The nth-Order Bias Optimality for Multi-chain Markov Decision Processes

Xi-Ren Cao and Junyu Zhang *

Department of Electrical and Electronic Engineering

The Hong Kong University of Science and Technology

(Submitted to *IEEE Transactions on Automatic Control*)

Abstract

The paper proposes a new approach to the theory of Markov decision processes (MDPs) with average performance criteria and finite state and action spaces. Using the average performance and bias difference formulas derived in this paper, we develop an optimization theory for average performance (or gain) optimality, bias optimality, and all the high-order bias optimality, in a unified way. The approach is simple, direct, natural and intuitive; it does not depend on Laurent series expansion and discounted MDPs. We also propose one-phase policy iteration algorithms for bias and high-order bias optimal policies, which are more efficient than the two-phase algorithms in the literature. Furthermore, we derive the high-order bias optimality equations. This research is a part of our effort in developing sensitivity-based learning and optimization theory. The new insights provided by this approach may lead to some new research directions such as on-line learning, performance derivative based optimization, and potential or high-order potential aggregations.

Key Words: Bias optimality, discrete event systems, gain optimality, Markov decision processes, nth potentials, nth-bias optimality, policy iteration.

*Supported by a grant from Hong Kong UGC. E-mail addresses: eecao@ust.edu.hk and eezhjy@ust.edu.hk (10/2005)

1 Introduction

The paper proposes a new approach to the theory of Markov decision processes (MDPs) with average performance criteria. We show that a complete theory for average performance MDPs with finite state and action spaces and multiple chains can be derived naturally from the average performance and bias difference formulas. We derive policy iteration algorithms and develop optimization theory for gain (called average performance in this paper) optimality, bias optimality, and all the high-order bias optimality (defined in this paper), in a unified way.

This paper is a continuation of the recent research on sensitivity-based performance optimization of discrete event dynamic systems [3, 4, 5]. It is motivated by the previously established results. In particular, it is shown in [3, 4, 5] that the policy iteration algorithms and the theory for gain-optimality problems in Markov decision processes follow naturally from the performance difference formula.

Our work is closely related to [12, 13, 17, 18]. [12, 13] provide a solution to the bias optimality of the uni-chain case and leave the multi-chain case and high-order bias optimality untouched. Veinott's pioneering works [17, 18] provide a parallel solution to the problem from a different framework called n-discount optimality. Other works in MDPs include the linear programming approach [10]. We will not provide a comprehensive list of references in the MDP literature.

The contributions of this paper are as follows. First, this work provides a new approach to the performance optimization problems, including the average performance, bias, and the high-order bias optimality. Compared with previous works on MDPs and the n-discount optimality theory, this approach is simpler: the theory for bias and high-order bias optimality is almost the same as that for the average performance; the proof for the convergence of the policy iteration algorithms is the simplest, to the best of our knowledge. The approach is more direct: the approach is completely independent of the discounted MDP formulation and does not depend on Laurent series expansion; and the approach is more intuitive: it provides a different view for the optimization problem directly based on comparison of performance. Second, in our

approach, policy iteration is based on the performance and bias difference formulas and the optimality equations are only secondly. With this vision, we developed one-phase policy iteration algorithms for the bias and high-order bias optimal policies, which may save computation compared with the two-phase algorithms presented in Puterman's awarding winning book [16]. Third, this work is a part of our effort in developing the sensitivity-based performance optimization theory. The sensitivity-based view provides a unified framework for perturbation analysis and policy iteration for problems with different performance criteria including the discounted performance, average performance, and biases [3, 4, 5]; it also provides some new insights to the optimization problem and will lead to new research directions such as performance derivative-based optimization, on-line learning, aggregation, and problems that do not fit the standard MDP framework [6].

The paper is organized as follows. In Section 2, we define the nth bias and nth potential of a policy and the nth-bias optimality criterion. In Section 3, we derive the nth-bias difference formulas for multi-chain finite-state Markov chains. In Section 4, we derive policy iteration algorithms by using the nth-bias difference formulas in a clear and intuitive way. In Section 5, we give the nth-bias optimality equations. In Section 6, we provide some additional results and discussions comparing our results with those in the literature, especially with the n-discount optimality theory. Section 7 concludes the paper.

2 Nth-Bias Optimality

Consider a discrete-time multi-chain Markov decision process (MDP) [1, 16] with a finite state space $S = \{1, 2, \ldots, M\}$. Let A be the finite action space consisting of all available actions, and $A_i \subseteq A$ be the set of actions that are available when the Markov system is in state $i \in S$. When the system is in state i and action $a \in A_i$ is taken, the system transits to state j at the next decision epoch with transition probability $p_a(i, j)$, and a finite reward $r(i, a)$ is received.

A decision rule prescribes a procedure for action selection. A deterministic Marko-

vian decision rule is a function $d : S \to A$ that specifies the action $d(i) \in A_i$ taken when the system is in state $i \in S$. *A policy* $\pi = (d_0, d_1, d_2, \ldots)$ is a sequence of decision rules with d_k denoting the decision rule applied at decision epoch k, $k = 0, 1, \ldots$. A policy is said to be *stationary* if $d_k = d$ for all $k \geq 0$. Hence, a stationary policy has the form of $(d, d, \ldots)$. Since we only consider stationary deterministic Markovian policies in this paper, we will simply use d to denote a stationary policy $(d, d, \ldots)$, and use D to denote the set of all stationary deterministic Markovian policies. If policy d is adopted, the state transition probability matrix is denoted as $P_d = [p_{d(i)}(i, j)]_{i,j=1}^{M}$, and the reward function becomes $r(i, d(i))$, $i \in S$. We have $P_d e = e$, with $e = (1, \ldots, 1)^T$ being a column vector whose all components are one, where superscript "T" denotes transpose.

Consider a Markov chain $\{X_k, k = 0, 1, \ldots\}$ under a policy $d \in D$, where X_k denotes the system state at time k. The long-run *average performance* in [5, 9] (or the *gain* in [16]) of policy d, denoted as g_0^d, is defined as a vector with components

$$g_0^d(i) = \lim_{N\to\infty} \frac{1}{N} E^d \left[\sum_{k=0}^{N-1} r(X_k, d(X_k)) | X_0 = i \right], \quad i \in S, \tag{1}$$

where E^d denotes the expectation corresponding to the probability measure determined by the Markov chain $\{X_k, k = 0, 1, \ldots\}$ under policy d. We will see later that the limit exists. We can rewrite it in vector form:

$$g_0^d = \lim_{N\to\infty} \frac{1}{N} \left[\sum_{k=0}^{N-1} (P_d)^k r_d \right] = (P_d)^* r_d, \tag{2}$$

where $r_d = [r(1, d(1)), \ldots, r(M, d(M))]^T$, and

$$(P_d)^* := \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} (P_d)^k, \tag{3}$$

which was called the Cesaro limit and and its existence was proved in, e.g., [8], $(P_d)^0 = I$ with I being an $M \times M$ identity matrix. We can easily prove that $(P_d)^* e = e$ and

$$P_d (P_d)^* = (P_d)^* P_d = (P_d)^* (P_d)^* = (P_d)^*. \tag{4}$$

From (2) and (4), we obtain

$$P_d g_0^d = (P_d)^* g_0^d = g_0^d. \tag{5}$$

Now we define the *average performance (gain) optimality*. The *optimal average performance* is defined as $g_0^*(i) := \max_{d\in D} g_0^d(i)$, for all $i \in S$. A policy d^* is called *average performance (gain) optimal* if

$$g_0^{d^*}(i) = g_0^*(i) \quad \forall\, i \in S.$$

Let $D_0 := \{d \in D : g_0^d = g_0^*\}$ be the set of all average performance optimal policies. We will get that D_0 is not empty in Section 4.

From (2), it is clear that the average performance optimality criterion focuses on the long-run average or the steady-state behavior of a system; it ignores the transient performance in the initial period of the sample path. Therefore, the average performance optimality criterion is under-selective. We need to introduce a more selective optimality criterion - the *bias optimality* in [16] that is concerned with the transient performance.

For a policy $d \in D$, if P_d is aperiodic, its *bias* in [16] is defined as a vector g_1^d with components

$$g_1^d(i) = \lim_{N\to\infty} E^d\left\{\sum_{k=0}^{N}[r(X_k, d(X_k)) - g_0^d(i)]|X_0 = i\right\}, \quad i \in S, \tag{6}$$

which exists as shown below. By (2), (5) and (9), we can rewrite (6) in the vector form as follows,

$$\begin{aligned} g_1^d &= \sum_{k=0}^{\infty}[(P_d)^k r_d - g_0^d] \\ &= [I - P_d + (P_d)^*]^{-1}(r_d - g_0^d). \end{aligned} \tag{7}$$

From Theorem 4.3.1 of [11], the matrix $I - P_d + (P_d)^*$ is nonsingular, and we have

$$[I - P_d + (P_d)^*]^{-1}(P_d)^* = (P_d)^*[I - P_d + (P_d)^*]^{-1} = (P_d)^*, \tag{8}$$

and

$$[I - P_d + (P_d)^*]^{-1} = \sum_{k=0}^{\infty}[P_d - (P_d)^*]^k = \sum_{k=0}^{\infty}[(P_d)^k - (P_d)^*] + (P_d)^*. \tag{9}$$

For the periodic case, the bias (6) need to be defined with the Cesaro limit

$$g_1^d(i) = \lim_{N\to\infty} \frac{1}{N}\sum_{k=0}^{N-1} E^d\left\{\sum_{l=0}^{k}[r(X_l, d(X_l)) - g_0^d(i)]|X_0 = i\right\}, \quad i \in S, \tag{10}$$

From Theorem 2.14 of [9], we have

$$[I - P_d + (P_d)^*]^{-1} = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \sum_{l=0}^{k} [(P_d)^l - (P_d)^*] + (P_d)^*.$$

From this equation, we can see that (10) leads to the same equation (7).

For simplicity in expression, in this paper we assume that P_d is aperiodic for all $d \in D$. If P_d is periodic, we just need to replace the normal limit ($\lim_{N\to\infty}[\cdot]$) with the Cesaro limit ($\lim_{N\to\infty} 1/N \sum_{k=0}^{N-1}[\cdot]$) in the expression of $[I - P_d + (P_d)^*]^{-1}$ and then all results in this paper are true for the periodic case.

