Automatica 45 (2009) 1628–1638

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Continuous-time Markov decision processes with *n*th-bias optimality criteria☆

Junyu Zhang [a], Xi-Ren Cao [b,*]

[a] School of Mathematics and Computational Science, Sun Yat-sen University, Guangzhou, 510275, PR China

[b] Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong

ARTICLE INFO

Article history:
Received 16 January 2008
Received in revised form
28 October 2008
Accepted 7 March 2009
Available online 16 April 2009

Keywords:
Continuous-time systems
Markov decision processes
Multichain model
*n*th-bias optimality criteria
Policy iteration algorithms
Performance analysis
Sensitivity analysis

ABSTRACT

In this paper, we study the *n*th-bias optimality problem for finite continuous-time Markov decision processes (MDPs) with a multichain structure. We first provide *n*th-bias difference formulas for two policies and present some interesting characterizations of an *n*th-bias optimal policy by using these difference formulas. Then, we prove the existence of an *n*th-bias optimal policy by using *n*th-bias optimal policy iteration algorithms, and show that such an *n*th-bias optimal policy can be obtained in a finite number of policy iterations.

1. Introduction

Continuous-time Markov decision processes (MDPs) have received considerable attention because many real-world systems such as communication networks and logistics systems evolve in continuous time (Anderson, 1991; Guo, Hernández-Lerma, & Prieto-Rumeau, 2006). One of the most common optimality criteria in continuous-time MDPs is the *long-run average criterion*; see, for instance, Guo, Song, and Zhang (2009), Howard (1960), Miller (1968), and Puterman (1994) for the case of finite state and action spaces, Guo and Cao (2005), Guo and Hernández-Lerma (2003), Guo and Liu (2001), Haviv and Puterman (1998), Kakumanu (1972), Kitaev and Rykov (1995) and Puterman (1994) for the case of denumerable state spaces, and (Guo & Rieder, 2006) for the case of Polish spaces. In particular, an original idea, called the optimality two-inequality approach, is proposed in Guo and Rieder (2006); this approach allows weaker conditions and can deal with unbounded reward/cost functions which cannot be treated by both the standard optimality inequality approach (Guo & Liu, 2001; Sennott, 1999) and the optimality equation method (Puterman, 1994).

It is well known that the long-run average criterion focuses on the asymptotic or limit behavior of a system and ignores its transient performance. There may exist some policies that yield the same long-run average reward but have quite different finite-horizon rewards. Thus, the long-run average criterion is extremely *underselective* because it does not distinguish such policies that have different finite-horizon total rewards, as long as they have the same long-run average reward. Therefore, it is natural to propose and study more selective optimality criteria. To this end, bias optimality, *n*-discount optimality, and Blackwell optimality were introduced. These criteria are usually referred to as *sensitive discount optimality criteria*; see, for instance, Miller (1968) and Veinott (1969) for *finite* continuous-time MDPs,[1] Guo et al. (2006) and Prieto-Rumeau and Hernández-Lerma (2005, 2006) for denumerable but ergodic continuous-time MDPs. It should be noted that the approaches in these papers depend heavily on both the Laurent series expansion of a discounted reward (cf. (8) in Section 2) and the corresponding results for discrete-time MDPs.

In this paper, we deal with the finite continuous-time MDPs from a different perspective with the concept of *n*th-bias

☆ This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Editor Ian Petersen under the direction of Associate Editor George Yin. Supported by a grant from Hong Kong UGC. The research of the first author was also supported by Sun Yat-sen University Science Foundation.

* Corresponding author. Tel.: +852 2358 7048; fax: +852 2358 1485.
E-mail addresses: mcszhjy@mail.sysu.edu.cn (J. Zhang), eecao@ust.hk (X.-R. Cao).

[1] *A finite continuous-time MDP* is a continuous-time MDP with finite state and finite action spaces.

0005-1098/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2009.03.009

optimality. The nth-bias optimality, $n \geq 0$, is introduced in Cao (2007) and Cao and Zhang (2008) for discrete-time MDPs. The 0th-bias optimality is the long-run average optimality, and the 1st-bias optimality is the bias optimality. In general, the nth bias is defined as the "bias" of the $(n-1)$th bias for $n \geq 1$. For finite MDPs, when n is large enough the nth-bias optimality becomes the Blackwell optimality. The bigger the n is, the more selective the nth-bias optimality is. Essentially, the approach of nth-bias optimality is equivalent to that of n-discount optimality, but it is proposed from a totally different perspective with a sensitivity-based view. A complete theory for finite MDPs with nth-bias optimality criteria can be developed with no discounting. The nth-bias optimality problem was solved in Cao (2007) and Cao and Zhang (2008) for the finite discrete-time MDPs. In this paper, we extend these results to the finite continuous-time MDPs.

Although idea and the primary technique are similar between the discrete-time MDP (DTMDP) and the continuous-time MDP (CTMDP), there still exist the differences as follows.

First, the continuous-time MDP does not have the period problem as the discrete-time MDP. In the case of the discrete-time MDP, we assume that P_d is aperiodic for all $d \in D$. If P_d is periodic, we need to replace the normal limit ($\lim_{L\to\infty}[\cdot]$) with the Cesaro limit ($\lim_{L\to\infty}\frac{1}{L}\sum_{l=0}^{L-1}[\cdot]$) (refer to Cao and Zhang (2008)).

Second, since the nth-bias difference formulas are different in the form (the summation in the DTMDP and the integral in the CTMDP, refer to Lemma 2 in Cao and Zhang (2008) and Theorem 1 in this paper correspondingly), the proof techniques are also different. For example, the proof in Theorem 4 in Cao and Zhang (2008) for the case of DTMDP is direct while the corresponding result in Lemma 5 for CTMDP is proved by contradiction since the former technique does not work.

In this paper, we proceed as follows.

I. We first define nth biases and derive the nth-bias difference formulas of any two policies which have the same $(n-1)$th bias ($n \geq 0$). We prove that an $|S|$th-bias optimal policy is nth-bias optimal for all $n \geq 0$, where $|S|$ is the number of states.

II. By using the nth-bias difference formulas and some simple facts followed from the canonical form of a transition probability function, we derive the necessary and sufficient conditions for the nth-bias optimal policies in Theorems 3 and 4, respectively.

III. Again, by using the nth-bias difference formulas, we develop the nth-bias optimal policy iteration algorithm, for all $n \geq 0$, and prove that it stops at an nth-bias optimal policy in a finite number of iterations and this nth-bias optimal policy satisfies the sufficient conditions.

In summary, starting from any policy we can apply the (0th-bias) policy iteration algorithm, which reaches a (0th-bias) optimal policy; starting from this 0th-bias optimal policy, we can apply the (1st-) bias policy iteration algorithm, which reaches a (1st-) bias optimal policy; and starting from this $(n-1)$th-bias optimal policy, we can apply the nth-bias policy iteration algorithm, which reaches an nth-bias optimal policy; continuing this process, we can obtain an $|S|$th-bias optimal policy, which is a Blackwell optimal policy (Cao, 2007).

Our arguments are based on the nth-bias difference formulas of two policies which are derived in Theorem 1 and on a simple observation from the canonical form of the transition probability function. These formulas are all *new* in the literature and allow us to compare the biases of two different policies based on only one policy's biases under some conditions. Therefore, the policy iteration algorithms follow naturally from the bias difference formulas which actually form the basis of the optimization theory of the nth biases.

There are a number of advantages of our approach based on the performance difference formulas. First, compared with the previous works on MDPs and the n-discount optimality theory (Miller, 1968; Prieto-Rumeau & Hernández-Lerma, 2005, 2006; Veinott, 1969), the approach based on the bias difference formulas makes the presentation and derivation simpler, and is more intuitive and direct. It is completely independent of the discounted MDP formulation and does not depend on Laurent series expansion. Our proofs need no results for discrete-time MDPs or discounted continuous-time MDPs, while the approaches used in Miller (1968) and Veinott (1969) depend heavily on results about *discrete-time* MDPs and those in Prieto-Rumeau and Hernández-Lerma (2005, 2006) on the Laurent series expansion of discounted continuous-time MDPs.

Second, this research is a part of our effort in developing sensitivity-based learning and optimization theory for stochastic systems (see Cao (2007)). The development of this approach is based on the following "philosophical" thought: Optimization approaches such as policy iteration can be derived simply from performance difference formulas (the difference of the performance of any two policies); this sensitivity-based view on performance optimization was first proposed in Cao (2003), and many results in MDPs, perturbation analysis, and reinforcement learning, etc., can be derived and explained naturally with this sensitivity-based view (Cao, 2007). In this paper, we extend the theory of finite discrete-time multichain Markov decision processes with different performance optimization criteria (cf. Arapostathis, Borkar, Fernandez-Gaucherand, Ghosh, and Markus (1993), Cao and Guo (2004), Cao and Zhang (2008) and Lewis and Puterman (2002)) to that of finite continuous-time multichain MDPs. Our on-going research indicates that the sensitivity-based approach can also be applied to systems with continuous state spaces that are driven by Brownian motions and/or Levy processes (for a description of such processes, see Oksendal and Sulem (2007)). Thus, the research in this paper is one step towards a unified approach in optimization of stochastic systems including discrete time and continuous time, discrete states and continuous states.

Third, the sensitivity-based approach also links the policy iteration-based methods to other subjects such as perturbation analysis, gradient-based methods, and reinforcement learning, see Cao (2007) for the discrete-time case. Therefore, the research in this paper opens up topics in these directions such as sample-path-based learning, event-based optimization etc.

In Section 2, we briefly describe the model of continuous-time MDPs and the nth-bias optimality criteria that we are concerned with. In Section 3, we give some technical preliminaries and then derive the nth-bias difference formulas, which are the core of our approach. In Section 4, we derive the necessary and sufficient conditions of nth-bias optimal policies, and finally, in Section 5, we develop the nth-bias optimal policy iteration algorithms and prove that they stop in a finite number of steps to an nth-bias optimal policy satisfying the sufficient conditions.

2. *n*th-bias optimality criteria

In this section, we first state the model of continuous-time MDPs and briefly review some related results, and we then define the nth biases and the nth-bias optimality criteria.

