MARKOV DECISION PROCESSES
Models, Methods, Directions, and Open Problems

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

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

Submitted, November 16, 1999

Chapter 1

Bias Optimality

Abstract

The use of the long-run average reward or the *gain* as an optimality criterion has received considerable attention in the literature. However, for many practical models the gain has the undesirable property of being *underselective*, that is, there may be several gain optimal policies. After finding the set of policies that achieve the primary objective of maximizing the long-run average reward one might search for that which maximizes the "short-run" or transient reward. This reward, called the *bias* aids in distinguishing among multiple gain optimal policies. This chapter focuses on the usefulness of the bias in distinguishing multiple gain optimal policies, its computation, and the implicit discounting captured by bias on recurrent states.

1.1 Introduction

The use of the long-run average reward or the *gain* as an optimality criterion has received considerable attention in the literature. However, for many practical models the gain has the undesirable property of being *underselective*, that is, there may be several gain optimal policies. Since gain optimality is only concerned with the long-run behavior of the system there is the possibility of many gain optimal policies. Often, this leads decision-makers to seek more sensitive optimality criteria that take into account short-term system behavior.

Suppose the new manager of a warehouse has decided through market studies and a bit of analysis that when long-run average cost is the optimality criterion an "(s, S)" ordering policy is optimal. That is to say that past demand patterns suggest it is optimal to reorder when the inventory falls below the level s and that it should be increased to S units when orders are made. Furthermore, suppose that there are many such limits that achieve long-run average optimality. With this in mind, the manager has arbitrarily chosen the long-run average optimal policy (s', S'). In fact, in this example the manager could choose any ordering policy for any (finite) amount of time, and then start using any one of the optimal average cost policies and still achieve the optimal average cost. However, the decision-maker should be able to discern which of the optimal average cost policies is best from a management perspective and use that policy for all time. The use of the transient reward or the *bias* can assist in making such decisions.

In essence, after finding the set of policies that achieve the primary objective of maximizing the long-run average reward we search for that which maximizes the transient reward. This reward, called the *bias* is the obvious next step among optimality criterion since it appears as the second term of the Laurent series expansion of the discount reward function. In very simple models with a single absorbing state and multiple policies to choose from on transient states the concept of bias optimality is easy to understand. In these models all policies are average optimal and the bias optimal policy is the one which maximizes the expected total reward before reaching the absorbing state. However, in models in which all states are recurrent or models in which different policies have different recurrent classes, the meaning of bias optimality is not as transparent. It is one of our main objectives in this chapter to provide some insight on this point by developing a "transient" analysis for recurrent models based on relative value functions. We present an algorithmic and a probabilistic analysis of bias optimality and motivate the criterion with numerous examples. The reader of this chapter should keep the following questions in mind:

- How is bias related to average, total, and discounted rewards?
- How do we compute the bias?
- How are bias and gain computation related?
- In a particular problem, what intuition is available to identify bias optimal policies?
- Can we use sample path arguments to identify bias optimal policies?
- How is bias related to the timing of rewards?
- What does bias really mean in recurrent models?

1.2 Historical references

Most Markov Decision Process (MDP) research has regarded bias as a theoretical concept. It was viewed as one of many optimality criteria that is more sensitive than long-run average optimality, but its application has received little attention. In many applications when there are multiple gain optimal policies there is only one bias optimal policy. Hence, the *bias based* decision-maker need not look any further to decide between a group of gain optimal policies. Recently, Haviv and Puterman bias_opt:hav_put [4] showed in a queueing admission control model, with one server and a holding cost, that one can distinguish between two average optimal solutions by appealing to their bias. Their work was extended by Lewis, et. al article:bias1 [9] to a finite capacity, multi-class system with the possibility of multiple gain optimal policies. Further, Lewis and Puterman article:bias2 [11] showed that in the Haviv-Puterman model, the timing of rewards impacts bias optimality. Whereas the Haviv-Puterman paper showed that when rewards are received upon admitting a customer and there are two consecutive gain optimal control limits, say L and $L+1$, only $L+1$ is bias optimal, the Lewis-Puterman paper showed that if the rewards are received upon departure, only control limit L is bias optimal. This suggests that bias may implicitly discount rewards received later. Lewis and Puterman article:bias3 [10] present a new approach to compute the bias directly from the average optimality equations. This leads to sample path arguments that provide alternative derivations of the above mentioned results. In addition to the previously mentioned papers ([4], [9], [11]), the use of bias to distinguish between gain optimal policies has only been discussed in a short section of an expository chapter by Veinott [19]. Methods of computing optimal bias were considered for the finite state and action space case by Denardo [3] and Veinott [17] and on countable state and compact action spaces by Mann [13]. The extension of bias to general state spaces has not received much attention with the exception of section 10.3 of Hernandez-Lerma and Lasserre [7] where under certain assumptions it is shown to be equivalent to other sensitive optimality criterion.