Pre-multiplying (7) by $(P_d)^*$, by (8), (2) and (5) we obtain

$$(P_d)^* g_1^d = (P_d)^*(r_d - g_0^d) = 0. \tag{11}$$

Combining the aforementioned equation with (7), we get the *Poisson equation* [4, 6]

$$(I - P_d)g_1^d + g_0^d = r_d. \tag{12}$$

The solution to (12) is not unique; i.e., if g_1^d is a solution, then for any vector u satisfying $P_d u = u$, $g_1^d + u$ is also a solution. The g_1^d defined in (6) in fact is a unique solution to (12) with the normalizing condition (11), $(P_d)^* g_1^d = 0$. If g_1^d only satisfies the Poisson equation (12), we call g_1^d the *potential* of policy d in [6]. Thus, the bias of policy d is a special potential of policy d which also satisfies the normalizing condition (11). There exist many different versions of potentials for a policy in the form of $g_1^d + u$. To simplify the notations, we use the same notation g_1^d to denote both the bias and the potential of policy d. But unless otherwise noted, g_1^d is the bias of policy d.

Now we define the *bias optimality*. The *optimal bias* is denoted as $g_1^*(i) := \max_{d\in D_0} g_1^d(i)$, for all $i \in S$. A policy $d^* \in D_0$ is called *bias optimal* if

$$g_1^{d^*}(i) = g_1^*(i) \quad \forall\, i \in S.$$

Let $D_1 := \{d \in D_0 : g_1^d = g_1^*\} = \{d \in D : g_0^d = g_0^*, g_1^d = g_1^*\}$ be the set of all bias optimal policies. We will also see in Section 4 that D_1 is not empty.

In (6), $g_0^d(i)(= g_0^*(i))$ is the same for all $d \in D_0$. Thus, maximizing the bias in D_0 is equivalent to maximizing the sum of the mean rewards $E\{\sum_{k=0}^{\infty} r(X_k, d(X_k))\}$ in the initial period.

Next, we will see later that to optimize the average performance, we need to use the bias. Similarly, to optimize the bias, we need to use the "bias of the bias", i.e., the second order bias. In general, we can define, by induction, the $(n+1)$*th-order bias* (or simply the $(n+1)$th bias) g_{n+1}^d, $n \geq 1$, of a policy $d \in D$ as a vector with components

$$g_{n+1}^d(i) = -E^d \left\{ \sum_{k=0}^{\infty} \{g_n^d(X_k) - [(P_d)^* g_n^d](i)\} | X_0 = i \right\}, \quad i \in S,\ n \geq 1, \tag{13}$$

where $(P_d)^* g_n^d$ is the steady-state value of the nth bias. In vector form, we have

$$\begin{aligned} g_{n+1}^d &= -\sum_{k=0}^{\infty} [(P_d)^k - (P_d)^*] g_n^d \\ &= -[I - P_d + (P_d)^*]^{-1} [g_n^d - (P_d)^* g_n^d], \quad n \geq 1. \end{aligned} \tag{14}$$

Pre-multiplying the aforementioned equation by $(P_d)^*$, and by (8) and (4) we have

$$(P_d)^* g_{n+1}^d = -(P_d)^* [I - P_d + (P_d)^*]^{-1} [g_n^d - (P_d)^* g_n^d] = 0, \quad n \geq 1.$$

Together with $(P_d)^* g_1^d = 0$, we have

$$(P_d)^* g_n^d = 0, \quad n \geq 1. \tag{15}$$

Thus, (13) can be simplified as

$$g_{n+1}^d(i) = -E^d \left[\sum_{k=0}^{\infty} g_n^d(X_k) | X_0 = i \right], \quad i \in S,\ n \geq 1, \tag{16}$$

and (14) becomes

$$g_{n+1}^d = -[I - P_d + (P_d)^*]^{-1} g_n^d, \quad n \geq 1. \tag{17}$$

Combining (15) with (17), we obtain

$$(I - P_d) g_{n+1}^d = -g_n^d, \quad n \geq 1. \tag{18}$$

This is *Poisson equation* for the high-order biases. Again, the solution to (18) is not unique; i.e., if g_{n+1}^d is a solution, then for any vector satisfying $P_d u = u$, $g_{n+1}^d + u$

is also a solution. g_{n+1}^d defined in (13) or (16) in fact is a unique solution to (18) with the normalizing condition (15), $(P_d)^* g_{n+1}^d = 0$. If g_{n+1}^d only satisfies the Poisson equation (18), we call g_{n+1}^d the $(n+1)$*th potential* of policy d. Thus, the $(n+1)$th bias of policy d is a special $(n+1)$th potential of policy d which also satisfies the normalizing condition (15). To simplify the notations, we use the same notation g_{n+1}^d to denote both the $(n+1)$th bias and the $(n+1)$th potential of policy d. But unless otherwise noted, g_{n+1}^d is the $(n+1)$th bias of policy d.

From (17), g_{n+1}^d is the bias of $-g_n^d$, $n = 1, 2, \cdots$. In general, we can derive

$$g_{n+1}^d = (-1)^n [I - P_d + (P_d)^*]^{-(n+1)} (r_d - g_0^d), \quad n \geq 0. \tag{19}$$

Further, with (9) we can obtain

$$g_{n+1}^d = (-1)^n \sum_{k=0}^{\infty} \begin{pmatrix} n+k \\ n \end{pmatrix} [(P_d)^k r_d - g_0^d], \qquad n \geq 0,$$

or

$$g_{n+1}^d(i) = (-1)^n \sum_{k=0}^{\infty} \begin{pmatrix} n+k \\ n \end{pmatrix} E^d \left\{ [r(X_k, d(X_k)) - g_0^d(i)] | X_0 = i \right\}, \ n \geq 0. \tag{20}$$

This equation or (16) can be used to develop algorithms to estimate g_{n+1}^d on a single sample path without knowing P_d.

We now define the *nth-bias optimality*, $n \geq 0$. 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

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

where $D_{n-1} := \{d \in D_{n-2} : g_{n-1}^d = g_{n-1}^*\} = \{d \in D : g_k^d = g_k^*, k = 0, 1, \ldots, n-1\}$ is the set of all $(n-1)$th-bias optimal policies, and $D_{-1} := D$. We will prove that an nth-bias optimal policy, $n \geq 0$, always exists. By definition, we have $D_n \subseteq D_{n-1}$, $n \geq 0$. That is, the bigger n is, the more selective the nth-bias optimality is. The long-run average performance g_0^d and the bias g_1^d will also be called the 0*th-order bias* and the *1st-order bias*, respectively.

For the 2nd bias ($n = 2$), from (20) we have

$$g_2^d(i) = -\sum_{k=0}^{\infty}(k+1)E^d\{[r(X_k, d(X_k)) - g_0^d(i)]|X_0 = i\}, \quad i \in S. \tag{21}$$

By (6), we see that the bias g_1 gives the same weight on the received reward at the different decision epoch. But, g_2 exercises a large penalty (negative reward) on lately received reward $r(X_k, d(X_k)) - g_0^d(i)$ by multiplying it with a factor $(k+1)$. In other words, to maximize g_2 implies that we prefer to receive the reward early or to receive the temporary reward.

The following lemma will be used often in the remaining of this paper. For two vectors (or functions) x and y defined on state space S, we define $x = y$ if $x(i) = y(i)$ for all $i \in S$; $x \geq y$ if $x(i) \geq y(i)$ for all $i \in S$; $x \succeq y$ if $x \geq y$ and $x(i) > y(i)$ for at least some $i \in S$.

Lemma 1 *Let u be an M-dimensional vector. If $u \geq 0$ (or $u \leq 0$) and $(P_d)^* u = 0$, then $u(i) = 0$ for all recurrent states i under policy d.*

Proof: The result follows directly from the fact that $(P_d)^*(i,j) = 0$ for all $i \in S$ if j is a transient state under policy d and $(P_d)^*(i,j) > 0$ if both i and j are recurrent states in the same sub-closed set under policy d. □

3 Nth-Bias Difference Formulas

We first derive formulas for the performance differences and the differences of the $(n+1)$th biases of two different policies with the same nth bias, $n \geq 0$. These formulas form the basis of the optimization theory of the nth biases.

Lemma 2 *For any $h, d \in D$, let g_n^h and g_n^d be the nth biases of policies h and d, $n = 0, 1, \ldots$, respectively. Then*

(a) $g_0^h - g_0^d = (P_h)^*[(r_h + P_h g_1^d) - (r_d + P_d g_1^d)] + [(P_h)^* - I]g_0^d$.

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

$$g_1^h - g_1^d = [I - P_h + (P_h)^*]^{-1}[(r_h + P_h g_1^d) - (r_d + P_d g_1^d)] + (P_h)^*(P_h - P_d)g_2^d.$$

(c) If $g_n^h = g_n^d$, then

$$g_{n+1}^h - g_{n+1}^d = [I - P_h + (P_h)^*]^{-1}(P_h - P_d)g_{n+1}^d + (P_h)^*(P_h - P_d)g_{n+2}^d, \quad n \geq 1.$$

In addition, the above equations hold even if the g_1^d, g_2^d, and g_{n+2}^d on the right-hand sides of (a), (b) and (c), respectively, are the potential, 2nd potential, and $(n+2)$th potential of policy d, respectively; that is, they satisfy only the corresponding Poisson equations without the normalizing conditions.

Proof. (a) From (2) and (4), we have

$$\begin{aligned} & g_0^h - g_0^d = (P_h)^* r_h - g_0^d \\ = \; & (P_h)^*(r_h - g_0^d) + [(P_h)^* - I]g_0^d \\ = \; & (P_h)^*(r_h + P_h g_1^d - g_0^d - g_1^d) + [(P_h)^* - I]g_0^d. \end{aligned}$$

Then (a) follows directly from Poisson equation (12).

