A Probabilistic Analysis of Bias Optimality in Unichain Markov Decision Processes[1][2][3]

Mark E. Lewis
Department of Industrial and Operations Engineering
University of Michigan, 1205 Beal Avenue, Ann Arbor, MI 48109-2117
melewis@engin.umich.edu
734-763-0519

Martin L. Puterman
Faculty of Commerce and Business Administration
University of British Columbia, 2053 Main Mall, Vancouver, BC Canada V6T 1Z2
marty@coe.ubc.ca
604-822-8388

submitted August 12, 1999, refereed revision May 17, 2000, accepted August 8, 2000

[1] *AMS Subject Classifications:* **primary-90C40:** Markov decision processes, **secondary-60K25** queueing theory

[2] *IAOR Subject Classifications:* **primary-3160:** Markov processes, **secondary-3390:** dynamic programming: theory

[3] This work is partially supported by NSF grant: DMI-9908321

Abstract

This paper focuses on bias optimality in unichain, finite state and action space Markov Decision Processes. Using relative value functions, we present new methods for evaluating optimal bias. This leads to a probabilistic analysis which transforms the original reward problem into a minimum average cost problem. The result is an explanation of **how** and **why** bias implicitly discounts future rewards.

1 Introduction

Bias optimality has previously been regarded as a theoretical concept in Markov Decision Process (MDP) theory. It was viewed as one of many optimality criteria that is more sensitive than long-run average optimality, but its usefulness in application had not been considered. We show through probabilistic arguments that bias can be used to make decisions rather easily. Furthermore, we find numerous similarities between finding bias optimal policies and finding average optimal policies. This relates bias optimality to the vast literature on average optimality.

Recently, Haviv and Puterman [3] and Lewis, et. al [5] studied the usefulness of bias optimality to distinguish between multiple gain optimal policies in controlled queueing systems. Lewis and Puterman [6] then showed that in the Haviv-Puterman model, when rewards are received impacts the bias. In particular, given two gain optimal control levels, when rewards are received at arrival the larger control level is bias optimal, while when rewards are received upon departure the reverse holds. This suggests that bias may implicitly discount rewards received later.

Discount and average optimality have been considered extensively in the literature, cf. Puterman [11]. In contrast, bias optimality has received little direct attention. In fact, to our knowledge, in addition to the previously mentioned papers ([3], [5], [6]), the use of bias to distinguish between gain optimal policies has only appeared in a short section of an expository chapter by Veinott [14]. Methods of computing optimal bias were considered for the finite state and action space case by Denardo [1] and on countable state and compact action spaces by Mann [9].

Let v_λ^π represent the total discounted reward of a policy π. A policy π^* is $0-$discount optimal if $\lim_{\lambda\uparrow 1}(v_\lambda^{\pi^*} - v_\lambda^\pi) \geq 0$ for all policies π. Veinott [13] (see Theorem 10.1.6 of Puterman [11]) asserts that in the finite state and action space case bias optimality is **equivalent** to $0-$discount optimality. Furthermore, for general state space models, Hernandez-Lerma and Lasserre [4] (see Chapter 10) discuss the equivalence of bias optimality with opportunity cost optimality, weakly overtaking optimality, and optimality according to Dutta's criterion under some light conditions. There is also a vast literature on sensitive optimality that indirectly addresses bias optimality (cf. Veinott [13]). However, these works do not give an intuitive explanation for *what* the bias-based decision-maker prefers and *why*. This paper addresses these points.

2 Model Formulation

Our notation and formulation follows Puterman [11]. Consider an infinite horizon, discrete time, Markov decision process (MDP) with finite state space S. Let A_s be the finite set of actions available to a decision-maker when in state s. If the decision-maker chooses action $a \in A_s$ when in

