May 2015 **Report LIDS-P-3174**

Value and Policy Iteration 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 solution of Bellman's equation, and we provide convergence results for value and policy iteration.

1. INTRODUCTION

In this paper we consider a deterministic discrete-time optimal control problem involving the system

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

where x_k and u_k are the state and control at stage k, lying in sets X and U, respectively. 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 is $g(x_k, u_k)$, and is assumed nonnnegative and real-valued:

$$0 \leq g(x_k, u_k) < \infty, \qquad x_k \in X,\ u_k \in U(x_k). \tag{1.2}$$

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 classical shortest path problems involving a graph, or they may be continuous spaces as in classical problems of control to the origin or some other terminal set.

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

$$J_\pi(x_0) = \lim_{k\to\infty} \sum_{t=0}^{k} g\big(x_t, \mu_t(x_t)\big), \qquad x_0 \in X, \tag{1.3}$$

[the limit exists thanks to the nonnegativity assumption (1.2)]. 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

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

† 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.

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

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

In the context of dynamic programming (DP for short), one hopes to prove that J^* satisfies Bellman's equation:

$$J^*(x) = \inf_{u \in U(x)} \big\{ g(x,u) + J^*\big(f(x,u)\big) \big\}, \qquad \forall\ x \in X, \tag{1.4}$$

and that an optimal stationary policy may be obtained through the minimization in the right side of this equation. One also hopes to obtain J^* by means of value iteration (VI for short), which starting from some function $J_0 : X \mapsto [0, \infty]$, generates a sequence of functions $\{J_k\}$ according to

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

Another possibility to obtain J^* and an optimal policy is through policy iteration (PI for short), which starting from a stationary policy μ^0, generates a sequence of stationary policies $\{\mu^k\}$ via a sequence of policy evaluations to obtain J_{μ^k} from the equation

$$J_{\mu^k}(x) = g\big(x, \mu^k(x)\big) + J_{\mu^k}\big(f\big(x, \mu^k(x)\big)\big), \qquad x \in X, \tag{1.6}$$

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

$$\mu^{k+1}(x) \in \arg\min_{u \in U(x)} \big\{ g(x,u) + J_{\mu^k}\big(f(x,u)\big) \big\}, \qquad x \in X. \tag{1.7}$$

We implicitly assume here is that J_{μ^k} satisfies Eq. (1.6), which is true under the cost nonnegativity assumption (1.2) (cf. Prop. 2.1 in the next section). Also the minimum in Eq. (1.7) should be attained for each $x \in X$, which is true under some compactness condition on either $U(x)$ or the level sets of the function $g(x,\cdot) + J_k\big(f(x,\cdot)\big)$, or both.

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

$$g(x,u) = 0, \qquad x = f(x,u), \qquad \forall\ x \in X_s,\ u \in U(x). \tag{1.8}$$

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

$$X_s = \big\{ x \in X \mid J^*(x) = 0 \big\}. \tag{1.9}$$

In the applications of primary interest, g is 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

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

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 approached with policies that terminate from x. In particular, throughout the paper we assume the following.

Assumption 1.1: The cost nonnegativity condition (1.2) and stopping set conditions (1.8)-(1.9) hold. Moreover, for every $x \in X_f$ and $\epsilon > 0$, there exists a policy π that terminates starting from x and satisfies $J_\pi(x) \leq J^*(x) + \epsilon$.

