KYBERNETIKA — VOLUME 29 (1993), NUMBER 1, PAGES 1–17

EXISTENCE OF AVERAGE OPTIMAL POLICIES IN MARKOV CONTROL PROCESSES WITH STRICTLY UNBOUNDED COSTS

ONÉSIMO HERNÁNDEZ–LERMA

This paper deals with discrete-time Markov control processes on *Borel* spaces and *strictly unbounded* one-stage costs, i. e. costs that grow without bound on the complement of compact sets. Under mild assumptions, the existence of a ***minimum pair*** for the average cost problem is ensured, as well as the existence of ***stable*** optimal and pathwise-optimal control policies. It is shown that the existence of a minimum pair is *equivalent* to the existence of a solution to an "optimality inequality", which is a weaker version of the dynamic programming (or optimality) equation.

1. INTRODUCTION

This paper is concerned with the problem of minimizing the *average cost* (AC) for discrete-time Markov control processes (MCPs) with *Borel* state and control processes, and *strictly unbounded* one-stage costs, i. e. costs that grow without bound on the complement of compact sets. The most conspicuous *example* of the MCPs we have in mind is the *linear-quadratic* (or LQ) problem, which consists of the linear system equation

$$x_{t+1} = \alpha x_t + \beta a_t + \xi_t, \quad t = 0, 1, \ldots \tag{1.1}$$

and the quadratic one-stage cost

$$c(x, a) := x'\gamma x + a'\theta a, \tag{1.2}$$

where "prime" denotes transpose. In (1.1)–(1.2), the state and control (or action) spaces are $X := \mathrm{R}^p$ and $A := \mathrm{R}^q$ respectively, and the ξ_t are i. i. d. (independent and identically distributed) random disturbances. α, β, γ and θ are matrices of appropriate dimensions, with γ and θ symmetric and positive definite. However, the feature we are interested in of the LQ problem is *not* the linearity — we may as well take a general nonlinear system equation

$$x_{t+1} = G(x_t, a_t, \xi_t), \quad t = 0, 1, \ldots. \tag{1.3}$$

What we are interested in is the fact that the one-stage cost c is ***strictly unbounded***, in the sense that (cf. Assumption 2.1 (c) and Remark 2.4 (a))

$$\inf_{|x|>n} \inf_a c(x, a) \to \infty \text{ as } n \to \infty. \tag{1.4}$$

zapsat do jmen

Our objective is to show, under mild assumptions (Assumptions 2.1 and 3.3), the existence of a "minimum pair" and of AC-optimal policies for a class of MCPs that includes (1.1) - (1.2) and (1.3) - (1.2).

As noted by several authors (e. g. Hartley [12], Kushner [22]), there are virtually *no results in the MCP literature* directly applicable to the AC problem for (1.1) - (1.2) [or (1.3) - (1.2)], for, to begin with, most of this literature is concentrated on problems with (i) *denumerable* state space, and/or (ii) *compact* control constraint sets, and/or (iii) *bounded* one-stage costs: [1,6,7,11,14,20,30,...]. Thus one has to resort to "indirect" approaches or plainly to non-MCP techniques. For instance, among the latter, to solve the AC problem for (1.1) - (1.2) one uses ad hoc concepts from linear systems theory, such as controllability, observability, stabilizability [1,12,22,28]. Among the former, "indirect" approaches one may use compactness/compactification methods [2, p. 210; 21,29]; the "vanishing discount factor" approach [15,16]; the linear programming approach [17]; or combinations of these [12,29].

The approach we adopt in this paper, on the other hand, is a direct one, based on the fact that an AC problem with strictly unbounded costs is *necessarily* well-behaved. More precisely, if an arbitrary control policy yields a finite AC, then there exists a (possibly randomized) *stationary* policy that yields a better (i. e. lower) AC and, moreover, the latter policy is "stable" (in the sense of Definition 4.1). Thus we are able to show the existence of stable AC-optimal policies and that, furthermore, the existence of one such policy is essentially "equivalent" to the existence of a solution to the "optimality inequality" (see Theorem 5.3 and Corollary 5.4) — unlike the standard result, which shows that such an inequality is *sufficient* for optimality, see e. g. [15,16].

Organization of the paper: In § 2 we introduce the basic Markov control model and assumptions, and in § 3 we present the AC optimality criteria we are interested in (see (3.1) - (3.4) and Definition 3.4). § 4 deals with the definition and some important properties of *stable* relaxed (or randomized stationary) policies. Our main results are presented in § 5, and their proofs are collected in § 6. Finally, in § 7 we present two examples, one of which is the LQ problem (1.1) - (1.2), and conclude with some brief remarks on the "optimality inequality" versus the "optimality equation".

Remark 1.1. We use the following notation and terminology. Given a *Borel space* Y (i. e. a Borel subset of a complete and separable metric space), its Borel σ-algebra is denoted by $\mathcal{B}(Y)$; "measurable" always means "Borel-measurable". $\mathcal{P}(Y)$ stands for the space of probability measures (p. m.'s) on Y, and $C(Y)$ denotes the space of real-valued, continuous and bounded functions on Y. If Y and Z are Borel spaces, then a *stochastic kernel* (or conditional probability) *on* Y *given* Z is a function $P(\cdot\,|\,\cdot)$ such that $P(\cdot\,|\,z)$ is a p. m. on Y for each fixed $z \in Z$, and $P(B\,|\,\cdot)$ is a measurable function on Z for each fixed $B \in \mathcal{B}(Y)$. The family of all stochastic kernels on Y given Z is denoted by $\mathcal{P}(Y\,|\,Z)$. We also use standard abbreviations, such as p. m. = probability measure, a. s. = almost surely, a. a. = almost all.

2. THE CONTROL MODEL

We consider the usual Markov control model (X, A, Q, c) with state space X, control (or action) set A, transition law Q, and one-stage cost function c, which are assumed to satisfy the following. Both X and A are Borel spaces. The set of admissible control actions in state $x \in X$ is a nonempty set $A(x) \in \mathcal{B}(A)$ (see Remark 1.1 for notation). The set K of admissible state-action pairs, i.e.

