BIAS OPTIMALITY FOR MULTICHAIN MARKOV DECISION PROCESSES

Xi-Ren Cao and Junyu Zhang [1]

Department of Electrical and Electronic Engineering
The Hong Kong University of Science and Technology

Abstract: In recent research we find that the policy iteration algorithm for Markov decision processes (MDPs) is a natural consequence of the performance difference formula that compares the difference of the performance of two different policies. In this paper, we extend this idea to the bias-optimal policy of MDPs. We first derive a formula that compares the biases of any two policies which have the same gains, and then we show that a policy iteration algorithm leading to a bias-optimal policy follows naturally from this bias difference formula. Our results extend those in (Lewis & Puterman, 2001) to the multichain case and provide a simple and intuitive explanation for the mathematics in (Veinott, 1966; Veinott, 1969). The results also confirm the idea that the solutions to performance (including bias) optimal problems can be obtained from performance sensitivity formulas.

Keywords: Markov decision processes, Gain optimal, Bias optimal

1. INTRODUCTION

The research of this paper is a continuation of the recent research on performance optimization of discrete event dynamic systems with a sensitivity point of view (Cao, 2000; Cao & Guo, 2004). It is motivated by the previously established results. In particular, it is shown in (Cao, 2000; Cao, 2004) that the policy iteration algorithm for gain-optimal problems in Markov decision processes (MDPs) follows naturally from the performance difference formula.

While the gain-optimal policies optimize the steady-state performance, they ignore the system's transient performance. Therefore, further performance criteria such as the bias optimality need to be studied. A bias-optimal policy not only optimizes the average reward, but also maximizes the total expected reward starting from any initial state. In this paper, we show that following the same approach as we did for the gain-optimal problem (Cao, 2004), a bias-optimal policy iteration algorithm for multichain MDPs can be easily derived from the bias difference formula.

The existing works on bias optimality include (Lewis, 2001; Veinott, 1966; Veinott, 1969). While Lewis provided a solution to the unichain case and left the multichain case unsolved, Veinott's early works (Veinott, 1966; Veinott, 1969) did provide a nice solution to the problem. However, for some reasons Veinott's work did not receive its deserved attention in the literature.

The contributions of this paper are as follows. First, this work extends our sensitivity-based optimization approach to bias optimality. It confirms our belief that policy iteration algorithms follow naturally from the performance (gain or bias) difference formulas and therefore provides a new

[1] Supported by a grant from Hong Kong UGC. E-mail addresses: eecao@ust.edu.hk and eezhjy@ust.edu.hk (9/2004)

insight to the optimization problem. Second, this work provides a simple (almost the same as the gain-optimal problem) and intuitive approach to the bias optimality problem, leading to a clear explanation of Veinott's work; we hope that our work may help to popularize Veinott's results.

The rest of the paper is organized as follows. In Section 2, we briefly review the concepts and the results of bias for multichain MDPs. In Section 3, we derive the bias difference formula for multichain Markov processes. In Section 4, we show the standard policy iteration algorithm can be derived using the bias difference formula in a clear and intuitive way. In Section 5, we treat the ergodic chains as a special case and obtain some simple and neat results. Section 6 concludes the paper with some discussions.

2. FUNDAMENTAL THEORY

Consider a multichain Markov process $\{X_n, n = 0, 1, \cdots\}$ defined on a finite state space $S = \{1, 2, \cdots, M\}$. Let $P = [p(i,j)]$ be the state transition probability matrix. The reward at state s is $r(s)$. We have $Pe = e$, with $e = (1, \cdots, 1)^T$ being a vector whose all components are one.