state s, an immediate (expected) reward of $r(s,a)$ is received and the system enters state j with probability $P(j|s,a)$. Let $A = \times_{s\in S} A_s$ be the action space. A deterministic, Markovian *decision rule* d maps S into A and specifies which action the decision-maker will take when the system is in state s. We will often use r_d to denote the vector of expected rewards when using decision rule d and either $r_d(s)$ or $r(s,d(s))$ to denote an element of that vector. A sequence of decision rules $\pi = \{d_1, d_2, \ldots\}$ is called a deterministic, Markovian *policy* and specifies the decision-maker's actions for each state, for all time. We say that a policy is *stationary* if it uses the same decision rule at each decision epoch. It should cause no confusion that we also use d to denote such a policy which always uses decision rule d. The set of such policies is denoted D^∞. Each policy generates a sequence of random variables $\{(X_n, Y_n); n = 1, 2, \ldots\}$ where X_n denotes the state of the system and Y_n denotes the action chosen by policy π at decision epoch n given X_n. Unless otherwise noted, we assume the Markov decision process is *unichain*. That is, all stationary policies generate Markov chains that consist of a single ergodic class and possibly some transient states. We now formalize the definitions of gain and bias.

Definition 1 *The long-run average reward or* **gain** *of a policy π given that the system starts in state $s \in S$ denoted $g_\pi(s)$ is given by*

$$g_\pi(s) \quad = \quad \lim_{N\to\infty} \mathbb{E}_s^\pi \left(\frac{1}{N} \sum_{n=0}^{N-1} r(X_n, Y_n) \right).$$

where the expectation is conditioned on the state at time zero and taken with respect to the probability measure generated by π. Furthermore, a policy, π^*, is called **gain optimal** *if $g_{\pi^*}(s) \geq g_\pi(s)$ for all $s \in S$, for all π.*

Definition 2 *The* **bias** *of a stationary policy d, given that the system started in state s, denoted $h_d(s)$, is defined to be*

$$h_d(s) \quad = \quad \sum_{n=0}^{\infty} \mathbb{E}_s^d [r(X_n, d(X_n)) - g_d(X_n)]. \tag{1}$$

We say that a policy, d^* is **bias optimal** *if it is* **gain optimal**, *and $h_{d^*}(s) \geq h_d(s)$ for all $s \in S$, for all gain optimal d.*

If the Markov chain generated by d is aperiodic this sum is convergent, otherwise; replace the above sum with sums in the Cesaro sense. It is well-known that in finite state and action space MDP's stationary gain and bias optimal policies exist.

For a stationary policy, d we call $r_d(s) - g_d(s)$ the *excess reward* of d. If one defines a new system in which the excess reward replaces the reward function, then the bias is the expected (finite) total

reward in the modified system. Alternatively, the bias represents the expected difference in total reward under policy d between two different initial conditions; when the process begins in state s and when the process begins with the state selected according to the probability distribution defined by the s^{th} row of the *limiting matrix* $P_d^* = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} P_d^i$. Since we assume that the process under d is unichain, this initial distribution is the stationary distribution of the chain. When the process is multichain, the distributions specified by the rows of P_d^* may vary with the initial state. For other interpretations of the bias (see Puterman [11] Chapter 8).

3 Computing the Gain and Bias

We begin with some well-known results. Any results not specifically referenced may be found in Puterman [11]. Since P_d is unichain,

1. g_d is a constant, which we express as $g_d 1$ where 1 is a vector of 1's of dimension $|S|$.

2. If (g, h) satisfies

$$h = r_d - g1 + P_d h, \tag{2}$$

   $g = g_d$ and h is unique up to a constant. We refer to (2) as the *average evaluation equation* (AEE)

3. (g_d, h_d) is the unique solution of equation (2) and the additional condition $P_d^* h = 0$.

4. Let the first passage time to a recurrent state α be denoted τ_α, that is, $\tau_\alpha = \min_{n>0}\{X_n = \alpha\}$. Then g_d satisfies

$$g_d = (P_d^* r_d)(s) = \frac{\mathbb{E}_\alpha^d \sum_{n=0}^{\tau_\alpha - 1} r(X_n, d(Y_n))}{\mathbb{E}_\alpha^d \tau_\alpha} \tag{3}$$

   for any $s \in S$.

With these observations in mind, we have the following definition.