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

is a Borel subset of $X \times A$. The transition law Q is a stochastic kernel on X given K, i.e. $Q \in \mathcal{P}(X \mid \mathsf{K})$. Finally, c is a real-valued measurable function on K.

The above Markov control model is standard: [2,10,11,14,19]. Here we will also assume the following.

Assumption 2.1. (a) Q is weakly continuous, i.e., $\int u(y)Q(\mathrm{d}y \mid x,a)$ is a continuous and bounded function in $(x,a) \in \mathsf{K}$ for every $u \in C(X)$ (recall Remark 1.1 for the meaning of $C(X)$).

(b) $c(x,a)$ is l.s.c. (lower semicontinuous) and nonnegative;

(c) c is strictly unbounded (equivalently, a *moment*; see Remark 2.4 (a)), i.e., there exists an increasing sequence of compact sets $K_n \uparrow \mathsf{K}$ such that

$$\lim_n \inf \{c(x,a) \mid (x,a) \notin K_n\} = +\infty. \tag{2.2}$$

In the remainder of this section we briefly discuss Assumption 2.1.

Example 2.2. Let S be a Borel space, and let ξ_t be a sequence of i.i.d S-valued random variables with a common distribution μ. Let $G : \mathsf{K} \times S \to X$ be a given measurable function, where K is the set in (2.1), and consider a stochastic control system of the form

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

The corresponding transition law Q satisfies, for any nonnegative measurable function u on X,

$$\begin{aligned} \int_X u(y)Q(\mathrm{d}y \mid x,a) &= E\left[u(x_{t+1}) \mid x_t = x, \ \ a_t = a\right] \\ &= \int_S u\left[G(x,a,s)\right]\mu(\mathrm{d}s). \end{aligned}$$

Clearly, Q satisfies Assumption 2.1 (a) if $G(x,a,s)$ is continuous in $(x,a) \in \mathsf{K}$ for every $s \in S$. In particular, Assumption 2.1 (a) holds for the linear system (1.1). On the other hand, it is plain that the quadratic cost in (1.2) satisfies Assumption 2.1 (b) and 2.1 (c) even if θ is only *nonnegative* definite (as opposed to positive definite).

Example 2.3. Suppose that X is *compact*, and that there is an increasing sequence of compact sets $A_n \uparrow A$. Then Assumption 2.1 (c) trivially holds: define $K_n := X \times A_n$ and recall that (by convention) the infimum over the empty set is $+\infty$.

The following remarks are crucial for later developments.

Remark 2.4. (a) A nonnegative measurable function v on a Borel space Y is said to be a *moment* on Y [13, 24] if there is an increasing sequence of compact sets $Y_n \uparrow Y$ such that

$$\lim_n \inf_{y \notin Y_n} v(y) = +\infty.$$

Thus, Assumption 2.1 (c) states, in other words, that the one-stage cost $c(x, a)$ is a *moment* on K.

(b) Let $M \subset \mathcal{P}(Y)$ be a family of p. m.'s (probability measures) on Y. If there exists a moment v on Y such that

$$\sup_{\mu \in M} \int v \, \mathrm{d}\mu < \infty,$$

then M is tight (i. e. [3, p. 37] for each $\varepsilon > 0$ there is a compact set C in Y such that $\mu(C) \geq 1 - \varepsilon \ \ \forall \mu \in M$). The proof is trivial.

(c) In §§ 4 and 6, the above remark (b) will be used in conjunction with *Prohorov's Theorem* [3, p.37], which states the following: If $M \subset \mathcal{P}(Y)$ is tight, then it is relatively compact, i. e. every sequence in M contains a weakly convergent subsequence. More explicitly, every sequence $\{\mu_n\}$ in M contains a subsequence $\{\mu_{n_i}\}$ such that, for some p. m. μ on Y,

$$\lim_i \int u \, \mathrm{d}\mu_{n_i} = \int u \, \mathrm{d}\mu \quad \forall \, u \in C(Y). \tag{2.4}$$

(d) Let μ_n and μ be p. m.'s on a Borel space Y, such that $\mu_n \longrightarrow \mu$ weakly, and let $v : Y \longrightarrow \mathsf{R}$ be l. s. c. and bounded from below. Then

$$\liminf_n \int v \, \mathrm{d}\mu_n \geq \int v \, \mathrm{d}\mu. \tag{2.5}$$

Indeed, by the assumption on v, there is a sequence of functions $v^k \in C(Y)$ such that $v^k \uparrow v$. Therefore, for all k,

$$\liminf_n \int v \, \mathrm{d}\mu_n \geq \liminf_n \int v^k \, \mathrm{d}\mu_n = \int v^k \, \mathrm{d}\mu.$$

Letting $k \longrightarrow \infty$ we obtain (2.5).

3. PERFORMANCE CRITERIA

Definition 3.1. F denotes the set of all measurable functions $f : X \to A$ such that $f(x) \in A(x)$ for all $x \in X$, and Φ stands for the set of all stochastic kernels $\varphi \in \mathcal{P}(A|X)$ such that $\varphi(A(x)\,|\,x) = 1$ for all $x \in X$.

A function $f \in \mathsf{F}$ will be identified with the stochastic kernel $\varphi \in \Phi$ such that $\varphi(\cdot\,|\,x)$ is the p.m. concentrated at $f(x)$ $\forall\, x \in X$. Thus $\mathsf{F} \subset \Phi$. By Assumption 2.1(c), the set K in (2.1) contains the graph of a function $f \in \mathsf{F}$ (see e.g. [27] Example 2.6). In other words, the set F (hence Φ and the set of policies defined next) is nonempty.

Definition 3.2. As usual, a *control policy* (more briefly a *policy*) is a sequence $\delta = \{\delta_t\}$ such that, for each $t = 0, 1, \ldots$, $\delta_t(\cdot\,|\,h_t)$ is a conditional probability on A given the history $h_t := (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t)$, and which satisfies the constraint $\delta_t(A(x_t)\,|\,h_t) = 1$. The class of all policies is denoted by Δ. A policy $\delta = \{\delta_t\}$ is said to be a:

(i) *relaxed* (or randomized stationary) *policy* if there exists $\varphi \in \Phi$ such that $\delta_t(\cdot\,|\,h_t) = \varphi(\cdot\,|\,x_t)$ $\forall\, h_t$, $t \geq 0$;

(ii) (nonrandomized or) *deterministic stationary policy* if there exists $f \in \mathsf{F}$ such that $\delta_t(\cdot\,|\,h_t)$ is concentrated at $f(x_t)$ $\forall\, h_t$, $t \geq 0$.

