KYBERNETIKA — VOLUME *28* (1992), NUMBER 3, PAGES 191 – 212

DISCRETE–TIME MARKOV CONTROL PROCESSES WITH DISCOUNTED UNBOUNDED COSTS: OPTIMALITY CRITERIA

Onésimo Hernández–Lerma and Myriam Muñoz de Ozak

We consider discrete–time Markov control processes with *Borel* state and control spaces, *unbounded* costs per stage, and *not necessarily compact* control constraint sets. The basic control problem we are concerned with is to minimize the infinite–horizon, expected total *discounted* cost. Under easily verifiable assumptions, we provide characterizations of the optimal cost function and optimal policies, including all previously known optimality criteria, such as Bellman's Principle of Optimality, and the martingale and discrepancy function criteria. The convergence of value iteration, policy iteration and other approximation procedures is also discussed, together with criteria for asymptotic optimality.

1. INTRODUCTION

This paper deals with discrete–time Markov control processes (or MCPs for short) with Borel state and control spaces. The basic optimal control problem (formalized in § 3) is to minimize the total expected *discounted* cost. Given that the cost–per–stage function is *unbounded*, and that the control constraint sets are *not necessarily compact,* the main questions we are concerned with are:

1. If V^* denotes the optimal (i. e., minimum) cost function, what are the conditions for V^* to be a solution to the optimality equation (OE)? (See equations (3.4) and (4.1).)
2. If v is a function that satisfies the OE, how are v and V^* related?
3. How can we "approximate" V^*?
4. What are the conditions for a control policy to be optimal? In other words, may we characterize an optimal control policy?
5. Is it possible to decide when a control policy is "close" to being optimal?

All these questions have been dealt with in the literature, in one form or other, but usually separately, and under very restrictive conditions (such as conditions C_0, C_1 and C_2 in § 4), which exclude some important control problems – e. g. the "linear regulator" problem, which has (quadratic) *unbounded* costs and an *unbounded* control set (see Example 2.5). Thus our *main objective* in this paper is to study questions 1 to 5 from a unified viewpoint, under a set of easily verifiable assumptions that includes – to the best of our knowledge – virtually *all* the previous works on MCPs with *Borel* state and control spaces *and unbounded* costs–per–stage.

We begin with some preliminaries in §§ 2 and 3: § 2 discusses the basic Markov control (or decision) model we will be dealing with, and § 3 introduces the corresponding control problem. The main developments are presented in §§ 4 to 7. In § 4 we discuss the optimality equation and provide some answers to questions 1 to 4 above. § 5 is mainly concerned with question 3, whereas § 6 is mainly related to

question 4; the main result in that section (Theorem 6.1) relates several well–known optimality criteria, including Bellman's principle of optimality, and a martingale criterion. Finally, in § 7 an answer to question 5 is given in terms of "asymptotic optimality".

Related literature. The stochastic control problem we are interested in is quite standard – see any of the textbooks in the references –; but the studies on questions 1 to 5 appear scattered in the literatures on stochastic control, operations research, and applied probability. Thus there is no "main reference" for §§ 4 to 7 and, therefore, each of these sections is provided with its own set of Comments and related references.

Notation. Given a *Borel space,* i. e., a Borel subset of a complete separable metric space, its Borel sigma–algebra is denoted by $\mathcal{B}(X)$, and "measurable" always means "Borel–measurable". $L(X)$ stands for the family of l.s.c. (lower semicontinuous) functions on X, bounded from below, and $L(X)_+$ denotes the subclass of nonnegative functions in $L(X)$.

2. THE CONTROL MODEL

Let (X, A, Q, c) be a Markov control (or decision) model with state space X, control (or action) set A, transition law Q, and cost–per–stage c satisfying the following conditions. Both X and A are Borel spaces. To each $x \in X$ it is associated a non–empty set $A(x) \in \mathcal{B}(A)$ whose elements are the feasible control actions when the system is in the state x. The set

$$\mathbb{K} := \{(x, a) \,|\, x \in X, a \in A(x)\} \tag{2.1}$$

of admissible state–action pairs is assumed to be a Borel subset of $X \times A$. The transition law $Q(B \,|\, x, a)$, where $B \in \mathcal{B}(X)$ and $(x, a) \in \mathbb{K}$ is a stochastic kernel on X given $\mathbb{K}$ [3], [11]; that is, for each pair $(x, a) \in \mathbb{K}$, $Q(\cdot \,|\, x, a)$ is a probability measure on X, and for each $B \in \mathcal{B}(X)$, $Q(B \,|\, \cdot)$ is a measurable function on $\mathbb{K}$. Finally the cost–per–stage $c(x, a)$ is a measurable function on $\mathbb{K}$ bounded from below. In fact, without loss of generality, we will assume that c is *nonnegative*. To state one of main hypotheses (Assumption 2.1 (a) below) we require the following *definition*: A real–valued function v on $\mathbb{K}$ is said to be *inf–compact on* $\mathbb{K}$ if the set

$$\{a \in A(x) \,|\, v(x, a) \leq r\} \quad \text{is compact} \tag{2.2}$$

for every $x \in X$ and $r \in \mathbb{R}$. (For instance, if the sets $A(x)$ are compact and $v(x, a)$ is lower semicontinuous (l.s.c.) in $a \in A(x)$ for every $x \in X$, then v is inf–compact on $\mathbb{K}$. Conversely, if v is inf–compact on $\mathbb{K}$, then v is l.s.c. in $a \in A(x)$ for every $x \in X$.)

Assumption 2.1. (a) $c(x, a)$ is nonnegative, lower semicontinuous (l.s.c.) and inf–compact on $\mathbb{K}$;

(b) The transition law Q is *weakly continuous*; i. e. for any continuous and bounded function u on X, the map $(x, a) \to \int_X u(y) Q(dy \,|\, x, a)$ is continuous on $\mathbb{K}$;

(c) The multifunction (or set–valued map) $x \to A(x)$ is lower semicontinuous (l.s.c); that is, if $x_n \to x$ in X and $a \in A(x)$, then there are $a_n \in A(x_n)$ such that $a_n \to a$.

In the remainder of this section we will briefly discuss important facts related to Assumption 2.1.

Remark 2.2. Let $L(X)$ be the class of all functions on X that are l.s.c. and bounded from below. A function v belongs to $L(X)$ if and only if there is a sequence of continuous and bounded functions u_n on X such that $u_n \uparrow v$. Using this fact one can easily verify that *Assumption 2.1 (b) is equivalent to*: For any $v \in L(X)$, the map $(x,a) \to \int v(y)Q(dy\,|\,x,a)$ is l.s.c. and bounded from below on $\mathbb{K}$.

Example 2.3. Consider a stochastic control system of the form

$$x_{t+1} = F(x_t, a_t, \xi_t), \quad t = 0, 1, \ldots, \tag{2.3}$$

where $\{\xi_t\}$ is a sequence of independent and identically distributed (i. i. d) random vectors with valued in a Borel space S. In (2.3), $x_t \in X$ and $a_t \in A(x_t)$ denote the state of the system and the control variable at time t, respectively, and F is a given measurable function from $\mathbb{K} \times S$ to X. Denoting by μ the common distribution of the disturbances ξ_t, the transition law of the system can be written as

$$Q(B\,|\,x,a) = \int_S I_B[F(x,a,s)]\,\mu(ds), \quad B \in \mathcal{B}(X),$$

where I_B denotes the indicator function of B. It is then clear that if $(x,a) \to F(x,a,s)$ is continuous on $\mathbb{K}$ for every $s \in S$, then Assumption 2.1 (b) holds.

Example 2.4. Assumption 2.1 (c) holds if, e. g., $\mathbb{K}$ is convex (cf. [17, Lemma 3.2]). In turn, the latter convexity condition holds in many applied control problems: inventory/production systems, water resources management, etc.; see [1, 2, 9, 11].