Discount and average optimality have been considered extensively in the literature, therefore we will not provide a complete review here. For a comprehensive review refer to the survey paper of Arapostathis et. al [1] or Chapters 8 and 9 of Puterman [15]. Howard [8] introduced a policy iteration algorithm to solve the average reward model in the finite state space case. This has been considerably extended. For example, see the recent work of Hernández-Lerma and Lasserre [6] or Meyn [14]. Blackwell's [2] classic paper showed the existence of stationary optimal policies in the discounted finite state case and introduced a more sensitive optimality criterion now called *Blackwell* optimality. In essence, Blackwell optimal policies are discount optimal for all discount rates close to 1. It turns out that Blackwell optimality implies bias optimality, so that we have the existence of bias optimal policies in the finite state and action space case as well. There is also a vast literature on sensitive optimality that indirectly addresses bias optimality (cf. Veinott [18]). However, none of these works give an intuitive explanation for *which* policy the bias based decision-maker prefers and why.

1.3 Definitions

Assume that both the state space, $\mathbb{X}$, and the action space, $\mathbb{A}$, are finite. We offer several definitions of the bias of a stationary policy, the first of which leads to the interpretation of bias as the transient reward. Since the bias and the gain are so closely related a definition of the gain is included as well.

Definition 1 *The long-run average reward or* **gain** *of a policy π given that* *the system starts in state $x \in \mathbb{X}$, denoted $w(x, \pi)$, is given by*

$$w(x, \pi) \;=\; \lim_{N\to\infty} \mathbb{E}_x^{\pi}\left(\frac{1}{N}\sum_{n=0}^{N-1} r(x_n, a_n)\right). \tag{1.1}$$ def:gain

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*

$$w(x, \pi^*) \geq w(x, \pi) \textit{ for all } x \in \mathbb{X}, \textit{ for all } \pi \in \Pi.$$

Denote the set of stationary, deterministic (nonrandomized) policies by D^∞ and a particular element of that set by d^∞. We now formalize the definition of bias.

def:bias **Definition 2** *Suppose the Markov chain generated by a stationary, deterministic policy d^∞ is aperiodic. The* **bias** *of d^∞ given that the system started in state x, denoted $h(x, d^\infty)$, is defined to be*

$$h(x, d^\infty) \;=\; \sum_{n=0}^{\infty} \mathbb{E}_x^{d^\infty}[r(x_n, d(x_n)) - w(x_n, d^\infty)]. \tag{1.2}$$ def:bias1

Similarly, if the Markov chain generated by d^∞ is periodic, we define the bias to be

$$h(x, d^\infty) \;=\; \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \mathbb{E}_x^{d^\infty} \sum_{t=0}^{n}[r(x_t, d(x_t)) - w(x_t, d^\infty)]. \tag{1.3}$$ def:bias2

We say that a policy, $(d^*)^\infty$ is **bias optimal** *if it is* **gain optimal**, *and in addition*

$$h(x, (d^*)^\infty) \geq h(x, d^\infty) \textit{ for all } x \in \mathbb{X}, \textit{for all } d^\infty \in \Pi^S.$$

Note that although we have only defined the bias for stationary policies, when the state and action spaces are finite, this class of policies is large enough to guarantee existence of a policy that maximizes the transient reward. This will be discussed further momentarily. Furthermore, since we have assumed that the state space is finite the limit in (def:gain 1.1) exists. As we will soon see, (def:gain 1.1) allows for an interpretation of bias as the total reward for a slightly modified process. This requires another definition.

Definition 3 *For a particular stationary policy d^∞ and $x \in \mathbb{X}$ let*

$$e(x, d^\infty) \;=\; r(x, d(x)) - w(x, d^\infty) \qquad (1.4)$$

be called the **excess reward** *of d^∞.*

Assume for now that the Markov chain generated by a policy d^∞ is aperiodic. We will adopt the convention of using subscripts or superscripts when writing vectors or matrices corresponding to particular policies. Let v_d^N denote the vector of total expected rewards over the first N periods when using the policy d^∞ so that

$$v_d^N \;=\; \sum_{t=1}^{N} P_d^{t-1} r_d. \qquad (1.5)$$

From (def:bias1 1.2),

$$h_d \;=\; \sum_{t=1}^{N} P_d^{t-1} r_d - N w_d + \sum_{t=N+1}^{\infty} (P_d^{t-1} - P_d^*) r_d, \qquad (1.6)$$

threeterm

where $P_d^* = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} P_d^i$ is the *limiting matrix* of the Markov chain generated by d^∞. Note $w_d = P_d^* r_d$. Since h_d is finite, the third term in (threeterm 1.6) approaches zero as $N \to \infty$. Hence, we may write,

$$v_d^N \;=\; N w_d + h_d + o(1) \qquad (1.7)$$

where $o(1)$ denotes a vector with components that approach zero as $N \to \infty$. In component notation, we write $v_d^N(x) = v^N(x, d^\infty)$. As N becomes large, $v^N(x, d^\infty)$ approaches a line with slope $w(x, d^\infty)$ and intercept $h(x, d^\infty)$. Thus, for the process generated by the stationary policy d^∞ we have an interpretation of the gain as the asymptotic rate of increase relative to the horizon length of the total reward and the bias as the intercept or initial level.

An alternative interpretation can be realized from (def:bias1 1.2) if one defines a new system in which for each stationary policy, the reward function is replaced with the excess reward function. The bias is then 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 P_d^*. If 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. Under either scenario, it is well-known that the convergence to steady state occurs exponentially fast so that we can view bias as the total transient reward.

