin{aligned} J_\mu(x) &\geq \min_{u \in U(x)} \{g(x,u) + J_\mu(f(x,u))\} \\ &= \tilde{J}_1(x) \\ &\geq \tilde{J}_k(x) \end{aligned}$$

for all $x \in X$ and $k \geq 1$. This implies that

$$\begin{aligned} J_\mu(x) &\geq \min_{u \in U(x)} \{g(x,u) + J_\mu(f(x,u))\} \\ &\geq \lim_{k\to\infty} \tilde{J}_k(x) \\ &= \lim_{k\to\infty} \left\{ J_\mu(x_k) + \sum_{t=0}^{k-1} g(x_t, \bar{\mu}(x_t)) \right\} \\ &\geq \lim_{k\to\infty} \sum_{t=0}^{k-1} g(x_t, \bar{\mu}(x_t)) \\ &= J_{\bar{\mu}}(x) \end{aligned}$$

where the last inequality follows, since $J_\mu \geq 0$. In conclusion, we have

$$J_\mu(x) \geq \inf_{u \in U(x)} \{g(x,u) + J_\mu(f(x,u))\} \geq J_{\bar{\mu}}(x), \quad x \in X.$$

Using μ^k and μ^{k+1} in place of μ and $\bar{\mu}$ in the preceding relation, we obtain for all $x \in X$

$$J_{\mu^k}(x) \geq \inf_{u \in U(x)} \{g(x,u) + J_{\mu^k}(f(x,u))\} \geq J_{\mu^{k+1}}(x). \quad (19)$$

Thus, the sequence $\{J_{\mu^k}\}$ generated by PI monotonically converges to some function $J_\infty \in E^+(X)$, i.e., $J_{\mu^k} \downarrow J_\infty$. Moreover, by taking the limit as $k \to \infty$ in (19), we have the two relations

$$J_\infty(x) \geq \inf_{u \in U(x)} \{g(x,u) + J_\infty(f(x,u))\}, \quad x \in X$$

and

$$g(x,u) + J_{\mu^k}(f(x,u)) \geq J_\infty(x), \quad x \in X, \ u \in U(x).$$

We now take the limit in the second relation as $k \to \infty$, then the infimum over $u \in U(x)$, and then combine with the first relation, to obtain

$$J_\infty(x) = \inf_{u \in U(x)} \{g(x,u) + J_\infty(f(x,u))\}, \quad x \in X.$$

Thus, J_∞ is a solution of Bellman's equation, satisfying $J_\infty \in \mathcal{J}$ (since $J_{\mu^k} \in \mathcal{J}$ and $J_{\mu^k} \downarrow J_\infty$), so by the uniqueness result of Proposition 1, we have $J_\infty = J^*$. □

IV. Discussion, Special Cases, and Extensions

In this section, we elaborate on our main results and we derive easily verifiable conditions under which our assumptions hold.

A. Conditions That Imply Assumption 1

Consider Assumption 1. As noted in Section I, it holds when X and U are finite, a terminating policy exists from every x, and all the cycles of the state transition graph have positive length. For the case where X is infinite, let us assume that X is a normed space with norm denoted by $\|\cdot\|$, and say that π asymptotically terminates from x if the sequence $\{x_k\}$ generated starting from x and using π converges to X_s in the sense that

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

where $\text{dist}(x, X_s)$ denotes the minimum distance from x to X_s

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

The following proposition provides readily verifiable conditions that guarantee Assumption 1.

Proposition 5: Let the cost nonnegativity condition (2) and the stopping set conditions (8) and (9) hold, and assume further the following.

1) For every $x \in X_f$ and $\epsilon > 0$, there exists a policy π that asymptotically terminates from x and satisfies

$$J_\pi(x) \leq J^*(x) + \epsilon.$$

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

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

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

Then, Assumption 1 holds.