Following a standard convention, we will identify F (resp. Φ) with the set of all deterministic stationary (resp. relaxed) policies. Thus $\mathsf{F} \subset \Phi \subset \Delta$.

Let $(\Omega, \mathcal{F})$ be the measurable space consisting of the sample space $\Omega := (X \times A)^\infty$ and the corresponding product σ-algebra $\mathcal{F}$. Then for each policy $\delta \in \Delta$ and *initial distribution* $\nu \in \mathcal{P}(X)$, a probability measure P_ν^δ and a stochastic process $\{(x_t, a_t),\, t = 0, 1, \ldots\}$ are defined on Ω in a canonical way, where x_t and a_t represent the state and the control action at time t, respectively. The expectation operator with respect to P_ν^δ is written E_ν^δ. If ν is concentrated at (the *initial state*) $x_0 = x$, then we write P_ν^δ and E_ν^δ as P_x^δ and E_x^δ, respectively.

For each $\delta \in \Delta$ and $\nu \in \mathcal{P}(X)$, define

$$J_n(\delta, \nu) := E_\nu^\delta \left[\sum_{t=0}^{n-1} c(x_t, a_t) \right]. \tag{3.1}$$

Then the long-run expected *average cost* (AC) per unit time incurred by the policy δ, given the initial distribution ν, is given by

$$J(\delta, \nu) := \limsup_n n^{-1} J_n(\delta, \nu). \tag{3.2}$$

Similarly, the *pathwise* AC is given by

$$J^0(\delta, \nu) := \limsup_n n^{-1} \sum_{t=0}^{n-1} c(x_t, a_t). \tag{3.3}$$

Finally, let

$$j^* := \inf_{\nu} \inf_{\delta} J(\delta, \nu). \tag{3.4}$$

To ensure that the control problem is non-trivial, we suppose the following.

Assumption 3.3. $J(\widehat{\delta}, \widehat{\nu}) < \infty$ for some policy $\widehat{\delta}$ and some initial distribution $\widehat{\nu}$.

Assumptions 3.3 and 2.1 are supposed to hold throughout the following. (In § 7 we show two examples for which these assumptions hold.)

We are interested in several optimality criteria, one of which is the existence of a "minimum pair" introduced by Kurano [21].

Definition 3.4. Let δ^* be a policy and ν^* an initial distribution. Then:

(a) (δ^*, ν^*) is called a *minimum pair* if $J(\delta^*, \nu^*) = j^*$;

(b) δ^* is said to be *AC-optimal* if

$$J(\delta^*, \nu) = j^* \quad \forall \, \nu \in \mathcal{P}(X),$$

and *pathwise AC-optimal* if

$$J^0(\delta^*, \nu) = j^* \quad P_\nu^{\delta^*}\text{-a.s. } \forall \, \nu \in \mathcal{P}(X).$$

4. STABLE RELAXED POLICIES

As is well-known, when using a relaxed policy $\varphi \in \Phi$ the state process $\{x_t\}$ is an X-valued Markov chain with time-homogeneous transition kernel

$$Q(\cdot \,|\, x, \varphi) := \int_A Q(\cdot \,|\, x, a) \varphi(\mathrm{d}a \,|\, x), \quad x \in X. \tag{4.1}$$

We will also write

$$c(x, \varphi) := \int_A c(x, a) \varphi(\mathrm{d}a \,|\, x). \tag{4.2}$$

In particular, for a deterministic stationary policy $f \in \mathsf{F}$, (4.1) and (4.2) reduce to

$$Q(\cdot \,|\, x, f) := Q(\cdot \,|\, x, f(x)), \text{ and } c(x, f) := c(x, f(x)) \tag{4.3}$$

respectively.

Definition 4.1. A relaxed policy φ is said to be *stable* if:

(a) There exists an invariant p.m. $p^\varphi \in \mathcal{P}(X)$ for $Q(\cdot \,|\, \cdot, \varphi)$, i.e.

$$p^\varphi(\cdot) = \int_X Q(\cdot \,|\, x, \varphi) p^\varphi(\mathrm{d}x); \tag{4.4}$$

(b) the average cost $J(\varphi, p^\varphi)$ is finite and satisfies

$$J(\varphi, p^\varphi) = \int_X c(x, \varphi) p^\varphi(\mathrm{d}x). \tag{4.5}$$

The family of all stable relaxed policies is denoted by Φ_0.

Remark 4.2. (a) If $\varphi \in \Phi_0$, the invariant p.m. p^φ in Definition 4.1 is *not* required to be unique; if it is, then the transition kernel $Q(\cdot\,|\,\cdot,\varphi)$ is said to be *ergodic.* (Ergodicity holds if, e.g. $Q(\cdot\,|\,\cdot,\varphi)$ is *indecomposable* [31, pp. 389–390]. See also Remark 4.4 below.)

(b) If $\varphi \in \Phi_0$ and p^φ is as in (4.4)–(4.5), then by the Individual Ergodic Theorem [31, p. 388], the limit (cf. (3.1), (3.2))

$$J(\varphi,x) = \lim n^{-1} J_n(\varphi,x) \tag{4.6a}$$

exists for p^φ–a. a. (almost all) $x \in X$ and satisfies

$$\int J(\varphi,x)p^\varphi(\mathrm{d}x) = \int c(x,\varphi)p^\varphi(dx) = J(\varphi,p^\varphi) \tag{4.6b}$$

(where the second equality comes from (4.5)), and moreover [9, 31],

$$J^0(\varphi,p^\varphi) = \lim_n n^{-1} \sum_{t=0}^{n-1} c(x_t,\varphi) \quad \text{a.s. (almost surely).} \tag{4.7}$$

To state a stronger form of (4.7) we first recall the following [8, 18, 24, 26].