(b) From (12) and $g_0^h = g_0^d$, we have

$$g_1^h - g_1^d = r_h + P_h g_1^h - g_0^d - g_1^d.$$

Adding both sides with $-P_h(g_1^h - g_1^d)$, we have

$$(I - P_h)(g_1^h - g_1^d) = r_h + P_h g_1^d - g_0^d - g_1^d.$$

Since on the left-hand side $(P_h)^* g_1^h = 0$, we have

$$[I - P_h + (P_h)^*](g_1^h - g_1^d) = r_h + P_h g_1^d - g_0^d - g_1^d - (P_h)^* g_1^d.$$

From (8), we have

$$g_1^h - g_1^d = [I - P_h + (P_h)^*]^{-1}(r_h + P_h g_1^d - g_0^d - g_1^d) - (P_h)^* g_1^d.$$

With (4), we get

$$g_1^h - g_1^d = [I - P_h + (P_h)^*]^{-1}(r_h + P_h g_1^d - g_0^d - g_1^d) + (P_h)^*(P_h g_2^d - g_1^d - g_2^d).$$

Then (b) follows directly from Poisson equations (12) and (18) with $n = 1$.

In addition, from (9), and noting that $(P_h)^*(r_h + P_h g_1^d - g_0^d - g_1^d) = g_0^h - g_0^d = 0$, we have

$$\begin{aligned} g_1^h - g_1^d &= \sum_{k=0}^{\infty}(P_h)^k(r_h + P_h g_1^d - g_0^d - g_1^d) + (P_h)^*(P_h g_2^d - g_1^d - g_2^d) \quad (22) \\ &= \sum_{k=0}^{\infty}(P_h)^k(r_h + P_h g_1^d - r_d - P_d g_1^d) + (P_h)^*(P_h - P_d)g_2^d. \quad (23) \end{aligned}$$

(c) From (18) and $g_n^h = g_n^d$, we have

$$g_{n+1}^h - g_{n+1}^d = P_h g_{n+1}^d - g_n^d - g_{n+1}^d + P_h(g_{n+1}^h - g_{n+1}^d).$$

Thus,

$$(I - P_h)(g_{n+1}^h - g_{n+1}^d) = P_h g_{n+1}^d - g_n^d - g_{n+1}^d. \quad (24)$$

Because on the left-hand side we have $(P_h)^* g_{n+1}^h = 0$, by (24) we get

$$[I - P_h + (P_h)^*](g_{n+1}^h - g_{n+1}^d) = P_h g_{n+1}^d - g_n^d - g_{n+1}^d - (P_h)^* g_{n+1}^d.$$

Since $[I - P_h + (P_h)^*]^{-1}(P_h)^* = (P_h)^*$, we have

$$\begin{aligned} & g_{n+1}^h - g_{n+1}^d \quad (25) \\ = \; & [I - P_h + (P_h)^*]^{-1}(P_h g_{n+1}^d - g_n^d - g_{n+1}^d) - (P_h)^* g_{n+1}^d \\ = \; & [I - P_h + (P_h)^*]^{-1}(P_h g_{n+1}^d - g_n^d - g_{n+1}^d) + (P_h)^*(P_h g_{n+2}^d - g_{n+1}^d - g_{n+2}^d). \end{aligned}$$

Then (c) follows directly from Poisson equation (18) with n and $n+1$.

In addition, from (15) and $g_n^h = g_n^d$, we obtain

$$(P_h)^*(P_h g_{n+1}^d - g_n^d - g_{n+1}^d) = -(P_h)^* g_n^d = -(P_h)^* g_n^h = 0.$$

Thus, from (25) and using (9) we have

$$\begin{aligned} & g_{n+1}^h - g_{n+1}^d \\ = \; & \sum_{k=0}^{\infty}(P_h)^k(P_h g_{n+1}^d - g_n^d - g_{n+1}^d) + (P_h)^*(P_h g_{n+2}^d - g_{n+1}^d - g_{n+2}^d) \quad (26) \\ = \; & \sum_{k=0}^{\infty}(P_h)^k(P_h - P_d)g_{n+1}^d + (P_h)^*(P_h - P_d)g_{n+2}^d, \quad (27) \end{aligned}$$

for $n \geq 1$. From the process of the proof, we do not use the normalizing conditions, thus the "In addition" part of this lemma holds. □

The bias difference formulas in Lemma 2 allow us to compare the biases of two different policies based on only one policy's biases under some conditions. This is shown in Theorems 1, 2 and 3 below, for the average performance (gain), the 1st-order bias, and the higher-order biases, respectively. These theorems lead naturally to the performance optimization theory presented in the remaining of the paper.

Theorem 1 *Suppose that two policies d and h satisfy the following two conditions:*

(a) $P_h g_0^d \geq g_0^d$, *and*

(b) $r_h(i) + (P_h g_1^d)(i) \geq r_d(i) + (P_d g_1^d)(i)$ *when* $(P_h g_0^d)(i) = g_0^d(i)$ *for some* $i \in S$,

then $g_0^h \geq g_0^d$. *This theorem also holds if we change all the symbols "* $\geq$ *" to "* $\leq$ *". The result also holds if* g_1^d *is only the potential of policy* d.

Proof. For simplicity, we denote $y_d^h := r_h + P_h g_1^d - (r_d + P_d g_1^d)$ in this paper. Because $(P_h)^* P_h = (P_h)^*$, we have $(P_h)^*(P_h g_0^d - g_0^d) = 0$. Thus, by Lemma 1, with condition (a), we have $(P_h g_0^d)(i) = g_0^d(i)$ for all recurrent states i under policy h. Then it follows from condition (b) that $y_d^h(i) \geq 0$ for all recurrent states i under policy h. Observing that $(P_h)^*(i,j) = 0$ for all $i \in S$ and any transient state j under policy h, we have $(P_h)^* y_d^h \geq 0$. On the other hand, because $P_h g_0^d \geq g_0^d$, we have $(P_h)^k g_0^d \geq g_0^d$ for all $k \geq 1$. Therefore, by (3) we get $(P_h)^* g_0^d \geq g_0^d$. Finally, by the bias difference formulas in Lemma 2 (a), we have $g_0^h - g_0^d = (P_h)^* y_d^h + [(P_h)^* - I] g_0^d \geq 0$. The proof for the " $\leq$ " case is similar, we omit it. By Lemma 2, the theorem also holds if g_1^d is only the potential of policy d. □

Theorem 2 *Suppose that policy d is average performance optimal. If any $h \in D$ satisfies the following three conditions:*

(a) $P_h g_0^d = g_0^d$,

(b) $r_h + P_h g_1^d \geq r_d + P_d g_1^d$, *and*

(c) $(P_h g_2^d)(i) \geq (P_d g_2^d)(i)$ *when* $r_h(i) + (P_h g_1^d)(i) = r_d(i) + (P_d g_1^d)(i)$ *for some* $i \in S$,

then policy h is also average performance optimal and $g_1^h \geq g_1^d$.

In addition, if condition (a) is replaced by the following condition (a')

(a') *policy h is average performance optimal, i.e.,* $g_0^h = g_0^d$,

then $g_1^h \leq g_1^d$ *if we change all the symbols "* $\geq$ *" in (b) and (c) to "* $\leq$ *".*

The result also holds if g_2^d is only the 2nd potential of policy d.

Proof. From conditions (a) and (b), we have

$$g_0^h - g_0^d = (P_h)^* y_d^h + [(P_h)^* - I] g_0^d = (P_h)^* y_d^h \geq 0.$$

Since g_0^d is the optimal average performance, we have $g_0^h = g_0^d$, i.e., h is also average performance optimal. Thus, $(P_h)^* y_d^h = (P_h)^*(r_h + P_h g_1^d - r_d - P_d g_1^d) = 0$. Because $y_d^h \geq 0$, and by Lemma 1 $y_d^h(i) = 0$ for all recurrent states i under policy h, it follows from condition (c) that $[(P_h - P_d)g_2^d](i) \geq 0$ for all recurrent states i under policy h. Thus, from the structure of $(P_h)^*$,we have $(P_h)^*(P_h - P_d)g_2^d \geq 0$. From the bias difference formulas (23) in Lemma 2 (b), $g_1^h - g_1^d = \sum_{k=0}^{\infty}(P_h)^k y_d^h + (P_h)^*(P_h - P_d)g_2^d \geq \sum_{k=0}^{\infty}(P_h)^k y_d^h \geq 0$.

In addition, by condition (a'), $g_0^h = g_0^d$, we also have $g_0^h - g_0^d = (P_h)^* y_d^h = 0$. The additional part of the theorem follows in a similar way. By Lemma 2, the theorem also holds if g_2^d is only the 2nd potential of policy d. □

The proof of the following theorem is very similar and hence is omitted.

Theorem 3 *Suppose that policy d is $(n-1)$th-bias optimal, $n \geq 2$. If any $(n-2)$th-bias optimal policy h satisfies the following three conditions:*

(a) $P_h g_{n-1}^d = P_d g_{n-1}^d$, *for* $n \geq 3$, $r_h + P_h g_{n-1}^d = r_d + P_d g_{n-1}^d$, *for* $n = 2$,

(b) $P_h g_n^d \geq P_d g_n^d$, *and*

(c) $(P_h g_{n+1}^d)(i) \geq (P_d g_{n+1}^d)(i)$ *when* $(P_h g_n^d)(i) = (P_d g_n^d)(i)$ *for some* $i \in S$,

then policy h is also $(n-1)$th-bias optimal and $g_n^h \geq g_n^d$.

In addition, if condition (a) is replaced by the following condition (a')

(a') policy h is $(n-1)$th bias optimal, i.e., $g_{n-1}^h = g_{n-1}^d$,

then $g_n^h \leq g_n^d$ if we change all the symbols " $\geq$ " in (b) and (c) to " $\leq$ ".

The result also holds if g_{n+1}^d is only the $(n+1)$th potential of policy d.

4 Policy Iteration Algorithms

Theorems 1-3, which follow almost directly from the bias difference formulas, provide a clear picture for bias optimization: Given an nth-bias optimal policy d, we can find

another policy h in the space of $(n-1)$th-bias optimal policies that is nth-bias optimal and has a larger $(n+1)$th bias. We can continue this improvement procedure until it reaches a policy for which no improvement can be made by this procedure. This policy must be the $(n+1)$th-bias optimal. This procedure is called *policy iteration*. The remaining of this paper simply makes the above verbal description mathematically rigorous.

We will first find an average performance (also called the 0th-bias) optimal policy. For any non-average-performance optimal policy we can always construct a "better" policy by Theorem 1. That is, we can improve the average performance at each iteration. If there is no further improvement, we can prove that this policy is average performance optimal by the average performance and bias difference formulas. This process can be formally described as follows. The resulting policy iteration algorithm is the same as what in the literature, but the proof provided here is simpler.