A continuous-time MDP is defined as

$$\{S, A(i), q(j|i,a), r(i,a), a \in A(i), i, j \in S\}, \tag{1}$$

where S is a state space, $A(i)$ is the set of admissible actions at state i, $i \in S$, and we assume that S and $A(i)$, $i \in S$, are finite. Let $K := \{(i,a) : i \in S, a \in A(i)\}$ be the set of pairs of states and actions. The real-valued function $q(j|i,a)$ in (1) is the *transition rates* that satisfy:

(A) $0 \leq q(j|i,a) < \infty$, for all $(i,a) \in K$ and $i \neq j$; and
(B) $\sum_{j\in S} q(j|i,a) = 0$, $q(i|i,a) \leq 0$, for all $(i,a) \in K$.

The reward rate function $r(i, a)$ on K is real-valued. We set $A = \cup_{i\in S} A(i)$.

A continuous-time MDP evolves as follows. A decision-maker observes *continuously* the current state of a system. Whenever the system is in state $i \in S$, he/she chooses an action $a \in A(i)$ (depending on i) according to some rules. In general, if action a is chosen when the system is in state i, the state transition rate of the underlying Markov chain is $q(j|i, a)$ and the decision-maker receives a reward at a rate of $r(i, a)$ (say, \$/sec) (Suppose that $q(j|i, a)$ and $r(i, a)$ are all finite for all $a \in A(i), i, j \in S$). The goal of the decision-maker is to maximize some performance criteria, which in our present case are defined by (4) and (7), below.

We now introduce the class of (deterministic and stationary) policies.

Definition 1. A *policy* is a mapping d: $S \to A$ such that $d(i) \in A(i)$ for all $i \in S$.

We denote by D the family of all deterministic and stationary policies.

For each fixed $d \in D$, $Q(d) := [q(j|i, d(i))]_{i,j\in S}$ is also called an infinitesimal generator matrix; see Feller (1940), Guo and Cao (2005), Guo and Hernández-Lerma (2003), Guo and Liu (2001) and Kakumanu (1972) for instance. Moreover, by Feller (1940), Guo and Cao (2005), Guo and Hernández-Lerma (2003) and Guo and Liu (2001), there exists a unique homogeneous *transition function* $p(i, t, j, d) := P(X_t = j|X_0 = i, d)$ (where $\{X_t, t \geq 0\}$ is a continuous-time Markov chain) having the transition rates $q(j|i, d(i))$ and satisfying the Kolmogorov equation

$$\frac{\mathrm{d}}{\mathrm{d}t}p(i, t, j, d) = \sum_{k\in S} p(i, t, k, d)q(j|k, d(k)) = \sum_{k\in S} q(k|i, d(i))p(k, t, j, d), \tag{2}$$

for all $i, j \in S$ and $t \geq 0$, where t and d typically represent continuous time and policy respectively.

In what follows, we assume that all the relationships among matrices (or vectors) and the operators, such as "limit" and "max", hold componentwisely. Thus, for two vectors x and y defined on state space S, we define $x = y$ if $x(i) = y(i)$ for all $i \in S$, where $x(i)$ and $y(i)$ denote the ith component of x and y, respectively; $x > y$ if $x(i) > y(i)$ for all $i \in S$; $x \geq y$ if $x(i) \geq y(i)$ for all $i \in S$. Furthermore, we define $x \succeq y$ if $x \geq y$ and $x(i) > y(i)$ for at least one $i \in S$. The relation $\geq$ includes $=$, $\geq$, and $>$. Similar definitions are used for the relations $<$, $\leq$ and $\preceq$. Without any confusion, we denote by "0" the matrix and the vector with zero as all of their components, and by I the identity matrix.

We denote by $P(t, d) := [p(i, t, j, d)]_{i,j\in S}$ the (homogeneous) transition matrix of a Markov chain with the infinitesimal generator matrix $Q(d)$. Note that, $P(0, d) = I$. By these notations, (2) can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}P(t, d) = Q(d)P(t, d) = P(t, d)Q(d) \quad \forall t \geq 0, \forall d \in D. \tag{3}$$

By (3) and $P(0, d) = I$, we can obtain

$$P(t, d) = \exp\{Q(d)t\} = \sum_{n=0}^{\infty} \frac{t^n}{n!}(Q(d))^n,$$

where $(Q(d))^0 = I$.

For each $d \in D$, let $r(d)$ be a column vector with ith-component $r(i, d(i))$ for each $i \in S$. Then, the long-run average reward of policy d is defined as

$$g_0^d := \limsup_{T\to\infty} \frac{\int_0^T P(t, d)r(d)\mathrm{d}t}{T}, \tag{4}$$

where g_0^d is a column vector with ith-component $g_0^d(i)$ for each $i \in S$. The *optimal average reward* is defined as $g_0^*(i) := \max_{d\in D} g_0^d(i)$, for all $i \in S$. A policy d^* is said to be *average-reward optimal* if

$$g_0^{d^*} = g_0^*.$$

Let $D_0^* := \{d \in D : g_0^d = g_0^*\}$ be the set of all average-reward optimal policies. We will see that D_0^* is not empty in Section 5.

Thus, as is well known (see Guo et al. (2009), for instance), we have the following lemma.

Lemma 1. *For each $d \in D$, the following assertions hold:*

(a) *The limit $P^*(d) := \lim_{t\to\infty} P(t, d)$ exists, and $g_0^d = P^*(d)r(d)$;*
(b) *$Q(d)P^*(d) = P^*(d)Q(d) = 0$, and $Q(d)g_0^d = 0$;*
(c) *$\int_0^\infty |P(t, d) - P^*(d)|\mathrm{d}t < \infty$, where $|C|$ denotes the absolute value norm of any matrix $C = [c_{ij}]_{i,j\in S}$ defined by $|C| := \sum_{i,j\in S} |c_{ij}|$;*
(d) *$P(t, d)P^*(d) = P^*(d)P(t, d) = P^*(d)P^*(d) = P^*(d)$ for all $t \geq 0$.*

We denote by $B(S)$ the Banach space of all real-valued functions on S. Choosing an arbitrary policy $d \in D$, we can define an operator H_d from $B(S)$ to itself by

$$H_d u := \int_0^\infty [P(t, d) - P^*(d)]u\mathrm{d}t \quad \forall u \in B(S). \tag{5}$$

By Lemma 1(c) and the finiteness of S, we see that $H_d u$ are indeed in $B(S)$ for all $u \in B(S)$.

For each $d \in D$, as is well known, the bias g^d (or equivalently, the 1st bias g_1^d) of d is defined by

$$g^d \equiv g_1^d := \int_0^\infty [P(t, d) - P^*(d)]r(d)\mathrm{d}t = \int_0^\infty [P(t, d)r(d) - g_0^d]\mathrm{d}t, \tag{6}$$

where g^d is a column vector with ith-component $g^d(i)$ for each $i \in S$. The bias of d is the expected total difference between the immediate reward $P(t, d)r(d)$ and the long-run average reward g_0^d.

The *optimal bias* g^* is defined by its components $g^*(i) := \max_{d\in D_0^*} g^d(i)$ for all $i \in S$. A policy $d^* \in D_0^*$ is said to be *bias optimal* if

$$g^{d^*} = g^*.$$

Let $D_1^* := \{d \in D_0^* : g^d = g^*\}$ be the set of all bias optimal policies. With the notation for the 1st bias, we can also write $D_1^* = \{d \in D : g_0^d = g_0^*, g_1^d = g_1^*\}$, with $g_1^* \equiv g^*$. For each $d \in D$, by Lemma 1(a), (5) and (6), we have $g^d = H_d r(d)$. In general, for each $n \geq 1$, by (5), we define inductively that

$$g_n^d := (-1)^{n-1} H_d^n r(d). \tag{7}$$

By Lemma 1(c), g_n^d is finite.

Definition 2. For each $d \in D$ and $n \geq 0$, g_n^d is called the nth bias of policy d.

From (6) and (7), g_2^d is the bias of d when the reward rate is $-g_1^d$, and in general, g_n^d is the bias of d when the reward rate is $-g_{n-1}^d$.

An average-reward optimal policy $d \in D_0^*$ maximizes the long-run time average of the reward received. A bias optimal policy can be viewed as maximizing the total expected reward; since the steady-state reward is the same for all policies in D_0^*, a bias optimal policy in fact maximizes the initial part of the expected reward, or the transient rewards. Furthermore, as we can see from (6), bias places equal weight on rewards received at different times. We, however, sometimes prefer to receive the rewards earlier, this motivates us to study other optimality criteria that are more selective than the long-run average and bias optimality. They are the nth-bias optimality defined as follows.

Definition 3. The *optimal nth bias* is denoted as $g_n^*(i) := \max_{d\in D_{n-1}^*} g_n^d(i)$, for all $i \in S$. A policy $d^* \in D_{n-1}^*$ is called *nth-bias optimal* if

$$g_n^{d^*}(i) = g_n^*(i) \quad \forall i \in S, n \geq 0,$$

where $D_{n-1}^* := \{d \in D_{n-2}^* : g_{n-1}^d = g_{n-1}^*\} = \{d \in D : g_l^d = g_l^*, l = 0, 1, \ldots, n-1\}$ is the set of all $(n-1)$th-bias optimal policies, and $D_{-1}^* := D$.

From the definition, we can see that g_n^* always exists. We will prove that an nth-bias optimal policy, $n \geq 0$, also always exists in Section 5.

From Definition 3, the 0th-bias optimality and the 1st-bias optimality are the same as the average-reward optimality and the bias optimality, respectively. The nth-bias optimality has a clear physical meaning (see Cao and Zhang (2008)). For example, the 2nd-bias optimality gives a larger weight on the early rewards. From Definition 3, an $(n+1)$th-bias optimal policy is also nth-bias optimal, i.e., $D_{n+1}^* \subseteq D_n^*$ for all $n \geq 0$. That is, the bigger the n is, the more selective the nth-bias optimality is.

As a comparison, we note that the n-discount optimality in the existing literature (see Taylor (1976) and Veinott (1969)) is based on the following expansion

$$\frac{g_0^d}{\alpha} + \sum_{n=1}^{\infty} \alpha^{n-1} g_n^d = \int_0^{\infty} \mathrm{e}^{-\alpha t} P(t, d) r(d) \mathrm{d}t, \tag{8}$$

where $\alpha > 0$, which is the discounted factor.

