A SEMIMARTINGALE CHARACTERIZATION OF AVERAGE OPTIMAL STATIONARY POLICIES FOR MARKOV DECISION PROCESSES

QUANXIN ZHU AND XIANPING GUO

Received 30 November 2004; Revised 10 June 2005; Accepted 22 June 2005

This paper deals with discrete-time Markov decision processes with Borel state and action spaces. The criterion to be minimized is the average expected costs, and the costs may have neither upper nor lower bounds. In our former paper (to appear in Journal of Applied Probability), weaker conditions are proposed to ensure the existence of average optimal stationary policies. In this paper, we further study some properties of optimal policies. Under these weaker conditions, we not only obtain two necessary and sufficient conditions for optimal policies, but also give a “semimartingale characterization” of an average optimal stationary policy.

Copyright © 2006 Q. X. Zhu and X. P. Guo. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

The long-run average expected cost criterion in discrete-time Markov decision processes has been widely studied in the literature; for instance, see [3, 12–14], the survey paper [1], and their extensive references. As is well known, when the state and action spaces are both finite, the existence of average optimal stationary policies is indeed guaranteed [2, 3, 11, 12]. However, when a state space is countably infinite, an average optimal policy may not exist even though the action space is compact [3, 12]. Thus, many authors are interested in finding optimality conditions when a state space is not finite. We now simply describe some existing works. (I) When the costs/rewards are bounded, the minorant condition [3] or the ergodicity condition [5, 6, 8] ensures the existence of a bounded solution to the average optimality equation and of an average optimal stationary policy. Their common ways are via the Banach’s fixed point theorem. (II) When the costs are nonnegative (or bounded below), the optimality inequality approach [1, 9, 10] is used to prove the existence of average optimal stationary policies. A key character of this approach is via the Abelian theorem which requires that the costs have to be nonnegative (or bounded below). In particular, Hernández-Lerma and Lasserre [9] also get the average optimality
equation under the additional equi-continuity condition and give a “martingale character-
ization” of an average optimal stationary policy. (III) For the much more general case,
when the costs have neither upper nor lower bounds, in order to establish the average op-
timality equation and then prove the existence of an average optimal stationary policy,
the equi-continuity condition [4, 9] or the irreducibility condition (e.g., [10, Assump-
tion 10.3.5]) is required. But in [7], we propose weaker conditions under which we prove
the existence of average optimal stationary policies by two optimality inequalities rather
than the “optimality equality” in [4, 9, 10]. Moreover, we remove the equi-continuity
condition used in [4, 9, 10] and the irreducibility condition in [10]. In this paper, we
further study some properties of optimal policies. Under these weaker conditions, we not
only obtain two necessary and sufficient conditions for optimal policies, but also give a
semimartingale characterization of an average optimal stationary policy.

The rest of the paper is organized as follows. In Section 2, we introduce the control
model and the optimality problem that we are concerned with. After optimality condi-
tions and a technical preliminary lemma given in Section 3, we present a semimartingale
characterization of an average optimal stationary policy in Section 4.

2. The optimal control problem

Notation 1. If X is a Borel space (i.e., a Borel subset of a complete and separable metric
space), we denote by \( \mathcal{B}(X) \) its Borel \( \sigma \)-algebra.

In this section, we first introduce the control model

\[
\{ S, (A(x) \subseteq A, x \in S), \ Q(\cdot \mid x, a), \ c(x, a) \} ,
\]

where \( S \) and \( A \) are the state and the action spaces, respectively, which are assumed to be
Borel spaces, and \( A(x) \) denotes the set of available actions at state \( x \in S \). We suppose that
the set

\[
K := \{(x, a) : x \in S, \ a \in A(x) \}
\]

(2.2)
is a Borel subset of \( S \times A \). Furthermore, \( Q(\cdot \mid x, a) \) with \( (x, a) \in K \), the transition law,
is a stochastic kernel on \( S \) given \( K \).

Finally, \( c(x, a) \), the cost function, is assumed to be a real-valued measurable function
on \( K \). (As \( c(x, a) \) is allowed to take positive and negative values, it can also be interpreted
as a reward function rather than a “cost.”)