The interpretation of the bias as the total reward while correct, also has its limitations and can be misleading. In economic applications financial rewards received earlier, rather than later, are more valuable. In fact, earlier rewards translate into decision-making flexibility regardless of the volatility of the industry. Consider the discounted reward function where β is the discount rate,

$$v(x,\pi,\beta) \;=\; \mathbb{E}_x^\pi\left[\sum_{n=0}^{\infty}\beta^n r(x_n,a_n)\right]. \tag{1.8}$$

def:disc

Definition 4 *We say that a policy π^* is* **n-discount optimal** *for some integer $n \geq -1$ if*

$$\liminf_{\beta\uparrow 1}(1-\beta)^{-n}[v(x,\pi^*,\beta)-v(x,\pi,\beta)] \;\geq\; 0 \tag{1.9}$$

for all $x \in \mathbb{X}$ and $\pi \in \Pi$.

Since the state and action space are finite, we know that there exists stationary, deterministic $n-$discount optimal policies for all n (see Theorem 10.1.5 of Puterman mdp_book:put [15]). Furthermore, for stationary policies, $0-$discount optimal policies also maximize the bias among gain optimal policies so that the two criterion are equivalent (see Puterman mdp_book:put [15], Theorem 10.1.6). Thus, it suffices to find a policy $\pi^* \in \Pi^S$ such that

$$\liminf_{\beta\uparrow 1}[v(x,\pi^*,\beta)-v(x,\pi,\beta)] \;\geq\; 0 \tag{1.10}$$

eq:0-discount

for all $x \in \mathbb{X}$ and $\pi \in \Pi^S$. With this close relationship to discounting, it stands to reason that the bias retains some of the attributes of discounting. We will elaborate more on this subject later in the chapter.

The next definition that we present does not add intuition to the meaning of the bias. However, it does hint at the similarities between the computation of the gain and the bias. Let $H_P \equiv (I - P + P^*)^{-1}(I - P^*)$ be the deviation matrix of P.

def:bias_alt **Definition 5** *(Alternative definition of the bias) The bias of a stationary, policy d^∞ is*

$$h(x, d^\infty) \quad = \quad (H_{P_d} r_d)(x) \tag{1.11}$$

If we expand $(I - P_d + P_d^*)^{-1}$ in a power series the equivalence of Definitions def:bias 2 and def:bias_alt 5 becomes clear. Recall that if P_d^* is the stationary distribution of the Markov chain generated by the policy d^∞ the gain may be computed

$$w_{d^\infty} \quad = \quad P_d^* r_d. \tag{1.12}$$

compute:gain

Hence, the deviation matrix replaces the stationary distribution when computing the bias. This begs the question, can the bias be computed using some of the methods available in the vast literature on the long-run average reward problem? This is the subject of the next section.

A final interpretation of bias is available as the second term in the Laurent series expansion of the discount value function. The following appears in Puterman mdp_book:put [15], Theorem 8.2.3. Let $\rho = (1 - \beta)/\beta$ be the interest rate.

Theorem 6 *Assume that S is finite. Let ν denote the nonzero eigenvalue of $I - P_d$ with the smallest modulus. Then, for $0 < \rho < |\nu|$,*

$$v(d^\infty, \beta) \quad = \quad (1 + \rho)\left[\sum_{n=-1}^{\infty} \rho^n y_n^d\right] \tag{1.13}$$

where $y_{-1}^d = w_d$, $y_0^d = h_d$, and $y_n^d = (-1)^n H_{P_d}^{n+1} r_d$.

As was alluded to earlier, the bias appears as the second term in the Laurent series expansion of the total expected discounted reward function. The above observations lead to interpretations of the bias of a stationary policy as:

1. the intercept of the asymptotic total reward process (in the aperiodic case),
2. the total difference between the process beginning in a particular state and that which begins in stationarity,
3. the second term of the Laurent series expansion

We now turn to methods for computing the bias.

1.4 Computing the Bias from the Evaluation Equations

In most practical examples, computation of the bias directly from the above definitions is not feasible. We discuss some practical methods for the computation of the bias of a fixed stationary policy d^∞. These methods also lead to an intuitive understanding of bias. The gain and the bias of d^∞ may be computed by solving the following system of linear equations:

$$w = P_d w, \tag{1.14}$$

$$h = r_d - w + P_d h, \tag{1.15}$$

and

$$k = -h + P_d k \tag{1.16}$$

for vectors w, h, and k. Specifically, the gain and the bias of d^∞ satisfy (1.14) and (1.15) and there exists some vector k which together with the bias satisfies (1.16). To see this, note that if we multiply (1.15) by P_d^* we have

$$P_d^* h \;=\; P_d^* r_d - P_d^* w + P_d^* h,$$

thus,

$$P_d^* w \;=\; P_d^* r_d. \tag{1.17}$$

However, by repeated application of (1.14)

$$w \;=\; \frac{1}{N}[w + P_d w + \cdots + P_d^{N-1} w].$$

Taking limits as $N \to \infty$ we have $w = P_d^* w$. Combining this with (1.17) yields $w = P_d^* r_d$. To show that h_d satisfies (1.15), and (1.16) with $w = w_d$, multiply (1.16) by P_d^* to get $P_d^* h = 0$. Hence, from (1.15)

$$\begin{aligned} h - P_d h + P_d^* h \;&=\; r_d - w_d \\ &=\; r_d - P_d^* r_d. \end{aligned}$$