Example 2.5. (*The linear regulator problem.*) Instead of (2.3), consider the stochastic linear system

$$x_{t+1} = \gamma x_t + \beta a_t + \xi_t, \tag{2.4}$$

with $X = S = \mathbb{R}^n$, $A \equiv A(\cdot) = \mathbb{R}^m$; γ and β are matrices of appropriate dimensions. By the Examples 2.3 and 2.4, it is clear that the Assumptions 2.1 (b) and (c) are satisfied in this case. Moreover, the quadratic cost $c(x,a) = x'px + a'qa$ (where "prime" denotes transpose) satisfies Assumption 2.1 (a) if p and q are nonnegative and positive definite, respectively. For other specific control systems satisfying Assumptions 2.1, see e. g., the references in Example 2.4.

Definition 2.6. $\mathbb{F}$ denotes the family of measurable functions f from X to A such $f(x) \in A(x)$ for all $x \in X$.

The following lemma summarizes some important facts to be used in later sections.

Lemma 2.7. (a) If Assumption 2.1 (c) holds and v is inf–compact (cf. (2.2)), l.s.c. and bounded from below on $\mathbb{K}$, then the function $v^*(x) := \inf_{a \in A(x)} v(x,a)$ belongs to $L(X)$ and, furthermore, there is a function $f \in \mathbb{F}$ such that

$$v^*(x) = v(x, f(x)) \qquad \forall\, x \in X.$$

(b) If the Assumptions 2.1 (a), (b) and (c) hold, and $u \in L(X)$ is nonnegative, then the (nonnegative) function

$$u^*(x) := \inf_{a \in A(x)} \left[c(x,a) + \int_X u(y)Q(dy\,|\,x,a) \right]$$

belongs to $L(X)$, and there exists $f \in \mathbb{F}$ such that

$$u^*(x) = c(x, f(x)) + \int u(y)Q(dy\,|\,x, f(x)) \quad \forall\, x \in X.$$

(c) For each $n = 0, 1, \ldots$, let v_n be a l.s.c. function, bounded from below and inf–compact on $\mathbb{K}$. If $v_n \uparrow v_0$, then

$$\lim_{n\to\infty} \inf_{a\in A(x)} v_n(x,a) = \inf_{a\in A(x)} v_0(x,a) \quad \forall\, x \in X.$$

P r o o f . Part (a) is Lemma 3.2 (f) in [17].

(b) By Remark 2.2 and Assumption 2.1 (a), if $u \in L(X)$ is nonnegative, then

$$v(x,a) := c(x,a) + \int u(y)Q(dy \,|\, x,a)$$

is nonnegative, l.s.c. and inf–compact on $\mathbb{K}$. (Note that $u \geq 0$ implies that $\{a \in A(x) \,|\, v(x,a) \leq r\}$ is a closed subset of the compact set $\{a \in A(x) \,|\, c(x,a) \leq r\}$.) Thus (b) follows from part (a).

(c) Let us define, for $x \in X$,

$$l(x) := \lim_{n\to\infty} \inf_{a\in A(x)} v_n(x,a), \quad \text{and} \quad v_0^*(x) := \inf_{a\in A(x)} v_0(x,a).$$

Clearly, $l(x) \leq v_0^*(x)$. To prove the reverse inequality, fix an arbitrary $x \in X$, and for each $n \geq 0$, let (cf. (2.2))

$$A_n := \{a \in A(x) \,|\, v_n(x,a) \leq v_0^*(x)\}.$$

The inf–compactness hypothesis, together with $v_n \uparrow v_0$, implies that the A_n are compact sets such that $A_n \downarrow A_0$. On the other hand, by part (a), for each $n \geq 1$, there is $a_n \in A_n$ such that $v_n(x,a_n) = \inf_{a\in A(x)} v_n(x,a)$. Thus there exists a subsequence $\{a_{n_i}\}$ of $\{a_n\}$ and $a_0 \in A_0$ such that $a_{n_i} \to a_0$. Now, using again that v_n is monotone increasing, we have

$$v_{n_i}(x, a_{n_i}) \geq v_n(x, a_{n_i}) \qquad \forall\ n_i \geq n,$$

for any given $n \geq 1$. Letting $i \to \infty$, the lower semicontinuity assumption yields

$$l(x) \geq v_n(x, a_0).$$

This implies $l(x) \geq v_0(x,a_0) = v_0^*(x)$, for $v_n \uparrow v_0$. Since $x \in X$ was arbitrary, this completes the proof. □

Comments. It is worth noting that the main difference between our present assumptions and those in the previous literature lies in the inf–compactness in Assumption 2.1 (a) and the l.s.c. in Assumption 2.1 (c). Inf–compactness, allows *non–compact* constraint sets $A(x)$, but still it allows to use "compactness–like" arguments, as in the proof of Lemma 2.7 (c). Assumption 2.1 (c), on the other hand, is used to show that "minimal" functions, such as v^* and u^* in Lemma 2.7, are *lower semicontinuous*; without such an assumption, we can only ensure that v^* and u^* are measurable (cf. [17, Lemma 3.2], [27, Corollary 4.3]).

3. THE CONTROL PROBLEM

Let x_t and a_t denote, respectively, the state of the system and the control action applied at time $t = 0, 1, \ldots$. A rule to choose the control action a_t at each time t is called a control policy and is formally defined as follows.

A *control policy* π is a sequence $\{\pi_t\}$ such that for each $t = 0, 1, \ldots, \pi_t(\cdot \,|\, h_t)$ is a conditional probability on $\mathcal{B}(A)$, given the history $h_t := (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t)$,

that satisfies the constraint $\pi_t\,(A(x_t)\,|\,h_t) = 1$. The class of all policies is denoted by $\prod$.

Let $\mathbb{F}$ be the class of functions in Definition 2.6. A sequence $\{f_t\}$ of functions $f_t \in \mathbb{F}$ is called a *Markov policy*. A Markov policy $\{f_t\}$ is said to be a *stationary policy* if it is of the form $f_t = f$ for all $t = 0, 1, \ldots$ for some $f \in \mathbb{F}$; in this case we identify $\{f_t\}$ with $f \in \mathbb{F}$.

Let $(\Omega, \mathcal{F})$ be the measurable space consisting of the sample space $\Omega := X \times A \times X \times A \times \ldots$, and the corresponding product sigma–algebra $\mathcal{F}$. Then for an arbitrary policy $\pi \in \prod$ and (initial) state $x \in X$, a standard argument using a theorem of C. Ionescu Tulcea (see e. g. [18, p.80]) shows the existence of a unique probability measure P_x^π on $(\Omega, \mathcal{F})$, which is concentrated on the set of all sequences $(x_0, a_0, x_1, a_1, \ldots)$ with $(x_t, a_t) \in \mathbb{K}$ for all $t = 0, 1, \ldots$. ($\mathbb{K}$ is defined in (2.1).) Moreover, P_x^π satisfies that $P_x^\pi(x_0 = x) = 1$, and for every t = 0,1,...

$$P_x^\pi\,(a_t \in C\,|\,h_t) = \pi_t(C\,|\,h_t) \qquad \forall\, C \in \mathcal{B}(A) \tag{3.1}$$

$$P_x^\pi\,(x_{t+1} \in B\,|\,h_t, a_t) = Q\,(B\,|\,x_t, a_t) \qquad \forall\, B \in \mathcal{B}(X). \tag{3.2}$$

$(\Omega, \mathcal{F}, P_x^\pi, \{x_t\})$ is called a (discrete–time) *Markov control process*. The expectation operator with respect to P_x^π is denoted by E_x^π.

Remark 3.1. If $\pi = \{f_t\}$ is a *Markov* policy, then the state process $\{x_t\}$ is a Markov process with transition kernel $Q\,(\cdot\,|\,x, f_t(x))$; that is,

$$P_x^\pi\,(x_{t+1} \in B\,|\,x_0, \ldots, x_t) = P_x^\pi\,(x_{t+1} \in B\,|\,x_t) = Q\,(B\,|\,x_t, f_t(x_t))$$

for all $B \in \mathcal{B}(X)$ and $t = 0, 1, \ldots$. In particular, if $f \in \mathbb{F}$ is a *stationary* policy, then $\{x_t\}$ has a time–homogeneous transition kernel $Q\,(\cdot\,|\,x, f(x))$.

