J. Appl. Prob. **45,** 417–429 (2008)
Printed in England

BIAS AND OVERTAKING OPTIMALITY FOR CONTINUOUS-TIME JUMP MARKOV DECISION PROCESSES IN POLISH SPACES

QUANXIN ZHU,* ** *South China Normal University*
TOMÁS PRIETO-RUMEAU,* *** *Universidad Nacional de Educación a Distancia*

Abstract

In this paper we study the bias and the overtaking optimality criteria for continuous-time jump Markov decision processes in general state and action spaces. The corresponding transition rates are allowed to be *unbounded*, and the reward rates may have *neither upper nor lower bounds*. Under appropriate hypotheses, we prove the existence of solutions to the bias optimality equations, the existence of bias optimal policies, and an equivalence relation between bias and overtaking optimality.

Keywords: Continuous-time jump Markov decision process; expected average reward criterion; general state space; bias optimality; overtaking optimality

2000 Mathematics Subject Classification: Primary 90C40; 93E20

1. Introduction

The *long-run expected average reward* criterion is one of the most popular performance criteria for Markov decision processes (MDPs), and it has been extensively studied (see, e.g. [1], [9, Chapter 5], [10, Chapter 10], and [19, Chapter 8]). But, on the other hand, the average reward criterion turns out to be extremely *underselective* because an average reward optimal policy may have an arbitrarily bad behavior for large, but finite, lengths of time. To overcome this situation, more sensitive optimality criteria have been proposed. These include the variance-minimization criterion (which selects an average reward optimal policy with minimal variation; see, e.g. [11], [17], and [23]), the bias and the overtaking optimality criteria (that choose an average optimal policy with the maximal expected reward growth as the time horizon goes to ∞; see, e.g. [7], [8], [10, p. 132], [12], [16], and [19, Chapter 10]), and the so-called discount-sensitive criteria (which choose policies that are asymptotically optimal as the discount rate converges to 0; see [7], [13], [15], [19, Chapter 10], and [22]), among others.

The bias and the overtaking optimality criteria, which we study in this paper, have been widely studied for discrete-time MDPs [10], [19]. For continuous-time models, however, just a few references deal with this issue. For instance, Puterman [18] studied controlled diffusions on compact intervals and Jasso-Fuentes and Hernández-Lerma [12] considered general controlled diffusions. Regarding jump processes with nonfinite state space, Prieto-Rumeau and Hernández-Lerma [16] analyzed the case of a denumerable state space. In this paper we deal

Received 28 November 2007; revision received 10 April 2008.
* Research partially supported by the Natural Science Foundation of China (10626021), the Natural Science Foundation of Guangdong Province (06300957), and CONACYT grant 45693-F.
** Postal address: Department of Mathematics, South China Normal University, Guangzhou 510631, P. R. China. Email address: zqx22@126.com
*** Postal address: Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Educación a Distancia, Senda del Rey 9, Madrid 28040, Spain. Email address: tprieto@ccia.uned.es

with continuous-time controlled jump Markov processes on a Polish (nondenumerable) state space. The motivation is clear: the state space corresponding to many practical situations such as, for instance, inventory and water-regulation problems, is *not* denumerable.

The model under consideration in this paper is fairly general. We deal with a controlled jump Markov process with Polish state and action spaces. The transition rates and the reward rate are allowed to be *unbounded*. Our theoretical background consists of the papers [4], [6], [20], and [21], which analyze the discounted and the average reward optimality criteria for general continuous-time jump MDPs. Then, starting from the results in these papers, we refine the average reward optimality criterion and study bias and overtaking optimality.

With respect to reference [16], which analyzed the bias and the overtaking optimality criteria for denumerable-state jump MDPs, it is worth noting that the corresponding proofs are greatly simplified by the fact that the state space is denumerable. Therefore, the arguments in the present paper are different from those in [16], though we basically reach similar results. In this sense the present paper gives an answer to one of the open problems mentioned in [16, Section 6], regarding the extension of the results therein to the case of a general jump process.

The remainder of this paper is organized as follows. In Section 2 we introduce the control model that we are interested in and state our assumptions. In Section 3 we define the optimality criteria we will analyze, and we also state our main results, whose proofs are postponed to Section 4. In Section 4 we also give more insight into the relation between overtaking and bias optimality, and give some interesting results on bias optimality. Finally, we conclude in Section 5 with some general remarks.

2. The control model

In this section we define the control model we will be dealing with and state our main assumptions.

2.1. Model definition

If X is a Polish space (that is, a complete and separable metric space), we will denote by $\mathcal{B}(X)$ its Borel σ-algebra.

We are concerned with the following control model:

$$\{S, (A(x) \subseteq A, x \in S), q(\cdot \mid x, a), r(x, a)\}, \tag{2.1}$$

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

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

is a Borel subset of $S \times A$.

The $q(\cdot \mid x, a)$ in (2.1) denote the *transition rates*, and so they satisfy the following properties: for each $(x, a) \in K$ and $D \in \mathcal{B}(S)$,

(Q1) $D \mapsto q(D \mid x, a)$ is a signed measure on $\mathcal{B}(S)$, and $(x, a) \mapsto q(D \mid x, a)$ is Borel measurable on K;

(Q2) $0 \le q(D \mid x, a) < \infty$ whenever $x \notin D \in \mathcal{B}(S)$;

(Q3) $q(S \mid x, a) = 0$ and $0 \le -q(x \mid x, a) < \infty$.