To introduce the optimal control problem that we are concerned with, we need to
introduce the classes of admissible control policies.

For each \( t \geq 0 \), let \( H_t \) be the family of admissible histories up to time \( t \), that is, \( H_0 := S \),
and \( H_t := K \times H_{t-1} \) for each \( t \geq 1 \).

Definition 2.1. A randomized history-dependent policy is a sequence \( \pi := (\pi_t, \ t \geq 0) \) of
stochastic kernels \( \pi_t \) on \( A \) given \( H_t \) satisfying

\[
\pi_t(A(x) \mid h_t) = 1 \quad \forall h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x) \in H_t, \ t \geq 0.
\]

(2.3)
The class of all randomized history-dependent policies is denoted by $\Pi$. A randomized history-dependent policy $\pi := (\pi_t, t \geq 0) \in \Pi$ is called (deterministic) stationary if there exists a measurable function $f$ on $S$ with $f(x) \in A(x)$ for all $x \in S$, such that

$$\pi_t(\{f(x)\} \mid h_t) = \pi_t(\{f(x)\} \mid x) = 1 \quad \forall h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x) \in H_t, \, t \geq 0.$$  

(2.4)

For simplicity, denote this policy by $f$. The class of all stationary policies is denoted by $F$, which means that $F$ is the set of all measurable functions $f$ on $S$ with $f(x) \in A(x)$ for all $x \in S$.

For each $x \in S$ and $\pi \in \Pi$, by the well-known Tulcea’s theorem [3, 8, 10], there exist a unique probability measure space $(\Omega, \mathcal{F}, P^\pi_x)$ and a stochastic process $\{x_t, a_t, t \geq 0\}$ defined on $\Omega$ such that, for each $D \in \mathcal{B}(S)$ and $t \geq 0$,

$$P^\pi_x(x_{t+1} \in D \mid h_t, a_t) = Q(D \mid x_t, a_t),$$  

(2.5)

with $h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t) \in H_t$, where $x_t$ and $a_t$ denote the state and action variables at time $t \geq 0$, respectively. The expectation operator with respect to $P^\pi_x$ is denoted by $E^\pi_x$.

We now define the $\alpha$-discounted cost ($\alpha$-DC) and the long-run average expected cost (AEC) criteria, respectively, as follows: for each $\pi \in \Pi$, $x \in S$, and $\alpha \in (0, 1)$,

$$V^\alpha_a(x, \pi) := E^\pi_x\left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t)\right], \quad V^*_{\alpha a}(x) := \inf_{\pi \in \Pi} V^\alpha_a(x, \pi);$$  

$$V(x, \pi) := \limsup_{n \to \infty} \frac{E^\pi_x\left[\sum_{t=0}^{n-1} c(x_t, a_t)\right]}{n}, \quad \bar{V}^*(x) = \inf_{\pi \in \Pi} \bar{V}(x, \pi).$$  

(2.6)

Definition 2.2. A policy $\pi^* \in \Pi$ is said to be $\alpha$-DC-optimal if

$$V^\alpha_a(x, \pi^*) = V^*_{\alpha a}(x) \quad \forall x \in S.$$  

(2.7)

An AEC-optimal policy is defined similarly.

The main goal of this paper is to give conditions for a semimartingale characterization of an average optimal stationary policy.

3. Optimality conditions

In this section, we state conditions for a semimartingale characterization of an average optimal stationary policy, and give a preliminary lemma that is needed to prove our main results.
Semimartingale characterization of optimal policies

We will first introduce two sets of hypotheses. The first one, Assumption 3.1, is a combination of a “Lyapunov-like inequality” condition together with a growth condition on the one-step cost $c$.

**Assumption 3.1.** (1) There exist constants $b \geq 0$ and $0 < \beta < 1$ and a (measurable) function $w \geq 1$ on $S$ such that

$$\int_S w(y)Q(dy | x, a) \leq \beta w(x) + b \quad \forall (x, a) \in K. \quad (3.1)$$

(2) There exists a constant $M > 0$ such that $|c(x, a)| \leq Mw(x)$ for all $(x, a) \in K$.