Definition 4.3. Let λ be a σ-finite measure on X, and for each $\varphi \in \Phi$ define

$$L^\varphi(x,B) := P_x^\varphi(x_n \in B \text{ for some } n \geq 1), \quad x \in X,\ B \in \mathcal{B}(X).$$

Then the transition kernel $Q(\cdot\,|\,\cdot,\varphi)$ is said to be λ-*irreducible* if $\lambda(B) > 0$ implies

$$L^\varphi(x,B) > 0 \qquad \forall\, x \in X,$$

and λ-*recurrent* (of *Harris recurrent*) if $\lambda(B) > 0$ implies

$$L^\varphi(x,B) = 1 \qquad \forall\, x \in X. \tag{4.8}$$

If X is a *denumerable* set (with the discrete topology) and we take λ as the *counting* measure, then λ-irreducibility and λ-recurrence reduce to the standard, elementary notions of irreducibility and recurrence in the theory of Markov chains. In the examples in § 7 we take λ = *Lebesgue* measure on $X = \mathsf{R}^p$.

Remark 4.4. Laws of Large Numbers [24, 26]. Let $\varphi \in \Phi_0$ be a stable relaxed policy and let p^φ be as in (4.4)–(4.5). If, moreover, the transition kernel $Q(\cdot\,|\,\cdot,\varphi)$ is λ-recurrent (for some λ), then $Q(\cdot\,|\,\cdot,\varphi)$ is ergodic and for any initial distribution ν in $\mathcal{P}(X)$,

$$\begin{aligned} J^0(\varphi,\nu) &= \lim_n n^{-1} \sum_{t=0}^{n-1} c(x_t,\varphi) \\ &= \int c(x,\varphi)p^\varphi(\mathrm{d}x) \quad P_\nu^\varphi\text{-a.s.} \end{aligned} \tag{4.9a}$$

and

$$J(\varphi,\nu) = E_\nu^\delta J^0(\varphi,\nu) = \int c(x,\varphi)p^\varphi(\mathrm{d}x). \tag{4.9b}$$

[cf. (4.6)–(4.7).]

We conclude this section by noting two important facts.

Proposition 4.5 Let $\varphi^* \in \Phi_0$ be a stable relaxed policy with corresponding invariant p. m. p^{φ^*}. Then $(\varphi^*, p^{\varphi^*})$ is a minimum pair, i. e.

$$J\left(\varphi^*, p^{\varphi^*}\right) = j^* \tag{4.10}$$

if and only if

$$J(\varphi^*, x) = j^* \text{ for } p^{\varphi^*}\text{-a. a. } x \in X. \tag{4.11}$$

Proof. It is obvious that (4.11) implies (4.10): see (4.6). To prove the converse we need to show that if (4.10) holds, then the set $B := \{x \mid J(\varphi^*, x) > j^*\}$ has p^{φ^*}-measure zero. Now, by (3.4), the complement of B is $B^c = \{x \mid J(\varphi^*, x) = j^*\}$, so that, by (4.10) and (4.6),

$$j^* = \int_B J(\varphi^*, x) p^{\varphi^*}(\mathrm{d}x) + j^* p^{\varphi^*}(B^c)$$

or, equivalently,

$$\int_B J(\varphi^*, x) p^{\varphi^*}(\mathrm{d}x) = j^* p^{\varphi^*}(B).$$

This implies $p^{\varphi^*}(B) = 0$. □

The following proposition states an important property of MCPs with strictly unbounded costs: it says that, when dealing with the AC problem, we may restrict ourselves to work with *stable* relaxed policies — see (4.19).

Proposition 4.6. Suppose that Assumptions 2.1 and 3.3 hold. Then for any $\delta \in \Delta$ and $\nu \in \mathcal{P}(X)$ such that $J(\delta, \nu) < \infty$, there exists a stable relaxed policy $\varphi \in \Phi_0$ such that

$$J(\delta, \nu) \geq J(\varphi, p^{\varphi}). \tag{4.12}$$

Proof. The proof in fact uses standard arguments (see e. g. Kurano [21] Lemma 2.1), but is included here for completeness. It consists of the following steps:

(i) There exists a p. m. μ on $X \times A$ concentrated on K such that

$$J(\delta, \nu) \geq \int c \, \mathrm{d}\mu; \tag{4.13}$$

(ii) Decompose the p. m. μ in part (i) as $\mu(\mathrm{d}x, \mathrm{d}a) = \varphi(\mathrm{d}a \mid x)\widetilde{\mu}(\mathrm{d}x)$, where $\varphi \in \Phi$ and $\widetilde{\mu} \in \mathcal{P}(X)$ is the marginal of μ on X, i.e.,

$$\mu(B \times C) = \int_B \varphi(C \mid x)\widetilde{\mu}(\mathrm{d}x) \quad \forall\, B \in \mathcal{B}(X), \;\; C \in \mathcal{B}(A);$$

thus we may rewrite (4.13) as

$$J(\delta, \nu) \geq \int_X c(x, \varphi)\widetilde{\mu}(\mathrm{d}x);$$

(iii) The relaxed policy φ in (ii) is stable and $\widetilde{\mu} =: p^\varphi$ is an invariant p.m. for $Q(\cdot \mid \cdot, \varphi)$.

Proof of (i). For each $n = 1, 2, \ldots$, let μ_n be the p.m. on $X \times A$ defined as

$$\mu_n(\Gamma) := n^{-1} \sum_{t=0}^{n-1} P_\nu^\delta \left[(x_t, a_t) \in \Gamma\right], \quad \Gamma \in \mathcal{B}(X \times A). \tag{4.14}$$

By definition of control policy (Definition 3.2), μ_n is concentrated on K and, on the other hand, by (3.2),

$$J(\delta, \nu) = \limsup \int c \, \mathrm{d}\mu_n. \tag{4.15}$$

Thus for any given $\varepsilon > 0$, there exists N such that

$$\sup_{n \geq N} \int c \, \mathrm{d}\mu_n \leq J(\delta, \nu) + \varepsilon < \infty.$$

This implies, by Remarks 2.4 (a), (b), that $\{\mu_n\}$ is tight and, therefore, by Prohorov's Theorem (Remark 2.4 (c)), there is a subsequence $\{\mu_{n_i}\}$ of $\{\mu_n\}$ converging weakly to a p.m. μ on $X \times A$. Furthermore, since each μ_n is concentrated on K, so is μ. Finally, from (4.15),

$$J(\delta, \nu) \geq \liminf_i \int c \, \mathrm{d}\mu_{n_i} \geq \int c \, \mathrm{d}\mu, \tag{4.16}$$

where the latter inequality is due to the weak convergence and Assumption 2.1 (b); see Remark 2.4 (d).