The long-run average performance is defined as a vector η with components

$$\eta(s) = \lim_{N\to\infty} \frac{1}{N} E\left\{\sum_{n=0}^{N-1} r(X_n)|X_0 = s\right\} \quad (1)$$

Thus, we have

$$\eta = \lim_{N\to\infty} \frac{1}{N}\left\{\sum_{n=0}^{N-1} P^n r\right\} = P^* r \quad (2)$$

where $r = (r(1), \cdots, r(M))^T$, "T" denotes transpose, and $P^* = \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} P^n$ is the Cesaro limit. We can easily prove (A.4 in (Puterman, 1994)) that $P^*e = e$ and

$$PP^* = P^*P = P^*P^* = P^*. \quad (3)$$

From (2) and (3), we can get

$$P\eta = P^*\eta = \eta. \quad (4)$$

The potential $g = (g(1), \cdots, g(M))^T$ is defined by the Poisson equation

$$(I - P)g + \eta = r. \quad (5)$$

If g satisfies (5), then so does $g + u$, where u satisfies $(I-P)u = 0$. For example, we can choose $u = ce$ with c being any constant.

We may choose c such that $P^*g = 0$. Such a potential is called a bias. Since in this paper we use bias as the potential, we use the same notation g to denote both the bias and the potential. (5) becomes $(I - P + P^*)g = r - \eta$. From Theorem A.7 of (Puterman, 1994), the matrix $(I - P + P^*)$ is nonsingular and if P is aperiodic,

$$(I - P + P^*)^{-1} = \sum_{n=0}^{\infty}(P - P^*)^n. \quad (6)$$

If P is periodic, we have

$$(I - P + P^*)^{-1} = \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1}\sum_{i=0}^{n}(P - P^*)^i$$

which is also a Cesaro limit. For the simplicity in expression, we assume that P is aperiodic in this paper. If P is periodic, we just need to replace the normal limit with Cesaro limit and all results in this paper are hold for periodic case. Thus,

$$g = (I - P + P^*)^{-1}(r - \eta) \quad (7)$$

$$= \sum_{n=0}^{\infty}(P - P^*)^n(r - \eta) = \sum_{n=0}^{\infty} P^n(r - \eta) \quad (8)$$

And $g(i) = \sum_{n=0}^{\infty} E[r(X_n) - \eta|X_0 = i]$, we can estimate g by this equation on a single sample path without knowing P.

In MDPs (Puterman, 1994), there is an action space A consisting of all available actions. $A_s \subseteq A$ is the set of available actions in state $s \in S$. If the system is at state s, an action $a \in A_s$ can be taken and applied to the system. The action determines the state transition probabilities. When action a is taken at state s, the state transition probability distribution is denoted as $p_a(s,j), j \in S$ which determine the system state at the next decision epoch. The reward also depends on action and is denoted as $r(s,a)$.

A stationary deterministic policy is a mapping from S to A, denoted as $d : a = d(s)$, which determines the action taken at state s. We will only consider stationary deterministic policies. Denote D as the set of all stationary deterministic policies. If policy d is adopted, the state transition probability matrix is $P_d = [p_{d(i)}(i,j)]_{i,j=1}^M$. Since a policy corresponds to a state transition probability matrix, we sometimes refer to a state transition probability matrix as a policy.

We will use subscript "d" to denote all the quantities associated with policy d; e.g., P_d, r_d, η_d, g_d, w_d etc. In particular, the *gain* (*long-run average expected reward*) of policy $d \in D$ is defined as

$$\eta_d(s) = \lim_{N\to\infty} \frac{1}{N} E_s^d\left\{\sum_{n=0}^{N-1} r(X_n, d(X_n))|X_0 = s\right\}$$

$$= \lim_{N\to\infty} \frac{1}{N}\left\{\sum_{n=0}^{N-1}(P_d)^n r_d(s)\right\} = (P_d)^* r_d(s) \quad (9)$$

where

$$r_d = [r(1, d(1)), r(2, d(2)), \cdots, r(M, d(M))]^T$$

and $(P_d)^* r_d(s)$ denotes the sth component of vector $(P_d)^* r_d$.

We say policy d^* is a *gain-optimal* policy if

$$\eta_{d^*}(s) \geq \eta_d(s), \text{ for all } s \in S \text{ and } d \in D.$$

And $\eta^* = \eta_{d^*}$ is the *optimal gain*. Denote $D_{-1} \subseteq D$ as the set of all gain-optimal policies.