Given any policy $d \in D$, for any $i \in S$ and $a \in A_i$, let

$$H_d(i,a) := r(i,a) + \sum_{j \in S} p_a(i,j) g_1^d(j),$$

and

$$A_0^d(i) := \left\{ a \in A_i : \begin{array}{l} \sum_{j \in S} p_a(i,j) g_0^d(j) > g_0^d(i); \quad \text{or} \\ H_d(i,a) > H_d(i,d(i)) \\ \text{when } \sum_{j \in S} p_a(i,j) g_0^d(j) = g_0^d(i) \end{array} \right\}. \tag{28}$$

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

$$h(i) \in A_0^d(i) \text{ if } A_0^d(i) \neq \emptyset, \text{ and } h(i) = d(i) \text{ if } A_0^d(i) = \emptyset. \tag{29}$$

Note that such a policy may not be unique, since there may be more than one action in $A_0^d(i)$ for some state $i \in S$. Recall $y_d^h := r_h + P_h g_1^d - (r_d + P_d g_1^d)$. We have

$$y_d^h(i) = H_d(i,h(i)) - H_d(i,d(i)). \tag{30}$$

Theorem 4 *For any given $d \in D$, let h be defined as in (29). We have*

(a) $g_0^h \geq g_0^d$.

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

Proof. For any $i \in S$, if $A_0^d(i) = \emptyset$, then $h(i) = d(i)$ and we have $P_h(i,j) = P_d(i,j)$ for all $j \in S$. Thus, $\sum_{j\in S} P_h(i,j)g_0^d(j) = \sum_{j\in S} P_d(i,j)g_0^d(j) = g_0^d(i)$. Next, if $A_0^d(i) \neq \emptyset$, from the construction by (28), we have $\sum_{j\in S} P_h(i,j)g_0^d(j) \geq g_0^d(j)$. Thus, condition (a) in Theorem 1 holds. In addition, if $\sum_{j\in S} P_h(i,j)g_0^d(j) = g_0^d(i)$, then either $H_d(i,h(i)) = H_d(i,d(i))$ when $A_0^d(i) = \emptyset$, or $H_d(i,h(i)) > H_d(i,d(i))$ when $A_0^d(i) \neq \emptyset$. That is, condition (b) in Theorem 1 also holds. Thus, it follows from Theorem 1 that $g_0^h \geq g_0^d$.

For part (b), since $g_0^h = g_0^d$, $P_h g_0^d = g_0^d$ follows by (5). Then by (28), (29), and (30), we have either $h(i) = d(i)$ or $y_d^h(i) > 0$ for all $i \in S$. Because $h \neq d$, we have $y_d^h \succeq 0$, or $r_h + P_h g_1^d \succeq r_d + P_d g_1^d$. Noting that $g_0^h - g_0^d = (P_h)^*[r_h + P_h g_1^d - (r_d + P_d g_1^d)] = 0$ and by Lemma 1, we have $r_h(i) + (P_h g_1^d)(i) = r_d(i) + (P_d g_1^d)(i)$, for all recurrent states i under policy h. From (28) and (29),

$$h(i) = d(i), \quad \text{for all recurrent states } i \text{ under policy } h. \tag{31}$$

By (31) and $(P_h)^*(i,j) = 0$ when $i \in S$ and j is any transient state under policy h, we get $(P_h)^*(P_h - P_d) = 0$. Then by the bias difference formula (23) in Lemma 2 (b), $g_1^h - g_1^d = \sum_{k=0}^{\infty}(P_h)^k[r_h + P_h g_1^d - r_d - P_d g_1^d] \geq r_h + P_h g_1^d - r_d - P_d g_1^d \succeq 0$. □

Theorem 4 essentially claims that if h and d have the same average performance, and $r_h + P_h g_1^d \succeq r_d + P_d g_1^d$, then $g_1^h \succeq g_1^d$. With Theorem 4, we state the (standard) *Average Performance Optimality Policy Iteration Algorithm* as follows:

1. Select an arbitrary policy $d_0 \in D$, and set $k = 0$.

2. (Policy evaluation) Obtain $g_0^{d_k}$ and $g_1^{d_k}$ by solving

$$\begin{aligned} (P_{d_k} - I)g_0 &= 0, \\ r_{d_k} - g_0 + (P_{d_k} - I)g_1 &= 0 \end{aligned}$$

   subject to $(P_{d_k})^* g_1 = 0$.

3. (Policy improvement) Set d_k as policy d and obtain policy d_{k+1} as policy h in (28) and (29), setting $d_{k+1}(i) = d_k(i)$ if possible.

4. If $d_{k+1} = d_k$, stop and set $d^* = d_k$ and $g_0^* = g_0^{d_k}$, otherwise increase k by 1 and return to step 2.

Theorem 4 (a) guarantees that the average performance does not decrease at each iteration, and Theorem 4 (a) and (b) guarantee that the policies do not go cycling in the policy iteration procedure. This leads to the following convergence theorem.

Theorem 5 *The Average Performance Optimality Policy Iteration Algorithm stops at an average performance optimal policy in a finite number of iterations.*

Proof: By Theorem 4 (a), we have $g_0^{d_{k+1}} \geq g_0^{d_k}$. That is, as k increases, the average performance $g_0^{d_k}$ either increases or stays the same. Furthermore, by Theorem 4 (b), when $g_0^{d_k}$ stays the same, $g_1^{d_k}$ increases. Thus, any two policies in the sequence of d_k, $k = 0, 1, \ldots$, either have different average performance (g_0) or have different 1st bias (g_1). 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 that it stops at a policy denoted as d^*. This means that $A_0^{d^*}(i)$ is empty for all $i \in S$. Then d^* must satisfy the following equations,

$$P_d g_0^{d^*} \leq g_0^{d^*}, \quad \forall\, d \in D,$$

and if $(P_d g_0^{d^*})(i) = g_0^{d^*}(i)$ for some $i \in S$, we have $r_d(i) + (P_d g_1^{d^*})(i) \leq r_{d^*}(i) + (P_{d^*} g_1^{d^*})(i)$. (Otherwise for some i the set $A_0^{d^*}(i)$ in (28) is non-empty and the iteration continues at d^*.) Then by Theorem 1 for the " $\leq$ " case, $g_0^d - g_0^{d^*} = (P_d)^*(r_d + P_d g_1^{d^*} - r_{d^*} - P_{d^*} g_1^{d^*}) + [(P_d)^* - I] g_0^{d^*} \leq 0$, for all $d \in D$. Thus, policy d^* is average performance optimal. □

The existence of the average performance (or 0th-bias, or gain) optimal policy follows from Theorem 5 by construction with the policy iteration algorithm. Policy iteration works if we pick up any action in $A_0^d(i)$ by (28) in the policy improvement step. In real implementation, however, we usually choose the action with the largest value of $\sum_{j \in S} p_a(i,j) g_0^d(j)$ or $H_d(i,a)$ in (28). To the best of our knowledge, the proof presented here is the simplest (cf. [5, 16, 17]).

Following the same procedure as for the average performance optimal problem, by Theorem 2, from any average performance optimal policy we can construct another

average performance optimal policy that has a larger 1st bias, if such a policy exists. Ideally, we need only search the set of the average performance optimal policies, D_0, for a (1st) bias optimal policy. However, given an average performance optimal policy, it is difficult to specify D_0 for the multi-chain case. Fortunately, by Theorem 2, we can search the set $F_1 = \otimes_{i=1}^{M} F_1(i)$, where $\otimes$ denotes the Cartesian product, with

$$F_1(i) := \{a \in A_i \ : \ \sum_{j=1}^{M} p_a(i,j) g_0^*(j) = g_0^*(i)\}.$$

From (5), if $h \in D_0$, we have $P_h g_0^* = g_0^*$, and thus $h \in F_1$. That is, $D_0 \subseteq F_1$.

Now we develop the policy iteration theory for the bias optimality. Given $d \in D_0$, for any state $i \in S$, let

$$A_1^d(i) := \left\{ a \in F_1(i) : \begin{array}{l} H_d(i,a) > H_d(i,d(i)); \ \text{ or} \\ \sum_{j \in S} p_a(i,j) g_2^d(j) > \sum_{j \in S} p_{d(i)}(i,j) g_2^d(j) \\ \text{when } H_d(i,a) = H_d(i,d(i)) \end{array} \right\}. \tag{32}$$

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

$$h(i) \in A_1^d(i) \ \text{ if } A_1^d(i) \neq \emptyset, \ \text{ and } \ h(i) = d(i) \ \text{ if } A_1^d(i) = \emptyset. \tag{33}$$

Note that such a policy also may not be unique, since there may be more than one action in $A_1^d(i)$ for some state $i \in S$.

Theorem 6 *For any given $d \in D_0$, let h be defined as in (33). We have*

(a) $g_0^h = g_0^d = g_0^*$, i.e., $h \in D_0$, and

(b) $g_1^h \geq g_1^d$.

(c) If $g_1^h = g_1^d$ and $h \neq d$, then $g_2^h \succeq g_2^d$.

Proof. (a) By (32), $h(i) \in F_1(i)$ for all $i \in S$. Thus, $P_h g_0^* = g_0^*$. Again by (32), we have $H_d(i,h(i)) \geq H_d(i,d(i))$ for all $i \in S$. Thus, by the bias difference formulas in Lemma 2 (a), we have $g_0^h \geq g_0^d$. Since $g_0^d = g_0^*$, We must have $g_0^h = g_0^d = g_0^*$.

(b) By the construction in (32) and (33), the conditions (a), (b) and (c) in Theorem 2 hold. It follows from Theorem 2 that $g_1^h \geq g_1^d$.

(c) Since $g_1^h = g_1^d$, we obtain $r_h + P_h g_1^d = r_d + P_d g_1^d$ (by Poisson equation (12) and $g_0^h = g_0^d = g_0^*$). Then, by (32), (33) and $h \neq d$, $P_h g_2^d \succeq P_d g_2^d$ holds. Therefore,

the first term on the right-hand side of the 1st bias difference formulas in Lemma 2 (b) is zero, and we have $g_1^h - g_1^d = (P_h)^*(P_h - P_d)g_2^d = 0$. By Lemma 1, we have $(P_h g_2^d)(i) = (P_d g_2^d)(i)$ for all recurrent states i under policy h. From (32) and (33),

$$h(i) = d(i), \quad \text{for all recurrent states } i \text{ under policy } h. \tag{34}$$

By (34) and the structure of $(P_h)^*$, we get $(P_h)^*(P_h - P_d) = 0$. Then by (27) in Lemma 2 (c) $g_2^h - g_2^d = \sum_{k=0}^{\infty}(P_h)^k(P_h - P_d)g_2^d \geq (P_h - P_d)g_2^d \succeq 0$. □

The *(1st) Bias Optimality Policy Iteration Algorithm* then follows directly.