Specific and easily verifiable conditions that imply this assumption will be given in Section 4. 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 cycles of the state transition graph have positive length, all 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 1.1 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 $x \in X_f$. This is true for example in the classical linear quadratic optimal control problem, where $X = \Re^n$, $U = \Re^m$, g is positive semidefinite quadratic, and f represents a linear system of the form $x_{k+1} = Ax_k + Bu_k$, where A and B are given matrices. However, we will show in Section 4 that Assumption 1.1 is satisfied under more easily verifiable conditions. For example, it is satisfied assuming that g is strictly positive outside X_s in the sense that for each $\delta > 0$ there exists $\epsilon > 0$ such that

$$\inf_{u \in U(x)} g(x,u) \geq \epsilon, \qquad \forall\ x \in X \text{ such that } \mathrm{dist}(x, X_s) \geq \delta,$$

where for all $x \in X$, $\mathrm{dist}(x, X_s)$ denotes the minimum distance from x to X_s,

$$\mathrm{dist}(x, X_s) = \inf_{y \in X_s} \|x - y\|, \qquad x \in X,$$

and also that for every $\epsilon > 0$, there exists a $\delta_\epsilon > 0$ such that for each $x \in X_f$ with $\mathrm{dist}(x, X_s) \leq \delta_\epsilon$, there is a policy π that terminates from x and satisfies $J_\pi(x) \leq \epsilon$. The latter condition is a "local controllability" assumption 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, and can be easily checked in many applications.

Our main results are given in the following three propositions. In our terminology, all equations, inequalities, and convergence limits involving functions are meant to be pointwise. Regarding notation, we denote by $E^+(X)$ the set of all functions $J : X \mapsto [0, \infty]$, and by $\mathcal{J}$ the set of functions

$$\mathcal{J} = \big\{ J \in E^+(X) \mid J(x) = 0,\ \forall\ x \in X_s \big\}. \tag{1.10}$$

Note that in view of Eq. (1.8), $\mathcal{J}$ contains the cost function J_π of all policies π, as well as J^*.

Proposition 1.1: (Uniqueness of Solution of Bellman's Equation) The optimal cost function J^* is the unique solution of Bellman's equation (1.4) within the set of functions $\mathcal{J}$.

Examples where there are additional solutions $\hat{J}$ of Bellman's equation with $\hat{J} \geq J^*$ are well known. Particularly simple two-state shortest path examples of this type are given in [Ber13], Section 3.1.2, and in [BeY15], Example 1.1.

Proposition 1.2: (Convergence of VI)

(a) The VI sequence $\{J_k\}$ generated by Eq. (1.5) starting from any function $J_0 \in \mathcal{J}$ with $J_0 \geq J^*$ converges pointwise to J^*.

(b) Assume further that U is a metric space, and the sets $U_k(x, \lambda)$ given by

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

are compact for all $x \in X$, $\lambda \in \Re$, and k, where $\{J_k\}$ is the VI sequence $\{J_k\}$ generated by Eq. (1.5) starting from $J_0 \equiv 0$. Then the VI sequence $\{J_k\}$ generated by Eq. (1.5) converges pointwise to J^* starting from any function $J_0 \in \mathcal{J}$.

Easily verifiable assumptions implying the compactness assumption of part (b) above will be given later. Note that when there are solutions to Bellman's equations in addition to J^*, VI will not converge to J^* starting from any of these solutions. However, it is possible that Bellman's equation has J^* as its unique solution within the set of nonnegative functions, and yet VI does not converge to J^* starting from the zero function because the compactness condition of Prop. 1.2(b) is violated (there are several examples of this type in the literature, and Example 4.3.3 of [Ber13] is a deterministic problem for which Assumption 1.1 is satisfied).

Proposition 1.3: (Convergence of PI) The sequence $\{J_{\mu^k}\}$ generated by the PI algorithm (1.6), (1.7), 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 (1.7) can be carried out for every $x \in X$. Easily verifiable conditions that guarantee this also guarantee the compactness condition of Prop. 1.2(b), and will be noted following Prop. 2.1 in the next section. Moreover, in Section 4 we will prove a similar convergence result for a variant of the PI algorithm where the policy evaluation is carried out approximately through a finite number of VIs. There are simple two-state shortest path examples where the PI sequence J_{μ^k} does not converge to J^* if Assumption 1.1 is violated (see, e.g., [Ber13], Section 3.1.2, or [BeY15], Example 1.1).