Remark 3.2. If $\pi = \{f_t\}$ is a Markov policy, then expressions such as $Q\,(\cdot\,|\,x, f_t(x))$ and $c\,(x, f_t(x))$ will usually be written as $Q\,(\cdot\,|\,x, f_t)$ and $c(x, f_t)$, respectively.

Performance criterion. Given $\pi \in \prod$ and $x \in X$, let

$$V(\pi, x) := \mathsf{E}_x^\pi \sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \tag{3.3}$$

be the total expected *discounted cost* when using the policy π, given the initial state $x_0 = x$. The number $\alpha \in (0, 1)$ in (3.3) is called the *discount factor*.

The *optimal control problem* we are concerned with is to find an *optimal policy* $\pi^* \in \prod$, i. e., a policy π^* such that $V\,(\pi^*, x) = V^*(x)$ for all $x \in X$, where

$$V^*(x) := \inf_\pi V(\pi, x), \;\; x \in X, \tag{3.4}$$

is the *optimal cost* (or value) *function*.

The main objective of the following sections is to give several characterizations of an optimal policy, as well as of the optimal cost function V^*. We will also consider a concept of *asymptotic optimality*, which has proved to be very useful in e. g. *adaptive* control problems, i. e., problems in which the control model depends on *unknown* parameters.

4. THE OPTIMALITY EQUATION

If the cost per stage $c(x,a)$ is *bounded*, then the optimal cost function $V^*(x)$ is the *unique bounded* function that satisfies the *optimality equation* (abbreviated: OE)

$$V^*(x) = \min_{a\in A(x)} \left[c(x,a) + \alpha \int V^*(y) Q(dy \,|\, x,a) \right], \; x \in X, \tag{4.1}$$

and moreover, a policy π^* is optimal *if and only if* its cost $V(\pi^*,\cdot)$ satisfies (4.1). These are well–known results that go back to the earlier works in the field (e. g. [5]). It is also known, on the other hand, that if $c(x,a)$ is *unbounded*, then the OE (4.1) may not have a unique solution [1, 2], or an optimal policy may not exist [19]. Thus it is important to characterize the optimal policies or the solutions to (4.1) that coincide with V^*. To do this *we will suppose throughout the following that Assumption 2.1 and Assumption 4.1* (below) *hold*.

Assumption 4.1. There exists a policy $\widehat{\pi}$ such that $V(\widehat{\pi},x) < \infty$ for each $x \in X$.

For instance, each of the conditions C_0, C_1, C_2 in Definition 4.5 below implies Assumption 4.1. Another sufficient condition is the following: there exists a policy $\widehat{\pi}$ such that the long–run expected "average cost"

$$\limsup_{n\to\infty} n^{-1} \mathsf{E}_x^{\widehat{\pi}} \sum_{t=0}^{n-1} c(x_t, a_t)$$

is finite for each $x \in X$; see e. g. [13].

Assumption 4.1, together with (3.4), guarantees that the optimal cost function is finite–valued: $0 \le V^*(x) < \infty$ for each $x \in X$.

To state our next result we introduce some *notation*: Let $L(X)_+$ be the class of nonnegative and l.s.c. functions on X, and for each $u \in L(X)_+$ define a new function Tu by

$$Tu(x) := \min_{a\in A(x)} \left[c(x,a) + \alpha \int_X u(y) Q(dy \,|\, x,a) \right]. \tag{4.2}$$

By Lemma 2.7 (b), the operator T defined by (4.2) maps $L(X)_+$ into itself. We also consider the sequence $\{v_n\}$ of *value iteration* (VI) functions defined recursively by $v_0(\cdot) := 0$, and $v_n := Tv_{n-1} = T^n v_0$ for $n = 1, 2, \ldots$. That is, for $n \ge 1$ and $x \in X$,

$$v_n(x) := \min_{a\in A(x)} \left[c(x,a) + \alpha \int v_{n-1}(y) Q(dy \,|\, x,a) \right]. \tag{4.3}$$

Note that, by induction and Lemma 2.7 (b) again, $v_n \in L(X)_+$ for all $n \ge 0$. From elementary Dynamic Programming [2, 3, 9], $v_n(x)$ is the optimal cost function for an n-stage problem (with "terminal cost" $v_0(\cdot) = 0$) given $x_0 = x$; i. e.,

$$v_n(x) = \inf_{\pi} V_n(\pi, x), \tag{4.4}$$

where

$$V_n(\pi,x) := \mathsf{E}_x^{\pi} \left[\sum_{t=0}^{n-1} \alpha^t c(x_t, a_t) \right]. \tag{4.5}$$

Theorem 4.2. Suppose that Assumptions 2.1 and 4.1 hold. Then:

(a) $v_n \uparrow V^*$; hence

(b) V^* is the (pointwise) minimal function in $L(X)_+$ that satisfies the OE (4.1), or equivalently

$$V^* = TV^*. \tag{4.6}$$

(c) There exists a stationary policy $f^* \in \mathbb{F}$ such that $f^*(x) \in A(x)$ minimizes the r.h.s. (right–hand side) of (4.1) for all $x \in X$, i. e. (using the notation in Remark 3.2)

$$V^*(x) = c(x, f^*) + \alpha \int V^*(y) Q\,(dy \,|\, x, f^*)\,, \tag{4.7}$$

and f^* is optimal. Conversely, if $f^* \in \mathbb{F}$ is an optimal stationary policy, then it satisfies (4.7).

(d) If π^* is a policy such that $V\,(\pi^*, \cdot)$ is in $L(X)_+$ and it satisfies the OE and the condition

$$\lim_{n\to\infty} \alpha^n \mathsf{E}_x^{\pi} V\,(\pi^*, x_n) = 0 \qquad \forall\, \pi \in \prod \text{ and } x \in X, \tag{4.8}$$

then $V\,(\pi^*, \cdot) = V^*(\cdot)$; hence π^* is optimal.

Before proving Theorem 4.2 let us note the following.

Remark 4.3. (a) If V^* is not finite–valued, the convergence in Theorem 4.2 (a) may *not* hold; see e. g. [2, p. 233, problem 9].

(b) By part (b) of Theorem 4.2, if $\pi^* \in \prod$ is an optimal policy, then $V\,(\pi^*, \cdot) = V^*(\cdot)$ satisfies the OE (4.1)=(4.6). However, the converse is *not* true in general: In [2, p.215, Example 3] a policy π^* is given such that $V\,(\pi^*, \cdot)$ satisfies the OE, but π^* is *not* optimal. Such a policy π^* does not satisfy (4.8), of course.

(c) Observe that (4.8) trivially holds if $c(x, a)$ is *bounded*, for if $0 \le c(x, a) \le M$ $\forall\,(x, a) \in \mathbb{K}$, then, from (3.3), $0 \le V(\pi, \cdot) \le M/(1 - \alpha)$ $\forall\, \pi$. (Other conditions implying (4.8) are given in Theorem 4.6 below.)

Lemma 4.4. (a) If $v \in L(X)_+$ is such that $v \ge Tv$, then $v \ge V^*$.

(b) If v is a measurable function on X such that Tv is well defined and is such that $v \le Tv$ and

$$\lim_{n\to\infty} \alpha^n \mathsf{E}_x^{\pi} v\,(x_n) = 0 \qquad \forall\, \pi, x, \tag{4.9}$$

then $v \le V^*$.