**Remark 3.2.** Assumption 3.1(1) is well known as a Lyapunov-like inequality condition; see [10, page 121], for instance. Obviously, the constant $b$ in (3.1) can be replaced by a bounded nonnegative measurable function $b(x)$ on $S$ as in [10, Assumption 10.2.1(f)].

The second set of hypotheses we need is the following standard continuity-compactness conditions; see, for instance, [7, 12, 13, 15, 16] and their references.

**Assumption 3.3.** (1) For each $x \in S$, $A(x)$ is compact.

(2) For each fixed $x \in S$, $c(x, a)$ is lower semicontinuous in $a \in A(x)$, and the function $\int_S u(y)Q(dy | x, a)$ is continuous in $a \in A(x)$ for each bounded measurable function $u$ on $S$, and also for $u = w$ as in Assumption 3.1.

**Remark 3.4.** Assumption 3.3 is the same as in [10, Assumption 10.2.1]. Obviously, Assumption 3.3 holds when $A(x)$ is finite for each $x \in S$.

To ensure the existence of average optimal stationary policies, in addition to Assumptions 3.1 and 3.3, we give a weaker condition (Assumption 3.5 below). To state this assumption, we introduce the following notation.

For the function $w \geq 1$ in Assumption 3.1, we define the weighted supremum norm $\|u\|_w$ for real-valued functions $u$ on $S$ by

$$\|u\|_w := \sup_{x \in S} \left\{ w(x)^{-1} | u(x) | \right\}, \quad (3.2)$$

and the Banach space $B_w(S) := \{ u : \|u\|_w < \infty \}$.

**Assumption 3.5.** There exist two functions $v_1, v_2 \in B_w(S)$ and some state $x_0 \in S$ such that

$$v_1(x) \leq h_\alpha(x) \leq v_2(x) \quad \forall x \in S, \alpha \in (0, 1), \quad (3.3)$$

where $h_\alpha(x) := V_\alpha^*(x) - V_\alpha^*(x_0)$ is the so-called relative difference of the function $V_\alpha^*(x)$. **Remark 3.6.** Assumption 3.5 is from [7] and it is weaker than [9, Assumption 5.4.1(b)] and [13, Assumption (SEN2), page 132] because the function $v_1(x)$ may not be bounded below, whereas the difference $h_\alpha(x)$ is assumed to be bounded below in [9, 13].
Lemma 3.7. Suppose that Assumptions 3.1, 3.3, and 3.5 hold. Then the following hold.

(a) There exist a unique constant \( g^* \), two functions \( h_k^* \in B_w(S) \) \((k = 1, 2)\), and a stationary policy \( f^* \in F \) satisfying the two optimality inequalities

\[
\begin{align*}
g^* + h_1^*(x) &\leq \min_{a \in A(x)} \left\{ c(x, a) + \int_S h_k^+(y)Q(dy \mid x, a) \right\} \quad \forall x \in S; \quad (3.4) \\
g^* + h_2^*(x) &\geq \min_{a \in A(x)} \left\{ c(x, a) + \int_S h_k^+(y)Q(dy \mid x, a) \right\} \quad \forall x \in S. \quad (3.5)
\end{align*}
\]

(b) \( g^* = \inf_{\pi \in \Pi} V(x, \pi) \) for all \( x \in S \).

(c) Any stationary policy \( f^* \) in \( F \) realizing the minimum of (3.5) is average optimal, and so \( f^* \) in (3.6) is an average optimal stationary policy.

(d) In addition, from the proof of part (b), it yields that for each \( h \in B_w(S) \), \( x \in S \), and \( \pi \in \Pi \),

\[
\lim_{n \to \infty} \frac{1}{n} E_x^\pi [\|h(x_n)\|] = 0. \quad (3.7)
\]

Proof. See [7, Theorem 4.1]. \( \Box \)

4. A semimartingale characterization of average optimal stationary policies

In this section, we present our main results. To do this, we use the following notations.