Proof of (ii). This decomposition is well-known, e.g. [11, p. 89, Theorem 2], [19, Corollary 12.7].

Proof of (iii). From (4.1) and (4.4), it suffices to show that

$$\int Tv(x, a) \mu(\mathrm{d}x, \mathrm{d}a) = \int Tv(x, \varphi) \widetilde{\mu}(\mathrm{d}x) = 0 \;\; \forall \, v \in C(X), \tag{4.17}$$

where

$$Tv(x, a) := \int v(y) Q(\mathrm{d}y \mid x, a) - v(x).$$

To begin with, observe that for any bounded measurable function v on X, the sequence

$$M_n(v) := v(x_n) - \sum_{t=0}^{n-1} Tv(x_t, a_t), \;\; n \geq 0,$$

with $M_0(v) := v(x_0)$, is a P_ν^δ-martingale with respect to the σ-algebra generated by the history h_n (introduced in Definition 3.2). Thus, in particular, for all n, $E_\nu^\delta v(x_0) = E_\nu^\delta M_n(v)$, i.e.,

$$\int v \, \mathrm{d}\nu = E_\nu^\delta v(x_n) - n \int (Tv) \, \mathrm{d}\mu_n \;\; \forall n, \tag{4.18}$$

where μ_n is the p.m. in (4.14). Observe also that, by Assumption 2.1 (a), Tv is a continuous and bounded function on K if $v \in C(X)$. Finally, in (4.18), let $v \in C(X)$, replace $\{\mu_n\}$ by the weakly convergent subsequence $\{\mu_{n_i}\}$ in (4.16), and then divide by n_i and let $i \longrightarrow \infty$ to obtain (4.17). This completes the proof of Proposition 4.6. □

As a corollary of Proposition 4.6, the number j^* in (3.4) satisfies

$$j^* = \inf\{J(\varphi, p^\varphi) \mid \varphi \in \Phi_0\}, \tag{4.19}$$

where Φ_0 is the set of all *stable* relaxed policies.

5. MAIN RESULTS

In this section we state our main results; their proofs are collected in § 6. *We assume throughout that Assumptions* 2.1 *and* (3.3) *hold.*

Theorem 5.1. (a) There exists a stable relaxed policy $\varphi^* \in \Phi_0$ such that $(\varphi^*, p^{\varphi^*})$ is a minimum pair, i.e.

$$J\left(\varphi^*, p^{\varphi^*}\right) = j^*. \tag{5.1}$$

(b) If the policy $\varphi^* \in \Phi_0$ in (a) is such that $Q(\cdot \mid \cdot, \varphi^*)$ is λ-recurrent for some σ-finite measure λ on X, then φ^* is AC-optimal and pathwise AC-optimal, i.e. for any initial distribution $\nu \in \mathcal{P}(X)$,

$$J^0(\varphi^*, \nu) = j^* \quad P_\nu^{\varphi^*}\text{-a.s.}, \tag{5.2}$$

and

$$J(\varphi^*, \nu) = j^*. \tag{5.3}$$

The optimality inequality. The existence of an AC-optimal policy is sometimes based on the following well-known result [7, 15, 16], stated here for completeness and for comparison with our results.

Proposition 5.2. If there exists a relaxed policy φ and a measurable function h on X, bounded from below, and such that

$$j^* + h(x) \geq c(x, \varphi) + \int h(y) Q(\mathrm{d}y \mid x, \varphi) \quad \forall x, \tag{5.4}$$

then φ is AC-optimal; in fact, any policy $\varphi \in \Phi$ that satisfies (5.4) is AC-optimal.

Indeed, iteration of (5.4) yields

$$\begin{aligned} nj^* + h(x) &\geq E_x^\varphi \sum_{t=0}^{n-1} c(x_t, \varphi) + E_x^\varphi h(x_n) \\ &\geq J_n(\varphi, x) + L, \end{aligned} \tag{5.5}$$

where L is a lower bound for $h(\cdot)$. This implies $j^* \geq J(\varphi, x)\ \forall\, x$ and, therefore, the AC-optimality of φ follows from (3.4).

It turns out that (5.4) is "almost" equivalent to the existence of a stable minimum pair, in the following sense.

Theorem 5.3. (a) Let $\varphi \in \Phi_0$ be a stable relaxed policy with an invariant p. m. p^φ. Then (φ, p^φ) is a minimum pair if and only if there exists a nonnegative measurable function h on X such that h and φ satisfy (5.4) for p^φ-a. a. $x \in X$.

(b) If $\varphi \in \Phi_0$ is such that $Q(\cdot \mid \cdot, \varphi)$ is λ-recurrent, then φ is AC-optimal if and only if there exists a nonnegative measurable function h on X such that (5.4) holds for all $x \in X$.

Combining Theorems 5.1 and 5.3 we obtain the following.

Corollary 5.4. (a) There exists a stable relaxed policy φ^*, with invariant p. m. p^{φ^*}, and a nonnegative measurable function h on X such that

$$j^* + h(x) \geq c(x, \varphi^*) + \int h(y) Q(\mathrm{d}y|x, \varphi^*), \quad p^{\varphi^*}\text{-a. a. } x \in X. \tag{5.6}$$

Moreover, there is a deterministic stationary policy $f^* \in \mathsf{F}$ such that (using the notation (4.3))

$$j^* + h(x) \geq c(x, f^*) + \int h(y) Q(\mathrm{d}y|x, f^*), \quad p^{\varphi^*}\text{-a. a. } x \in X. \tag{5.7}$$

(b) If, in addition, φ^* is such that $Q(\cdot \mid \cdot, \varphi^*)$ is λ-recurrent, then (5.6)–(5.7) hold for all $x \in X$; hence (by Proposition 5.2) both φ^* and the deterministic policy $f^* \in \mathsf{F}$ in (5.7) are AC-optimal.