Proof: Fix $x \in X_f$ and $\epsilon > 0$. Let π be a policy that asymptotically terminates from x, and satisfies $J_\pi(x) \leq J^*(x) + \epsilon$, as per condition (1). Starting from x, this policy will generate a sequence $\{x_k\}$, such that for some index $\bar{k}$, we have

$$\text{dist}(x_{\bar{k}}, X_s) \leq \delta_\epsilon$$

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

$$J_{\pi'}(x) = J_{\pi,\bar{k}}(x) + J_{\bar{\pi}}(x_{\bar{k}}) \leq J_\pi(x) + J_{\bar{\pi}}(x_{\bar{k}}) \leq J^*(x) + 2\epsilon$$

where $J_{\pi,\bar{k}}(x)$ is the cost incurred by π starting from x up to reaching $x_{\bar{k}}$. □

Condition (1) of the preceding proposition requires that for states $x \in X_f$, the optimal cost $J^*(x)$ can be arbitrarily closely achieved with policies that asymptotically terminate from x. Problems for which condition (1) holds are those involving a cost per stage that is strictly positive outside of X_s.

More precisely, condition (1) holds if for each $\delta > 0$, there exists $\epsilon > 0$, such that

$$\inf_{u \in U(x)} g(x, u) \geq \epsilon \quad \forall\, x \in X \text{ such that } \operatorname{dist}(x, X_s) \geq \delta. \tag{20}$$

Then, for any x and policy π that does not asymptotically terminate from x, we will have $J_\pi(x) = \infty$, so that if $x \in X_f$, all policies π with $J_\pi(x) < \infty$ must be asymptotically terminating from x. In applications, condition (1) is natural and consistent with the aim of steering the state toward the terminal set X_s with finite cost. Condition (2) is a controllability condition implying that the state can be steered into X_s with arbitrarily small cost from a starting state that is sufficiently close to X_s.

Example 4 (Linear System Case): Consider a linear system

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

where A and B are given matrices, with the terminal set being the origin, i.e., $X_s = \{0\}$. We assume the following.

1) $X = \Re^n$, $U = \Re^m$, and there is an open sphere R centered at the origin, such that $U(x)$ contains R for all $x \in X$.
2) The system is controllable, i.e., one may drive the system from any state to the origin within at most n steps using suitable controls, or equivalently that the matrix $[B\ AB\ \cdots A^{n-1}B]$ has rank n.
3) g satisfies

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

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

Then, condition (2) of Proposition 5 is satisfied, while $x = 0$ is cost free and absorbing [see (8)]. Still, however, in the absence of additional assumptions, there may be multiple solutions to Bellman's equation within $\mathcal{J}$.

As an example, consider the scalar system $x_{k+1} = \gamma x_k + u_k$ with $X = U(x) = \Re$, and the quadratic cost $g(x, u) = u^2$. Then, Bellman's equation has the form

$$J(x) = \inf_{u \in \Re}\{u^2 + J(\gamma x + u)\}, \quad x \in \Re$$

and it is seen that the optimal cost function, $J^*(x) \equiv 0$, is a solution. Let us assume that $\gamma > 1$, so the system is unstable (the instability of the system is important for the purpose of this example). Then, it can be verified that the quadratic function $J(x) = (\gamma^2 - 1)x^2$, which belongs to $\mathcal{J}$, also solves Bellman's equation. This is a case where the algebraic Riccati equation associated with the problem has two nonnegative solutions, because there is no cost on the state, and a standard observability condition for the uniqueness of solution of the Riccati equation is violated.

If on the other hand, in addition to 1)–3), we assume that for some positive scalars q and p, we have $\inf_{u \in U(x)} g(x, u) \geq q\|x\|^p$ for all $x \in \Re^n$, then $J^*(x) > 0$ for all $x \neq 0$ [see (9)], while condition (1) of Proposition 5 is satisfied as well [see (20)]. Then by Proposition 5, Assumption 1 holds, and Bellman's equation has a unique solution within $\mathcal{J}$.

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

B. Optimistic Form of PI

Let us consider a variant of PI where the policies are evaluated inexactly, with a finite number of VIs. In particular, this algorithm starts with some $J_0 \in E(X)$, and generates a sequence of cost function and policy pairs $\{J_k, \mu^k\}$ as follows. Given J_k, we generate μ^k according to

$$\mu^k(x) \in \arg\min_{u \in U(x)}\{g(x, u) + J_k(f(x, u))\}, \quad x \in X \tag{21}$$

and then, we obtain J_{k+1} with $m_k \geq 1$ VIs using μ^k

$$J_{k+1}(x_0) = J_k(x_{m_k}) + \sum_{t=0}^{m_k - 1} g(x_t, \mu^k(x_t)), \quad x_0 \in X \tag{22}$$

where $\{x_t\}$ is the sequence generated using μ^k and starting from x_0, and m_k are arbitrary positive integers. Here, J_0 is a function in $\mathcal{J}$ that is required to satisfy

$$J_0(x) \geq \inf_{u \in U(x)}\{g(x, u) + J_0(f(x, u))\} \quad \forall\, x \in X,\ u \in U(x). \tag{23}$$