Since $I - P_d + P_d^*$ is invertible we get $h = (I - P_d + P_d^*)^{-1}(I - P_d^*)r_d = H_{P_d}r_d$ as desired. Moreover, the gain and the bias are the unique vectors with this property. We refer to (1.14) and (1.15) as the *average evaluation equations* (AEE) and to (1.14), (1.15), and (1.16) as the *bias evaluation equations* (BEE). If we restrict attention to (1.14) and (1.15) only, the gain is uniquely determined and the bias is determined up to m additive constants, where m is the number of closed, recurrent classes for the Markov chain generated by d^∞. Furthermore, each of the equations can be written in the form

$$(I - P_d)u = v. \tag{1.18}$$

It would be nice if $I - P_d$ was invertible. Of course it is not. However, H_{P_d} is often called the *Drazen inverse* of P_d (denoted $(I - P_d)^{\#}$) and exhibits many desirable properties of matrix inverses. Namely,

$$H_{P_d}^{\#}H_{P_d}H_{P_d}^{\#} = H_{P_d}^{\#}, \quad H_{P_d}H_{P_d}^{\#} = H_{P_d}^{\#}H_{P_d}, \quad \text{and} \quad H_{P_d}H_{P_d}^{\#}H_{P_d} = H_{P_d} \tag{1.19}$$

The Drazen inverse is used to derive the chain structure of a Markov chain. For more information on the Drazen inverse see Appendix A of Puterman [15]. In the following example we compute the bias using the BEE and the Drazen inverse of $I - P_d$.

Example 1

Let $\mathbb{X} = \{x_1, x_2, x_3\}$ and suppose the Markov chain generated by the policy d^∞ has transition structure

$$P_d = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}. \tag{1.20}$$

It is not hard to show that

$$P_d^* = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \tag{1.21}$$

and

$$H_{P_d} = \begin{bmatrix} 1/3 & 0 & -1/3 \\ -1/3 & 1/3 & 0 \\ 0 & -1/3 & 1/3 \end{bmatrix} \tag{1.22}$$

If $r'_d = \{1, -1, 0\}$ we get,

$$w_d = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

and

$$h_d = \begin{bmatrix} 1/3 \\ -2/3 \\ 1/3 \end{bmatrix}.$$

Furthermore, the BEE are satisfied since

$$\begin{aligned} 1/3 &= 1 - 0 - 2/3 \\ -2/3 &= -1 - 0 + 1/3 \\ 1/3 &= 0 - 0 + 1/3. \end{aligned}$$

and

$$\begin{aligned} k(x_1) &= -1/3 + k(x_2) \\ k(x_2) &= 2/3 + k(x_3) \\ k(x_3) &= -1/3 + k(x_1) \end{aligned}$$

has the solution, $k(x_1) = 0, k(x_2) = 1/3$, and $k(x_3) = -1/3$. Similar to the bias and the AEE, the vector k is not the unique vector that satisfies (1.16). Since the model consists of one recurrent class, $(k + c1, h_d)$ is also a solution to (1.16) for any constant c, where 1 denotes a vector with all components equal to 1. As we will see, a similar result can be used to simplify the computation of the bias of Markov decision processes.

total

1.4.1 Bias and total reward

In this section we describe some models in which the bias is equivalent to the expected total reward. This is important because the expected total reward criterion has received considerable attention in the reinforcement learning literature. This also gives us insight for alternative methods for computing the bias sample pathwise. Let $a^+ = \max\{a, 0\}$ be the positive part and

$a^- = \max\{-a, 0\}$ be the negative part of a real number a. Suppose we define

$$v_+(\pi, x) \;=\; \mathbb{E}_x^\pi \left\{ \sum_{n=0}^{\infty} r^+(x_n, y_n) \right\}, \tag{1.23}$$

pos_model

and

$$v_-(\pi, x) \;=\; \mathbb{E}_x^\pi \left\{ \sum_{n=0}^{\infty} r^-(x_n, y_n) \right\}. \tag{1.24}$$

neg_model

If either (pos_model 1.23) or (neg_model 1.24) is finite for all $x \in \mathbb{X}$, then $\lim_{N\to\infty} v_\pi^N$ exists. If both are finite then so is $\lim_{N\to\infty} v_\pi^N$. Furthermore, if (pos_model 1.23) and (neg_model 1.24) are finite for **all** $\pi \in \Pi$, then $w_\pi = 0$ for all $\pi \in \Pi$ (see Proposition 10.4.1, part (c) of Puterman mdp_book:put [15]). Using Definition def:bias 2, it stands to reason that if the total reward is finite, the two criterion should coincide. This is precisely the case.

prop:zero_gain **Proposition 7** *Let $d^\infty \in D^\infty$ and suppose $v_+(d^\infty, x)$ and $v_-(d^\infty, x)$ are finite for all $x \in \mathbb{X}$, then $\lim_{N\to\infty} v_{d^\infty}^N = v_{d^\infty} = h_{d^\infty}$.*

The importance of this result is that in models with expected total reward criterion whenever v_+ and v_- are finite, methods developed for determining bias optimal policies apply to compute optimal policies. This avoids many of the complexities that have developed in the theory of models with expected total reward criterion; especially the need to distinguish positive and negative models. See Puterman mdp_book:put [15], especially Chapter 7, for more on this issue.