P r o o f. (a) Suppose that $v \ge Tv$, and (see Lemma 2.7 (b)) let $f \in \mathbb{F}$ be a stationary policy that satisfies

$$v(x) \ge c(x, f) + \alpha \int v(y) Q\,(dy \,|\, x, f) \quad \forall\, x.$$

Iterating this inequality we obtain

$$v(x) \ge \mathsf{E}_x^f \sum_{t=0}^{n-1} \alpha^t c(x_t, f) + \alpha^n \mathsf{E}_x^f v(x_n), \quad \forall\, n, x,$$

where $\mathsf{E}_x^f v(x_n) = \int v(y) Q^n(dy \,|\, x, f)$, and $Q^n(B \,|\, x, f) = P_x^f(x_n \in B)$ denotes the n-step transition probability of the Markov chain $\{x_t\}$; see Remarks 3.1 and 3.2. Therefore, since v is nonnegative,

$$v(x) \geq \mathsf{E}_x^f \sum_{t=0}^{n-1} \alpha^t c(x_t, f) \quad \forall\, n, x,$$

and letting $n \to \infty$, (3.3) and (3.4) yield

$$v(x) \geq V(f, x) \geq V^*(x) \quad \forall\, x.$$

This proves (a).

(b) Let $\pi \in \prod$ and $x \in X$ be arbitrary. Then, from (3.2),

$$\begin{aligned}
&\mathsf{E}_x^\pi \left[\alpha^{t+1} v(x_{t+1}) \,|\, h_t, a_t\right] = \alpha^{t+1} \int v(y) Q(dy \,|\, x_t, a_t) \\
&= \alpha^t \left[c(x_t, a_t) + \alpha \int v(y) Q(dy \,|\, x_t, a_t) - c(x_t, a_t)\right] \\
&\geq \alpha^t \left[v(x_t) - c(x_t, a_t)\right],
\end{aligned}$$

since, by assumption, $Tv \geq v$. Hence

$$\alpha^t c(x_t, a_t) \geq -\mathsf{E}_x^\pi \left[\alpha^{t+1} v(x_{t+1}) - \alpha^t v(x_t) \,|\, h_t, a_t\right].$$

Thus taking expectations $\mathsf{E}_x^\pi(\cdot)$ and summing over $t = 0, \ldots, n-1$, we obtain

$$\sum_{t=0}^{n-1} \alpha^t \mathsf{E}_x^\pi c(x_t, a_t) \geq v(x) - \alpha^n \mathsf{E}_x^\pi v(x_n), \quad \forall\, n.$$

Letting $n \to \infty$, the latter inequality and (4.9) yield $V(\pi, x) \geq v(x)$, which implies (b), since π and x were arbitrary. □

Proof of Theorem 4.2. (a) – (b). To begin, note that the operator T in (4.2) is *monotone* on $L(X)_+$, i. e., $u \geq v$ implies $Tu \geq Tv$. Hence the VI functions v_n form a nondecreasing sequence in $L(X)_+$ and, therefore, there exists a function u in $L(X)_+$ such that $v_n \uparrow u$. This implies (by the Monotone Convergence Theorem) that

$$c(x, a) + \alpha \int v_{n-1}(y) Q(dy \,|\, x, a) \uparrow c(x, a) + \alpha \int u(y) Q(dy \,|\, x, a),$$

which combined with Lemma 2.7 (c) and (4.2) – (4.3) yields

$$u = Tu, \tag{4.10}$$

i. e. $u \in L(X)_+$ satisfies the OE (4.1) – (4.6). We will now show that $u = V^*$.

Indeed, from (4.10) and Lemma 4.4 (a), $u \geq V^*$. To prove the reverse inequality observe that, from (4.4) – (4.5),

$$v_n(x) \leq V_n(\pi, x) \leq V(\pi, x) \qquad \forall\, n, \pi, x,$$

and letting $n \to \infty$, we get $u(x) \leq V(\pi, x) \;\; \forall\; \pi, x$. This implies $u \leq V^*$. We have thus shown that $u = V^*$ satisfies part (a) and the OE (4.10)=(4.6).

Finally, to complete the proof of (a) – (b), note that $u = V^*$ is indeed the minimal solution to the OE, for if $u' \in L(X)_+$ is such that $u' = Tu'$, then Lemma 4.4 (a) yields $u' \geq V^*$.

(c) The existence of a stationary policy $f^* \in \mathbb{F}$ satisfying (4.7) follows from Lemma 2.7 (b). Now iteration of (4.7) shows (as in the proof of Lemma 4.4 (a)) that

$$\begin{aligned} V^*(x) &= \mathsf{E}_x^{f^*}\left[\sum_{t=0}^{n-1}\alpha^t c\,(x_t, f^*)\right] + \alpha^n \mathsf{E}_x^{f^*} V^*(x_n) \\ &\geq \mathsf{E}_x^{f^*}\left[\sum_{t=0}^{n-1}\alpha^t c\,(x_t, f^*)\right]. \end{aligned}$$

Letting $n \to \infty$, we obtain $V^*(x) \geq V\,(f^*, x)$, which combined with (3.4) yields $V^*(\cdot) = V\,(f^*, \cdot)$, i. e., f^* is optimal. Finally, the converse follows from the fact that, *for any stationary policy* $f \in \mathbb{F}$, *the cost* $V\,(f, \cdot)$ *satisfies* (by the Markov property; see Remarks 3.1 and 3.2)

$$V(f,x) = c(x,f) + \alpha \int_X V(f,y) Q\,(dy \,|\, x, f)\,. \tag{4.11}$$

(d) Apply Lemma 4.4 (b) to $v(\cdot) := V\,(\pi^*, \cdot)$. □

To close this section, we will show that each of the conditions C_0 to C_3 defined next implies (4.8).

Definition 4.5. C_i $(i = 1, 2, 3)$ stands for the following condition:

C_0. $c(x, a)$ is bounded (cf. Remark 4.3 (c)).

C_1. There exists a number $m > 0$ and a nonnegative measurable w on X such that, for all $(x, a) \in \mathbb{K}$,

(i) $c(x,a) \leq m w(x)$, and (ii) $\displaystyle\int w(y) Q\,(dy \,|\, x, a) \leq w(x)$.

C_2. $C(x) := \displaystyle\sum_{t=0}^{\infty} \alpha^t c_t(x) < \infty$ for every $x \in X$, where

$$c_t(x) := \sup_{a \in A(x)} \int c_{t-1}(y) Q\,(dy \,|\, x, a) \qquad \forall\ t = 1, 2, \ldots,$$

and $c_0(x) := \sup_{a \in A(x)} c(x, a)$.

C_3. $\lim_{n \to \infty} \alpha^n \mathsf{E}_x^{\pi} V\,(\pi', x_n) = 0 \quad \forall\, \pi, \pi' \in \prod$, and $x \in X$.

Theorem 4.6. (a) C_i implies C_{i+1} for $i = 0, 1, 2$ and C_3 implies (4.8). Hence:

(b) If any of the conditions C_0 to C_3 hold, then a policy π^* is optimal if and only if $V\,(\pi^*, \cdot)$ satisfies the OE.

P r o o f. (a) C_0 *implies* C_1. This is obvious: let $m > 0$ be an upper bound for $c(x, a)$ and take $w(\cdot) = 1$.

C_1 *implies* C_2. If C_1 holds, then a straightforward induction argument shows that $c_t(x) \leq m w(x)$ for all $x \in X$ and $t = 0, 1, \ldots$. Thus

$$C(x) \leq m w(x)/(1 - \alpha) < \infty \quad \text{for each} \quad x.$$

C_2 *implies* C_3. Suppose that C_2 holds, and let $\pi \in \prod$ and $x \in X$ be arbitrary. We will first show that

$$V(\pi, x) \leq C(x). \tag{4.12}$$

To begin, observe that, from (3.2),

$$\mathsf{E}_x^\pi \left[c_0\left(x_{t+1}\right) \mid h_t, a_t\right] = \int c_0(y) Q\left(dy \mid x_t, a_t\right) \leq c_1(x_t)$$

and, therefore, $\mathsf{E}_x^\pi c_0\left(x_{t+1}\right) \leq \mathsf{E}_x^\pi c_1(x_t)$. This kind of argument yields

$$\mathsf{E}_x^\pi c_0\left(x_t\right) \leq \mathsf{E}_x^\pi c_1\left(x_{t-1}\right) \leq \cdots \leq \mathsf{E}_x^\pi c_t(x_0) = c_t(x). \tag{4.13}$$

Thus, since $c(x_t, a_t) \leq c_0(x_t)$, we obtain

$$\mathsf{E}_x^\pi c\left(x_t, a_t\right) \leq \mathsf{E}_x^\pi c_0(x_t) \leq c_t(x) \qquad \forall\, t.$$

This inequality, together with (3.3) and the definition of $C(x)$ implies (4.12).