The paper is organized as follows. In Section 2 we provide background and references, which place in context our results and methods of analysis in relation to the literature. In Section 3 we give the proofs of Props. 1.1-1.3. In Section 4 we discuss special cases and easily verifiable conditions that imply our assumptions, and we provide extensions of our analysis.

We finally note that the ideas of this paper stem from more general ideas regarding the convergence of VI, which were developed in the context of abstract DP; see the recent book [Ber13]. The paper [Ber15] views the preceding propositions as special cases of more general abstract DP results.

2. BACKGROUND

The issues discussed in this paper have received attention since the 60's, originally in the work of Blackwell [Bla65], who considered the case $g \leq 0$, and the work by Strauch (Blackwell's PhD student) [Str66], who considered the case $g \geq 0$. For textbook accounts we refer to [BeS78], [Put74], [Ber12], and for a more abstract account, [Ber13]. These works showed that the cases where $g \leq 0$ (which corresponds to 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 for this to be guaranteed [see part (d) of the following proposition]. Moreover for the case $g \geq 0$, Bellman's equation may have solutions $\tilde{J} \neq J^*$ with $\tilde{J} \geq J^*$, and VI will not converge to J^* starting from such $\tilde{J}$. In addition it is well-known that in general, PI need not converge to J^* and may instead stop with a suboptimal policy (see for instance [BeY15], Example 1.1). The following proposition gives the standard results for our problem when $g \geq 0$ (see [BeS78], Props. 5.2, 5.4, and 5.10, [Ber12], Props. 4.1.1, 4.1.3, 4.1.5, 4.1.9, or [Ber13], Props. 4.3.3, 4.3.9, and 4.3.14).

Proposition 2.1:

(a) J^* satisfies Bellman's equation (1.4), and if $J \in E^+(X)$ is another solution, i.e., J satisfies

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

then $J^* \leq J$.

(b) For all stationary policies μ we have

$$J_\mu(x) = g\big(x, \mu(x)\big) + J_\mu\big(f\big(x, \mu(x)\big)\big), \qquad \forall\ x \in X. \tag{2.2}$$

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

$$\mu^*(x) \in \arg\min_{u \in U(x)} \big\{ g(x,u) + J^*\big(f(x,u)\big) \big\}, \qquad \forall\ x \in X. \tag{2.3}$$

(d) If U is a metric space and the sets

$$U_k(x, \lambda) = \big\{ u \in U(x) \mid g(x,u) + J_k\big(f(x,u)\big) \leq \lambda \big\} \tag{2.4}$$

are compact for all $x \in X$, $\lambda \in \Re$, and k, where $\{J_k\}$ is the sequence generated by the VI (1.5) starting from $J_0 \equiv 0$, then there exists at least one optimal stationary policy, and we have $J_k \to J^*$ for all $J \in E^+(X)$ with $J \leq J^*$.

Actually, only the assumption $g \geq 0$ is needed for the preceding proposition, and the other parts of Assumption 1.1 are not necessary. Compactness assumptions such as the one of part (d) above, were originally given in [Ber75], [Ber77], and in [Sch75]. They have been used in several other works, such as

[BeT91], [Ber12], Prop. 4.1.9. In particular, the condition of part (d) holds when $U(x)$ is a finite set for all $x \in X$. The condition of part (d) also holds when $X = \Re^n$, and for each $x \in X$, the set

$$\big\{u \in U(x) \mid g(x,u) \leq \lambda\big\}$$

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 from [YuB13], where it was shown that J^* is the unique fixed point of T within the class of all functions $J \in E^+(X)$ that satisfy

$$0 \leq J \leq cJ^* \qquad \text{for some } c > 0, \tag{2.5}$$

(we refer to [YuB13] 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

$$J^* \leq J \leq cJ^* \qquad \text{for some } c > 0,$$

and under the compactness conditions of Prop. 2.1(d), starting from any J that satisfies Eq. (2.5). The same paper and a related paper [BeY15] discuss extensively PI algorithms for stochastic nonnegative cost problems.