1.4.2 Unichain Markov decision processes

Given the previous discussion, one might conjecture that if there is but one recurrent class, we could replace the reward with the excess reward function and compute the bias of the transient states as the total reward until reaching the recurrent class. After all, this has the effect of treating the recurrent class as a single state with zero reward and confining the transient analysis to the transient states. The following example shows that this is not the case.

entry **Example 2**

Let $\mathbb{X} = \{1, 2, 3\}$. Suppose $A_1 = \{a, b\}$, $A_2 = \{a\}$ and $A_3 = \{a\}$, such that $P(2|1, a) = P(3|1, b) = 1$ and $P(3|2, a) = P(2|3, a) = 1$. Furthermore,

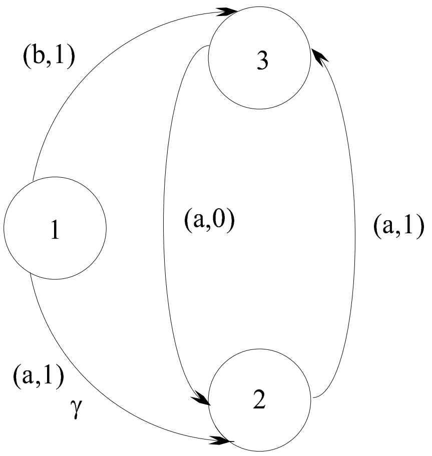

Figure 1.1: A deterministic example with finite gain, where the bias differs from the total reward until absorption into a recurrent class. fig:entry

assume that $r(1,a) = r(1,b) = r(2,a) = 1$ while $r(3,a) = 0$. See Figure 2. If γ chooses action a in state 1 and δ chooses action b, it is not hard to show that $h_\gamma = \{3/4, 1/4, -1, 4\}$ and $h_\delta = \{1/4, 1/4, -1/4\}$. Hence, despite the fact that the total rewards until reaching the recurrent class for both policies are the same, the biases differ. This hints at a point that we make later in our discussion; the bias based decision-maker distinguishes when during a recurrent cycle rewards are received.

Except where noted otherwise, assume for now that the process generated by any stationary policy is unichain; that is, the process has a single ergodic class and perhaps some transient states. Hence, the following hold

1. w is constant which we express as $w1$.
2. (1.14) is redundant and (1.15) becomes
$$h = r_d - w1 + P_d h. \qquad (1.25)$$ biaseq2
3. constant If (w, h) satisfies (1.25), $w = w_{d^\infty}$ and h is unique up to a constant.
4. If $(w, h_d^{rv(\alpha)})$ satisfies (1.25), and $h_d^{rv(\alpha)}(\alpha) = 0$, $h_d^{rv(\alpha)}$ is unique and is called the *relative value function* of d^∞ at α.

5. $(w_{d^\infty}, h_{d^\infty})$ is the unique solution of (biaseq2 1.25) and the additional condition $P^* h = 0$.

With these observations in mind, we have the following definition which was originally introduced in article:bias3 [10].

Definition 8 *Let $d^\infty \in D^\infty$ be a fixed stationary policy for which P_d is unichain. For each solution to the average evaluation equations (w_d, h), the constant difference between h and the bias of d^∞, h_{d^∞}, denoted $c_d(h)$, is called the* **bias constant** *associated with h.*

We now show how to use an arbitrary solution (w, h) of the AEE to compute the bias constant and therefore the bias. Suppose α is a recurrent state for the Markov chain generated by d^∞. Denote the first time the process enters the state α by τ_α. That is,

$$\tau_\alpha \quad = \quad \min\{n \geq 0 | X_n = \alpha\}. \tag{1.26}$$

Let

$$h_\alpha^d(s) \quad = \quad \mathbb{E}_s^d \left(\sum_{n=0}^{\tau_\alpha - 1} [r(X_n, Y_n) - w_d] \right). \tag{1.27} \quad \boxed{\texttt{eq:rel\_val\_0}}$$

Note

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

Hence, (w_d, h_α^d) satisfies (biaseq2 1.25) for policy d^∞. Furthermore, $h_\alpha^d(\alpha) = 0$. From point 4 above, h_α^d is the relative value function of the policy d^∞ at the reference state α; $h_\alpha^d = h_d^{rv(\alpha)}$. In addition, from (eq:rel_val_0 1.27) we interpretate the relative value function as *the expected total excess reward until the system reaches the recurrent state α.*