We refer to the following equations of η and g

$$\max_{a \in A_s} \{\sum_{j \in S} p_a(s,j)\eta(j) - \eta(s)\} = 0 \text{ and}$$

$$\max_{a \in B_s} \{r(s,a) - \eta(s) + \sum_{j \in S} p_a(s,j)g(j) - g(s)\} = 0,$$

where $B_s = \{a \in A_s : \sum_{j \in S} p_a(s,j)\eta(j) - \eta(s) = 0\}$, as the *multichain optimality equations*. In a matrix form, they are

$$\max_{d \in D} \{(P_d - I)\eta\} = 0 \text{ and} \tag{10}$$

$$\max_{d \in E} \{r_d - \eta + (P_d - I)g\} = 0, \tag{11}$$

where $E = \{d \in D : P_d \eta = \eta\}$.

It is well known that a policy whose corresponding η and g satisfy the multichain optimality equations is a gain-optimal policy, and there exists a gain-optimal policy satisfying the multichain optimality equations. This implies that the optimal gain always satisfies (10). However, it is also known that there may exist gain-optimal policies that do not satisfy the optimality equation (11) (Puterman, 1994).

From (9), we can see that the gain (average reward) criterion focuses on the limiting or steady-state behavior of a system and ignores transient performance. Therefore, the gain optimality criterion is under-selective. We need a more selective optimality criterion - bias optimality which can include the transient performance.

We say policy π^* is a *bias-optimal* policy if

$$g_{\pi^*}(s) \geq g_\pi(s), \text{ for all } s \in S \text{ and } \pi \in D_{-1}.$$

That is, a bias-optimal policy π^* is a policy with maximal bias among all the gain-optimal policies. And $g^* = g_{\pi^*}$ is the *optimal bias*. From (8), we can know that a bias-optimal policy is to maximize the total expected reward.

As we have emphasized, the policy iteration for gain-optimal policies is based on the performance difference equation:

$$\eta' - \eta = (P')^*(P'g + r' - Pg - r) + [(P')^* - I]\eta. \tag{12}$$

This can be obtained by left-multiplying $(P')^*$ on the both sides of (5).

3. BIAS DIFFERENCE

Throughout, we assume a model which satisfies the following assumptions:
1. stationary rewards and transition probabilities,
2. finite rewards, $|r(s,a)| < \infty \; \forall \, a \in A_s$ and $s \in S$,
3. state space S and action space A are both finite.

For two vectors u and v defined on state space S, we define $u = v$ if $u(i) = v(i)$ for all $i \in S$; $u \geq v$ if $u(i) \geq v(i)$ for all $i \in S$; $u \succeq v$ if $u \geq v$ and $u(i) > v(i)$ for at least one $i \in S$.

Lemma 1. Let policy d^* be gain optimal, and η^* be the corresponding optimal gain. Then for any stationary deterministic policy $d \in D$, we have
(a) $P_d \eta^* \leq \eta^*$.
(b) If $P_d \eta^* \preceq \eta^*$, then $\eta_d \preceq \eta^*$.
(c) If $d \in D_{-1}$, then $P_d \eta^* = (P_d)^* \eta^* = \eta^*$.

Proof. (a) is a direct consequence of (10). For (b), note that from (4) and $\eta_d \leq \eta^*$, we have $\eta_d = P_d \eta_d \leq P_d \eta^* \preceq \eta^*$. Now we prove (c). Assume $P_d \eta^* \neq \eta^*$. From part (a), we have $P_d \eta^* \leq \eta^*$. Thus, $P_d \eta^* \preceq \eta^*$. From part (b), we have $\eta_d \preceq \eta^*$. This conflicts with the fact that policy d is also gain optimal. Thus, $P_d \eta^* = \eta^*$. Hence, $(P_d)^n \eta^* = \eta^*$ for any integer n. Then $(P_d)^* \eta^* = \eta^*$ holds noting $(P_d)^* = \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} (P_d)^n$. □ In order to derive the bias difference formula, define the *bias offset* of policy P as