Let us now show that

$$\mathsf{E}_x^\pi C\left(x_n\right) \leq \sum_{t=n}^{\infty} \alpha^{t-n} c_t(x) \qquad \forall\, n = 0, 1, \ldots. \tag{4.14}$$

For $n = 0$, (4.14) follows from the definition of $C(x)$. For $n \geq 1$, (3.2) gives

$$\begin{aligned}
&\mathsf{E}_x^\pi \left[C(x_n) \mid h_{n-1}, a_{n-1}\right] = \int C(y) Q\left(dy \mid x_{n-1}, a_{n-1}\right) \\
&= \sum_{t=0}^{\infty} \alpha^t \int c_t(y) Q\left(dy \mid x_{n-1}, a_{n-1}\right) \\
&\leq \sum_{t=0}^{\infty} \alpha^t c_{t+1}\left(x_{n-1}\right).
\end{aligned}$$

Hence, taking expectation $\mathsf{E}_x^\pi(\cdot)$,

$$\mathsf{E}_x^\pi C(x_n) \leq \sum_{t=0}^{\infty} \alpha^t \mathsf{E}_x^\pi c_{t+1}\left(x_{n-1}\right).$$

However, as in (4.13), $\mathsf{E}_x^\pi c_{t+1}\left(x_{n-1}\right) \leq \mathsf{E}_x^\pi c_{t+2}\left(x_{n-2}\right) \leq \cdots \leq c_{t+n}(x)$, so that

$$\mathsf{E}_x^\pi C(x_n) \leq \sum_{t=0}^{\infty} \alpha^t c_{t+n}(x),$$

and (4.14) follows.

Finally, let π and π' be two arbitrary policies. Then from (4.12) with π' instead of π, and (4.14), we obtain

$$\mathsf{E}_x^\pi V\left(\pi', x_n\right) \leq \mathsf{E}_x^\pi C(x_n) \leq \sum_{t=n}^{\infty} \alpha^{t-n} c_t(x).$$

This in turn yields

$$\alpha^n \mathsf{E}_x^\pi V\left(\pi', x_n\right) \leq \sum_{t=n}^{\infty} \alpha^t c_t(x) \to 0 \quad \text{as} \quad n \to \infty,$$

since $C(x)$ is finite. Thus C_2 implies C_3.

C_3 *implies* (4.8). This is obvious, since π and π' in C_3 are arbitrary. This completes the proof of part (a).

(b) Follows from (a) and Theorem 4.2 (b), (d). □

Comments. 1. Theorems 4.2 (d) and 4.6 (b) extend all previous results relating an optimal "general" policy $\pi^* \in \prod$ (as opposed to an optimal *stationary* policy; see Theorem 4.2 (c)) to the OE (4.1), and they also clarify the role of the "growth condition" (4.8). For finite–state, finite–action MCPs, and dealing only with *Markov* policies, another characterization of optimal policies is given in [21]. Related results appear in [23].

2. As already noted at the beginning of this section (see also Remark 4.3 (c)) Theorem 4.2 is well–known in the bounded cost case (condition C_0). The condition C_1 was introduced by Lippman [22] to reduce the unbounded (in the supremum norm) cost problem to a bounded problem, which is done by defining a *weighted* supremum norm, where the "weight" is the function w in C_1. Lippman's approach has been used and extended by many authors; see e. g. [14, 29] and their references.

3. It is interesting to note that the condition C_1 (ii) on w implies that $\{w(x_n)\}$ is a P_x^π–super–martingale for any $\pi \in \prod$ and $x \in X$. That is, for any $n = 0, 1, \ldots$, (3.1) – (3.2) and C_1 (ii) yield

$$\begin{aligned} \mathsf{E}_x^\pi \left[w\left(x_{n+1}\right) \mid h_n \right] &= \int_A \int_X w(y) Q\left(dy \mid x_n, a_n\right) \pi_n\left(da_n \mid h_n\right) \\ &\leq w(x_n) \qquad P_x^\pi \text{ – a.s.} \end{aligned}$$

In Systems Theory, a function w satisfying C_1 (ii) is called a *Lyapunov function* and its relation to some "stability" and recurrence properties are well–known [6, 10, 16, 25]. It would be interesting to investigate what kind of information (if any) C_1(ii) gives on the "stability"properties of the controlled process $\{x_t\}$.

4. Condition C_2 has also been used by several authors, e. g. [1, 4, 7, 14].

5. Another sufficient condition for (4.8) can be obtained by analogy with related results for controlled *diffusion* processes. For instance, Kushner's [20] Theorem 3 can be restated in our context as follows:

(∗) Suppose that there is a nonnegative measurable function F on $\mathbb{R}$ such that $F(r)/r \uparrow \infty$ as $r \to \infty$, and

$$\int F\left(u(y)\right) Q\left(dy \mid x, a\right) \leq F\left(u(x)\right) \qquad \forall\, (x, a) \in \mathbb{K},$$

where $u(x) := V\left(\pi^*, x\right)$. Then (4.8) holds.

The proof of (∗) is similar to the proof for diffusions.

5. APPROXIMATIONS

The study of approximations to the optimal cost function V^* is important for both theoretical and computational purposes. For instance, in Theorem 4.2 (a) we have seen that V^* is the limit of the monotone *increasing* sequence of value iteration (VI) functions, from which we immediately conclude some properties of V^* (see Theorem 4.2 (b)). It is also worth noting that the VI approximation scheme is defined *recursively* and that it amounts to approximate V^* by problems with a *finite* number of stages (cf. (4.3) – (4.5)). In this section we consider three more types of approximations to V^*. The first one is via infinite–horizon problems with *bounded* (or "truncated") costs $c^n(x, a) \uparrow c(x, a)$, and the second one is a combination of bounded costs and finite–horizon (VI–like) approximations. These two are monotone *increasing* approximations to V^*. Finally, the third one is the standard *Policy Iteration* (PI), which provides *decreasing* approximations.

Assumptions 2.1 *and* 4.1 *are supposed to hold throughout the following.*

Bounded costs. Let $\{c^n(x,a),\, n = 0,1,\ldots\}$ be a sequence of nonnegative, *bounded* functions on $\mathbb{K}$ such that $c^n \uparrow c$ and, for each n, Assumption 2.1 (a) holds when c is replaced by c^n. (For instance, the truncated cost $c^n(x,a) := \min\{c(x,a), n\}$ satisfies Assumption 2.1 (a) if the sets $A(x)$ are compact.) Now, instead of (3.3) – (3.4), consider the corresponding cost functions

$$U_n(\pi,x) := \mathsf{E}_x^\pi \sum_{t=0}^{\infty} \alpha^t c^n(x_t,a_t), \quad \text{and} \quad U_n^*(x) := \inf_\pi U_n(\pi,x). \tag{5.1}$$

For each n, the optimal cost function $U_n^*(x)$ is the unique *bounded* function in $L(X)_+$ that satisfies the OE (cf. (4.1)=(4.6))

$$U_n^* = T_n U_n^*, \tag{5.2}$$

where, for $v \in L(X)_+$,

$$T_n v(x) := \min_{a\in A(x)} \left[c^n(x,a) + \alpha \int v(y) Q(dy \,|\, x,a) \right]. \tag{5.3}$$

Recursive bounded costs. The VI equation (4.3) suggests to introduce a sequence $\{u_n\}$ defined *recursively* as $u_0 := 0$, and $u_n := T_n u_{n-1}$ for $n \geq 1$; that is,

$$u_n(x) = \min_{a\in A(x)} \left[c^n(x,a) + \alpha \int u_{n-1}(y) Q(dy \,|\, x,a) \right]. \tag{5.4}$$