The results just noted for infinite horizon DP problems with nonnegative cost per stage have been shown in a stochastic setting, which does not take into account the favorable structure of deterministic problems or the presence of the stopping set X_s. For deterministic problems, there has been substantial research in the adaptive dynamic programming literature, regarding the validity of Bellman's equation and the uniqueness of its solution, as well as the attendant questions of convergence of value and policy iteration.

In particular, 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 the present paper involve substantial technical difficulties, since the analog of the (discrete-time) Bellman equation (1.4) 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 time-optimal control of second order linear systems to the origin. The analog of value iteration for continuous-time systems essentially involves the time integration of this equation, and its analysis must deal with difficult issues of stability and convergence to a steady-state solution. Nonetheless there have been proposals of continuous-time PI algorithms, in the early papers [Rek64], [Kle68], [SaL79], [Wer92], and the thesis [Bea95], as well as more recently in several works; see e.g., the book [VVL13], the survey [JiJ13], and the references quoted there. These works also address the possibility of value function approximation, similar to other approximation-oriented methodologies such as neurodynamic programming [BeT96] and reinforcement learning [SuB98], which consider primarily discrete-time systems. For example, among the restrictions of the PI method, is that it must be started with a stabilizing controller; see for example the paper [Kle68], 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 the present paper; questions of stability are addressed only indirectly through the finiteness of J^* and Assumption 1.1.

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 [VVL13], the papers [JiJ14], [Hey14a], [Hey14b], [LiW13], [WWL14], the survey papers in the edited volumes [SBP04] and [LeL13], and the special issue [LLL08]. 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 Euclidean spaces, as well as continuity and other conditions on g, special structure of the system, etc. It is beyond our scope to provide a detailed survey of the state-of-the-art of the VI and PI methodology 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 Props. 1.1-1.3. Of course, these works also address questions that we do not, such as issues of stability of the obtained controllers, the use of approximations, etc. Thus the results of the present work 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 the present paper, which is based on general results of Markovian decision problem theory and abstract forms of dynamic programming, 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.

3. PROOFS OF THE MAIN RESULTS

Let us denote for all $x \in X$,

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

and note the following key implication of Assumption 1.1:

$$J^*(x) = \inf_{\pi \in \Pi_{T,x}} J_\pi(x), \qquad \forall\ x \in X_f. \tag{3.1}$$

Proof of Prop. 1.1: Let $\hat{J} \in \mathcal{J}$ be a solution of the Bellman equation (2.1), so that

$$\hat{J}(x) \le g(x,u) + \hat{J}\big(f(x,u)\big), \qquad \forall\ x \in X,\ u \in U(x), \tag{3.2}$$

while by Prop. 2.1(a), $J^* \le \hat{J}$. 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 Eq. (3.2),

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

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

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

By combining the last two relations, we obtain

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

Taking the infimum over $\pi \in \Pi_{T,x_0}$ and using Eq. (3.1), it follows that $J^*(x_0) = \hat{J}(x_0)$ for all $x_0 \in X_f$. Since for $x_0 \notin X_f$, we have $J^*(x_0) = \hat{J}(x_0) = \infty$ [since $J^* \leq \hat{J}$ by Prop. 2.1(a)], we obtain $J^* = \hat{J}$. **Q.E.D.**

Proof of Prop. 1.2: (a) Starting with $J_0 \geq J^*$, let us apply the VI operation to both sides of this inequality. 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 have $J_k \in \mathcal{J}$ for all k. Hence for every $x_0 \in X_f$ and policy $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi_{T,x_0}$, we have

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

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 $J_0(x_k) = 0$ for all sufficiently large k, we have

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

By combining the last two relations, we obtain

$$J^*(x_0) \leq \liminf_{k\to\infty} J_k(x_0) \leq \limsup_{k\to\infty} J_k(x_0) \leq J_\pi(x_0), \qquad \forall\ x_0 \in X_f,\ \pi \in \Pi_{T,x_0}.$$

Taking the infimum over $\pi \in \Pi_{T,x_0}$ and using Eq. (3.1), 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 $J_k \to J^*$.

