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.

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

Hindawi Publishing Corporation
Journal of Applied Mathematics and Stochastic Analysis
Volume 2006, Article ID 81593, Pages 1–8
DOI 10.1155/JAMSA/2006/81593

equation under the additional equi-continuity condition and give a "martingale characterization" 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 optimality 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, Assumption 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 conditions 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 $\mathcal{B}(\mathrm{X})$ its Borel σ-algebra.

In this section, we first introduce the control model

$$\{S, (A(x) \subset A,\ x \in S),\ Q(\cdot \mid x,a),\ c(x,a)\}, \tag{2.1}$$

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

$$K := \{(x,a) : x \in S,\ a \in A(x)\} \tag{2.2}$$

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 π_t on A given H_t satisfying

$$\pi_t(A(x) \mid h_t) = 1 \quad \forall h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x) \in H_t,\ t \geq 0. \tag{2.3}$$

The class of all randomized history-dependent policies is denoted by Π. 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

$$\pi_t(\{f(x)\} \mid h_t) = \pi_t(\{f(x)\} \mid x) = 1 \quad \forall h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x) \in H_t,\ t \geq 0. \tag{2.4}$$

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 $(\Omega, \mathcal{F}, P_x^\pi)$ and a stochastic process $\{x_t, a_t, t \geq 0\}$ defined on Ω such that, for each $D \in \mathcal{B}(S)$ and $t \geq 0$,

$$P_x^\pi(x_{t+1} \in D \mid h_t, a_t) = Q(D \mid x_t, a_t), \tag{2.5}$$

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_x^π is denoted by E_x^π.

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

$$\begin{aligned} V_\alpha(x,\pi) &:= E_x^\pi\left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t)\right], \qquad V_\alpha^*(x) := \inf_{\pi \in \Pi} V_\alpha(x,\pi); \\ \bar{V}(x,\pi) &:= \limsup_{n\to\infty} \frac{E_x^\pi\left[\sum_{t=0}^{n-1} c(x_t, a_t)\right]}{n}, \qquad \bar{V}^*(x) = \inf_{\pi \in \Pi} \bar{V}(x,\pi). \end{aligned} \tag{2.6}$$

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

$$V_\alpha(x,\pi^*) = V_\alpha^*(x) \quad \forall x \in S. \tag{2.7}$$

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.

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

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

(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 $\int_S u(y)Q(dy \mid x,a)$ 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

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

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

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

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*

$$g^* + h_1^*(x) \le \min_{a \in A(x)} \left\{ c(x,a) + \int_S h_1^*(y) Q(dy \mid x, a) \right\} \quad \forall x \in S; \tag{3.4}$$

$$g^* + h_2^*(x) \ge \min_{a \in A(x)} \left\{ c(x,a) + \int_S h_2^*(y) Q(dy \mid x, a) \right\} \tag{3.5}$$

$$= c(x, f^*(x)) + \int_S h_2^*(y) Q(dy \mid x, f^*(x)) \quad \forall x \in S. \tag{3.6}$$

(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 $\pi \in \Pi$,*

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

Proof. See [7, Theorem 4.1]. □

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

$$\Delta_1(x,a) := c(x,a) + \int_S h_1^*(y) Q(dy \mid x, a) - h_1^*(x) - g^* \quad \forall (x,a) \in K; \tag{4.1}$$

$$\Delta_2(x,a) := c(x,a) + \int_S h_2^*(y) Q(dy \mid x, a) - h_2^*(x) - g^* \quad \forall (x,a) \in K; \tag{4.2}$$

$$M_n^{(1)} := \sum_{t=0}^{n-1} c(x_t, a_t) + h_1^*(x_n) - n g^* \quad \forall n \ge 1, \qquad M_0^{(1)} := h_1^*(x) \quad \forall x \in S; \tag{4.3}$$

$$M_n^{(2)} := \sum_{t=0}^{n-1} c(x_t, a_t) + h_2^*(x_n) - n g^* \quad \forall n \ge 1, \qquad M_0^{(2)} := h_2^*(x) \quad \forall x \in S. \tag{4.4}$$

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

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

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

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

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

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

$$\begin{aligned}\Delta_1(x_t,a_t) &:= c(x_t,a_t) \\ &\quad + \int_S h_1^*(y)Q(dy \mid x_t,a_t) - h_1^*(x_t) - g^* \quad \forall x_t \in S,\ a_t \in A(x_t),\ t \geq 0,\end{aligned} \tag{4.7}$$

which together with (2.5) yields

$$E_x^{\pi}\left[\Delta_1(x_t,a_t)\right] = E_x^{\pi}\left[c(x_t,a_t)\right] + E_x^{\pi}\left[h_1^*(x_{t+1})\right] - E_x^{\pi}\left[h_1^*(x_t)\right] - g^*, \tag{4.8}$$

and so

$$\sum_{t=0}^{n-1} E_x^{\pi}\left[\Delta_1(x_t,a_t)\right] = E_x^{\pi}\left[\sum_{t=0}^{n-1} c(x_t,a_t)\right] + E_x^{\pi}\left[h_1^*(x_n)\right] - h_1^*(x) - ng^* \quad \forall n \geq 1. \tag{4.9}$$

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. □

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^{π}-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 $P_x^{\pi^*}$-supermartingale, then π^* 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

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

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^{π}-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

$$E_x^{\pi}\left(M_{n+1}^{(2)} \mid h_n\right) = M_n^{(2)} + E_x^{\pi}\left(\Delta_2(x_n,a_n) \mid h_n\right) \quad \forall n \geq 0. \tag{4.12}$$

In particular, if $\pi = f^*$, then

$$E_x^{f^*}\left(M_{n+1}^{(2)} \mid h_n\right) = M_n^{(2)} + E_x^{f^*}\left(\Delta_2(x_n,a_n) \mid h_n\right) \quad \forall n \geq 0. \tag{4.13}$$

Also, combining (3.6) and (4.2), we get

$$\Delta_2(x, f^*(x)) \le 0 \quad \forall x \in S, \tag{4.14}$$

which together with (4.13) yields that $\{M_n^{(2)}\}$ is a $P_x^{f^*}$-supermartingale.