The right-hand side of (8) is the α-discounted expected reward of d, and the left-hand side of (8) is the Laurent series of the α-discounted expected reward of d. From (4) and (8), we see that the long-run average criterion (e.g., Guo and Hernández-Lerma (2003), Guo and Liu (2001), Guo et al. (2009), Haviv and Puterman (1998), Kakumanu (1972), Kitaev and Rykov (1995) and Puterman (1994)) is concerned with the optimality of the first term of the Laurent series, and the bias criterion (e.g., Miller (1968), Prieto-Rumeau and Hernández-Lerma (2005, 2006) and Veinott (1969)), with the optimality of the second term of the Laurent series given that its first term has been optimized. We also see that our nth-bias optimality is equivalent to the n-discount optimality in Prieto-Rumeau and Hernández-Lerma (2005).

The main goal of this paper is to provide a *new* and *self-contained* approach to the finite continuous-time MDPs with the nth-bias optimality criteria by using the performance difference formulas. We show the existence of the nth-bias optimal policies, propose policy iteration algorithms, and prove their convergence to the nth-bias optimal policies, $n \geq 0$. The approach is based on the nth-bias difference formulas from the sensitivity-based view.

3. *n*th-bias difference formulas

In this section, we provide the nth-bias difference formulas of any two different policies which have the same $(n-1)$th bias, $n \geq 1$ (we provide the 0th-bias difference formulas of any two different policies when the case $n = 0$). These formulas are used to prove the existence of an nth-bias optimal policy for all $n \geq 0$.

First note that, for each $d \in D$, by (5) and (7) we have

$$g_{n+1}^d = -\int_0^{\infty} [P(t, d) - P^*(d)] g_n^d \mathrm{d}t \quad \forall n \geq 1. \tag{9}$$

(9) is not convenient to compute g_{n+1}^d since it is an infinite integral. To solve g_{n+1}^d straightforward, we give a lemma below.

Lemma 2. *Let $d \in D$, then*

(a) *the Poisson equation holds:*

$$g_0^d = r(d) + Q(d) g_1^d. \tag{10}$$

(b) $P^*(d) g_n^d = 0$ *for all* $n \geq 1$.

(c) *For each fixed $n \geq 1$, g_{n+1}^d is the* unique *solution to the following equations:*

$$P^*(d) g_{n+1}^d = 0, \tag{11}$$

$$Q(d) g_{n+1}^d = g_n^d. \tag{12}$$

Proof. (a) By Lemma 1(a), Lemma 1(b), (3) and (6) we have

$$\begin{aligned} Q(d) g_1^d &= Q(d) \int_0^{\infty} [P(t, d) - P^*(d)] r(d) \mathrm{d}t \\ &= \int_0^{\infty} Q(d) P(t, d) r(d) \mathrm{d}t \\ &= \lim_{T\to\infty} \int_0^T \frac{\mathrm{d}}{\mathrm{d}t} P(t, d) r(d) \mathrm{d}t \\ &= \lim_{T\to\infty} [P(T, d) - I] r(d) \\ &= P^*(d) r(d) - r(d) \\ &= g_0^d - r(d), \end{aligned}$$

which gives (a).

(b) By Lemma 1(d) and (6), (b) holds for $n = 1$. In general, by Lemma 1(d) and (9), we have that, for each $n \geq 1$

$$\begin{aligned} P^*(d) g_{n+1}^d &= -P^*(d) \int_0^{\infty} [P(t, d) - P^*(d)] g_n^d \mathrm{d}t \\ &= -\int_0^{\infty} [P^*(d) P(t, d) - P^*(d) P^*(d)] g_n^d \mathrm{d}t = 0. \end{aligned} \tag{13}$$

Thus, (b) follows.

(c) By (9), Lemma 2(b) and (3), we have

$$\begin{aligned} Q(d) g_{n+1}^d &= -Q(d) \int_0^{\infty} [P(t, d) - P^*(d)] g_n^d \mathrm{d}t \\ &= -\int_0^{\infty} [Q(d) P(t, d)] g_n^d \mathrm{d}t \\ &= -[P^*(d) - I] g_n^d = g_n^d, \end{aligned}$$

which, together with (13), implies that g_{n+1}^d is a solution to (11) and (12). To prove the uniqueness, suppose that

$$P^*(d) x = 0 \quad \text{and} \quad Q(d) x = g_n^d.$$

Then, by Lemma 2(b) and Lemma 1(d), we have

$$\begin{aligned} P(t, d) Q(d) x &= P(t, d) g_n^d = P(t, d)[g_n^d - P^*(d) g_n^d] \\ &= [P(t, d) - P^*(d)] g_n^d, \end{aligned}$$

which, together with (3) and by a straightforward calculation, gives

$$[P(T, d) - I] x = \int_0^T [P(t, d) - P^*(d)] g_n^d \mathrm{d}t.$$

Letting $T \to \infty$, by $P^*(d) x = 0$ and Lemma 1(a), we get

$$x = -\int_0^{\infty} [P(t, d) - P^*(d)] g_n^d \mathrm{d}t = g_{n+1}^d,$$

so the uniqueness follows. □

By Lemma 2(b), (9) can be rewritten as

$$g_{n+1}^d = -\int_0^{\infty} P(t, d) g_n^d \mathrm{d}t \quad \forall n \geq 1. \tag{14}$$

We introduce the following notations. The group inverse of $Q(d)$ is defined as $Q(d)^{\#} = [Q(d) - P^*(d)]^{-1} + P^*(d)$ (cf. Cao (2007)). We have

$$Q(d) Q(d)^{\#} = Q(d)^{\#} Q(d) = I - P^*(d).$$

By $P^*(d)g_1^d = 0$ and pre-multiplying both sides of the Poisson equation (10) by $Q(d)^\#$, we obtain

$$\begin{aligned} g_1^d &= -Q(d)^\#[r(d) - g_0^d] \\ &= [-Q(d) + P^*(d)]^{-1}(r(d) - g_0^d). \end{aligned} \tag{15}$$

With the same reasoning, by (11) ($P^*(d)g_{n+1}^d = 0$) and pre-multiplying both sides of (12) by $Q(d)^\#$, we can rewrite (14) as

$$\begin{aligned} g_{n+1}^d &= Q(d)^\# g_n^d \\ &= -[-Q(d) + P^*(d)]^{-1} g_n^d \\ &= (-1)^n[-Q(d) + P^*(d)]^{-(n+1)}(r(d) - g_0^d) \quad \forall n \geq 1. \end{aligned} \tag{16}$$

Note that (15) and (16) for continuous-time MDPs have a similar form as those for discrete-time MDPs (see in Cao and Zhang (2008)).

Now we give our first main result about the nth-bias difference.

Theorem 1. *Suppose that d and h are both in D. Then,*

(a) $g_0^h - g_0^d = P^*(h)[r(h) + Q(h)g_1^d - g_0^d] + [P^*(h) - I]g_0^d$.

(b) *If $g_0^d = g_0^h$, then*

$$\begin{aligned} g_1^h - g_1^d &= \int_0^\infty P(t,h)[r(h) + Q(h)g_1^d - g_0^d]\mathrm{d}t \\ &\quad + P^*(h)[Q(h) - Q(d)]g_2^d \\ &= \int_0^\infty P(t,h)[r(h) + Q(h)g_1^d - g_0^d]\mathrm{d}t \\ &\quad + P^*(h)[g_1^h - g_1^d]. \end{aligned}$$

(c) *For any fixed $n \geq 1$, if $g_n^d = g_n^h$, then*

$$\begin{aligned} g_{n+1}^h - g_{n+1}^d &= \int_0^\infty P(t,h)[Q(h) - Q(d)]g_{n+1}^d\mathrm{d}t \\ &\quad + P^*(h)[Q(h) - Q(d)]g_{n+2}^d. \end{aligned}$$

Proof. (a) Since $P^*(h)Q(h) = 0$ (by Lemma 1(b)), from Lemma 2(a) we have

$$\begin{aligned} g_0^h - g_0^d &= P^*(h)r(h) - g_0^d \\ &= P^*(h)[r(h) + Q(h)g_1^d - r(d) - Q(d)g_1^d] \\ &\quad + P^*(h)r(d) - g_0^d + P^*(h)Q(d)g_1^d \\ &= P^*(h)[r(h) + Q(h)g_1^d - r(d) - Q(d)g_1^d] \\ &\quad + P^*(h)[r(d) + Q(d)g_1^d] - g_0^d \\ &= P^*(h)[r(h) + Q(h)g_1^d - g_0^d] + [P^*(h) - I]g_0^d. \end{aligned}$$

This implies (a).

(b) By (6), we have

$$\begin{aligned} g_1^h - g_1^d &= \int_0^\infty [P(t,h)r(h) - P(t,d)r(d)]\mathrm{d}t \\ &=: \int_0^\infty P(t,h)[r(h) + Q(h)g_1^d - g_0^d]\mathrm{d}t + \Delta, \end{aligned} \tag{17}$$

where, $\Delta := -\int_0^\infty P(t,h)Q(h)g_1^d\mathrm{d}t + \int_0^\infty [P(t,h)g_0^d - P(t,d)r(d)]\mathrm{d}t$.

Then, by (3) and a straightforward calculation, we have

$$\Delta = g_1^d - P^*(h)g_1^d + \int_0^\infty [P(t,h)g_0^d - P(t,d)r(d)]\mathrm{d}t.$$

Thus, by $P^*(d)r(d) = g_0^d = g_0^h = P^*(h)r(h)$ and Lemma 1(d), we have

$$\begin{aligned} \Delta &= g_1^d - P^*(h)g_1^d + \int_0^\infty [P(t,h)g_0^h - P(t,d)r(d)]\mathrm{d}t \\ &= g_1^d - P^*(h)g_1^d + \int_0^\infty [P^*(h)r(h) - P(t,d)r(d)]\mathrm{d}t \\ &= g_1^d - P^*(h)g_1^d - \int_0^\infty [P(t,d) - P^*(d)]r(d)\mathrm{d}t \\ &= -P^*(h)g_1^d, \end{aligned}$$

which, together with (17), Lemma 1(b) and (12), gives

$$\begin{aligned} g_1^h - g_1^d &= \int_0^\infty P(t,h)[r(h) + Q(h)g_1^d - g_0^d]\mathrm{d}t \\ &\quad + P^*(h)Q(h)g_2^d - P^*(h)Q(d)g_2^d \\ &= \int_0^\infty P(t,h)[r(h) + Q(h)g_1^d - g_0^d]\mathrm{d}t \\ &\quad + P^*(h)[Q(h) - Q(d)]g_2^d \\ &= \int_0^\infty P(t,h)[r(h) + Q(h)g_1^d - g_0^d]\mathrm{d}t \\ &\quad + P^*(h)[g_1^h - g_1^d]. \end{aligned}$$

This implies (b).