Also, the model is assumed to be *stable*, i.e.

$$q(x) := \sup_{a \in A(x)} \{-q(x \mid x, a)\} < \infty \quad \text{for all } x \in S. \tag{2.2}$$

Finally, $r(x, a)$, the *reward rate*, is assumed to be a real-valued measurable function on K. (As $r(x, a)$ is allowed to take positive and negative values, it can also be interpreted as a *cost rate*.)

This model is a standard continuous-time controlled jump Markov process; see, e.g. [4], [6], and [20].

2.2. Control policies

Now we introduce the class of admissible control policies. Let Π_m be the family of functions $\pi_t(B \mid x)$ such that

1. for each $x \in S$ and $t \geq 0$, $B \mapsto \pi_t(B \mid x)$ is a probability measure on $\mathcal{B}(A(x))$; and
2. for each $x \in S$ and $B \in \mathcal{B}(A(x))$, $t \mapsto \pi_t(B \mid x)$ is a Borel measurable function on $[0, \infty)$.

We say that $\pi = (\pi_t,\ t \geq 0) \in \Pi_m$ is a *randomized Markov policy*. In particular, if there exists a measurable function $f\colon S \to A$, with $f(x) \in A(x)$ for all $x \in S$, such that $\pi_t(\{f(x)\} \mid x) = 1$ for all $t \geq 0$ and $x \in S$, then π is called a (deterministic) *stationary policy*, and it is identified with f. The set of all stationary policies is denoted by F.

Given $\pi = (\pi_t,\ t \geq 0) \in \Pi_m$, we define the associated transition rates $q(D \mid x, \pi_t)$ and reward rates $r(x, \pi_t)$ as follows. For each $x \in S$, $D \in \mathcal{B}(S)$, and $t \geq 0$,

$$q(D \mid x, \pi_t) := \int_{A(x)} q(D \mid x, a)\pi_t(\mathrm{d}a \mid x), \tag{2.3}$$

$$r(x, \pi_t) := \int_{A(x)} r(x, a)\pi_t(\mathrm{d}a \mid x). \tag{2.4}$$

In particular, when $\pi = f \in F$, we will write $q(D \mid x, \pi_t)$ and $r(x, \pi_t)$ as $q(D \mid x, f)$ and $r(x, f)$, respectively. The integral in (2.3) is well defined and finite as a consequence of property (Q3). Later, we will impose conditions ensuring that (2.4) is finite.

Definition 2.1. A randomized Markov policy $\pi \in \Pi_m$ is said to be *admissible* if $q(D \mid x, \pi_t)$ is continuous in $t \geq 0$ for all $D \in \mathcal{B}(S)$ and $x \in S$.

We will denote by Π the family of admissible policies. Obviously, Π is nonempty since it contains the set of stationary policies F.

2.3. Assumptions

By Lemma 2.1 of [6], for each $\pi \in \Pi$, there exists a Q-process—that is, a possibly substochastic and nonhomogeneous transition function $\mathrm{P}^{\pi}(s, x, t, D)$—with transition rates $q(D \mid x, \pi_t)$. This Q-process, however, might not be regular. To ensure the regularity of the corresponding Q-process, we will borrow the following so-called drift condition from [6], [20], and [21].

Assumption A. *There exist a measurable function $w_1 \geq 1$ on S and constants $b_1 \geq 0$, $c_1 > 0$, $M > 0$, and $M' > 0$ such that*

(a) $\int_S w_1(y) q(\mathrm{d}y \mid x, a) \leq -c_1 w_1(x) + b_1$ *for all* $(x, a) \in K$;

(b) $q(x) \leq M w_1(x)$ *for all* $x \in S$, *with* $q(x)$ *as in (2.2);*

(c) $|r(x, a)| \leq M' w_1(x)$ *for all* $(x, a) \in K$.

Remark 2.1 of [6] gives a detailed discussion of Assumption A. In particular, Assumption A(b) is not required when the transition rates are uniformly bounded, i.e. $\sup_{x \in S} q(x) < \infty$.

For each initial state $x \in S$ at time $s \geq 0$ and $\pi \in \Pi$, we denote by $\mathrm{P}^\pi_{s,x}$ and $\mathrm{E}^\pi_{s,x}$ the respective probability measure and expectation operator determined by $\mathrm{P}^\pi(s, x, t, D)$. In particular, if $s = 0$, we write $\mathrm{E}^\pi_{0,x}$ and $\mathrm{P}^\pi_{0,x}$ as E^π_x and P^π_x, respectively. Therefore, for each $\pi \in \Pi$, there exists a Borel measurable S-valued Markov process with transition rates $q(D \mid x, \pi_t)$, which we will denote by $\{x^\pi_t\}$, or simply by $\{x_t\}$ when there is no risk of confusion.

