Onésimo Hernández-Lerma
Existence of average optimal policies in Markov control processes with strictly unbounded costs

*Kybernetika*, Vol. 29 (1993), No. 1, 1--17

Persistent URL: [http://dml.cz/dmlcz/124548](http://dml.cz/dmlcz/124548)

**Terms of use:**

© Institute of Information Theory and Automation AS CR, 1993

Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these Terms of use.

This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project [DML-CZ: The Czech Digital Mathematics Library](http://project.dml.cz)
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

$$x_{t+1} = ax_t + ba_t + \xi_t, \quad t = 0, 1, \ldots$$

(1.1)

and the quadratic one-stage cost

$$c(x, a) := x'\gamma x + a'\theta a,$$

(1.2)

where "prime" denotes transpose. In (1.1)–(1.2), the state and control (or action) spaces are $X := \mathbb{R}^p$ and $A := \mathbb{R}^q$ respectively, and the $\xi_t$ are i.i.d. (independent and identically distributed) random disturbances. $a$, $b$, $\gamma$ and $\theta$ are matrices of appropriate dimensions, with $\gamma$ and $\theta$ 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

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

(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))

$$\inf_{|x| > n} \inf_{a} \inf_{t} c(x, a) \to \infty \quad \text{as} \quad n \to \infty.$$  

(1.4)
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 $\sigma$-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 $P(\cdot | Z)$ such that $P(\cdot | Z)$ is a p.m. on $Y$ for each fixed $Z \in Z$, and $P(B | Z)$ 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.

\[ K := \{(x, a) | x \in X, a \in A(x)\} \quad (2.1) \]

is a Borel subset of \(X \times A\). The transition law \(Q\) is a stochastic kernel on \(X\) given \(K\), i.e. \(Q \in \mathcal{P}(X|K)\). 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.,

\[ \int u(y)Q(dy | x, a) \]

is a continuous and bounded function in \((x, a) \in K\) 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 \(K_n \uparrow K\) such that

\[ \liminf_n \{ c(x, a) | (x, a) \notin K_n \} = +\infty. \quad (2.2) \]

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

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

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

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

\[ \int_X u(y)Q(dy | x, a) = E[u(x_{t+1}) | x_t = x, a_t = a] \]

\[ = \int_S u[G(x, a, s)] \mu(ds). \]

Clearly, \(Q\) satisfies Assumption 2.1(a) if \(G(x, a, s)\) is continuous in \((x, a) \in K\) 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 \(\theta\) 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

$$
\liminf_{y \in Y_n} v(y) = +\infty.
$$

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

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

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

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 Prokhorov’s Theorem [3, p.37], which states the following: If $M \subseteq \mathcal{P}(Y)$ 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 $\{\mu_{n_k}\}$ such that, for some p.m. $\mu$ on $Y$,

$$
\lim_{k \to \infty} \int u \, d\mu_{n_k} = \int u \, d\mu \quad \forall u \in C(Y).
$$

(2.4)

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

$$
\liminf_{n} \int v \, d\mu_n \geq \int v \, d\mu.
$$

(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 $k$,

$$
\liminf_{n} \int v \, d\mu_n \geq \liminf_{n} \int v^k \, d\mu_n = \int v^k \, d\mu.
$$

Letting $k \to \infty$ 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 \( \Phi \) stands for the set of all stochastic kernels \( \varphi \in \mathcal{P}(A|X) \) such that \( \varphi(A|x) = 1 \) for all \( x \in X \).

A function \( f \in \Phi \) will be identified with the stochastic kernel \( \varphi \in \Phi \) such that \( \varphi(x|\cdot) \) is the p.m. concentrated at \( f(x) \) for all \( x \in X \). Thus \( F \subset \Phi \). By Assumption 2.1 (c), the set \( K \) in (2.1) contains the graph of a function \( f \in F \) (see e.g. [27] Example 2.6). In other words, the set \( F \) (hence \( \Phi \) 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 \( t = 0,1,\ldots \), \( \delta_t(h_t) \) 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 \( \Delta \). A policy \( \delta = \{\delta_t\} \) is said to be a:

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

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

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

Let \( (\Omega, \mathcal{F}) \) be the measurable space consisting of the sample space \( \Omega := (X \times A) \times \omega \) and the corresponding product \( \sigma \)-algebra \( \mathcal{F} \). Then for each policy \( \delta \in \Delta \) and an initial distribution \( \nu \in \mathcal{P}(X) \), a probability measure \( P_\nu^\delta \) and a stochastic process \( \{x_t, a_t, t = 0,1,\ldots\} \) are defined on \( \Omega \) 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 \( P_\nu^\delta \) is written \( E_\nu^\delta \). If \( \nu \) is concentrated at (the initial state) \( x_0 = x \), then we write \( P_\nu^\delta \) and \( E_\nu^\delta \) as \( P_x^\delta \) and \( E_x^\delta \), respectively.

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

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

(3.1)

Then the long-run expected average cost (AC) per unit time incurred by the policy \( \delta \), given the initial distribution \( \nu \), is given by

\[
J(\delta, \nu) := \limsup_n n^{-1} J_n(\delta, \nu).
\]  

(3.2)

Similarly, the pathwise AC is given by

\[
J^p(\delta, \nu) := \limsup_n n^{-1} \sum_{t=0}^{n-1} c(x_t, a_t).
\]  

(3.3)
Finally, let

\[ j^* := \inf_{\delta} \inf_{\nu} J(\delta, \nu). \]  

(3.4)

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

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

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 \(\delta^*\) be a policy and \(\nu^*\) an initial distribution. Then:

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

(b) \(\delta^*\) is said to be **AC-optimal** if 

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

and **pathwise AC-optimal** if 

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

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

\[ Q(\cdot | x, \varphi) := \int_A Q(\cdot | x, a) \varphi(da | x), \quad x \in X. \]  

(4.1)

We will also write 

\[ c(x, \varphi) := \int_A c(x, a) \varphi(da | x). \]  

(4.2)

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

\[ Q(\cdot | x, f) := Q(\cdot | x, f(x)), \quad \text{and} \quad c(x, f) := c(x, f(x)) \]  

(4.3)

respectively.

**Definition 4.1.** A relaxed policy \(\varphi \in \Phi\) 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.

\[ p^\varphi(\cdot) = \int_X Q(\cdot | x, \varphi)p^\varphi(dx); \]  

(4.4)

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

\[ J(\varphi, p^\varphi) = \int_X c(x, \varphi)p^\varphi(dx). \]  

(4.5)

The family of all stable relaxed policies is denoted by \(\Phi_0\).
Remark 4.2. (a) If $\varphi \in \Phi_0$, the invariant p.m. $p^\varphi$ 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 $p^\varphi$ as in (4.4)-(4.5), then by the Individual Ergodic Theorem [31, p. 388], the limit (cf. (3.1), (3.2))

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

exists for $p^\varphi$-a.e. (almost all) $x \in X$ and satisfies

$$\int J(\varphi, x)p^\varphi(dx) = \int c(x, \varphi)p^\varphi(dx) = J(\varphi, p^\varphi)$$

(4.6b)

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

$$J^0(\varphi, p^\varphi) = \lim n^{-1} \sum_{i=0}^{n-1} c(x_i, \varphi) \quad \text{a.s. (almost surely).}$$

(4.7)

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

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

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

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

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

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

$$L^\varphi(x, B) = 1 \quad \forall x \in X.$$  

(4.8)

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

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

$$J^0(\varphi, \nu) = \lim n^{-1} \sum_{i=0}^{n-1} c(x_i, \varphi)$$

$$= \int c(x, \varphi)p^\varphi(dx) \quad \text{P}^\varphi_n \text{-a.e.}$$

(4.9a)

and

$$J(\varphi, \nu) = E^\nu J^0(\varphi, \nu) = \int c(x, \varphi)p^\varphi(dx).$$

(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. $\nu^*$. Then $J(\varphi^*, \nu^*)$ is a minimum pair, i.e.