For example, J_0 may be equal to the cost function of some stationary policy, or be the function that takes the value 0 for $x \in X_s$ and ∞ at $x \notin X_s$. Note that when $m_k \equiv 1$, the method is equivalent to VI, while the case $m_k = \infty$ corresponds to the standard PI considered earlier. In practice, the most effective value of m_k may be found experimentally, with moderate values $m_k > 1$ usually working best. We refer to the textbooks [10] and [11] for discussions of this type of inexact PI algorithm (in [10] it is called the modified PI, while in [11] it is called the optimistic PI).

Proposition 6 (Convergence of Optimistic PI): Let Assumption 1 hold. For the PI algorithm (21) and (22), where J_0 belongs to $\mathcal{J}$ and satisfies the condition (23), we have $J_k \downarrow J^*$.

Proof: We have for all $x \in X$

$$\begin{aligned} J_0(x) &\geq \inf_{u \in U(x)}\{g(x, u) + J_0(f(x, u))\} \\ &= g(x, \mu^0(x)) + J_0(f(x, \mu^0(x))) \\ &\geq J_1(x) \\ &\geq g(x, \mu^0(x)) + J_1(f(x, \mu^0(x))) \\ &\geq \inf_{u \in U(x)}\{g(x, u) + J_1(f(x, u))\} \\ &= g(x, \mu^1(x)) + J_1(f(x, \mu^1(x))) \\ &\geq J_2(x) \end{aligned}$$

where the first inequality is the condition (23), the second and third inequalities follow because of the monotonicity of the m_0 VIs (22) for μ^0, and the fourth inequality follows from the policy improvement equation (21). Continuing similarly, we have

$$J_k(x) \geq \inf_{u \in U(x)} \{g(x,u) + J_k(f(x,u))\} \geq J_{k+1}(x)$$

for all $x \in X$ and k. Moreover, since $J_0 \in \mathcal{J}$, we have $J_k \in \mathcal{J}$ for all k. Thus, $J_k \downarrow J_\infty$ for some $J_\infty \in \mathcal{J}$, and similar to the proof of Proposition 3, it follows that J_∞ is a solution of Bellman's equation. Hence, by the uniqueness result of Proposition 1, we have $J_\infty = J^*$. □

C. Minimax Control to a Terminal Set of States

Our analysis can be readily extended to minimax problems with a terminal set of states. Here, the system is

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

where w_k is the control of an antagonistic opponent that aims to maximize the cost function. We assume that w_k is chosen from a given set W to maximize the sum of costs per stage, which are assumed nonnegative

$$0 \leq g(x,u,w) \leq \infty, \quad x \in X,\ U \in U(x),\ w \in W.$$

We wish to choose a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ to minimize the cost function

$$J_\pi(x_0) = \lim_{k \to \infty} \sup_{\substack{w_t \in W \\ t=0,1,\ldots}} \sum_{t=0}^{k} g(x_t, \mu_t(x_t), w_t)$$

where $\{x_t, \mu_t(x_t)\}$ is a state-control sequence corresponding to π and the sequence $\{w_0, w_1, \ldots\}$. We assume that there is a termination set X_s, the states of which are cost free and absorbing, that is

$$g(x,u,w) = 0, \quad x = f(x,u,w)$$

for all $x \in X_s$, $u \in U(x)$, $w \in W$, and that all states outside X_s have strictly positive optimal cost, so that