1. Select an arbitrary average performance optimal policy $d_0 \in D_0$, and set $k = 0$.

2. (Policy evaluation) Obtain $g_1^{d_k}$ and $g_2^{d_k}$ by solving

$$\begin{aligned} r_{d_k} - g_0^* + (P_{d_k} - I)g_1 &= 0, \\ g_1 + (P_{d_k} - I)g_2 &= 0 \end{aligned}$$

   subject to $(P_{d_k})^* g_2 = 0$.

3. (Policy improvement) Set d_k as policy d and obtain policy d_{k+1} as policy h in (32) and (33), setting $d_{k+1}(i) = d_k(i)$ if possible.

4. If $d_{k+1} = d_k$, stop and set $d^* = d_k$ and $g_1^* = g_1^{d_k}$, otherwise increase k by 1 and return to step 2.

The following convergence theorem follows from Theorem 6 immediately.

Theorem 7 *The (1st) Bias Optimality Policy Iteration Algorithm stops at a bias optimal policy in a finite number of iterations.*

Proof. The proof is essentially the same as that for Theorem 5. By Theorem 6 (a), all d_k, $k = 0, 1, \ldots$, produced by the algorithm are average performance optimal. By Theorem 6 (b), we have $g_1^{d_{k+1}} \geq g_1^{d_k}$. That is, as k increases, the bias $g_1^{d_k}$ either increases or stays the same. Furthermore, by Theorem 6 (c), when $g_1^{d_k}$ stays the same, $g_2^{d_k}$ increases. Thus, any two policies in the sequence of d_k, $k = 0, 1, \ldots$, either

have different g_1 or have different g_2. 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 that it stops at a policy denoted as d^*. We will prove that d^* is bias optimal. First, by Theorem 6 (a), d^* is average performance optimal. Next, for any $d \in D_0$, we have shown that $d \in F_1$. By construction from (32), we have

(i) $r_d + P_d g_1^{d^*} \leq r_{d^*} + P_{d^*} g_1^{d^*}$, and

(ii) $(P_d g_2^{d^*})(i) \leq (P_{d^*} g_2^{d^*})(i)$, when $r_d(i) + (P_d g_1^{d^*})(i) = r_{d^*}(i) + (P_{d^*} g_1^{d^*})(i)$ for some $i \in S$.

Then by the "In addition" part of Theorem 2, we get $g_1^d \leq g_1^{d^*}$ for all $d \in D_0$. Thus, policy d^* is bias optimal. □

The existence of the bias optimal policy can be proved by construction with policy iteration as sown in Theorem 7. In real implementation, we usually choose the action with the largest value of $H_d(i,a)$ or $\sum_{j \in S} p_a(i,j) g_2^d(j)$ in (32).

Now we extend the results to the nth biases, $n \geq 2$. As indicated by (21), maximizing g_2 means receiving the reward as early as possible, measured by a weighting factor $(k+1)$. Denote

$$F_2(i) := \{a \in F_1(i) : r(i,a) + \sum_{j=1}^{M} p_a(i,j) g_1^*(j) = g_0^*(i) + g_1^*(i)\},$$

and $F_n(i)$ recursively for $n \geq 2$,

$$F_{n+1}(i) := \{a \in F_n(i) : \sum_{j=1}^{M} p_a(i,j) g_n^*(j) = g_{n-1}^*(i) + g_n^*(i)\}. \tag{35}$$

Denote $F_n = \otimes_{i=1}^{M} F_n(i)$ for $n \geq 1$. Then we have the following lemma.

Lemma 3 $\quad F_{n+2} \subseteq D_n \subseteq F_{n+1}, \quad n \geq 0.$

Proof. Firstly, we will prove $D_n \subseteq F_{n+1}$, $n \geq 0$. For $n = 0$, we have already proved $D_0 \subseteq F_1$.

For $n = 1$, let $h \in D_1$ be a bias optimal policy. We have $g_0^h = g_0^*$ and $g_1^h = g_1^*$. Since h is also average performance optimal, we have $h \in F_1$. From (12), $r_h + P_h g_1^* = g_0^* + g_1^*$.

Together with $h \in F_1$, we have $h \in F_2$. Thus, $D_1 \subseteq F_2$. Next, suppose $D_k \subseteq F_{k+1}$, for a particular $k \geq 1$. Let $h \in D_{k+1}$ be a $(k+1)$th-bias optimal policy. We have $g_l^h = g_l^*$, where $l = 0, 1, \ldots, k+1$. Since h is also kth-bias optimal, we have $h \in F_{k+1}$ from assumption. From (18), $P_h g_{k+1}^* = g_k^* + g_{k+1}^*$. Together with $h \in F_{k+1}$, we have $h \in F_{k+2}$ by (35). Thus $D_{k+1} \subseteq F_{k+2}$. Therefore $D_n \subseteq F_{n+1}$ for all $n \geq 0$ by induction.

Secondly, we will prove $F_{n+2} \subseteq D_n$, $n \geq 0$. For $n = 0$, $h \in F_2$, we have $P_h g_0^* = g_0^*$ and $r_h + P_h g_1^* = g_0^* + g_1^*$. Pre-multiplying both sides of this equation by $(P_h)^*$, we have $g_0^* = (P_h)^* r_h = g_0^h$ noting $(P_h)^* g_0^* = g_0^*$. Obviously, policy h is average performance optimal, i.e., $h \in D_0$. Thus, $F_2 \subseteq D_0$.

For $n = 1$, $h \in F_3$. By (35) we have $P_h g_0^* = g_0^*$, $r_h + P_h g_1^* = g_0^* + g_1^*$ and $P_h g_2^* = g_1^* + g_2^*$. Since $h \in F_2$, we have $h \in D_0$. Pre-multiplying both sides of $P_h g_2^* = g_1^* + g_2^*$ by $(P_h)^*$, we have $(P_h)^* g_1^* = 0$. Combining this with $r_h + P_h g_1^* = g_0^* + g_1^*$ and (7), we get $g_1^* = [I - P_h + (P_h)^*]^{-1}(r_h - g_0^*) = g_1^h$. Obviously, policy h is bias optimal, i.e., $h \in D_1$. Thus, $F_3 \subseteq D_1$.

Next, suppose $F_{k+2} \subseteq D_k$, for a particular $k \geq 1$. For any $h \in F_{k+3}$, by (35) we have $h \in F_{k+2}$ and $P_h g_{k+2}^* = g_{k+1}^* + g_{k+2}^*$. From the assumption, $h \in D_k$. Pre-multiplying both sides of $P_h g_{k+2}^* = g_{k+1}^* + g_{k+2}^*$ by $(P_h)^*$, we have $(P_h)^* g_{k+1}^* = 0$. Combining this with $P_h g_{k+1}^* = g_k^* + g_{k+1}^*$ and (17), we get $g_{k+1}^* = -[I - P_h + (P_h)^*]^{-1} g_k^* = g_{k+1}^h$. Obviously, policy h is the $(k+1)$th-bias optimal, i.e., $h \in D_{k+1}$. Thus $F_{k+3} \subseteq D_{k+1}$. Therefore $F_{n+2} \subseteq D_n$ for all $n \geq 0$ by induction. □

Just as for the case of the bias, given an $(n-1)$th-bias optimal policy, $n \geq 2$, it is difficult to specify the set of all the $(n-1)$th-bias optimal policies, D_{n-1}. Thus, by Lemma 3, we may search in F_n for an nth-bias optimal policy.

Next, we devise a policy iteration algorithm for the nth-bias optimality by following the same procedure as for the average performance and the (1st) bias optimality problems. By Theorem 3, from any $(n-1)$th-bias optimal policy we can construct another $(n-1)$th-bias optimal policy which has a larger nth bias, if such a policy exists.

For a given $(n-1)$th-bias optimal policy $d \in D_{n-1}$ with the kth bias g_k^d, $k =$

$0, 1, \ldots, n+1$, $n \geq 2$, we first define

$$\left.\begin{array}{l} A_n^d(i) := \{a \in F_n(i) : \\ \sum_{j\in S} p_a(i,j) g_n^d(j) > \sum_{j\in S} p_{d(i)}(i,j) g_n^d(j); \quad \text{or} \\ \sum_{j\in S} p_a(i,j) g_{n+1}^d(j) > \sum_{j\in S} p_{d(i)}(i,j) g_{n+1}^d(j) \\ \text{when } \sum_{j\in S} p_a(i,j) g_n^d(j) = \sum_{j\in S} p_{d(i)}(i,j) g_n^d(j)) \end{array}\right\}, \; i \in S. \tag{36}$$

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

$$h(i) \in A_n^d(i) \;\text{ if } A_n^d(i) \neq \emptyset, \;\text{ and }\; h(i) = d(i) \;\text{ if } A_n^d(i) = \emptyset. \tag{37}$$

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

We omit the proofs of the following Theorems 8 and 9 since they are similar to the respective proofs of the $n = 1$ case.

Theorem 8 *For any given $(n-1)$th-bias optimal policy $d \in D_{n-1}$, $n \geq 2$, let h be defined as in (37). We have*

(a) $g_{n-1}^h = g_{n-1}^d = g_{n-1}^*$, *i.e.,* $h \in D_{n-1}$, *and*

(b) $g_n^h \geq g_n^d$.

(c) If $g_n^h = g_n^d$ *and* $h \neq d$, *then* $g_{n+1}^h \succeq g_{n+1}^d$.

The *nth-Bias Optimality Policy Iteration Algorithm* is then as follows:

1. Set $k = 0$ and select an arbitrary $(n-1)$th-bias optimal policy $d_0 \in D_{n-1}$, which may be obtained from the $(n-1)$th-bias optimality policy iteration algorithm.