Definition 3 *Let $d \in D^\infty$ be a fixed stationary policy. For each solution to the average evaluation equation (g_d, h), the constant difference between h and the bias of d, $c_d(h)$ is called the* **bias constant** *associated with h.*

This allows for a probabilistic interpretation of the bias. Let

$$h_\alpha^d(s) \quad = \quad \mathbb{E}_s^d \left(\sum_{n=0}^{\tau_\alpha - 1} [r(X_n, Y_n) - g_d] \right). \tag{4}$$

Note that

$$h_\alpha^d(s) = r(s,d(s)) - g_d + \mathbb{E}^d\left(\sum_{n=1}^{\tau_\alpha - 1}[r(X_n,Y_n) - g_d]\,\middle|\, X_0 = s\right) = r(s,d(s)) - g_d + (P_d h_\alpha^d)(s).$$

Hence, (g_d, h_α^d) satisfies (2) for d and represents the total excess reward earned until the process enters state α given that the process started in state s and uses policy d. Also note, by (3) that $h_\alpha^d(\alpha) = 0$.

We call the unique function h_d^{rv} which satisfies (2) together with $h_d^{rv}(s) = 0$ the **relative value function** of d with reference state s. Thus, there is a relative value function associated with each state s. By the previous argument, for each recurrent state α we have h_α^d is the relative value function associated with α. The fact that (g_d, h_α^d) satisfies the AEE was first shown in Derman and Veinott [2].

Since for a fixed policy d, the relative value functions and the bias satisfy the AEE, by our previous observations, they must differ by a constant. It is then simple to show there is a close relationship between the two. For example one may use the results on *Poisson's equation* discussed in Derman and Veinott [2] or more recently in Makowski and Shwartz [8] to obtain in the unichain case, for a stationary policy d that the bias of d is given by

$$h_d(s) = h_d^{rv(\alpha)}(s) - (P_d^* h_d^{rv(\alpha)})(s) = h_d^{rv(\alpha)}(s) - \frac{\mathbb{E}_\alpha^d \sum_{n=0}^{\tau_\alpha - 1} h_d^{rv(\alpha)}(X_n)}{\mathbb{E}_\alpha^d\, \tau_\alpha}. \tag{5}$$

Hence, in the same manner that we can compute the gain of a policy by (3), we can compute the bias using a relative value function in place of the reward function. In the next two sections, we show how this can be used to find bias optimal policies.

4 The Average Optimality Equation

Since the state and action space are finite, computation of gain optimal policies reduces to solving the *average optimality equation* (AOE)

$$h \;=\; \max_{d\in D}\{r_d - g1 + P_d h\} \tag{6}$$

for g and h. To begin our analysis of the average optimality equation, we consider a special case of a result of Schweitzer and Federgruen [12]. In essence, the result states that solutions of the AOE must differ by a constant just as they do for the AEE. We include a simple proof for this result to keep this paper self-contained. In addition to being useful for establishing several results below, we use it to show that the average optimality equation does not determine the set of gain

optimal solutions in unichain average reward models. Thus, the optimal gain and optimal policies are determined by the AOE, but the solution of the AOE is not unique. Note that if all states are recurrent under every policy, the gain and the bias of every optimal policy satisfies the AOE (cf. Lewis et. al [5]).

Proposition 1 *Suppose all stationary policies are unichain and let* (g_1, h_1) *and* (g_2, h_2) *be solutions to the AOE. Then* $g_1 = g_2$ *and*

$$h_1 \quad = \quad h_2 + c1 \tag{7}$$

for some constant c. In particular, if $h_1 = h^*$ *is the optimal bias, then*

$$h^* \quad = \quad h_2 + c(h_2)1. \tag{8}$$

Proof. Suppose (g_1, h_1) is a solution of (6). Let $G(h_1)$ denote the set of decision rules that attain the maximum in (6) for g_1 and h_1. Then there exists a $\delta \in G(h_1)$ for which

$$h_1 = \max_{d \in D}\{r_d - g1 + P_d h_1\} = r_\delta - g_1 1 + P_\delta h_1 \tag{9}$$

Since P_δ is unichain, the second equality of (9) uniquely determines g_1 and determines h_1 up to a constant. Since (g_1, h^*) also satisfies (6), (8) follows. The case for a general solution of (6) is analogous. ∎

Note that this result does not require that S be finite, only that the gain is constant.