$$J(\varphi^*, \nu^*) = j^*$$

if and only if

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

$$j^* = \int_B J(\varphi^*, x)\nu^*(dx) + j^*\nu^*(B^c)$$

or, equivalently,

$$\int_B J(\varphi^*, x)\nu^*(dx) = j^*\nu^*(B).$$

This implies $\nu^*(B) = 0$. \hfill \Box

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

$$J(\delta, \nu) \geq J(\varphi, \nu^*).$$  \hfill (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. $\mu$ on $X \times A$ concentrated on $K$ such that

$$J(\delta, \nu) \geq \int c \, d\mu;$$  \hfill (4.13)

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

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

thus we may rewrite (4.13) as

$$J(\delta, \nu) \geq \int_X c(x, \varphi)\bar{\mu}(dx);$$
(iii) The relaxed policy $\varphi$ in (ii) is stable and $\bar{\mu} := \mu^\varphi$ is an invariant p.m. for $Q(\cdot | \cdot, \varphi)$.

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

$$\mu_n(\Gamma) := n^{-1} \sum_{t=0}^{n-1} P^t_{\varphi}(x_t, a_t) \in \Gamma, \quad \Gamma \in \mathcal{B}(X \times A).$$

(4.14)

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

$$J(\delta, \nu) = \limsup_n \int c \, d\mu_n.$$

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

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

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 $\{\mu_{n_j}\}$ of $\{\mu_n\}$ converging weakly to a p.m. $\mu$ on $X \times A$. Furthermore, since each $\mu_n$ is concentrated on $K$, so is $\mu$.

Finally, from (4.15),

$$J(\delta, \nu) \geq \liminf_n \int c \, d\mu_n \geq \int c \, d\mu,$$

(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

$$\int T_v(x, a) \mu(dx, da) = \int T_v(x, \varphi) \bar{\mu}(dx) = 0 \quad \forall v \in C(X),$$

(4.17)

where

$$T_v(x, a) := \int v(y)Q(dy | x, a) - v(x).$$

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

$$M_n(v) := v(x_0) - \sum_{t=0}^{n-1} T_v(x_t, a_t), \quad n \geq 0,$$

with $M_0(v) := v(x_0)$, is a $P_\varphi^t$-martingale with respect to the $\sigma$-algebra generated by the history $h_n$ (introduced in Definition 3.2). Thus, in particular, for all $n$, $E^\varphi_n v(x_0) = E^\varphi_n M_n(v)$, i.e.,

$$\int v \, d\mu = E^\varphi_n v(x_n) - n \int (Tv) \, d\mu_n \quad \forall n,$$

(4.18)
where \( \mu_n \) is the p.m. in (4.14). Observe also that, by Assumption 2.1(a), \( T\nu \) is a continuous and bounded function on \( K \) if \( \nu \in C(X) \). Finally, in (4.18), let \( \nu \in C(X) \), replace \( \{\mu_n\} \) by the weakly convergent subsequence \( \{\mu_{n_i}\} \) in (4.16), and then divide by \( n_i \) and let \( i \to \infty \) 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

\[
j^* = \inf \{ J(\phi, p^{\phi}) \mid \phi \in \Phi_0 \},
\]

where \( \Phi_0 \) 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 \( \phi^* \in \Phi_0 \) such that \((\phi^*, p^{\phi^*})\) is a minimum pair, i.e.

\[
J(\phi^*, p^{\phi^*}) = j^*.
\]

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

\[
J^0(\phi^*, \nu) = J^* \quad \text{P}^{\phi^*} - \text{a.s.},
\]

and

\[
J(\phi^*, \nu) = j^*.
\]

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 \( \phi \) and a measurable function \( h \) on \( X \), bounded from below, and such that

\[
j^* + h(x) \geq c(x, \phi) + \int h(y) Q(dy \mid x, \phi) \quad \forall x,
\]

then \( \phi \) is AC-optimal; in fact, any policy \( \phi \in \Phi \) that satisfies (5.4) is AC-optimal. Indeed, iteration of (5.4) yields

\[
nj^* + h(x) \geq E_x^{n-1} \sum_{i=0}^{n-1} c(x_i, \phi) + E^\nu h(x_n)
\]

\[
\geq J_n(\phi, x) + L,
\]
where $L$ is a lower bound for $h(\cdot)$. This implies $j^* \geq J(\varphi, x) \forall x$ and, therefore, the AC-optimality of $\varphi$ 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. $p^\varphi$. Then $(\varphi, p^\varphi)$ is a minimum pair if and only if there exists a nonnegative measurable function $h$ on $X$ such that $h$ and $\varphi$ satisfy (5.4) for $p^\varphi$-a.a. $x \in X$.