If Assumption A holds then we obtain

$$\mathrm{E}^\pi_x[w_1(x_t)] \leq \exp(-c_1 t) w_1(x) + \frac{b_1}{c_1} \quad \text{for all } \pi \in \Pi,\ x \in S,\ t \geq 0. \tag{2.5}$$

For a proof, see [4, Theorem 3.1]. In particular, the integral in (2.4) is finite.

In addition to Assumption A, we need to impose further conditions. Assumptions B(a) and (b), below, contain standard continuity-compactness hypotheses; see, e.g. [6], [20], [21], and the references therein. Assumption B is also a standard assumption for discrete-time models; see, e.g. [10, p. 44] and [22]. Assumption B(c) is used to ensure the application of *Dynkin's formula*. Obviously, Assumption B(c) is not required when $\sup_{x \in S} q(x)$ is finite.

Assumption B. *For each $x \in S$,*

(a) $A(x)$ *is compact;*

(b) $r(x, a)$ *is continuous in $a \in A(x)$, and the function $\int_S u(y) q(\mathrm{d}y \mid x, a)$ is continuous in $a \in A(x)$ for each bounded measurable function u on S, and also for $u := w_1$, as in Assumption A;*

(c) *there exist a nonnegative measurable function w_2 on S and constants $b_2 \geq 0$, $c_2 > 0$, and $M_2 > 0$ such that*

$$q(x) w_1(x) \leq M_2 w_2(x) \quad \text{and} \quad \int_S w_2(y) q(\mathrm{d}y \mid x, a) \leq c_2 w_2(x) + b_2$$

for all $(x, a) \in K$.

For the function w_1 in Assumption A, we define the weighted supremum norm $\|\cdot\|_{w_1}$ as follows. Given a real-valued measurable function u on S,

$$\|u\|_{w_1} := \sup_{x \in S} \left\{ \frac{|u(x)|}{w_1(x)} \right\},$$

and let $B_{w_1}(S)$ be the Banach space of functions with finite w_1-norm. Using the weighted supremum norm when dealing with unbounded reward and transition rates is a standard technique; see, e.g. [6], [10, Chapter 8], [16], [20], and [22].

Assumption C. *For each $f \in F$, the Markov process $\{x_t\}$ with transition rates $q(\cdot \mid x, f)$ is Harris recurrent and uniformly w_1-exponentially ergodic (that is, there exists an invariant probability measure μ_f on S such that*

$$\sup_{f\in F} |\,\mathrm{E}_x^f[u(x_t)] - \mu_f(u)| \le R\mathrm{e}^{-\rho t}\|u\|_{w_1} w_1(x)$$

for all $x \in S$, $u \in B_{w_1}(S)$, and $t \ge 0$, where the positive constants R and ρ do not depend on f, and where $\mu_f(u) := \int_S u(y)\mu_f(\mathrm{d}y)$).

Sufficient conditions for Assumption C as well as some examples can be found in [6] and [14]. These are generalizations of the stochastic monotonicity and the 'Lyapunov-like inequality' conditions. For a discrete-state space, a uniform integrability condition is given in [7] and [16].

3. Main results

In the following we assume that Assumptions A, B, and C are satisfied. First of all, we define the main optimality criteria we are concerned with.

Given an admissible policy $\pi \in \Pi$, an initial state $x \in S$, and a time horizon $T \ge 0$, the expected total reward of the policy π on $[0, T]$ is defined as

$$V_T(x, \pi) := \mathrm{E}_x^\pi\left[\int_0^T r(x_t, \pi_t)\,\mathrm{d}t\right].$$

As a consequence of Assumption A(c) and (2.5), $V_T(x, \pi)$ is finite.

We say that the admissible policy π *overtakes* $\pi' \in \Pi$ if, for every $\varepsilon > 0$ and every $x \in S$, there exists $T_0 \ge 0$ such that $V_T(x, \pi) \ge V_T(x, \pi') - \varepsilon$ whenever $T \ge T_0$. Accordingly, we give our next definition.

Definition 3.1. A policy $f^* \in F$ is said to be overtaking optimal in F if it overtakes every $f \in F$, that is,

$$\liminf_{T\to\infty}[V_T(x, f^*) - V_T(x, f)] \ge 0$$

for all $f \in F$ and $x \in S$.

We also need to define the expected average reward optimality criterion. Given $x \in S$ and $\pi \in \Pi$, the corresponding *expected average reward* is defined as

$$V(x, \pi) := \liminf_{T\to\infty} \frac{1}{T} V_T(x, \pi).$$

Observe that, by Assumption A(c) and (2.5), $|V(x, \pi)| \le b_1 M'/c_1$ for every $x \in S$ and $\pi \in \Pi$. Also, by Assumption C, if $f \in F$ then

$$V(x, f) = \lim_{T\to\infty} \frac{1}{T} V_T(x, f) = \int_S r(y, f)\mu_f(\mathrm{d}y) =: g(f), \tag{3.1}$$

which does not depend on the initial state $x \in S$. The constant $g(f)$ is usually referred to as the *gain* of f.

Definition 3.2. A policy $\pi^* \in \Pi$ is said to be expected average reward optimal (or average optimal, in short) if $V(x, \pi^*) \ge V(x, \pi)$ for all $\pi \in \Pi$ and $x \in S$.