Definition 4 *We refer to* $c(h_2)$ *as the* **optimal bias constant** *associated with* h_2*.*

The following example uses the above result to show that there are average optimal policies that do not yield solutions to the AOE.

Example 1

Suppose $S = \{1, 2\}$, $A_1 = \{a, b\}$ and $A_2 = \{c\}$, $r(1, a) = 2$, $r(1, b) = 3$ $r(2, c) = 1$ and $p(2|1, a) = p(2|1, b) = p(2|2, c) = 1$. Let δ be the decision rule which chooses action a in state 1 and let γ be the decision rule which chooses action b in state 1. Clearly this model is unichain and $g_\delta = g_\gamma = 1$, $h_\delta(1) = 1$, $h_\gamma(1) = 2$, $h_\delta(2) = h_\gamma(2) = 0$. Since h_δ and h_γ do *not* differ by a constant, it follows from Proposition 1, that (g_δ, h_δ) and (g_γ, h_γ) cannot both satisfy the optimality equation. ∎

Since solutions to the AOE differ by a constant, for all $d, d' \in G(h)$ we have $h_d^{rv(\alpha)} = h_{d'}^{rv(\alpha)}$. Hence, in the sequel we suppress the dependence on d.

5 The Bias Optimality Equation

Suppose (g^*, h) satisfies the AOE and in addition h satisfies

$$w = \max_{d \in G}\{-h + P_d w\} \tag{10}$$

for some vector w, where G denotes the set of decision rules that attain the maximum in the AOE (6) for g^* and h, then h is the optimal bias. We refer to the combined set (10) and the AOE as the *bias optimality equations* (BOE).

Upon substituting (8) into (10) where h_2 is the relative value function with reference state α, we have the following result,

Proposition 2 *Suppose $h^{rv(\alpha)}$ is a relative value function with reference state α such that $(g^*, h^{rv(\alpha)})$ is a solution to the AOE. The BOE (10) can be rewritten*

$$w = max_{d \in G}\{-h^{rv(\alpha)} - c(h^{rv(\alpha)})1 + P_d w\}. \tag{11}$$

Observe that (11) has exactly the same form as the AOE (6). That is to say, setting $r_d = -h^{rv(\alpha)}$ and $g = c(h^{rv})1$ we have again the AOE. Thus, in a unichain model, (11) uniquely determines $c(h^{rv})$ and determines w up to a constant. Furthermore, all solution methods and theory for the AOE apply directly in this case. In particular, (11) can be solved by value iteration or policy iteration.

Alternatively, as in the case of the AEE, if (g_d, h) satisfy the AOE and $P_d^* h = 0$ then h is the optimal bias. Neglecting the trivial case $r_d(s) = 0$ for all $s \in S$ and all $d \in D$, it is interesting to note that since P_d^* is positive on the recurrent class generated by d, the optimal bias must have both positive and negative elements. We will show in the examples that follow that we can take advantage of this fact.

Since $h^{rv(\alpha)}$ is independent of d it follows from (5) that solving for the policy with maximum bias reduces to finding the policy that achieves the maximum bias constant, say c^*. That is,

$$c^* 1 = \max_{d \in G}\{-P_d^* h^{rv(\alpha)}\} = -\min_{d \in G}\{P_d^* h^{rv(\alpha)}\} \tag{12}$$

Thus, under the assumption that there exists a state α that is recurrent for all decision rules in G we can compute the optimal bias by

$$c^* = -\min_{d \in G}\left(\frac{\mathbb{E}_\alpha^d \sum_{n=0}^{\tau_\alpha - 1} h^{rv(\alpha)}(X_n)}{\mathbb{E}_\alpha^d \tau_\alpha}\right) \tag{13}$$

Since we are minimizing, $h^{rv(\alpha)}$ can be interpreted as a cost function. **Thus, finding a bias optimal policy corresponds to solving a modified minimum average cost problem**. We emphasize the importance of these observations in the following example.