$$w = -g + Pw. \tag{13}$$

Again, the solution to (13) is not unique and we may set $P^* w = 0$. Thus we can rewrite (13) as

$$(I - P + P^*)w = -g.$$

Therefore,

$$w = -(I - P + P^*)^{-1} g = -\sum_{n=0}^{\infty} (P - P^*)^n g$$
$$= -\sum_{n=0}^{\infty} P^n g = -\sum_{n=0}^{\infty} (n+1) P^n (r - \eta). \tag{14}$$

And $w(i) = -\sum_{n=0}^{\infty} (n+1) E[r(X_n) - \eta | X_0 = i]$. Comparing (8) and (14), (14) is almost the same as (8) if we replace $(r - \eta)$ in (7) with $-g$. This point is very important in estimating w. Therefore we can also estimate w on a single sample path without knowing P.

Now we derive the bias difference formula.

Lemma 2. For policies d and π, if $\eta_d = \eta_\pi$, then

$$g_d - g_\pi = \sum_{n=0}^{\infty} (P_d)^n (P_d g_\pi + r_d - P_\pi g_\pi - r_\pi)$$
$$+ (P_d)^* (P_d - P_\pi) w_\pi. \tag{15}$$

Proof. Denote $x_\pi^d = P_d g_\pi + r_d - P_\pi g_\pi - r_\pi$. From (5) and $\eta_d = \eta_\pi$, we have

$$\begin{aligned} g_d - g_\pi &= P_d g_d + r_d - \eta_d - (P_\pi g_\pi + r_\pi - \eta_\pi) \\ &= P_d g_\pi + r_d - P_\pi g_\pi - r_\pi + P_d(g_d - g_\pi). \end{aligned}$$

From (13), we have $g_\pi = P_\pi w_\pi - w_\pi$. Left-multiply $(P_d)^*$ to both sides of this equation, we get

$$(P_d)^* g_\pi = (P_d)^*(P_\pi - P_d) w_\pi. \tag{16}$$

From (16) and $(P_d)^* g_d = 0$,

$(I - P_d + (P_d)^*)(g_d - g_\pi) = x_\pi^d + (P_d)^*(P_d - P_\pi) w_\pi$.

From (6) and $[I - P_d + (P_d)^*]^{-1}(P_d)^* = (P_d)^*$, we have

$$\begin{aligned} g_d - g_\pi &= [I - P_d + (P_d)^*]^{-1} x_\pi^d + (P_d)^*(P_d - P_\pi) w_\pi \\ &= \sum_{n=0}^{\infty} (P_d)^n x_\pi^d - \sum_{n=1}^{\infty} (P_d)^* x_\pi^d + (P_d)^*(P_d - P_\pi) w_\pi. \end{aligned} \tag{17}$$

Next, from (5), (3) and (9), we have

$$\begin{aligned} (P_d)^* x_\pi^d &= (P_d)^*(P_d g_\pi + r_d - \eta_\pi - g_\pi) \\ &= (P_d)^* r_d - (P_d)^* \eta_\pi = \eta_d - \eta_d = 0. \end{aligned}$$

The lemma then follows directly from (17). □

Now we will use the above lemma to prove the following theorem.

Theorem 1. Suppose d^* is any gain-optimal policy and the corresponding quantities are P, r, η^*, g and w. If another policy $d \in D$ satisfies the following three conditions:
(a) $P_d \eta^* = \eta^*$,
(b) $P_d g + r_d \geq P g + r$, and
(c) $P_d w(i) \geq P w(i)$ when $P_d g(i) + r_d(i) = P g(i) + r(i)$ for some $i \in S$,
then $\eta_d = \eta^*$ and $g_d \geq g$.