Since for a fixed policy d^∞ the relative value functions and the bias satisfy the AEE they must differ by a constant. Choose positive recurrent state α and let $c_d(h_d^{rv(\alpha)})$ be the bias constant associated with the relative value function. Then

$$h_d \quad = \quad h_d^{rv(\alpha)} + c_d(h_d^{rv(\alpha)})1, \tag{1.28} \quad \boxed{\texttt{eq:h}}$$

and

$$P_d^* h_d \;=\; P_d^* h_d^{rv(\alpha)} + P_d^* c_d(h_d^{rv(\alpha)})1. \tag{1.29}$$

However, since $P_d^* h_d = 0$,

$$\begin{aligned} P_d^* h_d^{rv(\alpha)} &= -P_d^* c_d(h_d^{rv(\alpha)})1 & (1.30)\\ &= -c_d(h_d^{rv(\alpha)})1, & (1.31) \end{aligned}$$

biasconst

where the last equality follows from the fact that the unichain assumption implies that the rows of P_d^* are equal. Making the appropriate substitution into (eq:h 1.28) yields the following proposition.

prop:c **Proposition 9** *Suppose a finite state and action space Markov decision process is unichain. Let d^∞ be a stationary policy. Denote the relative value function of d at a recurrent state α by $h_d^{rv(\alpha)}$. Let $c_d = (P_d^* h_d^{rv(\alpha)})(s)$ for any state $s \in \mathbb{X}$. Then the bias of d is given by $h_d = h_d^{rv(\alpha)} - c_d 1$.*

Remark 10 *The above result holds in the countable state case provided α is positive recurrent.*

Hence, while in Definition def:bias_alt 5 we replace the stationary distribution in the computation of the gain by the deviation matrix to compute the bias, we can instead replace the reward function in $w = P_d^* r_d$, by the relative value function to compute the bias, by computing $P_d^* h_d^{rv}$. Furthermore, applying a classic result in renewal theory we have,

$$h_d(s) \;=\; h_d^{rv(\alpha)}(s) - \frac{\mathbb{E}_\alpha^{d^\infty} \sum_{n=0}^{\tau_\alpha - 1} h_d^{rv(\alpha)}(x_n)}{\mathbb{E}_\alpha^{d^\infty} \tau_\alpha}. \tag{1.32}$$

This expression allows us to compute the bias of a stationary policy sample pathwise. The following simple example illustrates each of these methods for computing bias.

ex:main **Example 3**

Suppose $\mathbb{X} = \{0,1,2,3\}$, $A_0 = \{a,b,c\}$, $A_1 = \{a\}$, $A_2 = \{b\}$, and $A_3 = \{c\}$, $r(0,a) = r(2,b) = 1$, $r(0,b) = r(1,a) = -1$, $r(0,c) = r(3,c) = 0$ and $p(1|0,a) = p(3|0,c) = p(2|0,b) = p(3|1,a) = p(3|2,b) = p(0|3,c) = 1$. Let δ be the decision rule that chooses action a in state zero, γ be the decision

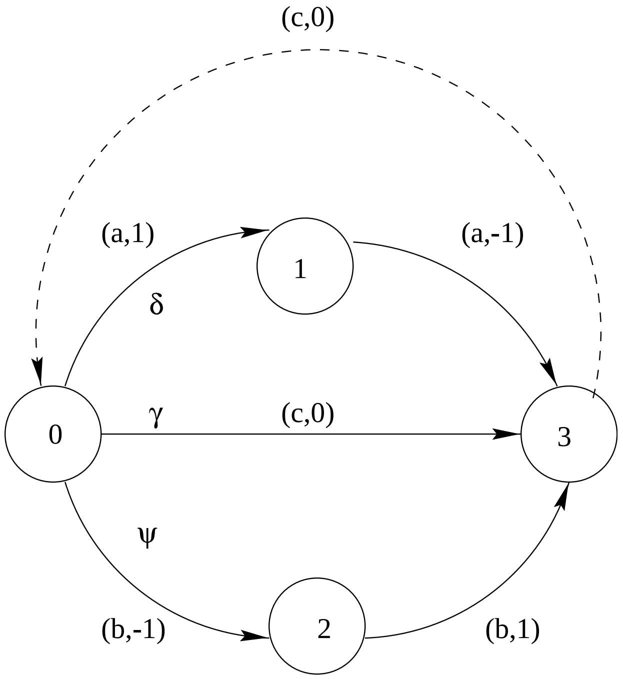

Figure 1.2: A deterministic example with average reward 0. fig:determ

rule that selects action c and ψ be that which chooses action b. Clearly, this model is unichain and $w_\delta = w_\gamma = w_\psi = 0$. Suppose we arbitrarily choose $\{0\}$ as the reference state (so $h_d^{rv(\alpha)}(0) = 0$ for each d). Since state zero is the only state that requires a decision, the relative value function at α is the same for all policies. Hence, we suppress the dependence on d. By examination of Figure fig:determ 1.2 we have $h^{rv(\alpha)}(1) = -1$, $h^{rv(\alpha)}(2) = 1$, and $h^{rv(\alpha)}(3) = 0$. The stationary distributions, β_d^* say, are $\beta_\gamma^* = \{1/2, 0, 0, 1/2\}$, $\beta_\delta^* = \{1/3, 1/3, 0, 1/3\}$, and $\beta_\psi^* = \{1/3, 0, 1/3, 1/3\}$. The bias constants are $-\beta_\gamma^* h^{rv(\alpha)} = 0$, $-\beta_\delta^* h^{rv(\alpha)} = 1/3$, and $-\beta_\psi^* h^{rv(\alpha)} = -1/3$. Hence, we have

$$h_\gamma = \begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix}, \quad h_\delta = \begin{bmatrix} 1/3 \\ -2/3 \\ 4/3 \\ 1/3 \end{bmatrix}, \text{ and } \quad h_\psi = \begin{bmatrix} -1/3 \\ -4/3 \\ 2/3 \\ -1/3 \end{bmatrix}.$$