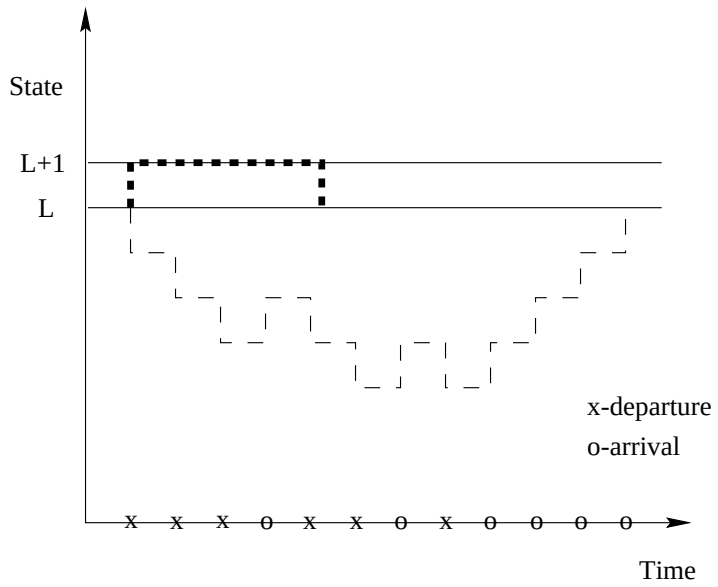

Figure 1: We would like to compare control limits L and $L+1$.

Example 2

Consider an admission controlled $M/M/1/k$ queueing system with Poisson arrival rate λ and exponential service rate μ. Assume that a holding cost is accrued at rate $f(s)$ while there are s customers in the system. If admitted the job enters the queue and the decision-maker immediately receives reward R. Rejected customers are lost. Assume that the cost is convex, increasing in s, and $f(0) = 0$. Furthermore, assume that we discretize the model by applying the standard uniformization technique in Lippman [7]. Without loss of generality assume that the uniformization constant is $\lambda + \mu = 1$. Since rejecting all customers yields $g = 0$, we assume customers are accepted in state zero. This example was previously considered in Haviv and Puterman [3] where it was shown that bias distinguishes between gain optimal policies.

Consider the set of policies T^∞ that accept customers until the number of customers in the system reaches some control limit $L > 0$ and rejects customers for all $s \geq L$. Let L denote the stationary policy that uses control limit L. The following lemma asserts that it is better to start with fewer customers in the system. The proof is similar to the analogous result in Lewis et. al [5] and is omitted. We will use this result in the sample path arguments to follow.

Lemma 1 *Suppose* (g^*, h) *satisfy the average optimality equation. For* $s \in S$, $h(s+1) < h(s)$.

Haviv and Puterman [3] show using alternative methods that if there are two gain optimal control limits, they occur consecutively, and only the higher one is bias optimal. Let L and $L+1$ be gain optimal control limits. Let c_L be the bias constant for control limit L. Similarly for c_{L+1}. Let $\alpha = L$ be the reference state. Clearly, L is recurrent for both policies. Suppose we start two

processes on the same probability space in state L. Process 1 uses control limit L and process 2 uses control limit $L+1$. It is easy to see if the first event is a departure, both processes move to state $L-1$. Since the policies are the same on all states below L, the costs accrued (measured by $h^{rv(\alpha)}$) until the return to state $L-1$ are the same. This is denoted by the lighter dashed line of Figure 1. Further, if the first event is an arrival, Process 1 rejects arriving customers and thus, immediately returns to L accruing cost $h^{rv(\alpha)}(L) = 0$ on the cycle. Process 2 accepts the arriving customer, accrues cost $h^{rv(\alpha)}(L)$, and moves to state $L+1$. The process then accrues cost $h^{rv(\alpha)}(L+1)$ a geometric number of times (with parameter μ) for false arrivals in the uniformization (recall $\lambda + \mu = 1$) before returning to state L. This is denoted by the bold line in Figure 1.

Hence, while the total cost is the same for each policy when a departure is the first event, when an arrival occurs first, process 2 accrues $h^{rv(\alpha)}(L+1)$ for each extra decision epoch in the cycle. From Lemma 1, $h^{rv(\alpha)}(L+1) < h^{rv(\alpha)}(s)$ for all $s \leq L$. Thus, each extra decision epoch in process 2 can only stand to decrease the average cost. That is to say, $c_{L+1} > c_L$, and the bias of control limit $L+1$ is larger than that of L. ∎

The previous example shows that by an astute choice of the reference state a simple sample path argument can be used to show the usefulness of bias in distinguishing among gain optimal policies. In the next section we discuss how bias implicitly discounts rewards received late in the cycle using the relative value functions.