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

The finite-state, finite-control version of this problem has been discussed in [4], under the name robust shortest path planning, for the case where g can take both positive and negative values. A problem that is closely related is reachability of a target set in minimum time, which is obtained for

$$g(x,u,w) = \begin{cases} 0 & \text{if } x \in X_s \\ 1 & \text{if } x \notin X_s \end{cases}$$

assuming also that the control process stops once the state enters the set X_s. Here, w is a disturbance described by set membership ($w \in W$), and the objective is to reach the set X_s in the minimum guaranteed number of steps. The set X_f is the set of states for which X_s is guaranteed to be reached in a finite number of steps. Another related problem is reachability of a target tube, where for a given set $\hat{X}$

$$g(x,u,w) = \begin{cases} 0 & \text{if } x \in \hat{X} \\ 1 & \text{if } x \notin \hat{X} \end{cases}$$

and the objective is to find the initial states starting from which we can guarantee to keep all future states within $\hat{X}$. These two reachability problems were first formulated and analyzed as part of the author's Ph.D. thesis research [33], and the subsequent paper [34]. In fact, the reachability algorithms given in these works are essentially special cases of the VI algorithm of this paper, starting with appropriate initial functions J_0. Moreover, the compactness condition of Proposition 2-2) draws its origin from corresponding compactness conditions first given in these references.

To extend our results to the general form of the minimax problem described above, we need to adapt the definition of termination. In particular, given a state x, in the minimax context, we say that a policy π terminates from x if there exists an index $\bar{k}$ [which depends on (π, x)], such that the sequence $\{x_k\}$, which is generated starting from x and using π, satisfies $x_{\bar{k}} \in X_s$ for all sequences $\{w_0, \ldots, w_{\bar{k}-1}\}$ with $w_t \in W$ for all $t = 0, \ldots, \bar{k}-1$. Then, Assumption 1 is modified to reflect this new definition of termination, and our results can be readily extended, with Propositions 1–3 and 6, and their proofs, essentially holding as stated. The main adjustment needed is to replace expressions of the forms

$$g(x,u) + J(f(x,u))$$

and

$$J(x_k) + \sum_{t=0}^{k-1} g(x_t, u_t)$$

in these proofs with

$$\sup_{w \in W} \{g(x,u,w) + J(f(x,u,w))\}$$

and

$$\sup_{\substack{w_t \in W \\ t=0,\ldots,k-1}} \left\{ J(x_k) + \sum_{t=0}^{k-1} g(x_t, u_t, w_t) \right\}$$

respectively; see also [2] for a more abstract view of such lines of argument.

V. Concluding Remarks

In this paper, we have considered problems of deterministic optimal control to a terminal set of states subject to very general assumptions. Under reasonably practical conditions, we have established the uniqueness of solution of Bellman's equation, and the convergence of VI and PI algorithms, even when there are states with infinite optimal cost. Our analysis bypasses the need for assumptions involving the existence of globally stabilizing controllers, which guarantee that the optimal cost function J^* is real-valued. This generality makes our results a convenient starting point for the analysis of problems involving additional assumptions, and perhaps cost function approximations.

While we have restricted attention to undiscounted problems, the line of analysis of this paper also applies to discounted problems with one-stage cost function g that may be unbounded from above. Similar but more favorable results can be obtained, thanks to the presence of the discount factor; see [2], which contains the related analysis for stochastic and minimax, discounted and undiscounted problems, with nonnegative cost per stage.

The results for these problems, and the results of this paper, have a common ancestry. They fundamentally draw their validity from the notions of regularity, which were developed in the author's abstract DP monograph [1] and were recently extended in [2]. Let us describe the regularity idea briefly, and its connection to the analysis of this paper. Given a set of functions $S \in E^+(X)$, we say that a collection $\mathcal{C}$ of policy-state pairs (π, x_0), with $\pi \in \Pi$ and $x_0 \in X$, is S-regular if for all $(\pi, x_0) \in \mathcal{C}$ and $J \in S$, we have

$$J_\pi(x_0) = \lim_{k\to\infty} \left\{ J(x_k) + \sum_{t=0}^{k-1} g(x_t, \mu_t(x_t)) \right\}.$$

In words, for all $(\pi, x_0) \in \mathcal{C}$, $J_\pi(x_0)$ can be obtained in the limit by VI starting from any $J \in S$ (rather than just from $J \equiv 0$). The favorable properties with respect to VI of an S-regular collection $\mathcal{C}$ can be translated into interesting properties relating to solutions of Bellman's equation and convergence of VI. In particular, the optimal cost function over the set of policies $\{\pi \mid (\pi, x) \in \mathcal{C}\}$

$$J^*_{\mathcal{C}}(x) = \inf_{\{\pi \mid (\pi,x)\in\mathcal{C}\}} J_\pi(x), \quad x \in X$$

under appropriate problem-dependent assumptions is the unique solution of Bellman's equation within the set $\{J \in S \mid J \geq J^*_{\mathcal{C}}\}$, and can be obtained by VI starting from any J within that set [2].