Note that if we neglect state 2, the transition structure and rewards of the Markov reward process generated by δ^∞ are identical to that discussed in

Example 1 and thus the computed bias vectors (without state 2) are the same. We will return to this example throughout the rest of this section.

1.4.3 The Average Optimality Equation (Unichain Case)

The definition of bias optimality requires that one know the set of gain optimal policies *a priori*. This is not usually the case, except in the total reward models of Section 1.4.1. In this section we review conditions for gain optimality which lead to algorithms for computing bias optimal policies in a similar way to computing gain optimal policies.

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

$$h \;=\; \max_{d\in D}\{r_d - w1 + P_d h\} \tag{1.33}$$

aoe

for w and h. Let $G(h)$ be the set of policies that achieve the maximum in (1.33). That is,

$$\delta \in argmax_{d\in D}\{r_d + P_d h\} \equiv G(h). \tag{1.34}$$

aose

We refer to (1.34) as the *average optimality selection equations* (AOSE). To begin our analysis of the average optimality equations, we consider a special case of a result of Schweitzer and Federgruen [16]. In essence, the result states that solutions of the AOE must differ by a constant just as they do for the AEE.

prop:sch_fed

Proposition 11 *Suppose all stationary policies are unichain and let (w_1, h_1) and (w_2, h_2) be solutions to the AOE. Then $w_1 = w_2$ and*

$$h_1 \;=\; h_2 + c1 \tag{1.35}$$

h1h2eq

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

$$h^* \;=\; h_2 + c^*(h_2)1. \tag{1.36}$$

heq

We refer to $c^*(h_2)$ as the **optimal bias constant** associated with h_2. Note that if for $(\delta^*)^\infty$, $h_{\delta^*} = h^*$, the optimal bias constant is the bias constant for h_{δ^*} and h^*. Further note that this result does not require that $\mathbb{X}$ be finite, only that the gain is constant. A nice discussion of the average optimality equations on Borel spaces can be found in Hernandez-Lerma and Laserre [5].

It is easy to see that all three policies considered in Example 3 satisfy the AOSE and that the bias of each differs by a constant as indicated by Proposition 11.

We now return to the question of whether there are decision rules that are gain optimal, but do not satisfy the AOSE. When all policies generate irreducible Markov chains it is known that the average optimality equations are indeed necessary and sufficient (see for Lewis, et. al [9]). The following example shows that this need not be the case in unichain models.

ex:unichain

Example 4

Suppose $\mathbb{X} = \{1,2\}$, $A_1 = \{a,b\}$ 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 that chooses action a in state 1 and let γ be the decision rule that chooses action b in state 1. This model is unichain and $w_\delta = w_\gamma = 1$, $h_\delta^{rv(2)}(1) = 1$, $h_\gamma^{rv(2)}(1) = 2$, $h_\delta^{rv(2)}(2) = h_\gamma^{rv(2)}(2) = 0$. Since $h_\delta^{rv(2)}$ and $h_\gamma^{rv(2)}$ do *not* differ by a constant, it follows from Proposition 11, that $(w_\delta, h_\delta^{rv(2)})$ and $(w_\gamma, h_\gamma^{rv(2)})$ cannot both satisfy the average optimality equations, even though both policies are average optimal. ■

In the sequel we show that our previous observations lead to simple sample path arguments as well as algorithmic solution methods for finding the optimal bias.

1.5 The Bias Optimality Equation

section:bias_opt

Suppose in addition to satisfying the AOE, there exists a vector k, such that h satisfies

$$k = \max_{d\in G(h)}\{-h + P_d k\} \tag{1.37}$$

boe

and δ satisfies

$$\delta \in argmax_{d\in G(h)}\{P_d k\} \tag{1.38}$$

bose

Then δ^∞ is bias optimal and h is the optimal bias. We refer to the combined set (1.37) and the AOE as the *bias optimality equations* (BOE) and to (1.38) as the *bias optimality selection equations* (BOSE).

We can take advantage of the result of Proposition 11; if (w, h_1) and (w, h_2) are solutions to the AOE then h_1 and h_2 differ by a constant. Upon substituting (1.36) into (1.37) for the relative value with reference state α when $(w, h^{rv(\alpha)})$ satisfies the AOE, we have the following important result.

rewrite

Theorem 12 *Suppose $h^{rv(\alpha)}$ is a relative value function with reference state α such that $(w^*, h^{rv(\alpha)})$ is a solution to the AOE. The BOE (1.37) can be rewritten*

$$k = max_{d \in G(h^{rv(\alpha)})}\{-h^{rv(\alpha)} - c1 + P_d k\}. \tag{1.39}$$

mboe

Further, suppose (k^*, c^*) satisfies (1.39). Then k^* is unique up to a constant and c^* is the optimal bias constant associated with $h^{rv(\alpha)}$.