(b) The VI iterates starting from any function $J \in \mathcal{J}$ lie between the VI iterates starting from the zero function and the VI iterates starting from $J_0 = \max\{J, J^*\}$. Both of these latter iterates converge to J^* by part (a) and Prop. 2.1(d). **Q.E.D.**

Proof of Prop. 1.3: If μ is a stationary policy and $\bar{\mu}$ satisfies the policy improvement equation

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

[cf. Eq. (1.7)], we have for all $x \in X$,

$$J_\mu(x) = g\big(x, \mu(x)\big) + J_\mu\big(f\big(x, \mu(x)\big)\big) \geq \min_{u\in U(x)} \big\{ g(x,u) + J_\mu\big(f(x,u)\big) \big\} = g\big(x, \bar{\mu}(x)\big) + J_\mu\big(f\big(x, \bar{\mu}(x)\big)\big), \quad (3.3)$$

where the first equality follows from the definition of J_μ and the second equality follows from the definition of $\bar{\mu}$. Repeatedly applying this relation, we see that the sequence $\big\{\tilde{J}_k(x_0)\big\}$ defined by

$$\tilde{J}_k(x_0) = J_\mu(x_k) + \sum_{t=0}^{k-1} g\big(x_t, \bar{\mu}(x_t)\big), \qquad k = 1, 2, \ldots,$$

is monotonically nonincreasing, where $\{x_k\}$ is the sequence generated starting from x_0 and using μ. Moreover, from Eq. (3.3) we have

$$J_\mu(x_0) \geq \min_{u\in U(x_0)} \big\{ g(x,u) + J_\mu\big(f(x,u)\big) \big\} = \tilde{J}_1(x_0) \geq \tilde{J}_k(x_0),$$

for all k. This implies that

$$J_\mu(x_0) \geq \min_{u \in U(x_0)} \big\{ g(x,u) + J_\mu\big(f(x,u)\big) \big\} \geq \lim_{k\to\infty} \tilde{J}_k(x_0) \geq \lim_{k\to\infty} \sum_{t=0}^{k-1} g\big(x_t, \bar{\mu}(x_t)\big) = J_{\bar{\mu}}(x_0),$$

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

$$J_\mu(x) \geq \inf_{u \in U(x)} \big\{ g(x,u) + J_\mu\big(f(x,u)\big) \big\} \geq J_{\bar{\mu}}(x), \qquad x \in X. \tag{3.4}$$

Using μ^k and $\bar{\mu}^k$ in place of μ and $\bar{\mu}$, we see that the sequence $\{J_{\mu^k}\}$ generated by PI converges monotonically to some function $J_\infty \in E^+(X)$, i.e., $J_{\mu^k} \downarrow J_\infty$. Moreover, from Eq. (3.4) we have

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

as well as

$$g(x,u) + J_{\mu_k}\big(f(x,u)\big) \geq J_\infty(x), \qquad x \in X,\ u \in U(x).$$

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

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

Thus J_∞ is a solution of Bellman's equation, satisfying $J_\infty \geq J^*$ (since $J_{\mu^k} \geq J^*$ for all k) and $J_\infty \in \mathcal{J}$ (since $J_{\mu^k} \in \mathcal{J}$), so by Prop. 2.1(a), it must satisfy $J_\infty = J^*$. **Q.E.D.**

4. 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. Consider first Assumption 1.1. As noted in Section 1, it holds when X and U are finite, a terminating policy exists from every x, and all 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 $\|\cdot\|$, 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

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

The following proposition provides readily verifiable conditions that guarantee Assumption 1.1.