Within the deterministic optimal control context of this paper, it works well to choose $\mathcal{C}$ to be the set of all (π, x), such that $x \in X_f$ and π is terminating starting from x, and to choose S to be $\mathcal{J}$, as defined by (10). Then, in view of Assumption 1, we have $J^*_{\mathcal{C}} = J^*$, and the favorable properties of $J^*_{\mathcal{C}}$ are shared by J^*. For other types of problems, different choices of $\mathcal{C}$ may be appropriate, and corresponding results relating to the uniqueness of solutions of Bellman's equation and the validity of VI and PI may be obtained; see the analysis of [2].

REFERENCES

[1] D. P. Bertsekas, *Abstract Dynamic Programming*. Belmont, MA, USA: Athena Scientific, 2013.

[2] D. P. Bertsekas, "Regular policies in abstract dynamic programming," Lab. Inf. Decision Syst., MIT, Cambridge, MA, USA, Tech. Rep. LIDS-P-3173, Apr. 2015.

[3] D. L. Kleinman, "On an iterative technique for Riccati equation computations," *IEEE Trans. Autom. Control*, vol. 13, no. 1, pp. 114–115, Feb. 1968.

[4] D. P. Bertsekas, "Robust shortest path planning and semicontractive dynamic programming," Lab. Inf. Decision Syst., MIT, Cambridge, MA, USA, Tech. Rep. LIDSP-2915, Jan. 2015.

[5] D. P. Bertsekas and H. Yu, "Stochastic shortest path problems under weak conditions," Lab. Inf. Decision Syst., MIT, Cambridge, MA, USA, Tech. Rep. LIDS-2909, Mar. 2015.

[6] H. Yu and D. P. Bertsekas, "A mixed value and policy iteration method for stochastic control with universally measurable policies," Lab. Inf. Decision Syst., MIT, Cambridge, MA, USA, Tech. Rep. LIDS-2905, Jul. 2013.

[7] D. Blackwell, "Positive dynamic programming," in *Proc. 5th Symp. Math. Statist. Probab.*, Berkeley, CA, USA, 1965, pp. 415–418.

[8] R. E. Strauch, "Negative dynamic programming," *Ann. Math. Statist.*, vol. 37, no. 4, pp. 871–890, Aug. 1966.

[9] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete-Time Case*. New York, NY, USA: Academic, 1978. [Online]. Available: http://web.mit.edu/dimitrib/www/home.html

[10] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: Wiley, 1994.

[11] D. P. Bertsekas, *Dynamic Programming and Optimal Control: Approximate Dynamic Programming*, vol. 2. Belmont, MA, USA: Athena Scientific, Aug. 2012.

[12] D. P. Bertsekas, "Monotone mappings in dynamic programming," in *Proc. 14th IEEE Conf. Decision Control*, Houston, TX, USA, Dec. 1975, pp. 20–25.

[13] D. P. Bertsekas, "Monotone mappings with application in dynamic programming," *SIAM J. Control Optim.*, vol. 15, no. 3, pp. 438–464, 1977.

[14] M. Schäl, "Conditions for optimality in dynamic programming and for the limit of *n*-stage optimal policies to be optimal," *Zeitschrift Wahrscheinlichkeitstheorie Verwandte Gebiete*, vol. 32, no. 3, pp. 179–196, 1975.

[15] D. P. Bertsekas and J. N. Tsitsiklis, "An analysis of stochastic shortest path problems," *Math. Oper. Res.*, vol. 16, no. 3, pp. 580–595, Aug. 1991.

[16] Z. V. Rekasius, "Suboptimal design of intentionally nonlinear controllers," *IEEE Trans. Autom. Control*, vol. 9, no. 4, pp. 380–386, Oct. 1964.

[17] G. N. Saridis and C.-S. G. Lee, "An approximation theory of optimal control for trainable manipulators," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 3, pp. 152–159, Mar. 1979.

[18] P. J. Werbos, "Approximate dynamic programming for real-time control and neural modeling," in *Handbook of Intelligent Control*, D. A. White and D. A. Sofge, Eds. London, U.K.: Multiscience Press, 1992, pp. 493–525.