6 Bias and Implicit Discounting

Neither the interpretation of the bias as the total excess reward before reaching stationarity nor as the average cost over a cycle give a complete picture. If either were so, one might conjecture that in a similar manner to total or average reward models, a decision-maker using bias as the optimality criterion would be indifferent to when in the cycle rewards were received. Suppose we consider Example 2 except that rewards are received upon service completion instead of upon acceptance to the system. Using the bias optimality equation (10), [6] showed that if there are two gain optimal control limits, the **lower** control limit is bias optimal. Thus, by changing when rewards are received, we have changed which control limit is preferred.

Example 3

Consider the $M/M/1/k$ queueing system of Example 2 except that the rewards are received upon service completion instead of acceptance. Assume L and $L+1$ are gain optimal control limits and again let L be the reference state. It is a simple task to show that $h^{rv}(s) < 0$ when $s < L$ while $h^{rv}(s) \geq 0$ for $s \geq L$. Recall, these inequalities are reversed in Example 2. Using the same argument

as above establishes that the lower control limit is bias optimal. The bias-based decision-maker prefers the negative relative values. ∎

This analysis allows us to interpret why bias prefers control limit L or $L+1$. In Example 2 rewards are received at arrivals and thus, the reward is received before the cost of having the customer in the system is accrued. On the other hand, in Example 3 since rewards are received at service completions the decision-maker must accrue the cost of having a customer in the system before receiving the reward. Thus, the decision-maker only chooses to increase the amount of waiting space if the reward is received before cost. Furthermore, since the optimal policies are not the same for both problems, it is clear that the bias-based decision-maker differentiates between receiving rewards on entrance or exit. The following result provides supporting evidence for this assertion.

Theorem 1 *Suppose that α is a positive recurrent state for a fixed policy $d \in D^\infty$. Further suppose that $h_d^{rv(\alpha)}$ is the relative value function of d with $h_d^{rv(\alpha)}(\alpha) = 0$. Let c_d be the bias constant associated with $h_d^{rv(\alpha)}$. Then*

$$c_d = -\frac{\mathbb{E}_\alpha^d \sum_{n=0}^{\tau_\alpha - 1}(n+1)[r(X_n, Y_n) - g]}{\mathbb{E}_\alpha^d \tau_\alpha} \tag{14}$$

Proof. From (5) it suffices to show that $P_d^* h_d^{rv(\alpha)}(\alpha) = \frac{\mathbb{E}_\alpha^d \sum_{n=0}^{\tau_\alpha - 1}(n+1)[r(X_n,Y_n)-g]}{\mathbb{E}_\alpha^d \tau_\alpha}$. Recall,

$$P_d^* h_d^{rv(\alpha)}(\alpha) = \frac{\mathbb{E}_\alpha^d \sum_{t=0}^{\tau_\alpha - 1} h_d^{rv(\alpha)}(X_t)}{\mathbb{E}_\alpha^d \tau_\alpha}.$$

Consider the numerator,

$$\mathbb{E}_\alpha^d \sum_{t=0}^{\tau_\alpha - 1} h_d^{rv(\alpha)}(X_t) = \mathbb{E}_\alpha^d \left(\sum_{t=0}^{\tau_\alpha - 1} \mathbb{E}_{X_t}^d \sum_{n=0}^{\tau_\alpha - 1} [r(X_n, Y_n) - g] \right) = \mathbb{E}_\alpha^d \left(\sum_{t=0}^{\tau_\alpha - 1} \mathbb{E}^d \left(\sum_{n=t}^{\tau_\alpha - 1} [r(X_n, Y_n) - g] \middle| X_t \right) \right),$$

where the second equality follows from the time-homogeneity of the process. Conditioning on the first passage time, given that the initial state is α we get,