Proof. Let $x = P_d g + r_d - (P g + r)$ and $y = P_d w - P w$. From condition (a), condition (b), Lemma 1(c) and (12), we can know

$$\eta_d - \eta^* = (P_d)^* x + [(P_d)^* - I]\eta^* = (P_d)^* x \geq 0.$$

Since η^* is the optimal gain, policy d is also a gain-optimal policy, i.e. $\eta_d = \eta^*$. Then $(P_d)^* x = 0$. Noting $x \geq 0$ and the components of $(P_d)^*$ are positive on all recurrent states, we can have $x(i) = 0$, $\forall$ recurrent state i under policy d, and so it follows from condition (c) that $y(i) \geq 0$, $\forall$ recurrent state under policy d. Noting that $(P_d)^*(i,j) = 0$ where $i \in S$ and j is a transient state under policy d, we can have $(P_d)^* y \geq 0$. Refer to Lemma 2,

$$g_d - g = \sum_{n=0}^{\infty} (P_d)^n x + (P_d)^* y \geq \sum_{n=0}^{\infty} (P_d)^n x.$$

Noting that P_d is a non-negative matrix and $x \geq 0$, $g_d \geq g$. □

4. POLICY ITERATION ALGORITHM

From Theorem 1, we can easily derive the bias optimality conditions:

Theorem 2. Let $\hat{P}$, $\hat{r}$, $\hat{\eta}$, $\hat{g}$ and $\hat{w}$ be the transition probability matrix, the reward, the gain, the bias and the bias offset of a policy $\hat{d} \in D$. Suppose the following " bias optimality conditions" hold:

$$\hat{\eta}(i) = \max_{a \in A_i} \{ \sum_{j \in S} p_a(i,j) \hat{\eta}(j) \}, \tag{18}$$

$$\hat{\eta}(i) + \hat{g}(i) = \max_{a \in B_i} \{ r(i,a) + \sum_{j \in S} p_a(i,j) \hat{g}(j) \}, \tag{19}$$

$$\hat{g}(i) + \hat{w}(i) = \max_{a \in C_i} \{ \sum_{j \in S} p_a(i,j) \hat{w}(j) \}, \tag{20}$$

$\forall\, i \in S$, where $B_i := \{ a \in A_i |\ \sum_{j \in S} p_a(i,j) \hat{\eta}(j) = \hat{\eta}(i) \}$, $C_i := \{ a \in B_i |\ \hat{\eta}(i) + \hat{g}(i) = r(i,a) + \sum_{j \in S} p_a(i,j) \hat{g}(j) \}$. Then $\hat{\eta} \geq \eta_d$ for all $d \in D$ and $\hat{g} \geq g_d$ for all $d \in D_{-1}$; that is, policy $\hat{d}$ is bias optimal.

Proof. (18), (19) and (20) can be restated in matrix forms as

$$\hat{\eta} \geq P_d \hat{\eta} \quad \forall\, d \in D, \tag{21}$$

$$\hat{\eta} + \hat{g} \geq r_d + P_d \hat{g} \quad \forall\, d \in E, \tag{22}$$

$$\hat{g} + \hat{w} \geq P_d \hat{w} \quad \forall\, d \in F, \tag{23}$$

where $E \doteq \{ d \in D : P_d \hat{\eta} = \hat{\eta} \}$ and $F \doteq \{ d \in E : \hat{\eta} + \hat{g} = P_d \hat{g} + r_d \}$. First, $\hat{d}$ is gain optimal because $\hat{\eta}$ and $\hat{g}$ satisfy the gain optimality equations (18) and (19) (same as (10) and (11))(Cao, 2004). Next, We will prove that $\hat{g} \geq g_d$ for all $d \in D_{-1}$ in the following. For $\forall\, d \in D_{-1}$, we have proved before $P_d \hat{\eta} = \hat{\eta}$ in Lemma 1 and $(P_d)^*(P_d \hat{g} + r_d - \hat{P} \hat{g} - \hat{r}) = 0$. Together with condition (22), we have $\hat{\eta} + \hat{g} = \hat{P} \hat{g} + \hat{r} \geq r_d + P_d \hat{g}$, $\forall\, d \in D_{-1}$ and $\hat{\eta}(i) + \hat{g}(i) = r_d(i) + P_d \hat{g}(i)$, $\forall$ recurrent state i under policy d. Together with condition (23), $(P_d)^*(P_d - \hat{P}) \hat{w} = (P_d)^*(P_d \hat{w} - \hat{g} - \hat{w}) \leq 0$. By Lemma 2,