2. (Policy evaluation) Obtain $g_n^{d_k}$ and $g_{n+1}^{d_k}$ by solving

$$\begin{aligned} -g_{n-1}^* + (P_{d_k} - I) g_n &= 0, \\ -g_n + (P_{d_k} - I) g_{n+1} &= 0 \end{aligned}$$

   subject to $(P_{d_k})^* g_{n+1} = 0$.

3. (Policy improvement) Set d_k as policy d and obtain policy d_{k+1} as policy h in (36) and (37), setting $d_{k+1}(i) = d_k(i)$, $i \in S$, if possible.

4. If $d_{k+1} = d_k$, stop and set $d^* = d_k$ and $g_n^* = g_n^{d_k}$; otherwise increase k by 1 and return to step 2.

Theorem 9 *The nth-Bias Optimality Policy Iteration Algorithm stops at an nth-bias optimal policy in a finite number of iterations.*

The existence of the nth-bias optimal policy can also by proved with Theorem 9 by construction.

As shown above, the nth-bias optimality policy iteration procedure for an nth-bias optimal policy consists of n steps. Each step is based on two biases g_l and g_{l+1}, and reaches an optimal lth bias, $l = 0, 1, \ldots, n$. In addition to this procedure, we can also develop an algorithm which works roughly as follows: at each iteration k, we choose an action that maximizes (myopically) all the expected lth biases $\sum_{j \in S} p_a(i,j) g_l^{d_k}(j)$, $l = 2, \ldots, n$, and $r(i,a) + \sum_{j \in S} p_a(i,j) g_1^{d_k}(j)$ for $l = 1$. Generally, this algorithm may take fewer policies to reach an nth-bias optimal policy. We will leave the details to the readers.

At this point, the theory for the nth-bias optimization is almost complete. In the next section, we will derive the optimality equations that the nth-bias optimal policies must satisfy. Unlike for ergodic systems, for multi-chain MDPs we cannot find a set of equations that are both necessary and sufficient. In fact, in our formulation these optimality equations are not the central piece of the theory; they provide additional information but are not essential.

5 Bias Optimality Equations

We need two lemmas, one for $n = 1$ and the other for $n \neq 1$.

Lemma 4 *For any $d \in D_{n-1}$, where $n \neq 1$,*

(*a*) *if $P_d g_n^* \preceq g_{n-1}^* + g_n^*$, then $g_n^d \preceq g_n^*$, where $g_{-1}^* := 0$.*

(*b*) *If d is nth-bias optimal, then $P_d g_n^* = g_{n-1}^* + g_n^*$.*

Proof. First, we prove the lemma for $n = 0$ with $D_{-1} = D$. For part (a), by (5) we have $g_0^d = P_d g_0^d \leq P_d g_0^* \preceq g_0^*$. For part (b), since policy d is average performance optimal, we have $g_0^d = g_0^*$. Thus, from (5) we obtain $P_d g_0^* = g_0^*$.

Next, we prove the lemma for $n > 1$. Since $d \in D_{n-1}$, we have $g_k^d = g_k^*$ for all $0 \leq k \leq n-1$. By (18) we have $g_n^d = -g_{n-1}^d + P_d g_n^d \leq -g_{n-1}^d + P_d g_n^*$. Then the condition in (a) leads to $g_n^d \preceq -g_{n-1}^d + g_{n-1}^* + g_n^* = g_n^*$. For part (b), since policy d is nth-bias optimal, we have $g_{n-1}^d = g_{n-1}^*$ and $g_n^d = g_n^*$. Thus, from (18) we have $P_d g_n^* = g_{n-1}^* + g_n^*$. □

Lemma 5 *For any $d \in D_0$,*

(a) *if $r_d + P_d g_1^* \preceq g_0^* + g_1^*$, then $g_1^d \preceq g_1^*$.*

(b) *If d is bias optimal, then $r_d + P_d g_1^* = g_0^* + g_1^*$.*

Proof. For part (a), since $d \in D_0$, we have $g_0^d = g_0^*$. By (12) and the condition of part (a), we have $g_1^d = r_d + P_d g_1^d - g_0^d \leq r_d + P_d g_1^* - g_0^d \preceq g_0^* + g_1^* - g_0^d = g_1^*$. Next we prove (b). Since policy d is bias optimal, we have $g_0^d = g_0^*$ and $g_1^d = g_1^*$. Thus, from (12) we obtain that $r_d + P_d g_1^* = g_0^* + g_1^*$. □

From Lemmas 4 and 5, we conclude that the nth-bias optimal policy must belong to the set of $\{d : d \in D_{n-1}$ and $P_d g_n^* = g_{n-1}^* + g_n^*\}$ for $n > 1$ and $\{d : d \in D_0$ and $r_d + P_d g_1^* = g_0^* + g_1^*\}$ for $n = 1$.

With Theorems 1-3 and Lemmas 4-5, we are ready to derive *the bias optimality equations*. Let $g_0, g_1, \ldots, g_{n+1}$ be a set of M-dimensional vectors. The following equations are called the bias optimality equations.

$$g_0(i) = \max_{a \in E_0(i)} \left\{ \sum_{j \in S} p_a(i,j) g_0(j) \right\}, \tag{38}$$

$$g_0(i) + g_1(i) = \max_{a \in E_1(g_0)(i)} \left\{ r(i,a) + \sum_{j \in S} p_a(i,j) g_1(j) \right\}, \tag{39}$$

$$g_k(i) + g_{k+1}(i) = \max_{a \in E_{k+1}(g_0, g_1, \ldots, g_k)(i)} \left\{ \sum_{j \in S} p_a(i,j) g_{k+1}(j) \right\}, \quad k = 1, 2, \ldots, n, \tag{40}$$

where $E_0(i) := A_i$,

$$E_1(g_0)(i) \;:=\; \arg \max_{a \in E_0(i)} \left\{ \sum_{j \in S} p_a(i,j) g_0(j) \right\},$$

$$E_2(g_0, g_1)(i) := \arg \max_{a \in E_1(g_0)(i)} \left\{ r(i,a) + \sum_{j \in S} p_a(i,j) g_1(j) \right\},$$

and

$$E_{k+1}(g_0, \ldots, g_k)(i) := \arg \max_{a \in E_k(g_0, \ldots, g_{k-1})(i)} \left\{ \sum_{j \in S} p_a(i,j) g_k(j) \right\}, \quad k = 2, \ldots, n.$$

For $k \geq 1$, we set

$$E_k(g_0, \ldots, g_{k-1}) := \otimes_{i=1}^{M} E_k(g_0, \ldots, g_{k-1})(i).$$

We denote $d \in E_k(g_0, \ldots, g_{k-1})$ if $(d(1), d(2), \ldots, d(M)) \in E_k(g_0, \ldots, g_{k-1})$. We have $D = \otimes_{i \in S} A_i = E_0$.

The bias optimality equations (38) - (40) take the vector form as follows.

$$g_0 = \max_{d \in E_0} \{P_d g_0\}, \tag{41}$$

$$g_0 + g_1 = \max_{d \in E_1(g_0)} \{r_d + P_d g_1\}, \tag{42}$$

$$g_k + g_{k+1} = \max_{d \in E_{k+1}(g_0, g_1, \ldots, g_k)} \{P_d g_{k+1}\}, \quad k = 1, \ldots, n. \tag{43}$$

If $g_0, g_1, \ldots, g_k$ satisfy the first $(k+1)$ bias optimality equations, then the set $E_{k+1}(g_0, g_1, \ldots, g_k)$ contains all the policies d such that the following equations hold: $g_0 = P_d g_0$, $g_0 + g_1 = r_d + P_d g_1$, and $g_{k-1} + g_k = P_d g_k$. That is

$$\begin{aligned} & E_{k+1}(g_0, g_1, \ldots, g_k) \\ = \; & \{d \in D : g_0 = P_d g_0, \; g_0 + g_1 = r_d + P_d g_1, \; g_{l-1} + g_l = P_d g_l, \; l = 2, \ldots, k\}. \end{aligned}$$

In particular, by (35) we have

$$F_n = E_n(g_0^*, \ldots, g_{n-1}^*). \tag{44}$$

It was well known that any average performance (0th-bias) optimal policy satisfies the first bias optimality equation (38) but may not satisfy the second bias optimality equation (39). On the other hand, if a policy satisfies both (38) and (39), then it must be average performance (0th-bias) optimal. These results can be extended to the nth-bias optimality with $n \geq 1$.

Theorem 10 *$g_k^*, k = 0, 1, \ldots, n$, satisfy the first $(n+1)$ bias optimality equations, $n \geq 0$.*

Proof: From Section 4, we have proved that there exists an nth-bias optimal policy, and we denote it as d_n^*, for $n \geq 0$.

We first consider the case $n = 0$. Let d_0^* be an average performance optimal policy with average performance g_0^* and bias $g_1^{d_0^*}$. From (5), $P_{d_0^*} g_0^* = g_0^*$. We need to prove that g_0^* satisfies the first bias optimality equation (41), i.e., $P_d g_0^* \leq g_0^*$ 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 $(P_h g_0^*)(i) > g_0^*(i)$. 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 $(P_{\hat{d}} g_0^*)(i) > g_0^*(i)$ and $(P_{\hat{d}} g_0^*)(j) = g_0^*(j)$ for $j \in S - \{i\}$. Thus,

$$P_{\hat{d}} g_0^* \succeq g_0^*. \tag{45}$$

Therefore, $(P_{\hat{d}})^l g_0^* \succeq g_0^*$ for all $l \geq 1$, so $(P_{\hat{d}})^* g_0^* \geq g_0^*$ follows. Because $(P_{\hat{d}})^*(P_{\hat{d}} g_0^* - g_0^*) = 0$, by Lemma 1 we have $(P_{\hat{d}} g_0^*)(k) = g_0^*(k)$ for all recurrent states k under policy $\hat{d}$. Then the particular state i must be transient under policy $\hat{d}$. By the construction of $\hat{d}$, we have $(P_{\hat{d}})^*[r_{\hat{d}} - r_{d_0^*} + (P_{\hat{d}} - P_{d_0^*}) g_1^{d_0^*}] = 0$. (The only nonzero component of the vector in bracket is at state i which is a transient state.) Finally, by the bias difference formulas in Lemma 2 (a), we have

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

If $g_0^{\hat{d}} = g_0^*$, then $P_{\hat{d}} g_0^* = P_{\hat{d}} g_0^{\hat{d}} = g_0^{\hat{d}} = g_0^*$. This conflicts with (45). Thus, we have $g_0^{\hat{d}} \succeq g_0^*$. This is impossible because g_0^* is the optimal average performance. Therefore, the theorem holds for $n = 0$.