$$\begin{aligned}
\mathbb{E}_\alpha^d \sum_{t=0}^{\tau_\alpha - 1} h_d^{rv(\alpha)}(X_t) &= \sum_{k=1}^{\infty} \mathbb{E}_\alpha^d \left(\sum_{t=0}^{k-1} \mathbb{E}^d \left(\sum_{n=t}^{\tau_\alpha - 1} [r(X_n, Y_n) - g] \middle| X_t \right) \middle| \tau_\alpha = k \right) P_\alpha(\tau_\alpha = k) \\
&= \sum_{k=1}^{\infty} \sum_{t=0}^{k-1} \mathbb{E}_\alpha^d \left(\sum_{n=t}^{\tau_\alpha - 1} [r(X_n, Y_n) - g] \middle| \tau_\alpha = k \right) P_\alpha(\tau_\alpha = k) \\
&= \mathbb{E}_\alpha^d \sum_{t=0}^{\tau_\alpha - 1} \left(\sum_{n=t}^{\tau_\alpha - 1} [r(X_n, Y_n) - g] \right)
\end{aligned}$$

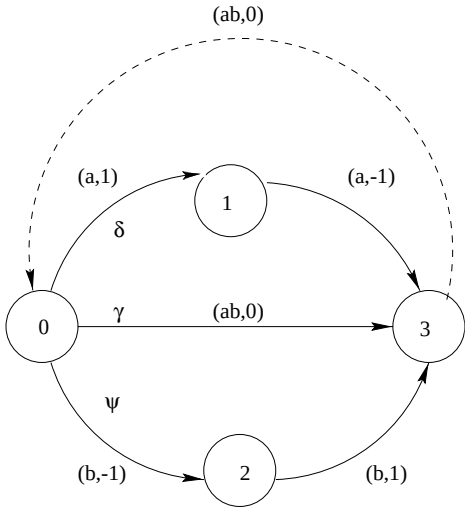

Figure 2: A deterministic example with average reward 0. Quantities in parentheses denote actions and reward, respectively.

A little algebra yields

$$\mathbb{E}_\alpha^d \sum_{t=0}^{\tau_\alpha - 1} h_d^{rv(\alpha)}(X_t) = \mathbb{E}_\alpha^d \sum_{n=0}^{\tau_\alpha - 1} \left(\sum_{t=0}^{n} [r(X_n, Y_n) - g] \right) = \mathbb{E}_\alpha^d \sum_{n=0}^{\tau_\alpha - 1} (n+1)[r(X_n, Y_n) - g] \tag{15}$$

Substituting this into (5) yields the result. ∎

Since we are maximizing this term, the bias-based decision-maker prefers to receive negative excess rewards later and positive excess rewards earlier. Example 4 below illustrates this point. The factor $n+1$ in the previous result was noted in Meyn [10], however, to our knowledge this is the first time that it has been used to explain implicit discounting captured by the bias.

Example 4

Suppose $S = \{0, 1, 2, 3\}$, $A_0 = \{a, b, ab\}$, $A_1 = \{a\}$, $A_2 = \{b\}$, and $A_3 = \{ab\}$, $r(0, a) = r(2, b) = 1$, $r(0, b) = r(1, a) = -1$, $r(0, ab) = r(3, ab) = 0$ and $p(1|0, a) = p(2|0, b) = p(3|1, a) = p(3|2, a) = p(3|0, ab) = p(0|3, ab) = 1$. Let δ be the decision rule that chooses action a in state zero, γ the one that chooses ab, and ψ be that which chooses action b. Clearly, this model is unichain and $g_\delta = g_\gamma = g_\psi = 0$. It is also easy to see that the bias constant, c_γ, of γ must be zero. Choose $\{0\}$ as the reference state (so $h^{rv(0)}(0) = 0$). By examination of Figure 2 we have $h^{rv(0)}(1) = -1$, $h^{rv(0)}(2) = 1$, and $h^{rv(0)}(3) = 0$. The stationary distributions, β_d^*, are $\beta_\delta^* = \{1/3, 1/3, 0, 1/3\}$ and $\beta_\psi^* = \{1/3, 0, 1/3, 1/3\}$. Thus, the bias constants are $-\beta_\delta^* h^{rv(0)} = 1/3$ and $-\beta_\psi^* h^{rv(0)} = -1/3$. By