(c) By (12) and (14), from $g_n^d = g_n^h$, we have

$$\begin{aligned} &g_{n+1}^h - g_{n+1}^d \\ &= \int_0^\infty [P(t,d) - P(t,h)]g_n^d\mathrm{d}t \\ &= \int_0^\infty [P(t,d) - P(t,h)]Q(d)g_{n+1}^d\mathrm{d}t \\ &= \int_0^\infty P(t,h)[Q(h) - Q(d)]g_{n+1}^d\mathrm{d}t \\ &\quad + \int_0^\infty P(t,d)Q(d)g_{n+1}^d\mathrm{d}t - \int_0^\infty P(t,h)Q(h)g_{n+1}^d\mathrm{d}t \\ &= \int_0^\infty P(t,h)[Q(h) - Q(d)]g_{n+1}^d\mathrm{d}t + P^*(d)g_{n+1}^d - P^*(h)g_{n+1}^d \\ &= \int_0^\infty P(t,h)[Q(h) - Q(d)]g_{n+1}^d\mathrm{d}t - P^*(h)g_{n+1}^d, \end{aligned}$$

which, together with (12) and Lemma 1(b), gives

$$\begin{aligned} g_{n+1}^h - g_{n+1}^d &= \int_0^\infty P(t,h)[Q(h) - Q(d)]g_{n+1}^d\mathrm{d}t \\ &\quad - P^*(h)Q(d)g_{n+2}^d \\ &= \int_0^\infty P(t,h)[Q(h) - Q(d)]g_{n+1}^d\mathrm{d}t \\ &\quad + P^*(h)[Q(h) - Q(d)]g_{n+2}^d. \end{aligned}$$

This implies (c). □

Theorem 1 gives the nth-bias difference formulas of two policies having the same $(n-1)$th bias. These formulas are all *new* in the literature and will be used to prove some interesting characterizations of the nth-bias optimal policies below.

Theorem 2. *Let $|S|$ be the number of states in S. If $d \in D$ is $|S|$th-bias optimal, then d is also nth-bias optimal for all $n \geq 0$.*

Proof. By Definition 3 and the uniqueness of $P(t,d)$ determined by $Q(d)$, it suffices to show that $g_n^h = g_n^d$ for all $n \geq |S|$ and $d, h \in D_{|S|}^*$ with $Q(h) \neq Q(d)$. In fact, for each $h \in D_{|S|}^*$, as $d \in D$ is $|S|$th-bias

optimal, we have $g_n^h = g_n^d$ for all $0 \le n \le |S|$. Then, by Lemma 1(a), Lemma 1(b) and (12), we have

$$[Q(h) - Q(d)]g_0^d = Q(h)g_0^h - Q(d)g_0^d = 0 \quad \text{and}$$
$$[Q(h) - Q(d)]g_n^d = Q(h)g_n^h - Q(d)g_n^d = 0, \quad \forall 2 \le n \le |S|,$$

which, together with (7), give

$$[Q(h) - Q(d)]g_0^d = 0 \quad \text{and} \quad [Q(h) - Q(d)]H_d^n r(d) = 0, \quad \forall 2 \le n \le |S|. \tag{18}$$

Therefore, the vectors g_0^d and $H_d^n r(d) (2 \le n \le |S|)$ belong to the null space of (the operator determined by) $Q(h) - Q(d)$. Since $Q(h) - Q(d) \ne 0$, the rank of $Q(h) - Q(d)$ must be at least 1, and so the dimension of the null space of $Q(h) - Q(d)$ is at most $|S| - 1$. Hence, g_0^d and $H_d^n r(d) (2 \le n \le |S|)$ are linearly dependent. Thus, there exists an integer $1 \le k \le |S| - 1$ such that $H_d^{k+1} r(d)$ is a linear combination of g_0^d and $H_d^n r(d)$ $(2 \le n \le k)$.

We now show by induction that, for each $m \ge 2$, $H_d^m r(d)$ is a linear combination of g_0^d and $H_d^n r(d)$ $(2 \le n \le k)$. To see this, suppose that this conclusion holds for some $m (\ge k + 1)$. That is, there exist k numbers λ_l such that

$$H_d^m r(d) = \lambda_1 g_0^d + \sum_{l=2}^{k} \lambda_l (H_d^l r(d)),$$

which, together with (5) and $H_d g_0^d = 0$, gives

$$H_d^{m+1} r(d) = \sum_{l=2}^{k-1} \lambda_l (H_d^{l+1} r(d)) + \lambda_k (H_d^{k+1} r(d)). \tag{19}$$

Since $H_d^{k+1} r(d)$ is a linear combination of g_0^d and $H_d^n r(d)$ $(2 \le n \le k)$, it follows from (19) that $H_d^{m+1} r(d)$ is also a linear combination of g_0^d and $H_d^n r(d)$ $(2 \le n \le k)$, and so the desired conclusion is proved. Therefore, by (7) and (18) we have

$$[Q(h) - Q(d)]g_0^d = 0 \quad \text{and} \quad [Q(h) - Q(d)]g_n^d = 0 \quad \forall n \ge |S|,$$

which, together with Theorem 1(c), gives that, for each $n \ge |S|$,

$$g_{n+1}^h - g_{n+1}^d = \int_0^\infty P(t, h)[Q(h) - Q(d)]g_{n+1}^d \mathrm{d}t + P^*(h)[Q(h) - Q(d)]g_{n+2}^d = 0.$$

The proof is completed. □

Theorem 2 is very interesting. It shows that, in order to obtain a policy which is nth-bias optimal for *all* $n \ge 0$, it suffices to find an $|S|$th-bias optimal policy. We can also easily prove that if $d \in D_{|S|}^*$, then d is also a Blackwell optimal policy. Thus, in what follows, we only need to focus on the existence and calculation of an $|S|$th-bias optimal policy. We have some simple but useful lemmas.

Lemma 3. (a) *For each $d \in D$, if $P^*(d)u = 0$ and $u \le 0$ (or $u \ge 0$), then $u(i) = 0$ for all recurrent states i under $Q(d)$.*
(b) *For each $d \in D$, if $u(i) \ge 0$ (or $u(i) \le 0$) for all recurrent states i under $Q(d)$, then $P^*(d)u \ge 0$ (or $P^*(d)u \le 0$).*

Proof. Let $C_k^d \subset S$, $k = 1, 2, \ldots, m$, be the disjoint closed irreducible sets of the recurrent states under $Q(d)$, where m is the number of such sets; and C_{m+1}^d is the set of transient states. First, it is well known (e.g., Anderson (1991)) that by reordering the states, $P^*(d)$ takes the *canonical form* of

$$P^*(d) = \begin{bmatrix} P_1^*(d) & 0 & 0 & \cdots & 0 \\ 0 & P_2^*(d) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & P_m^*(d) & 0 \\ W_1^*(d) & W_2^*(d) & \cdots & W_m^*(d) & 0 \end{bmatrix}, \tag{20}$$

in which $P_k^*(d)$ corresponds to the transitions among states in C_k^d, $k = 1, 2, \ldots, m$; $W_k^*(d)$ to the transitions from the transient states in C_{m+1}^d to the recurrent states in C_k^d, $k = 1, \ldots, m$. Since the columns in $P^*(d)$ corresponding to transient states are all zeros; thus, all $u(l)$'s with l being transient states contribute nothing to $P^*(d)u$. Further, $P_k^*(d) > 0$ for $k = 1, 2, \ldots, m$. Then the lemma follows directly from the canonical form of $P^*(d)$ (20). □

Lemma 4. *The following assertions hold for any $d, h \in D$:*
(a) *If $g_0^d = g_0^h$, then $Q(d)g_0^h = 0$, and $P^*(h)[r(h) + Q(h)g_1^d - g_0^d] = 0$.*
(b) *If $g_n^d = g_n^h$ with $n \ge 1$, then $P^*(h)[Q(h) - Q(d)]g_{n+1}^d = 0$.*

Proof. (a) Since $g_0^d = g_0^h$, by Lemma 1(a) and (b), we have

$$Q(d)g_0^h = Q(d)g_0^d = Q(d)P^*(d)r(d) = 0.$$

Thus, the first part of (a) follows. Moreover, by $g_0^d = g_0^h = P^*(h)r(h)$ and Lemma 1(a) and (d), we have

$$[P^*(h) - I]g_0^d = [P^*(h) - I]g_0^h = 0,$$

which, together with Lemma 1(b), gives the second part of (a).

(b) By (12) and $g_n^d = g_n^h$, we have

$$P^*(h)Q(d)g_{n+1}^d = P^*(h)g_n^d = P^*(h)g_n^h = 0,$$

which, together with Lemma 1(b), yields (b). □

4. Necessary and sufficient conditions for *n*th-bias optimal policies

From the right-hand sides of the nth-bias difference formulas in Theorem 1 and Lemma 3 in Section 3, we can derive the bias optimality Eqs. (24)–(26) in the following.