$$\begin{aligned} g_d - \hat{g} = &\sum_{n=0}^{\infty} (P_d)^n (P_d \hat{g} + r_d - \hat{P} \hat{g} - \hat{r}) \\ &+ (P_d)^*(P_d - \hat{P}) \hat{w} \leq 0. \end{aligned}$$

As a result, $\hat{g} \geq g_d$, $\forall\, d \in D_{-1}$. □

Following the same procedure as for the gain-optimal problem, by Theorem 2, from any gain-optimal policy we can construct another gain-optimal policy whose bias is larger if such a policy exists. For a given $d \in D_{-1}, i \in S$ and $a \in B_i$, let

$$H_d(i,a) := r(i,a) + \sum_{j \in S} p_a(i,j) g_d(j), \text{ and} \tag{24}$$

$$A_d(i) := \left\{ a \in B_i : \begin{array}{l} H_d(i,a) > H_d(i,d(i)); \text{ or} \\ \sum_{j\in S} p_a(i,j)w_d(j) > \sum_{j\in S} p_d(i,j)w_d(j) \\ \text{when } H_d(i,a) = H_d(i,d(i)) \end{array} \right\} \quad (25)$$

We then define an improvement policy h (depending on d) as follows:

$$\begin{cases} h(i) \in A_d(i) \text{ if } A_d(i) \neq \emptyset; \\ h(i) = d(i) \quad \text{if } A_d(i) = \emptyset. \end{cases} \quad (26)$$

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

$$x_d^h := r_h + P_h g_d - r_d - P_d g_d,$$
$$y_d^h := P_h w_d - P_d w_d.$$

Theorem 3. For any given $d \in D_{-1}$, let h be defined as in (26). Then

(a) $g_h \geq g_d$, and $y_d^h(i) \geq 0$ for all recurrent states i under P_h.

(b) If $y_d^h(i) > 0$ for some recurrent state i under P_h, then $g_h \succeq g_d$.

(c) If $r_h + P_h g_d \neq r_d + P_d g_d$, then $g_h \succeq g_d$.

(d) If $g_h = g_d$ and $h \neq d$, then $w_h \succeq w_d$.

Proof. (a) We take policy d and policy h as policy d^* and policy d in Theorem 1, respectively. Then by the construction in (25) and (26), the conditions (a), (b) and (c) in Theorem 1 hold. Thus, it follows from Theorem 1 that $g_h \geq g_d$. Moreover, as in the proof of Theorem 1, we have $y_d^h(i) \geq 0$ for all recurrent states i under policy h; thus, part (a) follows.
(b) Since $x_d^h \geq 0$, $y_d^h \geq 0$ and condition in (b), we have $(P_h)^* y_d^h \succeq 0$, and so $g_h - g_d = \sum_{n=0}^{\infty}(P_h)^n x_d^h + (P_h)^* y_d^h \succeq 0$. Then part (b) follows.
(c) By part (a), it suffices to prove that $g_h \neq g_d$. Suppose that $g_h = g_d$. Then

$$P_h g_d + r_h = P_h g_h + r_h = \eta^* + g_h = P_d g_d + r_d$$

which contradicts to the given condition. Therefore, part (c) is proved.
(d) Since $g_h = g_d$, $r_h + P_h g_d = r_d + P_d g_d$ follows by part (c). Then $P_h w_d \succeq P_d w_d$ holds by (25) and $h \neq d$. Noting $g_h - g_d = (P_h)^*(P_h - P_d)w_d = 0$, we have $P_h w_d(i) = P_d w_d(i)$ $\forall$ recurrent state i under policy h. From (25) and (26),

$$h(i) = d(i), \forall \text{ recurrent state } i \text{ under } h. \quad (27)$$

we define u as $u = -w + Pu$. Just the same as Lemma 2, we have

$$w_h - w_d = \sum_{n=0}^{\infty}(P_h)^n(P_h - P_d)w_d + (P_h)^*(P_h - P_d)u_d.$$