To obtain the deterministic stationary policy f^* in (5.7), starting from (5.6), it suffices to apply the following (slight) generalization of Blackwell's theorem [4] p. 864 (which is the same as the Lemma in [5] p. 228):

Lemma 5.5. (Blackwell). Let $v : \mathsf{K} \to \mathsf{R}$ be a measurable function, and $\varphi \in \Phi$ a relaxed policy such that the map

$$x \to v(x, \varphi) := \int_A v(x, a) \varphi(\mathrm{d}a \mid x)$$

if finite-valued. Then there exists a deterministic stationary policy $f \in \mathsf{F}$ such that

$$v(x, \varphi) \geq v(x, f(x)) \quad \forall\, x \in X.$$

Thus if we write the right-hand side of (5.6) as

$$\int_A \left[c(x, a) + \int_X h(y) Q(\mathrm{d}y|x, a) \right] \varphi^*(\mathrm{d}a|x) =: \int_A v(x, a) \varphi^*(\mathrm{d}a|x),$$

then the existence of $f^* \in \mathsf{F}$ satisfying (5.7) follows from Lemma 5.5. Finally, part (b) in Corollary 5.4 follows directly from the assumption of λ-recurrence, as in the proof below of Theorem 5.3 (b).

Remark 5.6. (5.4) implies

$$j^* + h(x) \geq \inf_{a \in A(x)} \left[c(x,a) + \int h(y) Q(\mathrm{d}y|x,a) \right]. \tag{5.8}$$

This is the so-called *optimality inequality*. If $f \in \mathsf{F}$ is such that $f(x) \in A(x)$ attains the minimum in (5.8) for all $x \in X$, then Proposition 5.2 yields that f is AC-optimal. However, the existence of such an f is *not* ensured in general, unless we strengthen the hypotheses on c, Q and $A(\cdot)$. This is the reason why to obtain $f^* \in \mathsf{F}$ satisfying (5.7) we had to resort to Blackwell's theorem (Lemma 5.5).

In the next section we prove Theorems 5.1 and 5.3; however, the reader may wish to read first the applications in § 7.

6. PROOFS

Proof of Theorem 5.1. (a) Recall that, by (4.19), the search for a minimum pair may be restricted to policies $\varphi \in \Phi_0$.

Let $0 < \varepsilon_n < 1$ be a sequence of numbers such that $\varepsilon_n \downarrow 0$ and, for each n, let $\varphi_n \in \Phi_0$ be such that

$$J(\varphi_n, p^{\varphi_n}) = \int c \, \mathrm{d}\gamma_n \leq j^* + \varepsilon_n \tag{6.1}$$

where γ_n is the p. m. on $X \times A$, concentrated on K, such that

$$\gamma_n(B \times C) := \int_B \varphi_n(C|x) p^{\varphi_n}(\mathrm{d}x) \;\; \forall B \in \mathcal{B}(X),\, C \in \mathcal{B}(A).$$

Thus, since $\sup_n \int c \, \mathrm{d}\gamma_n \leq j^* + 1$ and c is a moment (see Remarks 2.4 (a), (b), (c)), there is a subsequence $\{\gamma_{n_i}\}$ of $\{\gamma_n\}$ converging weakly to a p. m. γ^* on $X \times A$, concentrated on K, and such that

$$j^* \geq \liminf_i \int c \, \mathrm{d}\gamma_{n_i} \geq \int c \, \mathrm{d}\gamma*, \tag{6.2}$$

where the first inequality comes from (6.1) and the second from the Remark 2.4 (d). Finally, decompose γ^* as in the proof of Proposition 4.6, parts (ii) and (iii), i.e. $\gamma^*(\mathrm{d}x, \mathrm{d}a) = \varphi^*(\mathrm{d}a|x) p^{\varphi^*}(\mathrm{d}x)$, where $\varphi^* \in \Phi_0$, to obtain, from (6.2) and (4.6),

$$j^* \geq \int c(x, \varphi^*) p^{\varphi^*}(\mathrm{d}x) = J(\varphi^*, p^{\varphi^*}).$$

This yields (5.1).

(b) If the policy φ^* in part (a) is λ-recurrent, then (5.2)–(5.3) follow from (5.1) and (4.9). This completes the proof of Theorem 5.1. □

Proof of Theorem 5.3. (a) (*Sufficiency.*) Suppose that $\varphi \in \Phi_0$ and $h(\cdot)$ satisfy (5.4) for p^φ–a. a. $x \in X$; that is,

$$j^* + h(x) \geq c(x,\varphi) + \int h(y)Q(\mathrm{d}y|x,\varphi) \quad p^\varphi\text{–a. a. } x \in X. \tag{6.3}$$

Integrating with respect to p^φ yields, by (4.4),

$$j^* + \int h \,\mathrm{d}p^\varphi \geq \int c(x,\varphi)p^\varphi(\mathrm{d}x) + \int h\,\mathrm{d}p^\varphi,$$

which combined with (4.6) implies $j^* \geq J(\varphi, p^\varphi)$. Thus (φ, p^φ) is a minimum pair.

(*Necessity.*) Conversely, suppose that $\varphi \in \Phi_0$ is such that (φ, p^φ) is a minimum pair, so that, from (4.6) and Proposition 4.5,

$$J(\varphi,x) = \lim_n n^{-1}J_n(\varphi,x) = j^* \quad \text{for } p^\varphi\text{–a. a. } x \in X. \tag{6.4}$$

Now, define $h_0 := J_0 := 0$, and for $n = 1, 2, \ldots$, $x \in X$,

$$\begin{aligned} j_n(x) &:= J_n(\varphi,x) - J_{n-1}(\varphi,x), \\ M_n &:= \inf_x J_n(\varphi,x), \\ h_n(x) &:= J_n(\varphi,x) - M_n \ (\geq 0), \\ h(x) &:= \liminf_m m^{-1}\sum_{n=1}^{m-1} h_n(x). \end{aligned} \tag{6.5}$$