We now give another characterization of an nth-bias optimal policy. To do so, we need to introduce some notations:

For each $d \in D$ and $n \ge 1$, define

$$A_0^d(i) := \left\{ a \in A(i) : \sum_{j \in S} q(j|i, a) g_0^d(j) = 0 \right\}, \tag{21}$$
$$A_1^d(i) := \left\{ a \in A_0^d(i) : r(i, a) + \sum_{j \in S} q(j|i, a) g_1^d(j) = g_0^d(i) \right\},$$
$$\vdots$$
$$A_n^d(i) := \left\{ a \in A_{n-1}^d(i) : \sum_{j \in S} q(j|i, a) g_n^d(j) = g_{n-1}^d(i) \right\}, \quad \forall i \in S.$$

Apparently, by (10) and (12), we have $d(i) \in A_n^d(i)$, for all $i \in S$ and for all $n \ge 0$ and the action space $A_n^d(i)$ depends on policy d. When d is an nth-bias optimal policy d^*, we denote $A_n^d(i)$ as $A_n(i)$ (since it does not depend on d, hereby we have $d^*(i) \in A_n(i)$), for all $i \in S$. That is,

$$A_0(i) := \left\{ a \in A(i) : \sum_{j \in S} q(j|i, a) g_0^*(j) = 0 \right\}, \tag{22}$$
$$A_1(i) := \left\{ a \in A_0(i) : r(i, a) + \sum_{j \in S} q(j|i, a) g_1^*(j) = g_0^*(i) \right\},$$
$$\vdots$$
$$A_n(i) := \left\{ a \in A_{n-1}(i) : \sum_{j \in S} q(j|i, a) g_n^*(j) = g_{n-1}^*(i) \right\},$$
$$\forall i \in S. \tag{23}$$

Theorem 3. *Suppose that policy d^* satisfies the following $(n+2)$ $(n \geq 0)$ bias optimality conditions*

$$\max_{a\in A(i)}\left\{\sum_{j\in S} q(j|i,a)g_0^{d^*}(j)\right\} = 0, \tag{24}$$

$$\max_{a\in A_0^{d^*}(i)}\left\{r(i,a)+\sum_{j\in S} q(j|i,a)g_1^{d^*}(j)\right\} = g_0^{d^*}(i), \tag{25}$$

$$\max_{a\in A_k^{d^*}(i)}\left\{\sum_{j\in S} q(j|i,a)g_{k+1}^{d^*}(j)\right\} = g_k^{d^*}(i),$$
$$\forall i \in S \text{ and } 1 \leq k \leq n. \tag{26}$$

Then, d^* is nth-bias optimal.

Proof. First, consider the case $n = 0$. For each $d \in D$, since d^* satisfies (24), we have $u := Q(d)g_0^{d^*} \leq 0$. Then, by Lemma 3(a) and Lemma 1(b), we have $u(i) = 0$ for all recurrent states i under $Q(d)$. Thus, it follows from (25) and (21) that $v(i) := [r(d) + Q(d)g_1^{d^*} - g_0^{d^*}](i) \leq 0$ for all recurrent states i under $Q(d)$. By Lemma 3(b), we have $P^*(d)v \leq 0$. On the other hand, since $Q(d)g_0^{d^*} \leq 0$, we have $P(t,d)Q(d)g_0^{d^*} \leq 0$ for all $t \geq 0$. Thus, by (3) we have $[P(T,d)-I]g_0^{d^*} \leq 0$ for all $T \geq 0$; therefore, from Lemma 1(a) and letting $T \to \infty$, we have $[P^*(d)-I]g_0^{d^*} \leq 0$. Thus, by Theorem 1(a) we have $g_0^d - g_0^{d^*} = P^*(d)v + [P^*(d) - I]g_0^{d^*} \leq 0$, for all $d \in D$. This means that d^* is 0th-bias optimal, i.e., in D_0^*.

Second, consider the case $n = 1$. For each $d \in D_0^*$, we have $g_0^{d^*} = g_0^d$, and so it follows from Lemma 1(b) that $d(i)$ is in $A_0^{d^*}(i)$ for all $i \in S$. Thus, by (25) we have $g_0^{d^*} - r(d) - Q(d)g_1^{d^*} \geq 0$. Therefore, by Lemmas 3(a) and 4(a), we have

$$r(i,d(i)) + \sum_{j\in S} q(j|i,d(i))g_1^{d^*}(j) = g_0^{d^*}(i)$$

for all recurrent states i under $Q(d)$, which, together with the definition of $A_1^{d^*}(i)$ and (26) with $n = 1$, gives

$$\sum_{j\in S} q(j|i,d(i))g_2^{d^*}(j) \leq g_1^{d^*}(i) = \sum_{j\in S} q(j|i,d^*(i))g_2^{d^*}(j),$$

which holds for all recurrent states i under $Q(d)$, and so $P^*(d)[Q(d) - Q(d^*)]g_2^{d^*} \leq 0$ (by Lemma 3(b)). Since we have shown that $r(d) + Q(d)g_1^{d^*} - g_0^{d^*} \leq 0$, by $P(t,d) \geq 0$ and $P^*(d)[Q(d) - Q(d^*)]g_2^{d^*} \leq 0$, we have

$$\int_0^\infty P(t,d)[r(d) + Q(d)g_1^{d^*} - g_0^{d^*}]\mathrm{d}t + P^*(d)[Q(d) - Q(d^*)]g_2^{d^*} \leq 0,$$

which, together with Theorem 1(b), yields

$$g_1^d - g_1^{d^*} \leq 0,$$

and so d^* is 1st-bias optimal, i.e., in D_1^*.

Finally, consider the case $n \geq 2$. By induction, suppose that d^* is in D_m^* for some $1 \leq m \leq n-1$. Then, to show that d^* is in D_{m+1}^*, by Definition 3 we need to prove that $g_{m+1}^{d^*} \geq g_{m+1}^d$ for all $d \in D_m^*$. In fact, for each $d \in D_m^*$, by the definition of D_m^* and the induction hypothesis, we have $g_l^{d^*} = g_l^d$ for all $0 \leq l \leq m$. Then, we see that $d(i)$ is in $A_m^{d^*}(i)$ for all $i \in S$ (by (12) for $m \geq 2$ and by (10) for $m = 1$), and then it follows from (12) and (26) that

$$Q(d)g_{m+1}^{d^*} \leq g_m^{d^*} = Q(d^*)g_{m+1}^{d^*}.$$

Hence, by Lemmas 3(a) and 4(b), we have

$$\sum_{j\in S} q(j|i,d(i))g_{m+1}^{d^*}(j) = g_m^{d^*}(i),$$

for any recurrent state i under $Q(d)$, and so $d(i)$ is in $A_{m+1}^{d^*}(i)$ for all recurrent states i under $Q(d)$. Thus, by (26) we have

$$\sum_{j\in S} q(j|i,d(i))g_{m+2}^{d^*}(j) \leq g_{m+1}^{d^*}(i) = \sum_{j\in S} q(j|i,d^*(i))g_{m+2}^{d^*}(j)$$

for all recurrent states i under $Q(d)$, and so $P^*(d)[Q(d) - Q(d^*)]g_{m+2}^{d^*} \leq 0$ (by Lemma 3(b)). Since we have shown that $Q(d)g_{m+1}^{d^*} - g_m^{d^*} \leq 0$, by $P(t,d) \geq 0$ we have $P(t,d)[Q(d) - Q(d^*)]g_{m+1}^{d^*} \leq 0$, and so

$$\int_0^\infty P(t,d)[Q(d)g_{m+1}^{d^*} - g_m^{d^*}]\mathrm{d}t + P^*(d)[Q(d) - Q(d^*)]g_{m+2}^{d^*} \leq 0,$$

which, together with Theorem 1(c), gives

$$g_{m+1}^d - g_{m+1}^{d^*} \leq 0, \quad \forall d \in D_m^*,$$

and so d^* is $(m+1)$th-bias optimal, i.e., in D_{m+1}^*. Hence, Theorem 3 is proved. □

Theorem 3 provides a sufficient condition for the nth-bias optimal policies. We now give a necessary condition for the nth-bias optimal policies.

Theorem 4. *The optimal kth biases, g_k^*, $k = 0, 1, \ldots, n$, satisfy the first $(n+1)$ bias optimality conditions, $n \geq 0$.*

Proof. We will prove that, there exists an nth-bias optimal policy in Section 5 using the construction method, and we denote it as d_n^*, for $n \geq 0$.

We first consider the case $n = 0$. Let d_0^* be an average optimal policy with average reward g_0^* and bias $g_1^{d_0^*}$. From Lemma 1(b), $Q(d_0^*)g_0^* = 0$. We need to prove that g_0^* satisfies the first bias optimality condition (24), i.e., $Q(d)g_0^* \leq 0$ for all $d \in D$. Assume that this does not hold; that is, there exists a policy h and some state $i \in S$ such that $(Q(h)g_0^*)(i) > 0$. Based on this, we can construct another policy $\hat{d}$ by setting $\hat{d}(j) = d_0^*(j)$ for all $j \in S - \{i\}$ and $\hat{d}(i) = h(i)$. Consequently, $r(\hat{d})(j) = r(d_0^*)(j)$ for $j \in S - \{i\}$ and $r(\hat{d})(i) = r(h)(i)$. Then we have $(Q(\hat{d})g_0^*)(i) > 0$ and $(Q(\hat{d})g_0^*)(j) = 0$ for $j \in S - \{i\}$. Thus,

$$Q(\hat{d})g_0^* \geq 0. \tag{27}$$

Therefore, we have $P(t,\hat{d})Q(\hat{d})g_0^* \geq 0$ for all $t \geq 0$. Thus, by (3) we have $[P(T,\hat{d})-I]g_0^* \geq 0$ for all $T \geq 0$. From Lemma 1(a) and letting $T \to \infty$, we have $[P^*(\hat{d})-I]g_0^* \geq 0$. Assume that $[P^*(\hat{d})-I]g_0^* = 0$. We get $Q(\hat{d})g_0^* = Q(\hat{d})P^*(\hat{d})g_0^* = 0$, which contradicts (27). As a result,

$$[P^*(\hat{d}) - I]g_0^* \geq 0. \tag{28}$$

Since $P^*(\hat{d})Q(\hat{d}) = 0$, $P^*(\hat{d})Q(\hat{d})g_0^* = 0$ follows. By Lemma 3(a) we have $(Q(\hat{d})g_0^*)(j) = 0$ for all recurrent states j under policy $\hat{d}$. Then the particular state i must be a transient state under policy $\hat{d}$. By the construction of $\hat{d}$, we have $P^*(\hat{d})[r(\hat{d}) - r(d_0^*) + (Q(\hat{d}) - Q(d_0^*))g_1^{d_0^*}] = 0$. (The only nonzero component of the vector in bracket is in state i which is a transient state.) Finally, by the average-reward difference formula in Theorem 1(a) and (28), we have

$$g_0^{\hat{d}} - g_0^* = [P^*(\hat{d}) - I]g_0^* \geq 0.$$

This is impossible because g_0^* is the optimal average reward. Hence, the assumption does not hold. Therefore, the theorem holds for $n = 0$.

The case $n > 0$ can be proved in the same way by constructing counterexamples. Considering the length of this paper, we omit the proof of the case $n > 0$. □

Because the nth-bias optimal policy exists (in Section 5), the solution to the first $(n + 1)$ bias optimality equations also exists.