By (27), we can get $(P_h)^*(P_h - P_d)u_d = 0$. Then $w_h - w_d = \sum_{n=0}^{\infty}(P_h)^n(P_h - P_d)w_d \geq (P_h - P_d)w_d \succeq 0$. □

With Theorem 3, we can state the (standard) *Bias Optimality Policy Iteration Algorithm* as follows:
1. Set $n = 0$ and select an arbitrary gain-optimal policy $d_0 \in D_{-1}$.
2. (Policy evaluation) Obtain g_{d_n} and w_{d_n} by solving

$$r_{d_n} - \eta^* + (P_{d_n} - I)g = 0 \quad (28)$$
$$-g + (P_{d_n} - I)w = 0 \quad (29)$$

subject to $(P_{d_n})^* w = 0$.
3. (Policy improvement) Obtain policy d_{n+1} as the policy h in (25) and (26). setting $d_{n+1}(s) = d_n(s)$ if possible.
4.If $d_{n+1} = d_n$, stop and set $d^* = d_n$ and $g^* = g_{d_n}$; otherwise increment n by 1 and return to step 2.

Theorem 3 can be used to compare the biases of two gain-optimal policies and to prove the anti-cycling property in the policy iteration procedure. The existence of the solution to the optimality equations can be proved by construction as shown in Theorem 4.

Theorem 4. The Policy Iteration Algorithm stops at a bias-optimal policy in a finite number of iterations.

Proof. By Theorem 3(a), $g_{d_{n+1}} \geq g_{d_n}$. That is, as n increases, g_{d_n} either increases or stays the same. Furthermore, by Theorem 3(d), when g_{d_n} stays the same, w_{d_n} increases. Thus, any two policies in the sequence of d_n, $n = 0, 1, \cdots$, either have different bias or have different bias offset. That is, every policy in the iteration sequence is different. Since the number of policies is finite, the iteration must stop after a finite number of iterations. Suppose it stops at a policy denoted as d^*. Then d^* must satisfy the optimality conditions (18), (19) and (20) because otherwise for some i the set $A_{d^*}(i)$ in (25) is non-empty and we can find the next improved policy in the policy iteration. Thus, by Theorem 2, policy d^* is bias optimal. □

5. THE ERGODIC CASE

For an ergodic Markov chain, we define $\theta = (\theta(1), \cdots, \theta(M))$ be the (row) vector representing its steady-state probabilities. Then $\theta e = 1$. The steady-state probability flow balance equation is $\theta = \theta P$. The gain $\eta(i)$ is a constant function on the state space S and we write it as $\eta = \theta r$. We have $P^* = e\theta$, and the gain-difference equation (12) becomes

$$\eta' - \eta = \theta'(P'g + r' - Pg - r). \quad (30)$$

Lemma 3. Suppose Markov chain is ergodic and policy d^* is any gain-optimal policy. The corresponding quantities are P, θ, r, η^* and g with respect to d^*. Any another gain-optimal policy $d \in D_{-1}$ must satisfy $P_d g + r_d = Pg + r$.

Proof. Take η_d as η' and η^* as η in (30), we have

$$\eta_d - \eta^* = \theta_d(P_d g + r_d - Pg - r).$$

Because $\eta_d = \eta^*$, we have

$$\theta_d(P_d g + r_d - Pg - r) = 0.$$

Assume that this lemma does not hold, and without loss of generality, we assume that there exists a policy π and a state $i \in S$ such that

$$P_\pi g(i) + r_\pi(i) - Pg(i) - r(i) > 0.$$

Then we can construct another policy P_ϵ. P_ϵ is the same as P except ith row, which is $P_\epsilon(i,j) = P_\pi(i,j), j = 1, 2, \cdots, M$. Consequently, $r_\epsilon(k) = r(k), k \in S - \{i\}$, $r_\epsilon(i) = r_\pi(i)$. We can claim that