Policy iteration (PI). Let $f_0 \in \mathbb{F}$ be a stationary policy with a finite–valued discounted cost $V(f_0,\cdot) =: w_0(\cdot) \in L(X)_+$. As in (4.11), we may write

$$w_0(x) = c(x,f_0) + \alpha \int w_0(y) Q(dy \,|\, x,f_0) \qquad \forall\, x \in X. \tag{5.5}$$

Now, with T being the operator defined in (4.2), let $f_1 \in \mathbb{F}$ be such that

$$c(x,f_1) + \alpha \int w_0(y) Q(dy \,|\, x,f_1) = T w_0(x), \tag{5.6}$$

i. e. (cf. Lemma 2.7 (b)),

$$c(x,f_1) + \alpha \int w_0(y) Q(dy \,|\, x,f_1) = \min_{a\in A(x)} \left[c(x,a) + \alpha \int w_0(y) Q(dy \,|\, x,a) \right].$$

Write $w_1(\cdot) := V(f_1,\cdot)$. In general, given $f_n \in \mathbb{F}$, suppose that $w_n(\cdot) := V(f_n,\cdot)$ is in $L(X)_+$, and let $f_{n+1} \in \mathbb{F}$ be such that

$$\begin{aligned} c(x,f_{n+1}) \;+\; & \alpha \int w_n(y) Q(dy \,|\, x,f_{n+1}) = T w_n(x) \\ = \; & \min_{a\in A(x)} \left[c(x,a) + \alpha \int w_n(y) Q(dy \,|\, x,a) \right]. \end{aligned} \tag{5.7}$$

Theorem 5.1. (a) Each of the sequences U_n^* and u_n is monotone increasing and converges to V^*.

(b) There exists a measurable nonnegative function $w \geq V^*$ such that $w_n \downarrow w$, and w satisfies the OE $w = Tw$. If, moreover, w satisfies

$$\lim_{n\to\infty} \alpha^n \mathsf{E}_x^\pi w(x_n) = 0 \qquad \forall\, \pi, x, \tag{5.8}$$

then $w = V^*$.

P r o o f . (a) Let us first show that $U_n^* \uparrow V^*$. To begin with, note that, since $c^n \uparrow c$, it is clear from (5.1) that U_n^* is an increasing sequence in $L(X)_+$ and, therefore, there exists a function $u \in L(X)_+$ such that $U_n^* \uparrow u$. Moreover, from Lemma 2.7 (c), letting $n \to \infty$ in (5.2) we see that $u = Tu$, i. e., u satisfies the OE. This implies that $u \geq V^*$, since, by Theorem 4.2 (b), V^* is the minimal solution in $L(X)_+$ to the OE. On the other hand, it is clear from (5.1) that $U_n^* \leq V^*$ for all n, so that $u \leq V^*$. Thus $u = V^*$, i. e. $U_n^* \uparrow V^*$. Finally, a completely analogous argument shows that $u_n \uparrow V^*$.

(b) Let us now consider the sequence of PI functions w_n. We will first show that this sequence is decreasing. From (5.5),

$$\begin{aligned} w_0(x) &\geq \min_{a \in A(x)} \left[c(x,a) + \alpha \int w_0(y) Q(dy \,|\, x, a) \right] \\ &= Tw_0(x), \end{aligned}$$

so that, by (5.6),

$$w_0(x) \geq c(x, f_1) + \alpha \int w_0(y) Q(dy \,|\, x, f_1).$$

As in the proof of Lemma 4.4 (a), the latter inequality implies

$$w_0(x) \geq V(f_1, x) =: w_1(x).$$

In fact, a similar argument clearly holds for arbitrary n, so that, from (5.7),

$$w_n \geq Tw_n \geq w_{n+1} \qquad \forall\, n \geq 0. \tag{5.9}$$

Hence, by monotonicity, there is a nonnegative measurable function w such that $w_n \downarrow w$. Clearly, $w \geq V^*$, since $w_n \geq V^*$ for all n. Now, from [18, Lemma 3.4] (or [17, Lemma 3.3]) if h_n is a sequence of functions on $\mathbb{K}$ such that $h_n \downarrow h$, then

$$\lim_{n \to \infty} \inf_{a \in A(x)} h_n(x,a) = \inf_{a \in A(x)} h(x,a).$$

Thus applying this result to (5.9), we get $w \geq Tw \geq w$, i. e., w satisfies the OE $w = Tw$. Finally, the last statement in part (b), assuming (5.9), follows from Lemma 4.4 (b). □

Comments. 1. Each of the conditions C_0, C_1 and C_2 in Definition 4.5 implies (5.8), in which case $w = V^*$. In general, however, $w > V^*$. This kind of "abnormal" behavior of upper, decreasing approximations w_n (as opposed to the "nicely behaved" increasing approximations in Theorems 5.1 (a) or 4.2 (a), which do converge to V^*), has been noted by several authors in related contexts [1, 17, 30].

2. For MCPs satisfying C_0, C_1 or C_2, or with some particular structural property – e. g. convexity –, many other types of approximations are possible [2, 7 ,11, 12, 14, 15, 17, 28 – 30].

6. OTHER OPTIMALITY CRITERIA

Let us rewrite the OE (4.1) as

$$\min_{a \in A(x)} \Phi(x,a) = 0. \tag{6.1}$$

where

$$\Phi(x,a) := c(x,a) + \alpha \int V^*(y) Q(dy \,|\, x, a) - V^*(x) \tag{6.2}$$

is the so–called *discrepancy function*. This name for Φ comes from the fact that

$$V(\pi, x) - V^*(x) \geq \Phi(x, a) \qquad (\geq 0) \tag{6.3}$$

for any policy $\pi = \{\pi_t\}$ with initial action $\pi_0(x) = a \in A(x)$ when $x_0 = x$. Thus $\Phi(x, a)$ bounds from below the "deviation from optimality" of the policy π (see [7, §5], or Lemma 6.2 (c) below).

The objective in this section is to present optimality criteria in terms of Φ and also in terms of the sequence $\{M_n\}$ defined as

$$M_n := \sum_{t=0}^{n-1} \alpha^t c(x_t, a_t) + \alpha^n V^*(x_n) \quad \text{for} \quad n = 1, 2, \ldots, \tag{6.4}$$

with $M_0 := V^*(x_0)$.

To begin, let us note that if

$$V^n(\pi, x) := \mathsf{E}_x^\pi \sum_{t=n}^{\infty} \alpha^{t-n} c(x_t, a_t) \tag{6.5}$$

denotes the total expected discounted cost from stage n onward, when using the policy π and given $x_0 = x$, then from (3.3) and (4.5) we have

$$V(\pi, x) = V_n(\pi, x) + \alpha^n V^n(\pi, x). \tag{6.6}$$

On the other hand, using (6.4) and (6.5) we can also write $V(\pi, x)$ as

$$V(\pi, x) = \mathsf{E}_x^\pi (M_n) + \alpha^n [V^n(\pi, x) - \mathsf{E}_x^\pi V^*(x_n)]. \tag{6.7}$$

We now state the main result in this section.

Theorem 6.1. Let π be a policy such that $V(\pi, x) < \infty$ for each $x \in X$. Then the following statements are equivalent:

(a) π is an optimal policy.

(b) $V^n(\pi, x) = \mathsf{E}_x^\pi V^*(x_n) \quad \forall\, n, x.$

(c) $\mathsf{E}_x^\pi \Phi(x_n, a_n) = 0 \quad \forall\, n, x.$

(d) $\{M_n\}$ is a P_x^π–martingale $\forall\, x$.

To prove this theorem we will use the following result from Schäl [28].