Next, we state our main results in this paper. We start with the following theorem which establishes the *average reward optimality equation*. Its proof can be found in [21, Theorem 4.1].

Theorem 3.1. *Under Assumptions A, B, and C, the following statements hold.*

(a) *There exist a unique constant g^*, a function $h^* \in B_{w_1}(S)$, and a stationary policy $f^* \in F$ satisfying the following average reward optimality equation (AROE):*

$$g^* = \max_{a \in A(x)} \left\{ r(x,a) + \int_S h^*(y) q(\mathrm{d}y \mid x, a) \right\} \tag{3.2}$$

$$= r(x, f^*) + \int_S h^*(y) q(\mathrm{d}y \mid x, f^*) \quad \text{for all } x \in S. \tag{3.3}$$

(b) *For all $x \in S$, $g^* = \sup_{\pi \in \Pi} V(x, \pi) = \sup_{f \in F} g(f)$.*

(c) *Any stationary policy $f \in F$ reaching the maximum in (3.2) for every $x \in S$ is average optimal, and so f^* in (3.3) is average optimal.*

Obviously, it follows from (3.1), the definition of overtaking optimality, and Theorem 3.1(b), that an overtaking optimal policy in F is necessarily gain optimal. In this sense, overtaking optimality is indeed a refinement of average reward optimality. Therefore, an overtaking optimal policy is an average reward optimal stationary policy which, in addition, has the largest finite-horizon reward growth. Of course, it remains to show that overtaking optimal policies exist.

We introduce the following notation. For each $x \in S$, let $A^*(x) \subseteq A(x)$ be the set of actions $a^* \in A(x)$ that attain the maximum in (3.2), i.e.

$$A^*(x) := \left\{ a^* \in A(x) \colon g^* = r(x, a^*) + \int_S h^*(y) q(\mathrm{d}y \mid x, a^*) \right\}.$$

By Assumption B(a) and (b), the sets $A^*(x)$ are nonempty and compact.

Definition 3.3. We denote by F_{ao} the set of average optimal deterministic stationary policies. A stationary policy $f^* \in F$ is called canonical if it attains the maximum in (3.2), i.e. $f^*(x) \in A^*(x)$ for each $x \in S$. The set of canonical policies is denoted by F_{ca}.

By Theorem 3.1, the sets F_{ao} and F_{ca} are nonempty and, in addition, $F_{\mathrm{ca}} \subseteq F_{\mathrm{ao}}$. The inclusion is, in general, strict; see the counterexample in [5].

Remark 3.1. In principle, the sets $A^*(x)$, as well as F_{ca}, depend on the function h^* in (3.2). Thus, in what follows, we suppose that h^* in the solution of the AROE, (3.2), remains fixed. In fact, it will be shown later (see Remark 4.1, below) that h^* is unique up to additive constants and, therefore, neither $A^*(x)$ nor F_{ca} depend on the particular solution h^*.

Our main result is the following.

Theorem 3.2. *Suppose that Assumptions A, B, and C hold, and let $(g^*, h^*) \in \mathbb{R} \times B_{w_1}(S)$ be a solution of the AROE. Then the following statements hold.*

(a) *There exist a policy $f^* \in F_{\mathrm{ca}}$, a unique constant $\sigma^* \in \mathbb{R}$, and a function $V^* \in B_{w_1}(S)$ satisfying*

$$\sigma^* = \max_{a \in A^*(x)} \left\{ -h^*(x) + \int_S V^*(y) q(\mathrm{d}y \mid x, a) \right\} \tag{3.4}$$

$$= -h^*(x) + \int_S V^*(y) q(\mathrm{d}y \mid x, f^*) \quad \text{for all } x \in S. \tag{3.5}$$

(b) *Any policy $f^* \in F$ attaining the maximum in (3.4) for every $x \in S$ is overtaking optimal in F, and so f^* in (3.5) is overtaking optimal in F.*

Theorem 3.2 shows that we can determine an overtaking optimal policy by solving two nested AROE-like equations: first we solve the AROE, (3.2), and second we restrict ourselves to the sets $A^*(x)$ for $x \in S$ and then solve (3.4). The two nested equations (3.2) and (3.4) are known as the *bias optimality equations*. The reason why they are so named will be made clear in Section 4.

4. Proofs

In this section our plan is the following. First, we define the bias of a stationary policy and the bias optimality criterion, which we prove to be equivalent to the overtaking optimality criterion. Then we prove that there exist bias optimal policies which, in addition, are canonical. Finally, we prove our main result, Theorem 3.2, and propose another characterization of bias optimality in Theorem 4.2.

Throughout this section, we suppose that Assumptions A, B, and C are satisfied.

4.1. Bias optimality

In the spirit of potential concepts in [2] and [3], for a given $f \in F$ and the corresponding invariant probability measure μ_f, we define the potential or *bias* of $f \in F$ as

$$h_f(x) := \int_0^\infty (\mathrm{E}_x^f[r(x_t, f)] - g(f))\,\mathrm{d}t \quad \text{for all } x \in S.$$

By Assumption C, the bias of f is finite and, in addition, $|h_f(x)| \le RM'w_1(x)/\rho$. Moreover, since μ_f is an invariant probability measure, it follows that

$$\mu_f(h_f) = 0 \quad \text{for all } f \in F. \tag{4.1}$$

Our next lemma, which is taken from Lemma 3.2 of [21], characterizes the bias of f by means of the *Poisson equation*.