$$\begin{aligned}\eta_\epsilon - \eta &= \theta_\epsilon(P_\epsilon g + r_\epsilon - Pg - r) \\ &= \theta_\epsilon(i)[P_\pi g(i) + r_\pi(i) - Pg(i) - r(i)] > 0\end{aligned}$$

noting each component of θ_ϵ is positive for ergodic chain. This conflicts with that d^* is a gain-optimal policy. Thus, the lemma holds. □

With Lemma 3, the bias difference equation (15) becomes (recall $(P_d)^* g_d = 0$ and (16))

$$g_d - g = (P_d)^*(g_d - g) = -(\theta_d g)e. \tag{31}$$

This equation provides an interesting insight to the bias-optimal problem: the difference of the biases at all states for any two gain-optimal policies d and d^* is a constant $\theta_d g$, with θ_d being the steady-state probability of policy d and g being the bias of policy d^*. Furthermore, we may choose any gain-optimal policy d^*, then to optimize of g_d is to optimize $\theta_d(-g)$. That is, the bias optimization problem is equivalent to the gain optimization problem with $-g$ as the reward function r. With this in mind, we can translate many results for gain optimality to bias optimality. In particular, we have

$$g_d - g = e(\theta_d)(P_d - P)w. \tag{32}$$

From (32), we have the following lemma.

Lemma 4. For ergodic MDP, suppose P and P_d are both gain optimal. If $P_d w \succeq Pw$, then we have $g_d > g$. If $P_d w = Pw$, then we have $g_d = g$.

Let d_0 be any gain-optimal policy and set $D_0 = \arg\max_{d \in D}\{P_d \eta^* = \eta^*, P_d g_{d_0} + r_d = P_{d_0} g_{d_0} + r_{d_0}\}$.
We then have the *Bias Optimality Policy Iteration Algorithm* for ergodic case:
1., 2. and 4. are the same as the multichain case.
3. Choose $d_{n+1} \in \arg\max_{d \in D_0}\{P_d w_{d_n}\}$, setting $d_{n+1}(s) = d_n(s)$ if possible.

6. CONCLUSION

We have demonstrated that by the bias-difference formula we can derive the bias-optimal policy iteration; this provides a simple and intuitive way to establish the results; it is also in the same framework as the gain-optimal problem.

Since the bias g and the bias offset w both can be estimated on a single sample path without knowing the state transition probability matrix P. we can develop on-line policy iteration algorithms.

Using the same idea as this paper, we can find N-discount optimal policy for $n \geq -1$. Further research is need in this direction and for problems with infinite state spaces.

REFERENCES

Cao, Xi-Ren (2000). A Unified Approach to Markov Decision Problems and Performance Sensitivity Analysis. *Automatica*, Vol. 36, Issue 5, pp. 771-774.

Cao, Xi-Ren and Guo, Xianping (2004). A Unified Approach to Markov Decision Problems and Performance Sensitivity Analysis with Discounted and Average Criteria: Multichain Cases. *Automatica*, Vol. 40, Issue 10, pp. 1749-1759.

Lewis, Mark E. and Puterman, Martin L. (2001). A Probabilistic Analysis of Bias Optimality in Unichain Markov Decision Processes. *IEEE Transactions on Automatic Control,* Vol. 46, Issue 1, pp. 96-100.

Martin L. Puterman (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, New York: Wiley.

Veinott, Arthur F. (1966). On Finding Optimal Policies in Discrete Dynamic Programming With No Discounting. *The Annals of Mathematical Statistics,* Vol. 37, No. 5, pp. 1284-1294.

Veinott, Arthur F. (1969). Discrete Dynamic Programming with Sensitive Discount Optimality Criteria. *The Annals of Mathematical Statistics,* Vol. 40, No. 5, pp. 1635-1660.