(b) If $\varphi \in \Phi_0$ is such that $Q(\cdot \mid , \varphi)$ is $\lambda$-recurrent, then $\varphi$ 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 $\varphi^*$, with invariant p.m. $p^{\varphi^*}$, and a nonnegative measurable function $h$ on $X$ such that

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

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

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

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

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

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

$$x \rightarrow v(x, \varphi) := \int_A v(x,a)\varphi(da \mid x)$$

if finite-valued. Then there exists a deterministic stationary policy $\pi \in F$ such that

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

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

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

then the existence of $\pi^* \in F$ satisfying (5.7) follows from Lemma 5.5. Finally, part (b) in Corollary 5.4 follows directly from the assumption of $\lambda$-recurrence, as in the proof below of Theorem 5.3(b).
Remark 5.6. (5.4) implies
\[ j^* + h(x) \geq \inf_{a \in A(x)} \left[ c(x, a) + \int h(y) Q(dy|x, a) \right]. \] (5.8)

This is the so-called optimality inequality. If \( f \in F \) 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 \( f^* \in F \) 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
\[ J(\varphi_n, p^{\varphi_n}) = \int c \, d\gamma_n \leq j^* + \varepsilon_n \] (6.1)
where \( \gamma_n \) is the p.m. on \( X \times A \), concentrated on \( K \), such that
\[ \gamma_n(B \times C) := \int_B \varphi_n(C|x) p^{\varphi_n}(dx) \quad \forall B \in \mathcal{B}(X), C \in \mathcal{B}(A). \]

Thus, since \( \sup_n \int c \, d\gamma_n \leq j^* + 1 \) 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. \( \gamma^* \) on \( X \times A \), concentrated on \( K \), and such that
\[ j^* \geq \liminf_i \int c \, d\gamma_{n_i} \geq \int c \, d\gamma^*, \] (6.2)
where the first inequality comes from (6.1) and the second from the Remark 2.4(d).

Finally, decompose \( \gamma^* \) as in the proof of Proposition 4.6, parts (ii) and (iii), i.e.
\[ \gamma^*(dx, da) = \nu^*(da|x)p^{\nu^*}(dx), \]
where \( \nu^* \in \Phi_0 \), to obtain, from (6.2) and (4.6),
\[ j^* \geq \int c(x, \nu^*) p^{\nu^*}(dx) = J(\nu^*, p^{\nu^*}). \]

This yields (5.1).

(b) If the policy \( \nu^* \) in part (a) is \( \lambda \)-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 \( p^\varphi \)-a. a. \( x \in X \); that is,

\[
j^* + h(x) \geq c(x, \varphi) + \int h(y)Q(dy|x, \varphi) \quad p^\varphi \text{-a. a. } x \in X.
\] (6.3)

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

\[
j^* + \int h \, dp^\varphi \geq \int c(x, \varphi) \, p^\varphi(dx) + \int h \, dp^\varphi,
\]

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

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

\[
J(\varphi, x) = \lim_{n \to \infty} J_n(\varphi, x) = j^* \quad \text{for } p^\varphi \text{-a. a. } x \in X.
\] (6.4)

Now, define \( h_0 := J_0 := 0 \), and for \( n = 1, 2, \ldots, z \in X \),

\[
\begin{align*}
J_n(x) &:= J_n(\varphi, x) - J_{n-1}(\varphi, x), \\
M_n &:= \inf_{\varphi} J_n(\varphi, x), \\
h_n(x) &:= J_n(\varphi, x) - M_n \geq 0, \\
h(x) &:= \liminf_n m^{-1} \sum_{n=1}^{m-1} h_n(x).
\end{align*}
\] (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,

\[
J_n(\varphi, x) = c(x, \varphi) + \int J_{n-1}(y, \varphi)Q(dy|x, \varphi),
\]

which is equivalent to

\[
j_n(x) + h_{n-1}(x) = c(x, \varphi) + \int h_{n-1}(y)Q(dy|x, \varphi).
\]