Proposition 4.1: Let the following two conditions hold.

(1) For every $x \in X_f$, any policy π with $J_\pi(x) < \infty$ asymptotically terminates from x.

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

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

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

Then Assumption 1.1 holds.

Proof: Fix $x \in X_f$ and $\epsilon > 0$. Condition (1) guarantees that for any fixed $x \in X_f$ and $\epsilon > 0$, there exists a policy π that asymptotically terminates from x, and satisfies

$$J_\pi(x) \leq J^*(x) + \epsilon/2.$$

Starting from x, this policy will generate a sequence $\{x_k\}$ such that for some index $\bar{k}$ we have

$$\lim_{k\to\infty} \text{dist}(x_{\bar{k}}, X_s) \leq \delta_{\epsilon/2},$$

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

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

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

Cost functions for which condition (1) of the preceding proposition 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

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

Then for any x and policy π that does not asymptotically terminate from x, we will have $J_\pi(x) = \infty$. From an applications point of view, the condition is natural and consistent with the aim of steering the state towards 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. As an example, condition (2) is satisfied when $X_s = \{0\}$ and the following hold:

(a) $X = \Re^n$, $U = \Re^m$, and there is an open sphere R centered at the origin such that $U(x)$ contains R for all $x \in X$.

(b) f represents a controllable linear system of the form

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

where A and B are given matrices.

(c) g satisfies

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

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

There are straightforward extensions of the preceding conditions 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.

An Optimistic Form of PI

Let us consider a variant of PI where 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

$$\mu^k(x) \in \arg\min_{u \in U(x)} \big\{g(x,u) + J_k\big(f(x,u)\big)\big\}, \qquad x \in X,\ k = 0, 1, \ldots, \tag{4.1}$$

and then we obtain J_{k+1} with $m_k \geq 1$ value iterations using μ_k:

$$J_{k+1}(x_0) = J_k(x_{m_k}) + \sum_{t=0}^{m_k-1} g\big(x_t, \mu^k(x_t)\big), \qquad x_0 \in X, \tag{4.2}$$

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

$$J_0(x) \geq \inf_{u \in U(x)} \big\{g(x,u) + J_0\big(f(x,u)\big)\big\}, \qquad \forall\ x \in X,\ u \in U(x). \tag{4.3}$$

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 [Put94] and [Ber12] for discussions of this type of inexact PI algorithm (in [Put91] it is called "modified" PI, while in [Ber12] it is called "optimistic" PI).

Proposition 4.2: (Convergence of Optimistic PI) For the PI algorithm (4.1)-(4.2), where J_0 belongs to $\mathcal{J}$ and satisfies the condition (4.3), we have $J_k \downarrow J^*$.

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

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

where the first inequality is the condition (4.3), the second and third inequalities follow because of the monotonicity of the m_0 value iterations (4.2) for μ^0, and the fourth inequality follows from the policy improvement equation (4.1). Continuing similarly, we have

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

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 Prop. 1.3, we can show that J_∞ is a solution of Bellman's equation. Moreover, an induction proof shows that $J_k \geq J^*$, so that $J_\infty \geq J^*$ while $J\infty \in \mathcal{J}$. The result now follows similar to the case of the standard PI algorithm (cf. Prop. 1.3). **Q.E.D.**

Minimax Control to a Terminal Set of States

There is a straightforward extension of our analysis to minimax problems with a terminal set of states. Here the system is