In the following section, we propose iteration algorithms for nth-bias optimal policies; and with these algorithms, we prove that the existence of a policy d^* satisfying the optimality equations (24)–(26).

5. Policy iteration algorithms for nth-bias optimal policies

5.1. 0th-bias optimal policy iteration algorithm

In this subsection, we provide an iteration algorithm for searching a 0th-bias optimal policy.

We first introduce some notations. For a given $d \in D$, $i \in S$, and $a \in A(i)$, let

$$w_d(i, a) := r(i, a) + \sum_{j \in S} q(j|i, a) g_1^d(j), \tag{29}$$

and

$$B_0^d(i) := \left\{ a \in A(i) : \begin{array}{l} \sum_{j \in S} q(j|i, a) g_0^d(j) > 0; \quad \text{or} \\ w_d(i, a) > w_d(i, d(i)) \\ \text{when } \sum_{j \in S} q(j|i, a) g_0^d(j) = 0 \end{array} \right\}. \tag{30}$$

We then define an improvement policy $h \in D$ (depending on d) as below:

$$h(i) \in B_0^d(i) \quad \text{when } B_0^d(i) \neq \emptyset,$$

$$\text{and} \quad h(i) = d(i) \quad \text{if } B_0^d(i) = \emptyset. \tag{31}$$

Note that such a policy h may not be unique since there may be more than one action in $B_0^d(i)$ for some state $i \in S$. Let

$$u_d^h := Q(h) g_0^d, \quad v_d^h := r(h) + Q(h) g_1^d - g_0^d. \tag{32}$$

Lemma 5. *For any given $d \in D$, let h be defined as in (31). Then,*

(a) $g_0^h \geq g_0^d$.
(b) *If $g_0^h = g_0^d$ and $h \neq d$, then $g_1^h \geq g_1^d$.*

Proof. (a) By (30) and (31), we have $u_d^h = Q(h) g_0^d \geq 0$. Then, by Lemmas 1(b) and 3(a), we have $u_d^h(i) = 0$ for all recurrent states i under $Q(h)$, and so it follows from (30) and (32) that $v_d^h(i) \geq 0$ for all recurrent states i under $Q(h)$. Moreover, by $Q(h) g_0^d \geq 0$ and (3), we have $[P(T, h) - I] g_0^d \geq 0$ for all $T \geq 0$, which together with Lemma 1(a) gives $[P^*(h) - I] g_0^d \geq 0$. Thus, by Theorem 1(a) and Lemma 3(b), we have $g_0^h - g_0^d = P^*(h) v_d^h + [P^*(h) - I] g_0^d \geq 0$, and thus (a) follows.

(b) We first prove that $g_1^h \geq g_1^d$. By $g_0^h = g_0^d$, we have $Q(h) g_0^d = 0$ and $[P^*(h) - I] g_0^d = 0$. From (30)–(32), we have $v_d^h \geq 0$. Noting that $g_0^h - g_0^d = P^*(h) v_d^h + [P^*(h) - I] g_0^d = P^*(h) v_d^h = 0$ and by Lemma 3(a), we have $v_d^h(i) = 0$, for all recurrent states i under policy h. From (30)–(32),

$$h(i) = d(i), \quad \text{for all recurrent states } i \text{ under policy } h. \tag{33}$$

By (33), we get $P^*(h)[Q(h) - Q(d)] = 0$. Thus, by Theorem 1(b) and $v_d^h \geq 0$, we have $g_1^h - g_1^d = \int_0^\infty P(t, h) v_d^h dt + P^*(h)[Q(h) - Q(d)] g_2^d = \int_0^\infty P(t, h) v_d^h dt \geq 0$. The rest is to prove $g_1^h \neq g_1^d$. Suppose that $g_1^h = g_1^d$. By Lemma 2(a) and $g_0^h = g_0^d$, we have

$$\begin{aligned} r(h) + Q(h) g_1^d &= r(h) + Q(h) g_1^h = g_0^h \\ &= g_0^d = r(d) + Q(d) g_1^d. \end{aligned} \tag{34}$$

On the other hand, since $h \neq d$ and $Q(h) g_0^d = 0$, from (29)–(31), we have

$$r(h) + Q(h) g_1^d \geq r(d) + Q(d) g_1^d,$$

which leads to a contradiction with (34), therefore, $g_1^h \geq g_1^d$. □

With Lemma 5, we can state the *0th-bias optimal policy iteration algorithm* as follows:

(1) Let $k = 0$ and select an arbitrary policy $d_k \in D$.
(2) (Policy evaluation) Obtain (by Lemma 1(a) and (15) or Lemma 2) $g_0^{d_k}$ and $g_1^{d_k}$.
(3) (Policy improvement) Obtain an improvement policy d_{k+1} from (31).
(4) If $d_{k+1} = d_k$, then stop and d_{k+1} is 0th-bias optimal (by Theorem 5). Otherwise, increase k by 1 and return to step 2.

Lemma 5 can be used to prove the anti-cycling property in the policy iteration procedure. We now prove the existence of d^* satisfying (24) and (25) by using the 0th-bias optimal policy iteration algorithm.

Theorem 5. *In a finite number of iterations, the* 0th*-bias optimal policy iteration algorithm stops at a* 0th*-bias optimal policy, denoted as d_0^*, which satisfies (24) and (25).*

Proof. Let $\{d_k, k = 0, 1, 2, \ldots\}$ be the sequence of policies in the policy iteration algorithm above. Then, by Lemma 5(a), we have $g_0^{d_{k+1}} \geq g_0^{d_k}$. That is, as k increases, $g_0^{d_k}$ either increases or stays the same. Furthermore, by Lemma 5(b), when $g_0^{d_k}$ keeps the same, $g_1^{d_k}$ increases. Thus, any two policies in the sequence of $\{d_k, k = 0, 1, \ldots\}$, either have different long-run average rewards or have different 1st biases. Thus, every policy in the iteration sequence is different. Since the number of policies is finite, the iteration procedure must stop after a finite number of iterations. Suppose that it stops at a policy denoted as d_0^*. Then d_0^* must satisfy (24) and (25), because otherwise we can find the next improvement policy in the policy iteration, and the iteration procedure would not stop. Thus, by Theorem 3, d_0^* is 0th-bias optimal. □

5.2. 1st-bias optimal policy iteration algorithm

In this subsection, we provide an iteration algorithm for searching a 1st-bias optimal policy.

Lemma 6. (a) *For any $d^* \in D_0^*$, and $d \in D$, if the following two conditions hold*
(i) $Q(d) g_0^{d^*} = 0$,
(ii) $r(d) + Q(d) g_1^{d^*} \geq g_0^{d^*}$,
then $g_0^d = g_0^{d^*}$.

(b) *Under the conditions in* (a) *, if, in addition, $Q(d) g_2^{d^*}(i) \geq g_1^{d^*}(i)$ for all states i such that $[r(d) + Q(d) g_1^{d^*}](i) = g_0^{d^*}(i)$, then*

$$g_0^d = g_0^{d^*} \quad \text{and} \quad g_1^d \geq g_1^{d^*}.$$

Proof. (a) Let $u := [Q(d) - Q(d^*)]g_2^{d^*}$ and $w := r(d) + Q(d)g_1^{d^*} - g_0^{d^*} \geq 0$ (by condition (ii)). Then, by condition (i) we have $P(t, d)Q(d)g_0^{d^*} = 0$ for all $t \geq 0$, and so it follows from (3) and Lemma 1(a) as well as a straightforward calculation that $[P^*(d) - I]g_0^{d^*} = 0$. Thus, by Theorem 1(a), we have

$$g_0^d - g_0^{d^*} = P^*(d)w + [P^*(d) - I]g_0^{d^*} = P^*(d)w \geq 0. \tag{35}$$

Thus, by the long-run average optimality of d^*, we have

$$g_0^d = g_0^{d^*}, \tag{36}$$

and so (a) follows.

(b) From (35) and (36), we have $P^*(d)w = 0$. Since $w \geq 0$, by Lemma 3(a), we further have $w(i) = 0$ for all recurrent states i under $Q(d)$. Hence, by the conditions in (b) we have $u(i) \geq 0$ for all recurrent states i under $Q(d)$. Then, it follows from Lemma 3(b) that $P^*(d)u \geq 0$, and so by Theorem 1(b), we have

$$g_1^d - g_1^{d^*} = \int_0^\infty P(t, d)w\mathrm{d}t + P^*(d)u \geq 0.$$

This together with (36) proves (b). □

The main goal of this paper is to find a policy that is nth-bias optimal for all $n \geq 0$. From the proof of Theorem 3, a policy $d^* \in D$ satisfying (24)–(26) (with $n = 1$) is 1st-bias optimal. In what follows, we first provide a policy iteration algorithm for finding a 1st-bias optimal policy.

We use the following notations. For a given $d \in D_0^*$ (such as d_0^* in Theorem 5) and $i \in S$, let

$$B_1^d(i) := \left\{ a \in A_0(i) : \begin{array}{l} w_d(i, a) > w_d(i, d(i)); \quad \text{or} \\ \sum_{j\in S} q(j|i, a)g_2^d(j) > \\ \sum_{j\in S} q(j|i, d(i))g_2^d(j) \\ \text{when } w_d(i, a) = w_d(i, d(i)) \end{array} \right\}, \tag{37}$$

(cf. $A_0(i)$ in (22) and $w_d(i, a)$ in (29)). We then define an improvement policy $h \in D$ (depending on d) as follows:

$$h(i) \in B_1^d(i) \quad \text{when } B_1^d(i) \neq \emptyset,$$

$$\text{and} \quad h(i) = d(i) \quad \text{if } B_1^d(i) = \emptyset. \tag{38}$$

Note that such a policy may not be unique, since there may be more than one action in $B_1^d(i)$ for some state $i \in S$.

Lemma 7. *For any given $d \in D_0^*$, let h be defined as in (38). Then,*

(a) *$g_0^h = g_0^d$, and $g_1^h \geq g_1^d$; and*
(b) *if $g_1^h = g_1^d$ and $h \neq d$, then $g_2^h \geq g_2^d$.*

Proof. (a) We take d^* and d in Lemma 6 as d and h here, respectively. It follows from Lemma 2 that $w_d(i, d(i)) = g_0^d(i)$ and $\sum_{j\in S} q(j|i, d(i))g_2^d(j) = g_1^d(i)$ for all $i \in S$. Thus, by (22) and (38), the conditions in Lemma 6(b) holds. Therefore, (a) follows from Lemma 6(b).