Notice that $h(\cdot) \geq 0$, and $\sum_{n=1}^{m} j_n(x) = J_m(\varphi,x)$. On the other hand, by the Markov property,

$$J_n(\varphi,x) = c(x,\varphi) + \int J_{n-1}(y,\varphi)Q(\mathrm{d}y\,|\,x,\varphi),$$

which is equivalent to

$$j_n(x) + h_{n-1}(x) = c(x,\varphi) + \int h_{n-1}(y)Q(\mathrm{d}y\,|\,x,\varphi).$$

This yields, summing over $n = 1, \ldots, m$,

$$J_m(\varphi,x) + \sum_{n=1}^{m-1} h_n(x) = mc(x,\varphi) + \int \sum_{n=1}^{m-1} h_n(y)Q(\mathrm{d}y|x,\varphi).$$

Finally, divide by m and take liminf as $m \to \infty$; thus (6.4) and Fatou's Lemma yield

$$j^* + h(x) \geq c(x,\varphi) + \int h(y)Q(\mathrm{d}y|x,\varphi) \ p^\varphi\text{–a. a. } x. \tag{6.6}$$

This completes the proof of Theorem 5.3 (a).

(b) If (5.4) holds for all $x \in X$, then φ is AC-optimal: see Proposition 5.2. Conversely, suppose now that $\varphi \in \Phi_0$ is such that $Q(\cdot \mid \cdot, \varphi)$ is λ-recurrent. If, moreover, φ is AC-optimal, then, by (4.6) and (4.9), the limit in (6.4) holds for *all* $x \in X$. Therefore, the argument from (6.4) to (6.6) holds for all $x \in X$; thus (5.4) (or (6.6)) holds for all $x \in X$. □

7. APPLICATIONS AND FURTHER COMMENTS

Example 7.1. Consider the LQ system (1.1) - (1.2), which, for ease of reference, is repeated here:

$$x_{t+1} = \alpha x_t + \beta a_t + \xi_t, \quad t = 0, 1, \ldots; x_0 \text{ given}, \tag{7.1}$$

$$c(x, a) := x'\gamma x + a'\theta a. \tag{7.2}$$

As in § 1, we let $X := \mathsf{R}^p$, $A(\cdot) \equiv A := \mathsf{R}^q$; γ and θ are symmetric and positive definite matrices. If the initial state x_0 is random, then we assume that it is independent of the i. i. d. disturbances ξ_t. As already noted (see Example 2.2) Assumption 2.1 (a), (b), (c) trivially hold in this case. Thus, if Assumption 3.3 holds, then with *no further hypotheses* whatsoever, Theorem 5.1 (a) yields the existence of a *stable* relaxed policy φ^* such that $(\varphi^*, p^{\varphi^*})$ is a minimum pair; furthermore, Corollary 5.4 (a), together with Proposition 5.2, yields a deterministic stationary policy $f^* \in \mathsf{F}$ such that

$$J(f^*, x) = j^* \text{ for } p^{\varphi^*}\text{-a. a. } x \in X.$$

Sufficient conditions for Assumption 3.3 can be derived from Example 7.3 below (see Proposition 7.6).

Now, to get the stronger results in Theorems 5.1 (b), 5.3 (b) and Corollary 5.4 (b), we need $Q(\cdot \mid \cdot, \varphi^*)$ to be λ-recurrent, for some σ-finite measure λ on $X = \mathsf{R}^p$. So, let (e. g.) λ stand for the Lebesgue measure on X, and suppose:

Assumption 7.2. The random vectors ξ_t are absolutely continuous with a density μ which is positive λ-a.e. (such as, for instance, a Gaussian density).

Then, as is well-known [8, 10, 25], $Q(\cdot \mid \cdot, \varphi^*)$ is λ-recurrent and, therefore, *all* the results in § 5 are applicable.

Example 7.3. Let us consider again the quadratic cost (7.2), but (7.1) is now replaced by a nonlinear (autoregressive-like) system

$$x_{t+1} = G(x_t, a_t) + \xi_t. \tag{7.3}$$

The control constraint sets $A(x) \subset \mathsf{R}^q$ are assumed to be (nonempty) closed sets and such that K — defined in (2.1) — is convex. If, in addition, we suppose:

Assumption 7.4. $G : \mathsf{K} \to X$ is continuous,
then Assumption 2.1 holds.

To obtain Assumption 3.3 and λ-recurrence, let us suppose the following (essentially "growth") conditions.

Assumption 7.5. (a) $G(x, \varphi) := \int G(x, a)\varphi(\mathrm{d}a|x)$ is locally bounded for every $\varphi \in \Phi$;

(b) For some constant $m > 0$, $G(x, a)'\gamma G(x, a) \le mc(x, a)$ for all $(x, a) \in \mathsf{K}$, where γ in the coefficient matrix in (7.2);

(c) Assumption 7.2 holds and, also, $E(\xi_0) = 0$ and $E\,|\,\xi_0\,|^2 < \infty$;

(d) There is a relaxed policy $\widehat{\varphi} \in \Phi$ for which the following holds: There are positive constants $\rho < 1$, k_1, k_2 such that

(d$_1$) $E\,|\,G(x, \widehat{\varphi}) + \xi_0|^2 \le \rho|x|^2 \quad \forall\,|x| \ge k_1$, and

(d$_2$) $\int (a'\theta a)\widehat{\varphi}(\mathrm{d}a|x) \le k_2|x|^2 \quad \forall x.$

Then standard results on ergodicity of time series [8, 25] yield Assumption 3.3 and λ-recurrence. More precisely, we have:

Proposition 7.6 [8, 25]. (a) If Assumption 7.5 (a) and (c) hold, then $Q(\cdot\,|\,\cdot, \varphi)$ is (aperiodic and) λ-recurrent $\forall \varphi \in \Phi$;

(b) If, moreover, Assumptions 7.5 (b) and (d) hold, then when using the policy $\widehat{\varphi}$, the state (Markov) process $\{x_t\}$ is geometrically ergodic and its unique invariant p. m., say $\widehat{p}$, has a finite second moment, i. e. $\int |x|^2\widehat{p}(\mathrm{d}x) < \infty$.