Lemma 4.1. ([21, Lemma 3.2].) *Let $f \in F$ be any stationary policy. Then the following statements hold.*

(a) *The function h_f is in $B_{w_1}(S)$, where w_1 is as in Assumption A, and $\|h_f\|_{w_1} \le RM'/\rho$.*

(b) *The pair $(g(f), h_f)$ is the unique solution in $\mathbb{R} \times B_{w_1}(S)$ of the Poisson equation for f, i.e.*

$$g = r(x, f) + \int_S h(y)q(\mathrm{d}y \mid x, f) \quad \textit{for all } x \in S, \tag{4.2}$$

for which $\mu_f(h) = 0$.

Now we provide an interpretation of the bias. For each $x \in S$, $f \in F$, and $T \ge 0$, by the Poisson equation, (4.2), and Dynkin's formula, we have

$$\begin{aligned}
\mathrm{E}_x^f[h_f(x_T)] - h_f(x) &= \mathrm{E}_x^f\left[\int_0^T \int_S h_f(y)q(\mathrm{d}y \mid x_t, f)\,\mathrm{d}t\right] \\
&= Tg(f) - \mathrm{E}_x^f\left[\int_0^T r(x_t, f)\,\mathrm{d}t\right] \\
&= Tg(f) - V_T(x, f),
\end{aligned}$$

and so

$$V_T(x, f) = Tg(f) + h_f(x) - \mathrm{E}_x^f[h_f(x_T)]. \tag{4.3}$$

By Assumption C and since $\mu_f(h_f) = 0$ (recall (4.1)),

$$\lim_{T\to\infty} \mathrm{E}_x^f[h_f(x_T)] = 0.$$

This shows that the total expected reward $V_T(x, f)$ is, asymptotically as $T \to \infty$, a straight line with slope $g(f)$ and ordinate $h_f(x)$. Hence, intuitively, in order to find an overtaking optimal policy in F, we should try to maximize the bias $h_f(x)$ among the class of stationary policies with maximal gain, that is, maximize the bias in F_{ao}. This leads to our next definition.

Definition 4.1. A policy $\bar{f}$ in F_{ao} is called bias optimal if

$$h_{\bar{f}}(x) = \sup_{f\in F_{\mathrm{ao}}} h_f(x) \quad \text{for all } x \in S.$$

We call $\bar{h}(x) := \sup_{f\in F_{\mathrm{ao}}} h_f(x)$ for $x \in S$ the optimal bias function, which is finite as a consequence of Lemma 4.1(a).

The relation between bias and overtaking optimal policies is made clear in the next result.

Theorem 4.1. *Suppose that Assumptions A, B, and C hold. An average reward optimal policy $\bar{f} \in F_{\mathrm{ao}}$ is overtaking optimal in F if and only if it is bias optimal.*

Proof. Let $\bar{f} \in F_{\mathrm{ao}}$ be a bias optimal policy. Let us prove that $\bar{f}$ is overtaking optimal in F. To this end, fix an arbitrary $f \in F$. Then, for each $x \in S$ and $T \geq 0$, it follows from (4.3) that

$$\begin{aligned} V_T(x, \bar{f}) - V_T(x, f) = {} & T(g(\bar{f}) - g(f)) + h_{\bar{f}}(x) - h_f(x) \\ & - (\mathrm{E}_x^{\bar{f}}[h_{\bar{f}}(x_T)] - \mathrm{E}_x^f[h_f(x_T)]), \end{aligned} \tag{4.4}$$

where, by Assumption C and (4.1), for every $x \in S$,

$$\lim_{T\to\infty} \mathrm{E}_x^f[h_f(x_T)] = \mu_f(h_f) = 0 \quad \text{and} \quad \lim_{T\to\infty} \mathrm{E}_x^{\bar{f}}[h_{\bar{f}}(x_T)] = \mu_{\bar{f}}(h_{\bar{f}}) = 0. \tag{4.5}$$

On the other hand, by Theorem 3.1, we have $g(\bar{f}) = g^* \geq g(f)$. Therefore, one of the following statements hold:

(i) either $g(\bar{f}) > g(f)$; or

(ii) $g(\bar{f}) = g(f)$ and $h_{\bar{f}}(x) \geq h_f(x)$ for every $x \in S$ (by Definition 4.1).

In either case, letting $T \to \infty$ in (4.4) and recalling (4.5), we obtain

$$\lim_{T\to\infty} [V_T(x, \bar{f}) - V_T(x, f)] \geq 0 \quad \text{for all } x \in S,$$

which implies that $\bar{f}$ is overtaking optimal in F.