The case $n > 0$ can be proved in the same way by constructing counter examples, and we put it in Appendix A. □

Theorem 11 *If the M-dimensional vectors g_0, g_1, …, g_n, g_{n+1} satisfy the first $(n+2)$ bias optimality equations, then g_k is the optimal kth bias, $k = 0, 1, \ldots, n$.*

Proof: First, we prove the case $n = 0$. That is, if two vectors g_0 and g_1 satisfy the first two bias optimality equations (41) and (42) (equivalently (38) and (39)), then g_0 is the optimal average performance. Denote by d_0^* the policy that reaches the maximum in both (41) and (42). From (41), we have $(P_{d_0^*})^* g_0 = g_0$. From (42), we have $r_{d_0^*} + P_{d_0^*} g_1 = g_0 + g_1$. Pre-multiplying its both sides by $(P_{d_0^*})^*$, we get $g_0^{d_0^*} = (P_{d_0^*})^* r_{d_0^*} = (P_{d_0^*})^* g_0 = g_0$. Thus, g_0 is the average performance of d_0^*. Since g_1 only satisfies the Poisson equation $r_{d_0^*} + P_{d_0^*} g_1 = g_0 + g_1$, g_1 is the potential of policy d_0^*.

We prove that for any $d \in D$, $g_0 \geq g_0^d$. From (41) and (42), we have

(i) for all $d \in D$, $P_d g_0 \leq g_0$, and

(ii) $r_d(i) + (P_d g_1)(i) \leq g_0(i) + g_1(i)$ when $(P_d g_0)(i) = g_0(i)$ for some $i \in S$.

Then by Theorem 1 for the "$\leq$" case, we know $g_0^d \leq g_0$. Thus, g_0 is the optimal average performance and d_0^* is average performance optimal.

The proof for the case $n > 0$ are similar, and we put it in Appendix B. □

Theorem 10 and Theorem 11 provide a necessary and a sufficient condition, respectively, for the nth-bias optimal policies. Because the nth-bias optimal policy exists, the solution to the first $(n + 2)$ bias optimality equations also exists. From Theorem 11, if $(g_0, g_1, \ldots, g_{n+1})$ is one of the solutions of the first $(n + 2)$ bias optimality equations, then $g_0, g_1, \ldots, g_n$ are unique (the optimal values). But g_{n+1} is only the $(n+1)$th potential of a nth-bias optimal policy, therefore it may not be uniquely determined by these equations.

From the definition, $D \supseteq D_0 \supseteq D_1 \supseteq \cdots \supseteq D_{n-1} \supseteq D_n \supseteq \cdots$. That is, as n increases, the set D_n shrinks. Veinott [19] proved that if a policy is $(M-m+1)$th bias optimal, where M is the number of states and m the number of recurrent classes, then it is nth bias optimal for all $n \geq 0$. That is, $D_{M-m+1} = D_{M-m+2} = \cdots = D_n = \cdots$, for all $n \geq M - m + 1$.

6 Discussions

We first review some related results in [2, 15, 16, 18, 19]. A policy $d^* \in D$ is said to be *n-discount optimal* for some integer $n \geq -1$ if

$$\liminf_{\lambda \uparrow 1} (1-\lambda)^{-n} [v_\lambda^{d^*} - v_\lambda^d] \geq 0, \quad \text{for all } \pi \in D,$$

where $v_\lambda^d(s) = E^d \sum_{k=0}^{\infty} [\lambda^k r(X_k, d(X_k)) | X_0 = s]$. In a Markov decision process with transition matrix P_d and reward r_d, v_λ^d can be expanded in the Laurent series,

$$v_\lambda^d = (1+\rho) \left[\frac{y_{-1}^d}{\rho} + y_0^d + \sum_{k=1}^{\infty} \rho^k y_k^d \right],$$

where $\rho = \frac{1-\lambda}{\lambda}$ and y_k, $k = -1, 0, \ldots$, denote the coefficients of the Laurent series expansion. And y_k, $k = -1, 0, \ldots$ satisfy the following equations.

$$\begin{aligned} (I - P_d) y_{-1} &= 0, \\ y_{-1} + (I - P_d) y_0 &= r_d, \\ &\vdots \\ y_{n-1} + (I - P_d) y_n &= 0, \ n = 1, 2, \ldots. \end{aligned}$$

These equations correspond to (5) and Poisson equations (12) and (18), respectively, and we can prove that $g_{k+1}^d = y_k^d$, for $k \geq -1$.

An *n-discount optimal policy* maximizes the first nth derivative of v_λ^d with respect to ρ. Therefore, an nth-bias optimal policy is an $(n-1)$-discount optimal policy, and vise versa.

Historically, Veinott published his pioneering work in as early as 1969 [18] on the n-discount optimality, which laid the foundation for the n-discount optimality theory. Later in his award winning book [16], Puterman refined the method on the n-discount optimality. He derived the n-discount optimality equations and provides a policy iteration algorithm for the n-discount optimal policies.

We now briefly state Puterman's algorithm. To simplify presentation, for $d \in D$ we define

$$r_d^n = \begin{cases} r_d, & n = 0; \\ 0, & n = -1, 1, 2, 3, \ldots. \end{cases}$$

The N-discount Optimality Policy Iteration Algorithm in [16] is

1. Set $n = -1$, $D_{-1} = D$, $y^*_{-2} = 0$, $k = 0$, and select a $d_0 \in D$.

2. (Policy evaluation) Obtain y^k_n and y^k_{n+1} by solving

$$\begin{aligned} r^n_{d_k} - y^*_{n-1} + (P_{d_k} - I)y_n &= 0, \\ r^{n+1}_{d_k} - y_n + (P_{d_k} - I)y_{n+1} &= 0 \end{aligned}$$

   subject to $(P_{d_k})^* y_{n+1} = 0$.

3. (Policy improvement)

   (a) (n-improvement) Choose

$$d_{k+1} \in \arg\max_{d \in D_n} \{r^n_d + P_d y^k_n\},$$

   setting $d_{k+1}(s) = d_k(s)$ if possible. If $d_{k+1} = d_k$ go to (b); otherwise increase k by 1 and return to step 2.

   (b) ($(n+1)$-improvement) Choose

$$d_{k+1} \in \arg\max_{d \in D_n} \{r^{n+1}_d + P_d y^k_{n+1}\},$$

   setting $d_{k+1}(s) = d_k(s)$ if possible. If $d_{k+1} = d_k$, go to step 4; otherwise increase k by 1 and return to step 2.

4. Set

$$D_{n+1} = \arg\max_{d \in D_n} \{r^{n+1}_d + P_d y^k_{n+1}\}.$$

   If D_{n+1} contains a single decision rule or $n = N$, stop. Otherwise, set $y^*_n = y^k_n$, increase n by 1, set $k = 0$, $d_0 = d_k$ and return to step 2.

Puterman's algorithm iterates between two phases in step 3(a) and 3(b). With 3(a), y_n improves, and the algorithm keeps implementing 3(a) until no improvement can be achieved then shifts to 3(b) to improve y_{n+1}.

Compared with the n-discount optimality theory, our approach is completely independent of the discounted MDP formulation. It does not depend on Laurent series

expansion. The approach is based on the performance difference formula. It is simpler, more direct, and more intuitive. The n-discount optimality is equivalent to the $(n+1)$th bias optimality defined in this paper. However, because of its complicated theory, it does not gain its deserved popularity. We wish our simpler approach may help to popularize these results. Next, based on the performance difference formulas, we derived policy iteration algorithms that are generally more efficient than the two-phase algorithms in the literature. Finally, our approach fits the recently established framework of sensitivity-based optimization; the new sensitivity-based view may lead to new research directions, such the bias derivatives, on-line optimization, and potential aggregations.

7 Conclusion

With the nth-bias difference formulas, we have developed an optimization theory for MDPs that covers a complete spectrum from average performance optimality, bias optimality, to higher-order bias optimality. This approach is intuitively clear. Policy iteration algorithms can be easily developed with this approach.

The new approach fits the recently developed sensitivity-based learning and optimization framework for discrete even dynamic systems [3, 4, 6] and provides some new insights. For example, we first derive the performance difference formulas for problems that do not fit the standard MDP framework and develop policy iteration algorithms, see [7] for some examples. Also, sample-path-based estimation algorithms, or reinforcement learning type of algorithms for nth biases can be developed with (20) without knowing the state transition probability matrix P_d [14]. With such algorithms on-line optimization methods can be developed. In addition, if a policy space contains continuous parameters, the derivatives of the nth biases with respect to the parameters can be easily derived. These derivative formulas are similar to the difference formulas (see [3, 4, 5]). Then the gradient-based optimization approaches can be developed. Finally, with the sensitivity-based point of view, potential aggregation can be implemented to save computation by utilizing the special feature of a

particular problem, see [6].

Further research also includes to extend the results to continuous-time MDPs and MDPs with countable state spaces, and/or compact action sets.

Appendix A

The Proof of Theorem 10 for the Case $n > 0$.

(a) We prove the case $n = 1$. Let d_1^* be a bias optimal policy with average performance g_0^* and (1st) bias g_1^*. We have shown that g_0^* satisfies the first bias optimality equation (41) and need to prove that g_0^* and g_1^* satisfy the second bias optimality equation (42). By Poisson equation (12), we have $g_0^* + g_1^* = r_{d_1^*} + P_{d_1^*} g_1^*$. Now we need to prove

$$r_d + P_d g_1^* \leq g_0^* + g_1^* \quad \text{for all } d \in D \text{ satisfying } P_d g_0^* = g_0^*. \tag{46}$$

Assume that (46) does not hold. Then there exists a policy h and some state $i \in S$ such that $(P_h g_0^*)(i) = g_0^*(i)$ and

$$r_h(i) + (P_h g_1^*)(i) > g_0^*(i) + g_1^*(i).$$

Based on this, we can construct another policy $\hat{d}$ by setting $\hat{d}(j) = d_1^*(j)$ for all $j \in S - \{i\}$ and $\hat{d}(i) = h(i)$. Consequently, $r_{\hat{d}}(j) = r_{d_1^*}(j)$ for all $j \in S - \{i\}$ and $r_{\hat{d}}(i) = r_h(i)$. Therefore, by construction we have $P_{\hat{d}} g_0^* = g_0^*$, $r_{\hat{d}}(j) + (P_{\hat{d}} g_1^*)(j) = r_{d_1^*}(j) + (P_{d_1^*} g_1^*)(j)$ for all $j \in S - \{i\}$ and $r_{\hat{d}}(i) + (P_{\hat{d}} g_1^*)(i) > r_{d_1^*}(i) + (P_{d_1^*} g_1^*)(i)$. From Theorem 1, we have $g_0^{\hat{d}} \geq g_0^*$. Because g_0^* is the optimal average performance, so we must have $g_0^{\hat{d}} = g_0^*$. Next, by Theorem 4, we must have $g_1^{\hat{d}} \succeq g_1^*$. This conflicts to the fact that g_1^* is the optimal bias. Thus the theorem holds for $n = 1$.