This yields, summing over \( n = 1, \ldots, m \),

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

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

\[
j^* + h(x) \geq c(x, \varphi) + \int h(y)Q(dy|x, \varphi) \quad p^\varphi \text{-a. a. } x.
\] (6.6)

This completes the proof of Theorem 5.3(a).
(b) If (5.4) holds for all \( x \in X \), then \( \varphi \) is AC-optimal; see Proposition 5.2. Conversely, suppose now that \( \varphi \in \Phi_0 \) is such that \( Q(\cdot, \cdot, \varphi) \) is \( \lambda \)-recurrent. If, moreover, \( \varphi \) 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:

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

\[
c(x, a) := x'\gamma x + a'\theta a.
\]

As in §1, we let \( X := \mathbb{R}^p \), \( A(\cdot) \equiv A := \mathbb{R}^q \); \( \gamma \) and \( \theta \) 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 \( \xi_t \). 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 \( \varphi^* \) such that \( (\varphi^*, p^{\varphi^*}) \) is a minimum pair; furthermore, Corollary 5.4(a), together with Proposition 5.2, yields a deterministic stationary policy \( \pi^* \in \Pi \) such that

\[
J(\pi^*; x) = j^* \text{ for } \text{p}^{\varphi^*}\text{-a.a. } x \in X.
\]

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 \( Q(\cdot, \cdot, \varphi^*) \) to be \( \lambda \)-recurrernt, for some \( \sigma \)-finite measure \( \lambda \) on \( X = \mathbb{R}^p \). So, let (e.g.) \( \lambda \) stand for the Lebesgue measure on \( X \), and suppose:

Assumption 7.2. The random vectors \( \xi_t \) are absolutely continuous with a density \( \mu \) which is positive \( \lambda \)-a.e. (such as, for instance, a Gaussian density).

Then, as is well-known [8, 10, 25], \( Q(\cdot, \cdot, \varphi^*) \) is \( \lambda \)-recurrernt 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

\[
x_{t+1} = G(x_t, a_t) + \xi_t.
\]

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

O. HERNÁNDEZ-LEIRMA
Existence of Average Optimal Policies in Markov Control Processes

**Assumption 7.4.** \( G : K \rightarrow X \) is continuous, then Assumption 2.1 holds.

To obtain Assumption 3.3 and \( \lambda \)-recurrence, let us suppose the following (essentially “growth”) conditions.

**Assumption 7.5.**
(a) \( G(x, \varphi) := \int G(x,a) \varphi(da|x) \) is locally bounded for every \( \varphi \in \Phi \);
(b) For some constant \( m > 0 \), \( G(x,a) \gamma G(x,a) \leq mc(x,a) \) for all \( (x,a) \in K \), where \( \gamma \) 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 \( \tilde{\varphi} \in \Phi \) for which the following holds: There are positive constants \( \rho < 1, k_1, k_2 \) such that
\[
(d_1) \quad E \left[ G(x, \tilde{\varphi}) + \xi_0 \right]^2 \leq \rho |x|^2 \quad \forall x \geq k_1, \text{ and }
(d_2) \quad \int (a' \theta a) \tilde{\varphi}(da|x) \leq k_2 |x|^2 \quad \forall x.
\]

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

**Proposition 7.6** [8, 25].
(a) If Assumptions 7.5(a) and (c) hold, then \( Q(\cdot | \cdot, \varphi) \) is (aperiodic and) \( \lambda \)-recurrent \( \forall \varphi \in \Phi \);
(b) If, moreover, Assumptions 7.5(b) and (d) hold, then when using the policy \( \tilde{\varphi} \), the state (Markov) process \( \{x_t\} \) is geometrically ergodic and its unique invariant p.m., say \( \tilde{\pi} \), has a finite second moment, i.e. \( \int |x|^2 \tilde{\pi}(dx) < \infty \).

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 \( \lambda \)- (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 restricted 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]

\[
f^* + h(x) = c(x, \varphi) + \int h(y)Q(dy | x, \varphi) \quad \forall x \in X
\]
would be obtained. This would yield not only that $\varphi$ 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


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