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

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:

$$g(x, u, w) \geq 0, \qquad x \in X,\ U \in U(x),\ w \in W.$$

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

$$J_\pi(x_0) = \sup_{\substack{w_k \in W \\ k=0,1,\ldots}} \lim_{k\to\infty} \sum_{t=0}^{k} g\big(x_k, \mu_k(x_k), w_k\big),$$

where $\big\{x_k, \mu_k(x_k)\big\}$ 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, which is cost-free and absorbing,

$$g(x, u, w) = 0, \qquad f(x, u, w) \in X_s, \quad \forall\ x \in X_s,\ u \in U(x),\ w \in W,$$

and that all states outside X_s have strictly positive optimal cost, so that

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

The finite-state version of this problem has been discussed in [Ber14], under the name *robust shortest path planning*, for the case where g can take both positive and negative values. Another special case is the problem of *reachability of a target set*, which is obtained for

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

The objective in this problem is to reach the set X_s in the minimum guaranteed number of steps. The set X_f here is the set of states for which X_s is guaranteed to be reached in a finite number of steps. A related problem is the problem of reachability of a target tube, where for a given set $\hat{X}$,

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

and the objective is to find the initial states for which we can guarantee to keep all future states within $\hat{X}$ (note that the analysis of the present paper can be adapted to the case where g can take the value ∞, although for simplicity we have not done so). These two reachability problems were the subject of the author's Ph.D. thesis research [Ber71], and the subsequent papers [BeR71], [Ber72]. In fact the reachability

algorithms given in these works are simply the VI algorithm of the present paper, starting with appropriate initial functions J_0.

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 1.1 is modified to reflect this new definition of termination, and our results can be readily extended, with the Props. 1.1, 1.2, 1.3, and 4.2, and their proofs holding essentially as stated.

5. CONCLUDING REMARKS

The analysis of this paper considers deterministic optimal control problems under very general assumptions, including the possibilities of arbitrary state and control spaces, and infinite optimal cost from some initial states. The analysis of VI and PI even when there are infinite optimal cost states is unusual, and bypasses the need for assumptions involving the existence of globally stabilizing controllers that guarantee that the optimal cost function J^* is real-valued. This generality makes our results a convenient starting point for analysis of problems involving additional assumptions, and perhaps cost function approximations.

6. REFERENCES

[BeR71] Bertsekas, D. P., and Rhodes, I. B., 1971. "On the Minimax Reachability of Target Sets and Target Tubes," Automatica, Vol. 7, pp. 233-241.

[BeS78] Bertsekas, D. P., and Shreve, S. E., 1978. Stochastic Optimal Control: The Discrete Time Case, Academic Press, N. Y.; 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. Operations Research, Vol. 16, pp. 580-595.

[BeT96] Bertsekas, D. P., and Tsitsiklis, J. N., 1996. Neuro-Dynamic Programming, Athena Scientific, Belmont, MA.

[BeY15] Bertsekas, D. P., and Yu, H., 2015. "Stochastic Shortest Path Problems Under Weak Conditions," Lab. for Information and Decision Systems Report LIDS-2909, revision of March 2015.

[Ber71] Bertsekas, D. P., 1971. "Control of Uncertain Systems With a Set-Membership Description of the Uncertainty," Ph.D. Thesis, Dept. of EECS, MIT; may be downloaded from http://web.mit.edu/dimitrib/www/publ.html.

[Ber72] Bertsekas, D. P., 1972. "Infinite Time Reachability of State Space Regions by Using Feedback Control," IEEE Trans. Aut. Control, Vol. AC-17, pp. 604-613.

[Ber75] Bertsekas, D. P., 1975. "Monotone Mappings in Dynamic Programming," Proc. 1975 IEEE Conference on Decision and Control, Houston, TX, pp. 20-25.

[Ber77] Bertsekas, D. P., 1977. "Monotone Mappings with Application in Dynamic Programming," SIAM J. on Control and Optimization, Vol. 15, pp. 438-464.

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

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

[Ber15] Bertsekas, D. P., 2015. "Regular Policies in Abstract Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-3173, April 2015.

[Bla65] Blackwell, D., 1965. "Positive Dynamic Programming," Proc. Fifth Berkeley Symposium Math. Statistics and Probability, pp. 415-418.