Let \( h_1^*, h_2^*, g^* \) be as in Lemma 3.7, and define

\[
\begin{align*}
\Delta_1(x, a) &:= c(x, a) + \int_S h_1^+(y)Q(dy \mid x, a) - h_1^*(x) - g^* \quad \forall (x, a) \in K; \quad (4.1) \\
\Delta_2(x, a) &:= c(x, a) + \int_S h_2^+(y)Q(dy \mid x, a) - h_2^*(x) - g^* \quad \forall (x, a) \in K; \quad (4.2) \\
M_n^{(1)} &:= \sum_{t=0}^{n-1} c(x_t, a_t) + h_1^+(x_n) - ng^* \quad \forall n \geq 1, \quad M_0^{(1)} := h_1^*(x) \quad \forall x \in S; \quad (4.3) \\
M_n^{(2)} &:= \sum_{t=0}^{n-1} c(x_t, a_t) + h_2^+(x_n) - ng^* \quad \forall n \geq 1, \quad M_0^{(2)} := h_2^*(x) \quad \forall x \in S. \quad (4.4)
\end{align*}
\]

Theorem 4.1. Under Assumptions 3.1, 3.3, and 3.5, the following statements hold.

(a) A policy \( \pi^* \) is AEC-optimal and \( V(x, \pi^*) = V^*(x) = g^* \) for all \( x \in S \) if and only if

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{t=0}^{n-1} E_x^{\pi^*} [\Delta_1(x_t, a_t)] = 0 \quad \forall x \in S. \quad (4.5)
\]
6 Semimartingale characterization of optimal policies

(b) A policy \( \pi^* \) is AEC-optimal and \( V(x,\pi^*) = V^*(x) = g^* \) for all \( x \in S \) if and only if

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{t=0}^{n-1} E_x^\pi [\Delta_2 (x_t, a_t)] = 0 \quad \forall x \in S. \tag{4.6}
\]

Proof. (a) For each \( \pi \in \Pi \) and \( x \in S \), it follows from (4.1) that

\[
\Delta_1 (x_t, a_t) := c(x_t, a_t)
\]

\[
+ \int \sum_{S} h_t^* (y) Q(dy \mid x_t, a_t) - h_t^* (x_t) - g^* \quad \forall x_t \in S, a_t \in A(x_t), t \geq 0,
\]

which together with (2.5) yields

\[
E_x^\pi [\Delta_1 (x_t, a_t)] = E_x^\pi [c(x_t, a_t)] + E_x^\pi [h_t^* (x_{t+1})] - E_x^\pi [h_t^* (x_t)] - g^*,
\]

and so

\[
\sum_{t=0}^{n-1} E_x^\pi [\Delta_1 (x_t, a_t)] = E_x^\pi \left[ \sum_{t=0}^{n-1} c(x_t, a_t) \right] + E_x^\pi [h_t^* (x_{n})] - h_t^* (x) - ng^* \quad \forall n \geq 1. \tag{4.9}
\]

Multiplying by \( 1/n \) and letting \( n \to \infty \), from (3.7), we see that part (a) is satisfied. Similarly, combining (4.2) and (3.7), we see that part (b) is also true. \( \square \)

Theorem 4.2. Suppose that Assumptions 3.1, 3.3, and 3.5 hold. Then the following hold:

(a) \( \{M_n^{(1)}\} \) is a \( P_x^\pi \)-submartingale for all \( \pi \in \Pi \) and \( x \in S \);

(b) let \( f^* \) be the average optimal stationary policy obtained in Lemma 3.7, then \( \{M_n^{(2)}\} \) is a \( P_x^{f^*} \)-supermartingale for all \( x \in S \);

(c) if \( \{M_n^{(2)}\} \) is a \( P_x^{f^*} \)-supermartingale, then \( \pi^* \) is AEC-optimal and \( V(x,\pi^*) = g^* \) for all \( x \in S \).