Lemma 6.2. Let π be a policy such that $V(\pi, x) < \infty$ for each $x \in X$ (one such policy exists, by Assumption 4.1). Then:

(a) $V^n(\pi, x) \geq \mathsf{E}_x^\pi V^*(x_n) \quad \forall\, n.$

(b) $\sum_{t=n}^{\infty} \alpha^{t-n} \mathsf{E}_x^\pi \Phi(x_t, a_t) = V^n(\pi, x) - \mathsf{E}_x^\pi V^*(x_n) \quad \forall\, n, x$; in particular (for n = 0),

(c) $V(\pi, x) - V^*(x) = \sum_{t=0}^{\infty} \alpha^t \mathsf{E}_x^\pi \Phi(x_t, a_t).$

Parts (a) and (b) in Lemma 6.2 correspond to Schäl's [28] Theorem 2.13 and Lemma 2.16, respectively. Schäl uses a "Lyapunov condition", similar to C_1 in § 4, to obtain the growth condition (4.8) (see our Theorem 4.6), from which Lemma 6.2 (b) is immediately deduced. In our case, the latter conclusion follows from Lemma 6.2 (a), which implies

$$0 \leq \alpha^n \mathsf{E}_x^\pi V^*(x_n) \leq \alpha^n V^n(\pi, x) \to 0 \quad \text{as} \quad n \to \infty \tag{6.8}$$

where the latter convergence is obtained from (6.6) and the assumption that $V(\pi, x)$ is finite.

We also need the following elementary result.

Lemma 6.3. For any $\pi \in \prod$ and $x \in X$, $\{M_n\}$ is a P_x^π–sub–martingale, i. e.

$$\mathsf{E}_x^\pi \left(M_{n+1} \,|\, h_n \right) \geq M_n \quad P_x^\pi \text{ – a.s.} \quad \forall\, n.$$

Therefore

$$\mathsf{E}_x^\pi M_{n+1} \geq \mathsf{E}_x^\pi M_n \geq \cdots \geq \mathsf{E}_x^\pi M_0 = V^*(x) \quad \forall\, n. \tag{6.9}$$

P r o o f . From (6.2) and (3.2),

$$\Phi\left(x_n, a_n\right) = \mathsf{E}_x^\pi \left[c\left(x_n, a_n\right) + \alpha V^*\left(x_{n+1}\right) - V^*(x_n) \,|\, h_n, a_n \right],$$

whereas from (6.4),

$$M_{n+1} = M_n + \alpha^n \left[c\left(x_n, a_n\right) + \alpha V^*\left(x_{n+1}\right) - V^*(x_n) \right].$$

Therefore, by the properties of conditional expectations,

$$\mathsf{E}_x^\pi \left(M_{n+1} \,|\, h_n \right) = M_n + \alpha^n \mathsf{E}_x^\pi \left[\Phi\left(x_n, a_n\right) \,|\, h_n \right]. \tag{6.10}$$

This implies the desired result, since $\Phi \geq 0$. □

Proof of Theorem 6.1. First we show that (a) and (b) are equivalent.

(a) *implies* (b). Let π be an optimal policy, i. e., $V(\pi, x) = V^*(x)$ for all x. Then, from (6.7) and (6.9),

$$\begin{aligned} V^*(x) &= \mathsf{E}_x^\pi \left(M_n \right) + \alpha^n \left[V^n(\pi, x) - \mathsf{E}_x^\pi V^*(x_n) \right] \\ &\geq V^*(x) + \alpha^n \left[V^n(\pi, x) - \mathsf{E}_x^\pi V^*(x_n) \right]. \end{aligned}$$

This implies $V^n(\pi, x) \leq \mathsf{E}_x^\pi V^*(x_n)$ and, therefore, by Lemma 6.2 (a), we obtain part (b) in Theorem 6.1. Conversely, (b) *implies* (a): take $n = 0$.

The equivalence of (b) and (c) follows from Lemma 6.2 (b).

Finally the equivalence of (c) and (d) follows from (6.10), the properties of conditional expectations, and $\Phi \geq 0$. □

Comments. Theorem 6.1 puts together optimality criteria known separately for several classes of controlled processes. For instance, the implication (a) $\Longrightarrow$ (b) is the well–known Bellman's *Principle of Optimality*; see, e. g., [2, p. 12], [18, p. 109]. The equivalence of parts (a) and (d) is also well–known [26]; for continuous–time (e. g. diffusion) processes see, e. g., [8]; for average–cost problems see [24]. We also note that the discrepancy function Φ in (6.2) is the "discounted–cost analogue" of Mandl's [24] discrepancy function φ in the average–cost case. On the other hand, observe that (4.7) can be written as

$$\Phi\left(x, f^*(x)\right) = 0 \qquad \forall\, x. \tag{6.11}$$

In other words, from Theorem 4.2 (c) and equation (6.1), we may restate the equivalence of (a) and (c) in Theorem 6.1 as follows: A *stationary* policy f^* is optimal if and only if it satisfies (6.11).

7. ASYMPTOTIC OPTIMALITY

Sections 4 and 6 present several characterizations of an optimal policy; these results do not say, however, how one can compute or, at least, "estimate" one such policy. In this section we briefly discuss the notion of asymptotic optimality, which allows us to say when a given control policy is "close" to being optimal. The basic ideas were introduced by Schäl [28] in his analysis of *adaptive* control problems (see also [11] Chapter 2).

The following definition, in which $\Phi(x, a)$ is the discrepancy function in (6.2), is motivated by Theorem 6.1 (c) – see also Lemma 6.2 (c) and equation (6.11).

Definition 7.1. (a) A policy $\pi \in \prod$ is said to be *asymptotically optimal* (AO) if, for each $x \in X$,

$$\mathsf{E}_x^\pi \Phi\,(x_n, a_n) \to 0 \quad \text{as} \quad n \to \infty. \tag{7.1}$$

(b) A Markov policy $\pi = \{f_n\}$ is called *pointwise asymptotically optimal* (pointwise AO) if, for each $x \in X$,

$$\Phi\,(x, f_n(x)) \to 0 \quad \text{as} \quad n \to \infty. \tag{7.2}$$

Observe that, by Theorem 6.1 (a), (c), if a policy is optimal, then it is AO. On the other hand, from Lemma 6.2 and equation (6.7) we immediately obtain the following result.

Theorem 7.2. Let $\pi \in \prod$ be such that $V\,(\pi, x) < \infty$ for each x. Then the following statements are equivalent:

(a) π is AO.

(b) $\lim_{n\to\infty} [V^n(\pi, x) - \mathsf{E}_x^\pi V^*(x_n)] = 0$ for each x.

(c) $\lim_{n\to\infty} \sum_{t=n}^{\infty} \alpha^{t-n} \mathsf{E}_x^\pi \Phi\,(x_t, a_t) = 0$ for each x.

(d) $V(\pi, x) = \mathsf{E}_x^\pi(M_n) + o(\alpha^n)$ as $n \to \infty$, for each x.

Theorem 7.2 is the "asymptotic version" of Theorem 6.1. Observe also that if the cost per stage $c(x, a)$ is *bounded*, then (7.1) (hence each of (a) – (d) in Theorem 7.2) is equivalent to: For each $x \in X$,

$$\Phi\,(x_n, a_n) \to 0 \quad \text{in} \quad P_x^\pi\text{–probability as } n \to \infty. \tag{7.3}$$

This follows from the Dominated Convergence Theorem. In the bounded cost case again, and if $\pi = \{f_n\}$ is a Markov policy, then (7.3) holds whenever the convergence in (7.2) is uniform in $x \in X$.

For *pointwise* asymptotic optimality we do not have a general result such as Theorem 7.2, but very often it is easier to verify (7.2) than (7.1). Let us give an example.

Example 7.3. Let $\{v_n\}$ be the sequence of value iteration (VI) functions in (4.3), and let $\pi = \{f_n\}$ be the Markov policy defined as follows: $f_0 \in \mathbb{F}$ is arbitrary, and for $n = 1, 2, \ldots$, $f_n \in \mathbb{F}$ minimizes the r.h.s. of (4.3), i. e.,

$$v_n(x) = c\,(x, f_n) + \alpha \int v_{n-1}(y) Q\,(dy \,|\, x, f_n) \quad \forall\, x. \tag{7.4}$$

(Recall Remark 3.2.) We will show that π is pointwise AO.