Conversely, suppose that $\bar{f} \in F_{\mathrm{ao}}$ is overtaking optimal in F, i.e.

$$\liminf_{T\to\infty} [V_T(x, \bar{f}) - V_T(x, f)] \geq 0 \quad \text{for all } x \in S \text{ and } f \in F,$$

and let us prove that $\bar{f}$ is bias optimal. The proof will proceed by contradiction. Hence, suppose that there exist $x \in S$ and $f' \in F_{\text{ao}}$ such that $h_{\bar{f}}(x) < h_{f'}(x)$. Since both $\bar{f}$ and f' are average optimal, $g(\bar{f}) = g(f') = g^*$. Thus, it follows, from (4.4) and (4.5), that

$$\lim_{T\to\infty} [V_T(x, \bar{f}) - V_T(x, f')] < 0,$$

which contradicts the fact that $\bar{f}$ overtakes f. This completes the proof.

4.2. Bias optimality in the class of canonical policies

In our next results we show that, when dealing with bias optimal policies, we can in fact restrict ourselves to the class of canonical policies. In what follows recall that $h^* \in B_{w_1}(S)$ is taken from the AROE, (3.2).

Proposition 4.1. *The following results hold.*

(a) *If f is any stationary policy in F_{ao} then $h_f(\cdot) \le h^*(\cdot) + \mu_f(-h^*)$.*

(b) *Let f be any stationary policy in F_{ao}. Then, $h_f(\cdot) = h^*(\cdot) + \mu_f(-h^*)$ if and only if $f \in F_{\text{ca}}$.*

(c) *If f is any stationary policy in $F_{\text{ao}} \setminus F_{\text{ca}}$ then there exists a canonical policy $\bar{f} \in F_{\text{ca}}$ such that $\mu_{\bar{f}} = \mu_f$ and, consequently,*

$$\sup_{f\in F_{\text{ca}}} \mu_f(-h^*) = \sup_{f\in F_{\text{ao}}} \mu_f(-h^*).$$

Proof. (a) Since f is a stationary policy in F_{ao}, by Theorem 3.1(b), $g(f) = g^*$. Hence, the Poisson equation, (4.2), for f becomes

$$g^* = r(x, f) + \int_S h_f(y) q(\mathrm{d}y \mid x, f) \quad \text{for all } x \in S. \tag{4.6}$$

On the other hand, it follows from (3.2) that

$$\begin{aligned} g^* &= \max_{a\in A(x)} \left\{ r(x, a) + \int_S h^*(y) q(\mathrm{d}y \mid x, a) \right\} \\ &\ge r(x, f) + \int_S h^*(y) q(\mathrm{d}y \mid x, f) \quad \text{for all } x \in S, \end{aligned}$$

which together with (4.6) gives

$$\int_S (h_f(y) - h^*(y)) q(\mathrm{d}y \mid x, f) \ge 0 \quad \text{for all } x \in S.$$

Then, using Dynkin's formula, we obtain, for every $t \ge 0$,

$$\mathrm{E}_x^f [h_f(x_t) - h^*(x_t)] \ge h_f(x) - h^*(x) \quad \text{for all } x \in S.$$

Now, letting $t \to \infty$ and by Assumption C, we have

$$\mu_f(h_f - h^*) \ge h_f(x) - h^*(x) \quad \text{for all } x \in S.$$

Thus, recalling (4.1), it follows that

$$\mu_f(-h^*) \ge h_f(x) - h^*(x) \quad \text{for all } x \in S,$$

which yields statement (a).

(b) Suppose that $h_f(\cdot) = h^*(\cdot) + \mu_f(-h^*)$. Since $f \in F_{ao}$, it follows that (g^*, h^*) is a solution of the Poisson equation for f, that is, f satisfies (3.3) and, therefore, f is canonical.

Conversely, if $f \in F_{ca}$ then, from Definition 3.3,

$$g^* = r(x, f) + \int_S h^*(y) q(\mathrm{d}y \mid x, f) \quad \text{for all } x \in S.$$

Now, using an argument similar to that in the proof of statement (a) and replacing the inequalities with the corresponding equalities, we see that statement (b) is true.

(c) The proof of statement (c) is similar to the proof of Lemma 5.2(b) of [11] (or Lemma 11.3.12 of [10]). In fact, from the proof of [21, Theorem 4.2] we see that, for every stationary policy f in $F_{ao} \setminus F_{ca}$, there exist some constant k and a Borel set $N = N_f$ in S such that $\mu_f(N) = 0$ and

$$h_f(x) = h^*(x) + k \quad \text{for all } x \in N^c, \tag{4.7}$$

where N^c denotes the complement of N. Then, for the canonical policy $f^* \in F_{ca}$ in Theorem 3.1, we define a new policy $\bar{f}$ as

$$\bar{f} := f^* \quad \text{on } N \qquad \text{and} \qquad \bar{f} := f \quad \text{on } N^c.$$

Thus, from (4.7) we claim that $\bar{f}$ is canonical and

$$q(\cdot \mid x, \bar{f}) = q(\cdot \mid x, f) \quad \text{for all } x \in N^c.$$

This, together with the definition of the invariant probability measure, yields the desired result.

Combining Proposition 4.1 with Definition 4.1, we can easily obtain the following results, which we state without proof.

Proposition 4.2. *Under Assumptions A, B, and C, the following statements hold.*

(a) *The optimal bias function satisfies* $\bar{h}(x) = \sup_{f \in F_{ca}} h_f(x)$ *for every* $x \in S$.