(5) δ is bias optimal. Alternatively, we can use Theorem 1 to compute the constant

$$\begin{aligned} c_\delta &= -\frac{\{[r(0)-g]+[r(1)-g]+[r(3)-g]\}+\{[r(1)-g]+[r(3)-g]\}+\{[r(3)-g]\}}{3} \\ &= -\frac{\{[r(0)-g]+2[r(1)-g]+3[r(3)-g]\}}{3} = -(1+2\cdot(-1)+3\cdot(0))/3 = 1/3. \end{aligned}$$

Similarly, $c_\psi = -(-1+2\cdot(1)+3\cdot(0))/3 = -1/3$.

Note that when the positive excess reward of 1 is received earlier it costs less than when it is received later (-1 compared to -2), and conversely. If we compare δ to γ the decision-maker chooses to receive the immediate reward and accrue the cost later. If we compare ψ to γ the decision-maker prefers not to accrue the immediate cost, despite the fact that there is a reward to be received later. ∎

Precisely the same logic can be applied to the prior queueing example. Since the reward received later is discounted the decision-maker chooses not to accept the arriving customer. This makes explicit the fact that the bias also captures the desirable properties of discounting.

7 Conclusions

We have presented a probabilistic approach to interpreting bias optimality. This leads to simple sample path arguments for results that previously required detailed algebra and presented no intuition for why the bias optimizing decision-maker would prefer a particular policy. Furthermore, this probabilistic analysis leads to an explanation of implicit discounting in bias.

It is important to note that the discounting captured by bias is only valid for recurrent states. In fact, it is easy to construct examples in which the bias-based decision-maker is indifferent to receiving reward earlier or later in transient states (see Veinott [14]). To capture this one must turn to more sensitive optimality criterion.

Finally, we have restricted our attention to finite state, finite action space models. The authors hope that this paper makes apparent the need to develop these ideas on more general spaces and for multichain Markov decision processes.

8 Acknowledgements

We would like to thank Enrique Lemus for some preliminary discussions on bias optimality.

References

[1] E. V. Denardo. Computing a bias optimal policy in a discrete-time Markov decision problem. *Operations Research*, 18:279–289, 1970.

[2] C. Derman and A. F. Veinott, Jr. A solution to a countable system of equations arising in Markovian decision processes. *Annals of Mathematical Statistics*, 38(2), 1967.

[3] M. Haviv and M. L. Puterman. Bias optimality in controlled queueing systems. *Journal of Applied Probability*, 35:136–150, 1998.

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

[5] M. E. Lewis, H. Ayhan, and R. D. Foley. Bias optimality in a queue with admission control. *Probability in the Engineering and Informational Sciences*, 13:309–327, 1999.

[6] M. E. Lewis and M. L. Puterman. A note on bias optimality in controlled queueing systems. *Journal of Applied Probability*, 37(1), 2000.

[7] S. A. Lippman. Applying a new device in the optimization of exponential queueing systems. *Operations Research*, 23(4):687–712, 1975.

[8] A. M. Makowski and A. Shwartz. On the Poisson equation for Markov chains: Existence of solutions and parameter dependence by probabilistic methods. Technical report, Technion–Israel Institute of Technology, September 1994.

[9] E. Mann. Optimality equations and sensitive optimality in bounded Markov decision processes. *Optimization*, 16(5):767–781, 1985.

[10] S. Meyn. The policy iteration algorithm for average reward Markov decision processes with general state space. *IEEE Transactions on Automatic Control*, 42:1663–1680, December 1997.

[11] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, Inc., New York, 1994.

[12] P. Schweitzer and A. Federgruen. The functional equations of undiscounted Markov renewal programming. *Mathematics of Operations Research*, 3:308–321, 1977.

[13] A. F. Veinott, Jr. Discrete dynamic programming with sensitive discount optimality criteria. *Annals of Mathematical Statistics*, 40(5):1635–1660, October 1969.

[14] A. F. Veinott, Jr. Markov decision chains. In *Studies in Optimization*, volume 10 of *Studies in Mathematics*, pages 124–159. Mathematics Association of America, 1974.