[Hey14a] Heydari, A., 2014. "Revisiting Approximate Dynamic Programming and its Convergence," IEEE Transactions on Cybernetics, Vol. 44, pp. 2733-2743.

[Hey14b] Heydari, A., 2014. "Stabilizing Value Iteration With and Without Approximation Errors, available at arXiv:1412.5675.

[JiJ13] Jiang, Y., and Jiang, Z. P., 2013. "Robust Adaptive Dynamic Programming for Linear and Nonlinear Systems: An Overview, Eur. J. Control, Vol. 19, pp. 417425.

[JiJ14] Jiang, Y., and Jiang, Z. P., 2014. "Robust Adaptive Dynamic Programming and Feedback Stabilization of Nonlinear Systems, IEEE Trans. on Neural Networks and Learning Systems, Vol. 25, pp. 882-893.

[LLL08] Lewis, F. L., Liu, D., and Lendaris, G. G., 2008. Special Issue on Adaptive Dynamic Programming and Reinforcement Learning in Feedback Control, IEEE Trans. on Systems, Man, and Cybernetics, Part B, Vol. 38, No. 4.

[LeL13] Lewis, F. L., and Liu, D., (Eds), 2013. Reinforcement Learning and Approximate Dynamic Programming for Feedback Control, Wiley, Hoboken, N. J.

[LiW13] Liu, D., and Wei, Q., 2013. "Finite-Approximation-Error-Based Optimal Control Approach for Discrete-Time Nonlinear Systems, IEEE Transactions on Cybernetics, Vol. 43, pp. 779-789.

[Kle68] Kleinman, D. L., 1968. "On an Iterative Technique for Riccati Equation Computations," IEEE Trans. Aut. Control, Vol. AC-13, pp. 114-115.

[PaB99] Patek, S. D., and Bertsekas, D. P., 1999. "Stochastic Shortest Path Games," SIAM J. on Control and Opt., Vol. 36, pp. 804-824.

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

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

[RDK09] Rinehart, M., Dahleh, M., and Kolmanovsky, I., 2009. "Value Iteration for (Switched) Homogeneous Systems," IEEE Transactions on Automatic Control, Vol. 54, pp. 1290-1294.

[Ran06] Rantzer, A., 2006. "Relaxed Dynamic Programming in Switching Systems, Proc. Inst. Elect. Eng., Vol. 153, pp. 567-574.

[SBP04] Si, J., Barto, A., Powell, W., and Wunsch, D., (Eds.) 2004. Learning and Approximate Dynamic Programming, IEEE Press, N. Y.

[Sch75] Schal, M., 1975. "Conditions for Optimality in Dynamic Programming and for the Limit of n-Stage Optimal Policies to be Optimal," Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, Vol. 32, pp. 179-196.

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

[SuB98] Sutton, R. S., and Barto, A. G., 1998. Reinforcement Learning, MIT Press, Cambridge, MA.

[VVL13] Vrabie, D., Vamvoudakis, K. G., and Lewis, F. L., 2013. Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles, The Institution of Engineering and Technology, London.

[WWL14] Wei, Q., Wang, F. Y., Liu, D., and Yang, X., 2014. "Finite-Approximation-Error-Based Discrete-Time Iterative Adaptive Dynamic Programming," IEEE Transactions on Cybernetics, Vol. 44, pp. 2820-2833.

[Wer92] Werbos, P. J., 1992. "Approximate Dynamic Programming for Real-Time Control and Neural Modeling, in Handbook of Intelligent Control (D. A. White and D. A. Sofge, eds.), Multiscience Press.

[YuB13] Yu, H., and Bertsekas, D. P., "A Mixed Value and Policy Iteration Method for Stochastic Control with Universally Measurable Policies," Lab. for Information and Decision Systems Report LIDS-P-2905, MIT, July 2013; to appear in Math. of OR.