From (6.2),

$$\Phi\,(x, f_n(x)) = c(x, f_n) + \alpha \int V^*(y) Q\,(dy \,|\, x, f_n) - V^*(x),$$

so that, from (7.4),

$$\Phi\,(x, f_n(x)) = \alpha \int [V^*(y) - v_{n-1}(y)]\, Q\,(dy \,|\, x, f_n) - [V^*(x) - v_n(x)]\,.$$

Thus, since $v_n \uparrow V^*$ (Theorem 4.2 (a)),

$$\Phi\,(x, f_n(x)) \le a \int [V^*(y) - v_{n-1}(y)]\, Q\,(dy \,|\, x, f_n) \quad \forall\, n, x. \tag{7.5}$$

On the other hand, from (4.1),

$$V^*(x) \leq c(x, f_n) + \alpha \int V^*(y) Q(dy \,|\, x, f_n),$$

which combined with (7.4) yields

$$V^*(x) - v_n(x) \leq \alpha \int (V^*(y) - v_{n-1}(y)) \, Q(dy \,|\, x, f_n). \tag{7.6}$$

Iterating this inequality we obtain

$$V^*(x) - v_n(x) \leq \alpha^2 \int (V^*(y) - v_{n-2}(y)) \, Q^2(dy \,|\, x; f_n, f_{n-1}),$$

where

$$Q^2(\cdot \,|\, x; f_n, f_{n-1}) = \int_X Q(\cdot \,|\, y, f_{n-1}) \, Q(dy \,|\, x, f_n).$$

In general, further iteration of (7.6) yields (since $v_0(\cdot) := 0$)

$$\begin{aligned} V^*(x) - v_n(x) &\leq \alpha^n \int V^*(y) Q^n(dy \,|\, x; f_n, f_{n-1}, \ldots, f_1) \\ &= \alpha^n \mathsf{E}_x^\pi V^*(x_n) \qquad \forall\, n, x, \end{aligned} \tag{7.7}$$

where Q_n denotes the n-step transition probability of the Markov chain $\{x_n\}$; see Remark 3.1. Thus assuming that $V(\pi, x) < \infty$ for each $x \in X$, the inequalities (7.5) – (7.7) yield

$$\Phi(x, f_n(x)) \leq \alpha^n \mathsf{E}_x^\pi V^*(x_n) \leq \alpha^n V^n(\pi, x) \to 0$$

by Lemma 6.2 and (6.8). This proves (7.2); that is, the "VI policy" defined by (7.4) is pointwise AO.

Comments. Asymptotic optimality (AO) has been studied by several authors, but typically under conditions such as C_0, C_1 and C_2. For applications of AO to several adaptive control policies and approximation procedures – including state or disturbance space discretizations, and "rolling horizon" policies – see e. g. [7, 11, 12, 14, 15, 17, 28].

ACKNOWLEDGEMENTS

The authors wish to thank Rolando Cavazos–Cadena for very helpful remarks on a previous version of this paper. This research was partially supported by the TWAS (Trieste), by CONACyT (México), and by ICFES and Colciencias (Colombia).

(Received June 14, 1991.)

REFERENCES

[1] A. Bensoussan: Stochastic control in discrete time and applications to the theory of production. Math. Programm. Study *18* (1982), 43–60.
[2] D. P. Bertsekas: Dynamic Programming: Deterministic and Stochastic Models. Prentice–Hall, Englewood Cliffs, N.J. 1987.
[3] D. P. Bertsekas and S. E. Shreve: Stochastic Optimal Control: The Discrete Time Case. Academic Press, New York 1978.
[4] R. N. Bhattacharya and M. Majumdar: Controlled semi–Markov models – the discounted case. J. Statist. Plann. Inference *21* (1989), 365–381.
[5] D. Blackwell: Discounted dynamic programming. Ann. Math. Statist. *36* (1965), 226–235.
[6] R. S. Bucy: Stability and positive supermartingales. J. Diff. Eq. *1* (1965), 151–155.
[7] R. Cavazos–Cadena: Finite–state approximations for denumerable state discounted Markov decision processes. Appl. Math. Optim. *14* (1986), 1–26.

[8] M. H. A. Davis: Martingale methods in stochastic control. Lecture Notes in Control and Inform. Sci. *16* (1979), 85–117.
[9] E. B. Dynkin and A. A. Yushkevich: Controlled Markov Processes. Springer–Verlag, New York 1979.
[10] O. Hernández–Lerma: Lyapunov criteria for stability of differential equations with Markov parameters. Bol. Soc. Mat. Mexicana *24* (1979), 27–48.
[11] O. Hernández–Lerma: Adaptive Markov Control Processes. Springer–Verlag, New York 1989.
[12] O. Hernández–Lerma and R. Cavazos–Cadena: Density estimation and adaptive control of Markov processes: average and discounted criteria. Acta Appl. Math. *20* (1990), 285–307.
[13] O. Hernández–Lerma and J. B. Lasserre: Average cost optimal policies for Markov control processes with Borel state space and unbounded costs. Syst. Control Lett. *15* (1990), 349–356.
[14] O. Hernández–Lerma and J. B. Lasserre: Value iteration and rolling plans for Markov control processes with unbounded rewards. J. Math. Anal. Appl. (to appear).
[15] O. Hernández–Lerma and J. B. Lasserre: Error bounds for rolling horizon policies in discrete-time Markov control processes. IEEE Trans. Automat. Control *35* (1990), 1118–1124.
[16] O. Hernández–Lerma, R. Montes de Oca and R. Cavazos–Cadena: Recurrence conditions for Markov decision processes with Borel state space: a survey. Ann. Oper. Res. *28* (1991), 29–46.
[17] O. Hernández–Lerma and W. Runggaldier: Monotone approximations for convex stochastic control problems (submitted for publication).
[18] K. Hinderer: Foundations of Non–Stationary Dynamic Programming with Discrete Time Parameter. Springer–Verlag, Berlin – Heidelberg – New York 1970.
[19] A. Hordijk and H. C. Tijms: A counterexample in discounted dynamic programming. J. Math. Anal. Appl. *39* (1972), 455–457.
[20] H. J. Kushner: Optimal discounted stochastic control for diffusion processes. SIAM J. Control *5* (1967), 520–531.
[21] S. A. Lippman: On the set of optimal policies in discrete dynamic programming. J. Math. Anal. Appl. *24* (1968), 2, 440–445.
[22] S. A. Lippman: On dynamic programming with unbounded rewards. Manag. Sci. *21* (1975), 1225–1233.
[23] P. Mandl: On the variance in controlled Markov chains. Kybernetika *7* (1971), 1, 1–12.
[24] P. Mandl: A connection between controlled Markov chains and martingales. Kybernetika *9* (1973), 4, 237–241.
[25] S. P. Meyn: Ergodic theorems for discrete time stochastic systems using a stochastic Lyapunov function. SIAM J. Control Optim. *27* (1989), 1409–1439.
[26] U. Rieder: On optimal policies and martingales in dynamic programming. J. Appl. Probab. *13* (1976), 507–518.
[27] U. Rieder: Measurable selection theorems for optimization problems. Manuscripta Math. *24* (1978), 115–131.
[28] M. Schäl: Estimation and control in discounted stochastic dynamic programming. Stochastics *20* (1987), 51–71.
[29] J. Wessels: Markov programming by successive approximations with respect to weighted supremum norms. J. Math. Anal. Appl. *58* (1977), 326–335.
[30] W. Whitt: Approximations of dynamic programs, I. Math. Oper. Res. *4* (1979), 179–185.

Prof. Dr. Onésimo Hernández–Lerma, Departamento de Matemáticas, CINVESTAV–IPN, A. Postal 14–740, 07000 México, D.F., Mexico.
Prof. Myriam Muñoz de Ozak, Departamento de Matemáticas y Estadística, Universidad Nacional de Colombia, Bogotá. Colombia.