(c) If $\{M_n^{(2)}\}$ is a $P_x^{\pi^*}$-supermartingale, then

$$E_x^{\pi^*}[M_n^{(2)}] \le E_x^{\pi^*}[M_0^{(2)}] = h_2^*(x) \quad \forall n \ge 0,\ x \in S, \tag{4.15}$$

which together with (4.4) yields

$$E_x^{\pi^*}\left[\sum_{t=0}^{n-1} c(x_t, a_t)\right] + E_x^{\pi^*}[h_2^*(x_n)] - ng^* \le h_2^*(x) \quad \forall n \ge 1,\ x \in S. \tag{4.16}$$

Multiplying by $1/n$ and letting $n \to \infty$, we get

$$V(x, \pi^*) \le g^* \quad \forall x \in S. \tag{4.17}$$

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^{π}-submartingale. Thus,

$$E_x^{\pi}[M_n^{(1)}] \ge E_x^{\pi}[M_0^{(1)}] = h_1^*(x) \quad \forall \pi \in \Pi,\ x \in S. \tag{4.18}$$

From this inequality and (4.3), we get

$$E_x^{\pi}\left[\sum_{t=0}^{n-1} c(x_t, a_t)\right] + E_x^{\pi}[h_1^*(x_n)] - ng^* \ge h_1^*(x) \quad \forall n \ge 1,\ \pi \in \Pi,\ x \in S. \tag{4.19}$$

As in the proof of (4.17), similarly, we have

$$V(x, \pi) \ge g^* \quad \forall \pi \in \Pi,\ x \in S, \tag{4.20}$$

and so

$$\inf_{\pi \in \Pi} V(x, \pi) \ge g^* \quad \forall x \in S, \tag{4.21}$$

which together with (4.17) implies that π^* 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

[1] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus, *Discrete-time controlled Markov processes with average cost criterion: a survey*, SIAM Journal on Control and Optimization **31** (1993), no. 2, 282–344.

[2] C. Derman, *Finite State Markovian Decision Processes*, Mathematics in Science and Engineering, vol. 67, Academic Press, New York, 1970.

[3] E. B. Dynkin and A. A. Yushkevich, *Controlled Markov Processes*, Fundamental Principles of Mathematical Sciences, vol. 235, Springer, Berlin, 1979.

[4] E. Gordienko and O. Hernández-Lerma, *Average cost Markov control processes with weighted norms: existence of canonical policies*, Applicationes Mathematicae **23** (1995), no. 2, 199–218.

[5] X. P. Guo, J. Y. Liu, and K. Liu, *Nonhomogeneous Markov decision processes with Borel state space—the average criterion with nonuniformly bounded rewards*, Mathematics of Operations Research **25** (2000), no. 4, 667–678.

[6] X. P. Guo and P. Shi, *Limiting average criteria for nonstationary Markov decision processes*, SIAM Journal on Optimization **11** (2001), no. 4, 1037–1053.

[7] X. P. Guo and Q. X. Zhu, *Average optimality for Markov decision processes in Borel spaces: a new condition and approach*, to appear in Journal of Applied Probability.

[8] O. Hernández-Lerma, *Adaptive Markov Control Processes*, Applied Mathematical Sciences, vol. 79, Springer, New York, 1989.

[9] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes. Basic Optimality Criteria*, Applications of Mathematics (New York), vol. 30, Springer, New York, 1996.

[10] ———, *Further Topics on Discrete-Time Markov Control Processes*, Applications of Mathematics (New York), vol. 42, Springer, New York, 1999.

[11] R. A. Howard, *Dynamic Programming and Markov Processes*, John Wiley & Sons, New York, 1960.

[12] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, New York, 1994.

[13] L. I. Sennott, *Stochastic Dynamic Programming and the Control of Queueing Systems*, Wiley Series in Probability and Statistics: Applied Probability and Statistics, John Wiley & Sons, New York, 1999.

[14] ———, *Average reward optimization theory for denumerable state spaces*, Handbook of Markov Decision Processes (E. A. Feinberg and A. Shwartz, eds.), Internat. Ser. Oper. Res. Management Sci., vol. 40, Kluwer Academic, Massachusetts, 2002, pp. 153–172.

[15] Q. X. Zhu and X. P. Guo, *Another set of condition for strong n* $(n = -1, 0)$ *discount optimality in Markov decision processes*, Stochastic Analysis and Applications **23** (2005), no. 5, 953–974.

[16] ———, *Unbounded cost Markov decision processes with limsup and liminf average criteria: new conditions*, Mathematical Methods of Operations Research **61** (2005), no. 3, 469–482.

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