(b) *For every* $x \in S$,

$$\bar{h}(x) = h^*(x) + \sup_{f \in F_{ca}} \mu_f(-h^*), \tag{4.8}$$

and, therefore, $\bar{h} \in B_{w_1}(S)$.

(c) *A canonical policy* $\bar{f} \in F_{ca}$ *is bias optimal if and only if it attains the maximum in (4.8), i.e.* $\mu_{\bar{f}}(-h^*) = \sup_{f \in F_{ca}} \mu_f(-h^*)$.

Hence, Proposition 4.2 shows that, when looking for bias optimal policies, we can restrict ourselves to the class of canonical policies F_{ca} (in lieu of the class F_{ao}).

Remark 4.1. From Proposition 4.1(c) and (4.8), we deduce that if h_1^* and h_2^* in $B_{w_1}(S)$ are two solutions of the AROE, (3.2), then, for every $x \in S$,

$$h_1^*(x) - h_2^*(x) = \sup_{f \in F_{ao}} \mu_f(-h_2^*) - \sup_{f \in F_{ao}} \mu_f(-h_1^*).$$

Since the definition of F_{ao} does not depend on the solution of the AROE, it follows that the solution h^* of (3.2) is unique up to additive constants. Therefore, F_{ca} does not depend on the particular solution h^* (cf. Remark 3.1).

Finally, we prove the main result in this paper.

Proof of Theorem 3.2. (a) We consider the Markov control model

$$\bar{\mathcal{M}} := \{S, A, (A^*(x) : x \in S), q(\cdot \mid x, a), r^*(x, a)\},$$

where

$$r^*(x, a) := -h^*(x) \quad \text{for all } (x, a) \in K.$$

The control model $\bar{\mathcal{M}}$ is the same as $\mathcal{M}$ except that $A(x)$ and $r(x, a)$ have been replaced with $A^*(x)$ and h^*.

Recall that $A^*(x)$ is compact for each $x \in S$. On the other hand, since $h^* \in B_{w_1}(S)$, there exists a constant M^* such that $|r^*(x, a)| \le M^* w_1(x)$ for all $(x, a) \in K$. Hence, it is easy to check that the new control model $\bar{\mathcal{M}}$ satisfies Assumptions A, B, and C, replacing $r(x, a)$ and $A(x)$ with $r^*(x, a)$ and $A^*(x)$. Therefore, by Theorem 3.1, there exist a unique constant $\sigma^* \in \mathbb{R}$, a function $V^* \in B_{w_1}(S)$, and a canonical policy $f^* \in F_{\text{ca}}$ such that

$$\begin{aligned}\sigma^* &= \max_{a \in A^*(x)} \left\{ -h^*(x) + \int_S V^*(y) q(\mathrm{d}y \mid x, a) \right\} \\ &= -h^*(x) + \int_S V^*(y) q(\mathrm{d}y \mid x, f^*) \quad \text{for all } x \in S\end{aligned}$$

and

$$\sigma^* = \mu_{f^*}(-h^*) = \sup_{f \in F_{\text{ca}}} \mu_f(-h^*). \tag{4.9}$$

(b) If $f^* \in F_{\text{ca}}$ attains the maximum in (3.5) for every $x \in S$ then, by (4.9) and Proposition 4.2, it is bias optimal and it follows from Theorem 4.1 that it is also overtaking optimal.

In Theorem 3.2 we establish the bias optimality equations and the existence of a bias optimal (and, hence, overtaking optimal) policy. Moreover, from Theorem 3.2 we conclude that a canonical policy is bias optimal if it satisfies the bias optimality equations. It is natural to question if there exists an equivalent relation between bias optimality and the bias optimality equations. Actually, if this conclusion is true, we only need to prove that a bias optimal policy must satisfy the bias optimality equations. Obviously, when the state space is denumerable, from the argument in [16] we can easily obtain the conclusion. However, the state space in this paper is *not* denumerable, and so an attempt to answer this problem faces significant technical difficulties. In this sense the following result proposes another condition which is equivalent to bias optimality.

Theorem 4.2. *Suppose that Assumptions A, B, and C hold. A canonical policy $\bar{f} \in F_{\text{ca}}$ is bias optimal if and only if $\mu_{\bar{f}}(\bar{h}) = 0$, where $\bar{h}$ is the optimal bias function.*

Proof. Suppose that $\bar{f} \in F_{\text{ca}}$ is an arbitrary bias optimal policy. By Definition 4.1 we have

$$\bar{h}(x) = h_{\bar{f}}(x) \quad \text{for all } x \in S.$$

Then, integrating this equation with respect to $\mu_{\bar{f}}$ and by (4.1), we obtain $\mu_{\bar{f}}(\bar{h}) = 0$, as we wanted to prove.

Conversely, suppose that a canonical policy $\bar{f} \in F_{\text{ca}}$ satisfies $\mu_{\bar{f}}(\bar{h}) = 0$. It follows from Proposition 4.2(b) that

$$\bar{h}(x) = h^*(x) + \sup_{f \in F_{\text{ca}}} \mu_f(-h^*) \quad \text{for all } x \in S.$$

We integrate this equation with respect to $\mu_{\bar{f}}$ and then we obtain

$$\mu_{\bar{f}}(-h^*) = \sup_{f \in F_{\text{ca}}} \mu_f(-h^*).$$

Hence, from Proposition 4.2(c), $\bar{f}$ is bias optimal. This completes the proof.