(b) Since $g_1^h = g_1^d$, by (a) and Lemma 2(a), we have

$$\begin{aligned} r(h) + Q(h)g_1^d &= r(h) + Q(h)g_1^h = g_0^h \\ &= g_0^d = r(d) + Q(d)g_1^d. \end{aligned} \tag{39}$$

By Lemma 1(b) and Lemma 2(b), (c), we obtain

$$P^*(h)[Q(h) - Q(d)]g_2^d = 0. \tag{40}$$

By (39) and (29), we see that $w_d(i, h(i)) = w_d(i, d(i))$ for all $i \in S$. Hence, by (37) and (38) we have

$$[Q(h) - Q(d)]g_2^d \geq 0, \tag{41}$$

which, together with Lemma 3(a) and (40) as well as (38), implies that $B_1^d(i) = \emptyset$ for all recurrent states i under $Q(h)$, and so we have

$$h(i) = d(i) \quad \forall \text{ recurrent state } i \text{ under } Q(h).$$

Therefore, we have

$$P^*(h)[Q(h) - Q(d)] = 0.$$

From this equation, (41) and Theorem 1(c), we have

$$\begin{aligned} g_2^h - g_2^d &= \int_0^\infty P(t, h)[Q(h) - Q(d)]g_2^d\mathrm{d}t \\ &\quad + P^*(h)[Q(d) - Q(h)]g_3^d \\ &= \int_0^\infty P(t, h)[Q(h) - Q(d)]g_2^d\mathrm{d}t \geq 0. \end{aligned}$$

Now, the rest is to show that $g_2^h \neq g_2^d$. Suppose that $g_2^h = g_2^d$. Then, by $g_1^h = g_1^d$ and (12), we have

$$Q(d)g_2^d = g_1^d = g_1^h = Q(h)g_2^h = Q(h)g_2^d. \tag{42}$$

On the other hand, since $h \neq d$, by (37)–(39), we have

$$Q(h)g_2^d \succeq Q(d)g_2^d,$$

which contradicts (42). □

From Lemma 7, we may propose the following 1st-*bias optimal policy iteration algorithm*.

(1) Let $k = 0$, and take $d_k := d_0^*$ as in Theorem 5 and set $g_0^* = g_0^{d_0^*}$.
(2) (Policy evaluation) Obtain (by (15) and (16) or Lemma 2) $g_1^{d_k}$ and $g_2^{d_k}$.
(3) (Policy improvement) Obtain policy d_{k+1} from (38).
(4) If $d_{k+1} = d_k$, then stop and d_{k+1} is 1st-bias optimal (by Theorem 6). Otherwise, increase k by 1 and return to step 2.

Lemma 7 is then used to prove the anti-cycling property in the 1st-bias optimal policy iteration algorithm. The existence of a policy satisfying (24)–(26) (with $n = 1$) follows naturally from the 1st-bias optimal policy iteration algorithm.

Theorem 6. *Starting from a* 0th-*bias optimal policy $d_0^*(\in D_0^*)$, the* 1st-*bias optimal policy iteration algorithm stops at a* 1st-*bias optimal policy (denoted as d_1^*) satisfying (24)–(26) (with $n = 1$), in a finite number of iterations.*

Proof. Let $\{d_k, k = 0, 1, \ldots\}$ be the sequence of policies obtained by the 1st-bias optimal policy iteration algorithm. Then, from the construction of $\{d_k, k = 0, 1, \ldots\}$ and Lemma 7(a), we see that d_k's are all in D_0^* and $g_1^{d_{k+1}} \geq g_1^{d_k}$. Hence, as k increases, $g_1^{d_k}$ either increases or stays the same. Furthermore, by Lemma 7(b), when $g_1^{d_k}$ keeps the same, $g_2^{d_k}$ increases. Thus, any two policies in the sequence of $\{d_k, k = 0, 1, \ldots\}$ either have different 1st biases or have different 2nd biases. Thus, every policy in the iteration sequence is different. Since the number of policies in D_0^* is finite, the iteration must stop after a finite number of iterations. Suppose that it stops at a policy, denoted as d_1^*. Then d_1^* must satisfy the optimality conditions (25) and (26) (with $n = 1$) because, otherwise, we can find the next improvement policy in the policy iteration. On the other hand, since d_1^* is also in D_0^*, we have $g_0^{d_1^*} = g_0^{d_0^*}$. Thus, by Theorem 4 we see that d_1^* also satisfies (24), and so d_1^* satisfies (24)–(26) (with $n = 1$). Hence, by Theorem 3, d_1^* is 1st-bias optimal. □

5.3. nth-bias optimal policy iteration algorithm

As we have seen, starting from d in D, we can obtain a 0th-bias optimal policy satisfying (24) and (25) by using the 0th-bias optimal policy iteration algorithm. Then, with such a 0th-bias optimal policy, we can further obtain a 1st-bias optimal policy satisfying (24)–(26) with $n = 1$, by using the 1st-bias optimal policy iteration algorithm. Following the similar procedure, we now propose policy iteration algorithms for nth-bias optimal policies, $n > 1$, and prove the existence of a policy satisfying (24)–(26) for all $n \geq 1$. This is achieved by *induction* on n.

Suppose that d is an $(n-1)$th-bias optimal policy for some $n \geq 2$. We need to derive a policy iteration algorithm for an nth-bias optimal policy and show the existence of a policy h satisfying (24)–(26) for n.

Now, we recall (23) when n there is replaced by $n-1$ here, and, for each $i \in S$, let

$$B_n^d(i) := \left\{ a \in A_{n-1}(i) : \begin{array}{l} \sum_{j\in S} q(j|i,a) g_n^d(j) > g_{n-1}^d(i); \text{ or} \\ \sum_{j\in S} q(j|i,a) g_{n+1}^d(j) > g_n^d(j) \\ \text{when } \sum_{j\in S} q(j|i,a) g_n^d(j) = g_{n-1}^d(i) \end{array} \right\}. \quad (43)$$

We then define an improvement policy $h \in D$ (depending on d) as follows:

$$h(i) \in B_n^d(i) \quad \text{when } B_n^d(i) \neq \emptyset,$$

$$\text{and} \quad h(i) = d(i) \quad \text{if } B_n^d(i) = \emptyset. \quad (44)$$

Lemma 8. *For d and $n > 1$ as in the inductive hypothesis, let h be defined as in (44). Then,*

(a) *$g_k^h = g_k^d$ for all $0 \leq k \leq n-1$, and $g_n^h \geq g_n^d$.*
(b) *If $g_n^h = g_n^d$ and $h \neq d$, then $g_{n+1}^h \geq g_{n+1}^d$.*

Proof. (a) Since $d(i)$ is in $A_{n-1}(i)$ (cf. (22) and (23)) for all $i \in S$, and so in $A_k(i)$ for all $0 \leq k \leq n-1$ and $i \in S$. Hence, by (23) and (44) we see that $h(i)$ is also in $A_k(i)$ for $0 \leq k \leq n-1$ and $i \in S$, and so we have

$$Q(h)g_0^d = 0, \quad r(h) + Q(h)g_1^d = g_0^d,$$

$$Q(h)g_{k+1}^d = g_k^d = Q(d)g_{k+1}^d \quad \forall 1 \leq k \leq n-2, \text{ and} \quad (45)$$

$$Q(h)g_n^d \geq g_{n-1}^d = Q(d)g_n^d. \quad (46)$$

By (45)–(46) and Theorem 1 we have

$$g_k^h = g_k^d \quad \forall 0 \leq k \leq n-2, \quad \text{and} \quad g_{n-1}^h \geq g_{n-1}^d,$$

which, together with $g_{n-1}^d \geq g_{n-1}^h$ (by the $(n-1)$th-bias optimality of d), gives the first part of (a). Noting that $g_{n-1}^h = g_{n-1}^d$ and by Theorem 1 and Lemma 3(a), we obtain $(Q(h)g_n^d)(i) = g_{n-1}^d(i) = (Q(d)g_n^d)(i)$ for all recurrent states i under policy h. Moreover, from (43)–(44) and Lemma 3(b) we see

$$(Q(h)g_{n+1}^d)(i) \geq g_n^d(i) = (Q(d)g_{n+1}^d)(i)$$
$$\text{for all recurrent states } i \text{ under policy } h, \quad (47)$$

which, together with Theorem 1(c) and (46), gives $g_n^h \geq g_n^d$, and so (a) follows.

(b) Since $g_n^h = g_n^d$, by (a) and Lemma 2(c), we have

$$Q(h)g_n^d = Q(h)g_n^h = g_{n-1}^h = g_{n-1}^d = Q(d)g_n^d,$$

which, together with Theorem 1(c) and $g_n^h = g_n^d$, gives

$$P^*(h)[Q(h) - Q(d)]g_{n+1}^d = 0. \quad (48)$$

Thus, by (47) and (48) and Lemma 3(a), we have that $B_n^d(i) = \emptyset$ for all recurrent states i under $Q(h)$. Hence, it follows from (44) that

$$h(i) = d(i) \quad \forall \text{ recurrent state } i \text{ under } Q(h),$$

and so

$$P^*(h)[Q(h) - Q(d)] = 0,$$

which together with Theorem 1(c) and (47) gives

$$g_{n+1}^h - g_{n+1}^d = \int_0^\infty P(t,h)[Q(h) - Q(d)]g_{n+1}^d \mathrm{d}t \geq 0.$$

Thus, the rest is to show that $g_{n+1}^h \neq g_{n+1}^d$. Suppose that $g_{n+1}^h = g_{n+1}^d$. Since $g_n^h = g_n^d$, it follows from (12) that

$$Q(d)g_{n+1}^d = g_n^d = g_n^h = Q(h)g_{n+1}^h = Q(h)g_{n+1}^d. \quad (49)$$

On the other hand, since $h \neq d$, by (43)–(44) we have

$$Q(h)g_{n+1}^d \geq g_n^d = Q(d)g_{n+1}^d,$$

which contradicts (49). □

With Lemma 8, we can state an *nth-bias optimal policy iteration algorithm* as follows:

(1) Let $k = 0$ and $d_k \in D_{n-1}^*$ (with fixed $n \geq 2$).
(2) (Policy evaluation) Obtain (by (16) or Lemma 2) $g_n^{d_k}$ and $g_{n+1}^{d_k}$.
(3) (Policy improvement) Obtain policy d_{k+1} from (43) to (44).
(4) If $d_{k+1} = d_k$, then stop and d_{k+1} is nth-bias optimal (by Theorem 7). Otherwise, increase k by 1 and return to step 2.