Proof. (a) For each \( h_n = (x_0, a_0, \ldots, x_{n-1}, a_{n-1}, x_n) \), it follows from (4.3) that

\[
E_x^\pi (M_n^{(1)} \mid h_n) = M_n^{(1)} + E_x^\pi (\Delta_1 (x_n, a_n) \mid h_n) \quad \forall n \geq 0. \tag{4.10}
\]

On the other hand, combining (3.4) and (4.1), we get

\[
\Delta_1 (x, a) \geq 0 \quad \forall (x, a) \in K, \tag{4.11}
\]

which together with (4.10) implies that \( \{M_n^{(1)}\} \) is a \( P_x^\pi \)-submartingale. (b) For each \( h_n = (x_0, a_0, \ldots, x_{n-1}, a_{n-1}, x_n) \), similarly, it follows from (4.4) that

\[
E_x^\pi (M_n^{(2)} \mid h_n) = M_n^{(2)} + E_x^\pi (\Delta_2 (x_n, a_n) \mid h_n) \quad \forall n \geq 0. \tag{4.12}
\]

In particular, if \( \pi = f^* \), then

\[
E_x^{f^*} (M_n^{(2)} \mid h_n) = M_n^{(2)} + E_x^{f^*} (\Delta_2 (x_n, a_n) \mid h_n) \quad \forall n \geq 0. \tag{4.13}
\]
Also, combining (3.6) and (4.2), we get
\[
\Delta_2(x, f^*(x)) \leq 0 \quad \forall x \in S,
\] (4.14)
which together with (4.13) yields that \(\{M_n^{(2)}\}\) is a \(P_x^*\)-supermartingale.

(c) If \(\{M_n^{(2)}\}\) is a \(P_{\pi}^*\)-supermartingale, then
\[
E_{\pi}^x[M_n^{(2)}] \leq E_{\pi}^x[M_0^{(2)}] = h_2^*(x) \quad \forall n \geq 0, x \in S,
\] (4.15)
which together with (4.4) yields
\[
E_{\pi}^x\left[\sum_{t=0}^{n-1} c(x_t, a_t) + h_2^*(x_n) - ng^* \right] \leq h_2^*(x) \quad \forall n \geq 1, x \in S.
\] (4.16)

Multiplying by \(1/n\) and letting \(n \to \infty\), we get
\[
V(x, \pi^*) \leq g^* \quad \forall x \in S.
\] (4.17)

On the other hand, from part (a), we have that for all \(\pi \in \Pi\) and \(x \in S\), \(\{M_n^{(1)}\}\) is a \(P_x^\pi\)-submartingale. Thus,
\[
E_{\pi}^x[M_n^{(1)}] \geq E_{\pi}^x[M_0^{(1)}] = h_1^*(x) \quad \forall \pi \in \Pi, x \in S.
\] (4.18)

From this inequality and (4.3), we get
\[
E_{\pi}^x\left[\sum_{t=0}^{n-1} c(x_t, a_t) + h_1^*(x_n) - ng^* \right] \geq h_1^*(x) \quad \forall n \geq 1, \pi \in \Pi, x \in S.
\] (4.19)

As in the proof of (4.17), similarly, we have
\[
V(x, \pi) \geq g^* \quad \forall \pi \in \Pi, x \in S,
\] (4.20)
and so
\[
\inf_{\pi \in \Pi} V(x, \pi) \geq g^* \quad \forall x \in S,
\] (4.21)
which together with (4.17) implies that \(\pi^*\) is AEC-optimal and \(V(x, \pi^*) = g^*\) for all \(x \in S\).

Remark 4.3. Theorems 4.1–4.2 are our main results: Theorem 4.1 gives two necessary and sufficient conditions for AEC-optimal policies, whereas Theorem 4.2 further provides a semimartingale characterization of an average optimal stationary policy.

Acknowledgment

This research is partially supported by NCET and EYTP, and the Natural Science Foundation of China.
References


Quanxin Zhu: Department of Mathematics, South China Normal University, Guangzhou 510631, China
*E-mail address*: zqx22@126.com

Xianping Guo: The School of Mathematics and Computational Science, Zhongshan University, Guangzhou 510275, China
*E-mail address*: mcsgxp@mail.sysu.edu.cn