(b) Now we prove the general case $n > 1$ by induction. That is, if $g_0^*, \ldots, g_k^*$ of a kth-bias optimal policy satisfy the first $(k+1)$ bias optimality equations, then $g_0^*, \ldots, g_{k+1}^*$ of a $(k+1)$th-bias optimal policy (denoted as d_{k+1}^*) must satisfy the first $(k+2)$ bias optimality equations, $k \geq 1$. By assumption, we only need to check the $(k+2)$th bias optimality equation. In other words, we need to prove that

$$P_d g_{k+1}^* \leq g_k^* + g_{k+1}^*, \quad \text{for all } d \in E_{k+1}(g_0^*, g_1^*, \ldots, g_k^*). \tag{47}$$

Suppose that (47) does not hold. Then there must exist an $h \in E_{k+1}(g_0^*, g_1^*, \ldots, g_k^*)$ and some state $i \in S$ such that $(P_h g_k^*)(i) = g_{k-1}^*(i) + g_k^*(i) = (P_{d_{k+1}^*} g_k^*)(i)$ for $k > 1$, $r_h(i) + (P_h g_k^*)(i) = r_{d_{k+1}^*}(i) + (P_{d_{k+1}^*} g_k^*)(i)$ for $k = 1$, and

$$(P_h g_{k+1}^*)(i) > g_k^*(i) + g_{k+1}^*(i) = (P_{d_{k+1}^*} g_{k+1}^*)(i).$$

Again, we can construct another policy $\hat{d}$ by setting $\hat{d}(j) = d_{k+1}^*(j)$ for all $j \in S - \{i\}$ and $\hat{d}(i) = h(i)$. Consequently, $r_{\hat{d}}(j) = r_{d_{k+1}^*}(j)$ for all $j \in S - \{i\}$ and $r_{\hat{d}}(i) = r_h(i)$. Because $h \in E_{k+1}(g_0^*, \ldots, g_k^*) \subseteq E_k(g_0^*, \ldots, g_{k-1}^*) = F_k$ (cf. (44)), by construction, the policy $\hat{d}$ can be viewed as constructed from d_{k+1}^* by (36) and (37) for $k > 1$ and by (32) and (33) for $k = 1$. Therefore, $g_k^{\hat{d}} \geq g_k^*$ follows from Theorem 8 (b) for $k > 1$ and from Theorem 6 (b) for $k = 1$. Because d_{k+1}^* is $(k+1)$th-bias optimal, we have $g_k^{\hat{d}} = g_k^*$. Then from Theorem 8 (c) for $k > 1$ and Theorem 6 (c) for $k = 1$, we have $g_{k+1}^{\hat{d}} \succeq g_{k+1}^*$. This is impossible because g_{k+1}^* is the optimal $(k+1)$th bias. Thus $g_0^*, \ldots, g_{k+1}^*$ satisfy the first $(k+2)$ bias optimality equations. We complete the proof of the theorem. □

Appendix B

The Proof of Theorem 11 for the Case $n > 0$.

(a) We prove the case $n = 1$. That is, if three vectors g_0, g_1 and g_2 satisfy the first three bias optimality equations, then g_0 is the optimal average performance and g_1 is the optimal 1st bias. Denote by d_1^* the policy that reaches the maximum in the first three equations. We have shown that g_0 is the average performance of d_1^*. Pre-multiplying both sides of $P_{d_1^*} g_2 = g_1 + g_2$ by $(P_{d_1^*})^*$ we get $P_{d_1^*} g_1 = 0$. By $r_{d_1^*} + P_{d_1^*} g_1 = g_0 + g_1$, we have $g_1 = [I - P_{d_1^*} + (P_{d_1^*})^*]^{-1}(r_{d_1^*} - g_0) = g_1^{d_1^*}$. That is, g_1 is the 1st bias of d_1^*. Since g_2 only satisfies the Poisson equation $P_{d_1^*} g_2 = g_1 + g_2$, g_2 is the 2nd potential of policy d_1^*.

Since we have proved that $g_0 = g_0^*$, now we just need to prove $g_1 = g_1^*$. For any $d \in D_0$, we know that $d \in F_1$. By the second and the third bias optimality equations, we have

(i) $r_d + P_d g_1^{d_1^*} \leq r_{d_1^*} + P_{d_1^*} g_1^{d_1^*} = g_0 + g_1$, and

(ii) $(P_d g_2^{d_1^*})(i) \le g_1(i) + g_2(i)$ when $r_d(i) + (P_d g_1^{d_1^*})(i) = g_0(i) + g_1(i)$ for some $i \in S$.

By the "In addition" part of Theorem 2, we have $g_1^d \le g_1$ for all $d \in D_0$. Thus, g_1 is the optimal (1st) bias and d_1^* is (1st) bias optimal.

(b) Now we prove the general case $n > 1$ by induction. Assume that the theorem holds for the case of $(n-1)$ (we have proved the case $n-1=1$). That is, if the vectors $g_0, \ldots, g_n$ satisfy the first $(n+1)$ bias optimality equations, then g_k is the optimal kth bias, $k = 0, 1, \ldots, n-1$. We wish to prove that the theorem holds for the case of n. That is, if the vectors $g_0, \ldots, g_{n+1}$ satisfy the first $(n+2)$ bias optimality equations, then $g_k = g_k^*$, $0 \le k \le n$. Again, we denote by d_n^* the policy that reaches the maximum in the first $(n+2)$ bias optimality equations. Then by the assumption of induction, we have proved that $g_0, \ldots, g_{n-1}$ are the 0th to the $(n-1)$th biases of d_n^*, respectively. For g_n, by the $(n+1)$th and $(n+2)$th bias optimality equations we have $g_n = -[I - P_{d_n^*} + (P_{d_n^*})^*]^{-1} g_{n-1} = g_n^{d_n^*}$. That is, g_n is the nth bias of d_n^*. Since g_{n+1} only satisfies the Poisson equation $P_{d_n^*} g_{n+1} = g_n + g_{n+1}$, g_{n+1} is the $(n+1)$th potential of policy d_n^*.

Again, from the assumption, we know $g_k = g_k^*$, $0 \le k \le n-1$. Now we just need to prove that $g_n = g_n^*$. For any $d \in D_{n-1}$, we know that $d \in F_n$. By the $(n+1)$th and the $(n+2)$th bias optimality equations, we have

(i) $P_d g_n^{d_n^*} \le P_{d_n^*} g_n^{d_n^*} = g_{n-1} + g_n$, and

(ii) $(P_d g_{n+1}^{d_n^*})(i) \le g_n(i) + g_{n+1}(i)$ when $(P_d g_n^{d_n^*})(i) = g_{n-1}(i) + g_n(i)$ for some $i \in S$.

By the "In addition" part of Theorem 3, we have $g_n^d \le g_n$ for all $d \in D_{n-1}$. Thus, g_n is the optimal nth bias and d_n^* is nth-bias optimal.

□

References

[1] Dimitri P. Bertsekas, *Dynamic Programming and optimal control*, volume I and II. Athena Scientific, Belmont, MA, 1995

[2] David Blackwell, "Discrete Dynamic Programming," *The Annals of Mathematical Statistics,* Vol. 33, No. 2 (June 1962), 719-726.

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

[4] Xi-Ren Cao, "From Perturbation Analysis to Markov Decision Processes and Reinforcement Learning," *Discrete Event Dynamic Systems: Theory and Applications*, 13 (2003), 9-39.

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

[6] Xi-Ren Cao, "The Potential Structure of Sample Paths and Performance Sensitivities of Markov Systems," *IEEE Transactions on Automatic Control*, Vol. 49, No. 12, pp. 2129-2142, December 2004.

[7] X. R. Cao, "Basic Ideas for Event-Based Optimization of Markov Systems", Discrete Event Dynamic Systems: Theory and Applications, Vol. 15, pp. 169-197, 2005.

[8] K. L. Chung, *Markov Chains with stationary Transition Probabilities*, Springer-Verlag, New York, 1960.

[9] Eugene A. Feinberg and Adam Shwartz, *Handbook of Markov Decision Processes: Methods and Application*, Kluwer Academic Publishers, 2002.

[10] L. C. M. Kallenberg, *Linear Programming and Finite Markovian Control Problems*, Mathematisch Centrum, Amsterdam, 1983.

[11] John G. Kemeny and J. Laurie Snell, *Finite Markov Chains*, D. Van Nostrand Company, Inc. New York, 1960.

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

[13] Mark E. Lewis and Martin L. Puterman, Bias Optimality, *The Handbook of Markov Decision Processes: Methods and Applications,* Edited by Eugene Feinberg and Adam Shwartz. Kluwer, 89-111. 2001.

[14] Sridhar Mahadevan, "Sensitive Discount Optimality: Unifying Discounted and Average Reward Reinforcement Learning," ICML 1996: 328-336.

[15] Bruce L. Miller and Arthur F. Veinott, "Discrete Dynamic Programming with a Small Interest rate," *The Annals of Mathematical Statistics,* Vol. 40, No. 2 (April 1969), 366-370.

[16] Martin L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, New York: Wiley, 1994.

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

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

[19] Arthur F. Veinott, Markov Decision Chains, in G. B. Dantzig and B. C. Eaves, *Studies in Optimization*, MAA Studies in Mathematics 10 (1974),124-159.