Lemma 8 can be used to prove the anti-cycling property in the nth-bias optimal policy iteration algorithm. The existence of a policy satisfying (24)–(26) can be proved by the nth-bias optimal policy iteration algorithm, see below.

Theorem 7. *Starting from an $(n-1)(n \geq 2)$th-bias optimal policy, the nth-bias optimal policy iteration algorithm stops at an nth-bias optimal policy satisfying the first $(n+2)$ bias optimality conditions from (24) to (26), in a finite number of iterations.*

Proof. For the fixed $n \geq 2$, let $\{d_k, k = 0, 1, \ldots\}$ be the sequence of policies obtained by the nth-bias optimal policy iteration algorithm, with d_0 being $(n-1)$th-bias optimal. Hence, as k increases, from the construction of $\{d_k, k = 0, 1, \ldots\}$ and Lemma 8(a), we see that $g_n^{d_k}$ either increases or stays the same. Furthermore, by Lemma 8(b), when $g_n^{d_k}$ keeps the same, $g_{n+1}^{d_k}$ increases. Thus, any two policies in the sequence of $\{d_k, k = 0, 1, \ldots\}$ either have different nth biases or have different $(n+1)$th biases. Thus, every policy in the iteration sequence is different. Since the number of policies in D_{n-1}^* is finite, the iteration must stop after a finite number of iterations. Suppose that it stops at a policy denoted as d^*. Thus, by (23) (with n being replaced by $n-1$ here) and (43), we see that d^* must satisfy

$$\max_{a\in A_{n-1}(i)} \sum_{j\in S} q(j|i,a) g_n^{d^*}(j) = g_{n-1}^{d^*}(i) \quad \forall i \in S, \quad (50)$$

$$\max_{a\in A_n^{d^*}(i)} \sum_{j\in S} q(j|i,a) g_{n+1}^{d^*}(j) = g_n^{d^*}(i) \quad \forall i \in S, \quad (51)$$

because, otherwise, we can find the next improvement policy in the policy iteration. Moreover, from the construction of d^* above and Lemma 8 we also have

(A) $g_l^{d^*} = g_l^{d_0}$ for all $0 \leq l \leq n-1$, which together with (23) gives
(B) $A_l^{d^*}(i) = A_l(i)$ for all $0 \leq l \leq n-1$ and $i \in S$, and
(C) d^* also satisfies the first n bias optimality conditions by Theorem 4.

Thus, by (A), (B) and (C) above as well as (50)–(51), we see that d^* satisfies the first $(n+2)$ bias optimality conditions from (24) to (26). Thus, by Theorem 3, the policy d^* is nth-bias optimal. □

Inductively, by Theorems 2 and 5–7 we conclude that we can use nth-bias optimal policy iteration algorithms to obtain a policy that is nth-bias optimal. In particular, we can obtain a policy that is nth-bias optimal for all $n \geq 0$ by using the 1st- to $|S|$th-bias policy iteration algorithms, and each algorithm takes a finite number of iterations.

6. Conclusion

In this paper, we deal with the finite continuous-time MDPs with a multichain structure from a sensitivity-based perspective with the concept of nth-bias optimality.

First, we derive the nth-bias difference formulas of two policies which have the same $(n-1)$th-bias. Then, we give a sufficient condition and a necessary condition for nth-bias optimal policies. Finally, we prove the existence of an nth-bias optimal policy by using nth-bias optimal policy iteration algorithms, and show that such an nth-bias optimal policy can be obtained in a finite number of policy iterations.

Our approach is based on nth-bias difference formulas of two policies which are derived from the sensitivity point of view. Our proofs need neither any result for discrete-time MDPs nor one for discounted continuous-time MDPs, while the approaches in the literature depend heavily on results about discrete-time MDPs or on the Laurent series expansion of discounted continuous-time MDPs. This research is a part of our effort in developing sensitivity-based learning and optimization theory. The nth-bias optimality approach presented in this paper is more intuitive and provides a clearer view: the optimization procedure is simply based on comparison of performance (or nth bias) of two policies. Policy iteration algorithms can be easily developed with this approach.

References

Anderson, W. J. (1991). *Continuous-time Markov chains*. New York: Springer-Verlag.

Arapostathis, A., Borkar, V. S., Fernandez-Gaucherand, E., Ghosh, M. K., & Markus, S. I. (1993). Discrete-time controlled Markov processes with average cost criterion: A survey. *SIAM Journal on Control and Optimization*, *31*(2), 282–344.

Cao, X.-R. (2003). From perturbation analysis to Markov decision processes and reinforcement learning. *Discrete Event Dynamic Systems: Theory and Applications*, *13*(1–2), 9–39.

Cao, X.-R. (2007). *Stochastic learning and optimization - a sensitivity-based approach*. New York: Springer.

Cao, X.-R., & Guo, X. P. (2004). A unified approach to Markov decision problems and performance sensitivity analysis with discounted and average criteria: Multichain cases. *Automatica*, *40*(9), 1749–1759.

Cao, X.-R., & Zhang, J. Y. (2008). The nth-order bias optimality for multichain Markov decision processes. *IEEE Transactions on Automatic Control*, *53*(2), 496–508.

Feller, W. (1940). On the integro-differential equations of purely discontinuous Markoff processes. *Transactions of the American Mathematical Society*, *48*, 488–515.

Guo, X. P., & Cao, X.-R. (2005). Optimal control of ergodic continuous-time Markov chains with average sample-path rewards. *SIAM Journal on Control and Optimization*, *44*(1), 29–48.

Guo, X. P., & Hernández-Lerma, O. (2003). Drift and monotonicity conditions for continuous-time controlled Markov chains an average criterion. *IEEE Transactions on Automatic Control*, *48*(2), 236–245.

Guo, X. P., Hernández-Lerma, O., & Prieto-Rumeau, T. (2006). A survey of recent results on continuous-time Markov decision processes. *Top*, *14*(2), 177–261.

Guo, X. P., & Liu, K. (2001). A note on optimality conditions for continuous-time Markov decision processes with average cost criterion. *IEEE Transactions on Automatic Control*, *46*(12), 1984–1989.

Guo, X. P., & Rieder, U. (2006). Average optimality for continuous-time Markov decision processes in Polish spaces. *The Annals of Applied Probability*, *16*(2), 730–756.

Guo, X. P., Song, X. Y., & Zhang, J. Y. (2009). Bias optimality for multichain continuous-time Markov decision processes. Preprint.

Haviv, M., & Puterman, M. L. (1998). Bias optimality in controlled queueing systems. *Journal of Applied Probability*, *35*(1), 136–150.

Howard, R. A. (1960). *Dynamic programming and Markov processes*. New York: Wiley.

Kakumanu, P. (1972). Nondiscounted continuous-time Markov decision processes with countable state and action spaces. *SIAM Journal on Control*, *10*, 210–220.

Kitaev, M. Y., & Rykov, V. V. (1995). *Controlled queueing systems*. CRC Press.

Lewis, M. E., & Puterman, M. L. (2002). Bias optimality. In E. A. Feinberg, & A. Shwartz (Eds.), *Handbook of Markov decision processes* (pp. 89–111). Boston: Kluwer.

Miller, B. L. (1968). Finite state continuous time Markov decision processes with an infinite planning horizon. *Journal Of Mathematical Analysis and Applications*, *22*, 552–569.

Oksendal, B., & Sulem, A. (2007). *Applied stochastic control of jump diffusions*. Springer.

Prieto-Rumeau, T., & Hernández-Lerma, O. (2005). The Laurent series, sensitive discount and Blackwell optimality for continuous-time controlled Markov chains. *Mathematical Methods of Operations Research*, *61*(1), 123–145.

Prieto-Rumeau, T., & Hernández-Lerma, O. (2006). Bias optimality for continuous-time controlled Markov chains. *SIAM Journal on Control and Optimization*, *45*(1), 51–73.

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. New York: Wiley.

Sennott, L. I. (1999). *Stochastic dynamic programming and the control of queueing systems*. New York: Wiley.

Taylor, H. M. (1976). A Laurent series for the resolvent of a strongly continuous stochastic semi-group. *Mathematical Programming Studies*, *6*, 258–263.

Veinott, A. F. (1969). Discrete dynamic programming with sensitive discount optimality criteria. *The Annals of Mathematical Statistics*, *40*(5), 1635–1660.

Junyu Zhang received the B.S. Degree in statistics and probability from Peking University, China, in 1999, the M.S. Degree from Academy of Mathematics and Systems Science, Chinese Academic of Science, in 2002, and the Ph.D. Degree in electrical and electronic engineering from the Hong Kong University of Science and Technology in 2006.

Since 2006, she has been an assistant Professor in Mathematics at the Sun Yat-sen University, Guangzhou, China. Her current research interests include Markov decision processes, reinforcement learning, and queueing theory.

Xi-Ren Cao received the M.S. and Ph.D. Degrees from Harvard University, in 1981 and 1984, respectively, where he was a research fellow from 1984 to 1986. He then worked as consultant engineer/engineering manager at Digital Equipment Corporation, USA, until October 1993. Then he joined the Hong Kong University of Science and Technology (HKUST), where he is currently Chair Professor and Director of the Research Center for Networking.

Dr. Cao owns three patents in data- and telecommunications and published three books in the area of performance optimization and discrete event dynamic systems. He received the Outstanding Transactions Paper Award from the IEEE Control System Society in 1987, the Outstanding Publication Award from the Institution of Management Science in 1990, and the Outstanding Service Award from IFAC in 2008. He is a Fellow of IEEE, a Fellow of IFAC, and is/was the Chairman of IEEE Fellow Evaluation Committee of IEEE Control System Society, Editor-in-Chief of *Discrete Event Dynamic Systems: Theory and Applications*, Associate Editor at Large of *IEEE Transactions of Automatic Control*, and Board of Governors of IEEE Control Systems Society and on the Technical Board of IFAC. His current research areas include discrete event dynamic systems, stochastic learning and optimization, performance analysis of communication systems, signal processing, and financial engineering.