[19] R. W. Beard, "Improving the closed-loop performance of nonlinear systems," Ph.D. dissertation, Dept. Elect. Eng., Rensselaer Polytech. Inst., Troy, NY, USA, Oct. 1995.

[20] D. Vrabie, K. G. Vamvoudakis, and F. L. Lewis, *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*. London, U.K.: IET, Jan. 2013.

[21] Z.-P. Jiang and Y. Jiang, "Robust adaptive dynamic programming for linear and nonlinear systems: An overview," *Eur. J. Control*, vol. 19, no. 5, pp. 417–425, Sep. 2013.

[22] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic Programming*. Belmont, MA, USA: Athena Scientific, 1996.

[23] R. S. Sutton and A. G. Barto, *Reinforcement Learning*. Cambridge, MA, USA: MIT Press, 1998.

[24] F. L. Lewis, D. Liu, and G. G. Lendaris, "Special issue on adaptive dynamic programming and reinforcement learning in feedback control," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 4, pp. 896–897, Aug. 2008.

[25] Y. Jiang and Z.-P. Jiang, "Robust adaptive dynamic programming and feedback stabilization of nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 882–893, May 2014.

[26] A. Heydari, "Revisiting approximate dynamic programming and its convergence," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2733–2743, Dec. 2014.

[27] A. Heydari. (2014). "Stabilizing value iteration with and without approximation errors." [Online]. Available: http://arxiv.org/abs/1412.5675

[28] D. Liu and Q. Wei, "Finite-approximation-error-based optimal control approach for discrete-time nonlinear systems," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 779–789, Apr. 2013.

[29] Q. Wei, F.-Y. Wang, D. Liu, and X. Yang, "Finite-approximation-error-based discrete-time iterative adaptive dynamic programming," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2820–2833, Dec. 2014.

[30] J. Si, A. Barto, W. Powell, and D. Wunsch, Eds., *Learning and Approximate Dynamic Programming*. New York, NY, USA: Wiley, 2004.

[31] F. L. Lewis and D. Liu, Eds., *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken, NJ, USA: Wiley, 2013.

[32] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," in *Proc. IEEE Amer. Control Conf.*, vol. 3. Jul. 1994, pp. 3475–3479.

[33] D. P. Bertsekas, "Control of uncertain systems with a set-membership description of the uncertainty," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 1971. [Online]. Available: http://web.mit.edu/dimitrib/www/publ.html

[34] D. P. Bertsekas, "Infinite time reachability of state-space regions by using feedback control," *IEEE Trans. Autom. Control*, vol. 17, no. 5, pp. 604–613, Oct. 1972.

Dimitri P. Bertsekas received the Degree in engineering from the National Technical University of Athens, Athens, Greece, the M.S. degree in electrical engineering from George Washington University, Washington, DC, USA, in 1969, and the Ph.D. degree in system science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 1971.

He has held faculty positions with the Engineering Economic Systems Department, Stanford University, Stanford, CA, USA, from 1971 to 1974, and the Electrical Engineering Department, University of Illinois at Urbana–Champaign, Urbana, IL, USA, from 1974 to 1979. Since 1979, he has been teaching at the Electrical Engineering and Computer Science Department, MIT, where he is currently a McAfee Professor of Engineering. He consults regularly with private industry and has held editorial positions in several journals. His research has spanned several fields, including optimization, control, largescale and distributed computation, and data communication networks, and is closely tied to his teaching and book authoring activities. He has written numerous research papers, and sixteen books and research monographs, several of which are used as textbooks in MIT. classes, including *Introduction to Probability* (2008, co-authored with J. Tsitsiklis).

Dr. Bertsekas received the INFORMS 1997 Prize for Research Excellence in the Interface Between Operations Research and Computer Science for his book *NeuroDynamic Programming* (co-authored with J. Tsitsiklis), the 2001 ACC John R. Ragazzini Education Award, the 2009 INFORMS Expository Writing Award, the 2014 ACC Richard E. Bellman Control Heritage Award, the 2014 Khachiyan Prize for LifeTime Accomplishments in Optimization, and the SIAM/MOS 2015 George B. Dantzig Prize. In 2001, he was elected to the U.S. National Academy of Engineering. His recent books are *Dynamic Programming and Optimal Control: 4th Edition* (2012), *Abstract Dynamic Programming* (2013), and *Convex Optimization Algorithms* (2015), all published by Athena Scientific.