5. Concluding remarks

In the previous sections we have studied continuous-time jump MDPs in general state and action spaces under the bias and the overtaking optimality criteria. We have proved the existence of a solution to the bias optimality equations, and we have also shown, within the class of canonical policies, the equivalence between bias and overtaking optimal policies (see Theorems 3.2, 4.1, and 4.2).

As mentioned throughout the paper, we have not characterized the whole family of bias optimal policies because, in general, the sets of canonical and average optimal policies do not coincide. In this sense one of the main contributions of this paper is to prove that there always exist bias optimal policies that are canonical (see Proposition 4.2), a subtle result far from being evident.

To conclude, we believe that the results in this paper give a satisfactory answer to the open question proposed in [16] regarding the generalization of bias and overtaking optimality from jump MDPs with denumerable state space to jump MDPs with general state space. Finally, proving the existence of overtaking optimal policies for classes of policies larger than F still remains an open issue.

References

[1] Arapostathis, A. *et al.* (1993). Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM J. Control Optimization* **31,** 282–344.

[2] Cao, X. R. (1998). The relations among potentials, perturbation analysis and Markov decision processes. *Discrete Event Dyn. Syst.* **8,** 71–87.

[3] Cao, X. R. and Chen, H. F. (1997). Potentials, perturbation realization and sensitivity analysis of Markov processes. *IEEE Trans. Automatic Control* **42,** 1382–1397.

[4] Guo, X. P. (2007). Continuous-time Markov decision processes with discounted rewards: the case of Polish spaces. *Math. Operat. Res.* **32,** 73–87.

[5] Guo, X. P. and Liu, K. (2001). A note on optimality conditions for continuous-time Markov decision processes with average cost criterion. *IEEE Trans. Automatic Control* **46,** 1984–1989.

[6] Guo, X. P. and Rieder, U. (2006). Average optimality for continuous-time Markov decision processes in Polish spaces. *Ann. Appl. Prob.* **16,** 730–756.

[7] Guo, X. P., Hernández-Lerma, O. and Prieto-Rumeau, T. (2006). A survey of recent results on continuous-time Markov decision processes. *Top* **14,** 177–261.

[8] Haviv, M. and Puterman, M. L. (1998). Bias optimality in controlled queueing systems. *J. Appl. Prob.* **35,** 136–150.

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

[10] Hernández-Lerma, O. and Lasserre, J. B. (1999). *Further Topics on Discrete-Time Markov Control Processes.* Springer, New York.

[11] Hernández-Lerma, O., Vega-Amaya, O. and Carrasco, G. (1999). Sample-path optimality and variance-minimization of average cost Markov control processes. *SIAM J. Control Optimization* **38,** 79–93.

[12] Jasso-Fuentes, H. and Hernández-Lerma, O. (2008). Characterizations of overtaking optimality for controlled diffusion processes. *Appl. Math. Optimization* **57,** 349–369.
[13] Jasso-Fuentes, H. and Hernández-Lerma, O. (2008). Ergodic control, bias, and sensitive discount optimality for Markov diffusion processes. To appear in *Stoch. Ann. Appl.*
[14] Lund, R. B., Meyn, S. P. and Tweedie, R. L. (1996). Computable exponential convergence rates for stochastically ordered Markov processes. *Ann. Appl. Prob.* **6,** 218–237.
[15] Prieto-Rumeau, T. and Hernández-Lerma, O. (2005). The Laurent series, sensitive discount and Blackwell optimality for continuous-time controlled Markov chains. *Math. Meth. Operat. Res.* **61,** 123–145.
[16] Prieto-Rumeau, T. and Hernández-Lerma, O. (2006). Bias optimality for continuous-time controlled Markov chains. *SIAM J. Control Optimization* **45,** 51–73.
[17] Prieto-Rumeau, T. and Hernández-Lerma, O. (2006). Variance minimization and the overtaking optimality approach to continuous-time controlled Markov chains. Submitted.
[18] Puterman, M. L. (1974). Sensitive discount optimality in controlled one-dimensional diffusions. *Ann. Prob.* **2,** 408–419.
[19] Puterman, M. L. (1994). *Markov Decision Process*. John Wiley, New York.
[20] Zhu, Q. X. (2007). Average optimality inequality for continuous-time Markov decision processes in Polish spaces. *Math. Meth. Operat. Res.* **66,** 299–313.
[21] Zhu, Q. X. (2008). Average optimality for continuous-time Markov decision processes with a policy iteration approach. *J. Math. Analysis Appl.* **339,** 691–704.
[22] Zhu, Q. X. and Guo, X. P. (2005). Another set of conditions for strong n ($n = -1, 0$) discount optimality in Markov decision processes. *Stoch. Anal. Appl.* **23,** 953–974.
[23] Zhu, Q. X. and Guo, X. P. (2007). Markov decision processes with variance minimization: a new condition and approach. *Stoch. Anal. Appl.* **25,** 577–592.