It goes without saying that Assumption 7.5 was specially designed for the additive-noise (or autoregressive-like) system (7.3). If we have instead a general MCP, say as in (2.3), sufficient conditions for Assumption 3.3 and λ- (or Harris-) recurrence, may be obtained in a number of ways [10, 18, 24, 26].

We conclude with a few remarks on the "optimality inequality" (5.8), which was derived from (5.4). As shown in Theorem 5.3, the inequality (5.4) is virtually equivalent to the existence of a minimum pair or an AC-optimal policy. The question is: is it possible to have *equality* in (5.4) or in (5.8)? Some authors have found the answer to be affirmative in settings much more *restrictive* than our Assumptions 2.1 and 3.3: even for linear systems [12] or for denumerable state/compact action sets [7], *additional* hypotheses are required to obtain the "optimality equation". It would be interesting to investigate conditions under which, in our general MCP, equality holds in (5.4), i. e. the *Poisson equation* [10, 26]

$$j^* + h(x) = c(x, \varphi) + \int h(y)Q(\mathrm{d}y\,|\,x, \varphi) \quad \forall\, x \in X$$

would be obtained. This would yield not only that φ is AC-optimal, but also that the transition kernel $Q(\cdot\,|\,\cdot,\varphi)$ satisfies nice recurrence properties, such as Doeblin's condition.

(Received March 19, 1992.)

REFERENCES

[1] D. P. Bertsekas: Dynamic Programming: Deterministic and Stochastic Models. Prentice-Hall, Englewood Cliffs, N. J. 1987.

[2] D. P. Bertsekas and S. E. Shreve: Stochastic Optimal Control: The Discrete Time Case. Academic Press, New York 1978.

[3] P. Billingsley: Convergence of Probability Measures. Wiley, New York 1968.

[4] D. Blackwell: Memoryless strategies in finite-stage dynamic programming. Ann. Math. Statist. *35* (1964), 863–865.

[5] D. Blackwell: Discounted dynamic programming. Ann. Math. Statist. *36* (1965), 226–235.

[6] V. S. Borkar: Control of Markov chains with long-run average cost criterion: the dynamic programming equations. SIAM J. Control Optim. *27* (1989), 642–657.

[7] R. Cavazos-Cadena: Solution to the optimality equation in a class of average Markov decision chains with unbounded costs. Kybernetika *27* (1991), 23–37.

[8] J. Diebolt and D. Guégan: Probabilistic properties of the general nonlinear markovian process of order one and applications to time series modelling. Rapport Technique No. 125, Laboratoire de Statistique Théorique et Appliquée, CNR-URA 1321, Université Paris VI, 1990.

[9] J. L. Doob: Stochastic Processes. Wiley, New York 1953.

[10] M. Duflo: Méthodes Récursives Aléatoires. Masson, Paris 1990.

[11] E. B. Dynkin and A. A. Yushkevich: Controlled Markov Processes. Springer-Verlag, Berlin 1979.

[12] R. Hartley: Dynamic programming and an undiscounted, infinite horizon, convex stochastic control problem. In: Recent Developments in Markov Decision Processes (R. Hartley, L. C. Thomas and D. J. White, eds.). Academic Press, London 1980, pp. 277–300.

[13] O. Hernández-Lerma: Lyapunov criteria for stability of differential equations with Markov parameters. Boletín Soc. Mat. Mexicana *24* (1979), 27–48.

[14] O. Hernández-Lerma: Adaptive Markov Control Processes. Springer-Verlag, New York 1989.

[15] O. Hernández-Lerma: Average optimality in dynamic programming on Borel spaces — unbounded costs and controls. Syst. Control Lett. *17* (1991), 237–242.

[16] O. Hernández-Lerma and J. B. Lasserre: Average cost optimal policies fot Markov control processes with Borel state space and unbounded costs. Syst. Control Lett. *15* (1990), 349–356.

[17] O. Hernández-Lerma and J. B. Lasserre: Linear programming and average optimality of Markov control processes on Borel spaces — unbounded costs. Rapport LAAS, LAAS-CNRS, Toulouse 1992. To appear in SIAM J. Control Optim.

[18] 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.

[19] K. Hinderer: Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter. Springer-Verlag, Berlin 1970.

[20] M. Yu. Kitayev: Semi-Markov and jump Markov control models: average cost criterion. Theory Probab. Appl. *30* (1985), 272–288.

[21] M. Kurano: The existence of a minimum pair of state and policy for Markov decision processes under the hypothesis of Doeblin. SIAM J. Control Optim. *27* (1989), 296–307.

[22] H. J. Kushner: Introduction to Stochastic Control. Holt, Rinehart and Winston, New York 1971.

[23] A. Leizarowitz: Optimal controls for diffusions in $\mathbb{R}^n$. J. Math. Anal. Appl. *149* (1990), 180–209.

[24] S. P. Meyn: Ergodic theorems for discrete time stochastic systems using a stochastic Lyapunov function. SIAM J. Control Optim. *27* (1989), 1409–1439.

[25] A. Mokkadem: Sur un modèle autorégressif nonlinéaire. Ergodicité et ergodicité géométrique. J. Time Series Anal. *8* (1987), 195–205.

[26] D. Revuz: Markov Chains. Second edition. North-Holland, Amsterdam 1984.

[27] U. Rieder: Measurable selection theorems for optimization problems. Manuscripta Math. *24* (1978), 507–518.

[28] V. I. Rotar' and T. A. Konyuhova: Two papers on asymptotic optimality in probability and almost surely. Preprint, Central Economic Mathematical Institute (CEMI), Moscow 1991.

[29] R. H. Stockbridge: Time-average control of martingale problems: a linear programming formulation. Ann. Probab. *18* (1990), 206–217.

[30] J. Wijngaard: Existence of average optimal strategies in markovian decision problems with strictly unbounded costs. In: Dynamic Programming and Its Applications (M. L. Puterman, ed.), Academic Press, New York 1978, pp. 369–386.

[31] K. Yosida: Functional Analysis. Fifth edition. Springer-Verlag, Berlin 1978.

Prof. Dr. Onésimo Hernández-Lerma, Departamento de Matemáticas, CINVESTAV-IPN, A. Postal 14-740, 07000 México, D.F. Mexico.