To see that the second part of the theorem follows directly from the first, observe that (1.39) has exactly the same form as the AOE (1.33). That is to say, setting $r_d = -h^{rv(\alpha)}$ and $w = -c1$ we have again the AOE. Thus, in a unichain model, the result is immediate from existing theory on the AOE. Furthermore, all solution methods and theory for the AOE apply directly in this case. In particular, (1.39) can be solved by the same value iteration or policy iteration algorithms used to find gain optimal policies in unichain Markov Decision Processes, however, now we base them on (1.39) .

iteration

Example 5

Consider the model of Example 3. Suppose we begin policy iteration with policy $\rho_0^\infty = \gamma^\infty$. Recall $(h^{rv(0)})' = \{0, -1, 1, 0\}$. We must find (c_0, k_0) to satisfy

$$k_0 \;=\; -h^{rv(0)} - c_0 1 + P_\gamma k_0. \tag{1.40}$$

One can easily show that $k_0' = \{0, 1, -1, 0\}$ and $c_0 = 0$ is a solution to this system. To choose the next policy, we find a policy, ρ_1 such that

$$\rho_1 \in argmax\{-h^{rv(0)} + P_\gamma k_0\}. \tag{1.41}$$

Since we need only make a decision in state 0, note that

$$\begin{aligned} \rho_1(0) &= argmax\{0+1, 0, 0-1\} && (1.42)\\ &= a. && (1.43) \end{aligned}$$

Thus, $\rho_1 = \delta$. For the second iteration of the algorithm we must solve,

$$\begin{aligned} k_1(0) &= 0 - c + k_1(1), \\ k_1(1) &= 1 - c + k_1(3), \\ k_1(2) &= -1 - c + k_1(3), \\ k_1(3) &= 0 - c + k_1(0). \end{aligned}$$

One can verify that $k_1 = \{0, 1/3, -5/3, -1/3\}$ and $c = 1/3$ satisfies the above equations. Furthermore, no further improvements can be made. Notice that the bias of ρ_1 is $1/3$ higher than the relative value function. That is, c is the bias constant of ρ_1. This agrees with the solution found in Example ex:main 3. ■

Alternatively, as in the AEE, if (w_d, h) satisfy the AOE and

$$P_d^* h = 0 \tag{1.44}$$

eq:bias2

where P_d^* is the stationary distribution of the chain generated by d, then h is the optimal bias. Neglecting the trivial case $r_d = 0$ for 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. Suppose that d^∞ is bias optimal. From Proposition prop:c 9

$$h^* = h^{rv(\alpha)} - (P_d^* h^{rv(\alpha)}). \tag{1.45}$$

In essence, solving for the policy with maximum bias reduces to finding the policy that achieves the maximum bias constant, say c^*. That is,

$$c^* 1 = \max_{d \in G(h^{rv(\alpha)})} \{-P_d^* h^{rv(\alpha)}\} = -\min_{d \in G(h^{rv(\alpha)})} \{P_d^* h^{rv(\alpha)}\} \tag{1.46}$$

where $h^{rv(\alpha)}$ is any relative value function of a gain optimal policy. Thus, under the assumption that there exists a state α that is recurrent for all decision rules in $G(h^{rv(\alpha)})$ we can alternatively compute the optimal bias by solving

$$c^* = -\min_{d \in G(h^{rv(\alpha)})} \left(\frac{\mathbb{E}_\alpha \sum_{n=0}^{\tau_\alpha - 1} h^{rv(\alpha)}(X_n, d(X_n))}{\mathbb{E}_\alpha^d \tau_\alpha} \right) \tag{1.47}$$

eq:sample_path

where the expectation is taken with respect to the probability transition function conditioned on starting in state α. Since we are minimizing, $h^{rv(\alpha)}$

can be interpreted as a cost function. Thus, finding a bias optimal policy corresponds to a minimum average cost problem. Furthermore, one might notice that given a relative value function we can solve for the gain, by noting $w_d = r_d + P_d h_d^{rv(\alpha)} - h_d^{rv(\alpha)}$. That is, the relative value function can be used to obtain both the gain and the bias. The crux of the analysis then for finding gain and bias, lies in understanding the relative value functions. We emphasize the importance of these observations in the following examples.

Since we will often be interested in the difference in the cost starting in states s and $s+1$, define for a function b on $\mathbb{N}$, $\Delta b(s) \equiv b(s+1) - b(s)$.

ex:m/m/1/k

Example 6

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 and increasing in s and $f(0) = 0$. Furthermore, assume that we discretize the model by applying the standard uniformization technique of Lippman unif:lipp [12]. Without loss of generality let the uniformization constant $\lambda + \mu = 1$. Since rejecting all customers yields $w = 0$, we assume customers are accepted in state zero. This example was previously considered in Haviv and Puterman bias_opt:hav_put [4] where it was shown algebraically that bias distinguishes between gain optimal policies. For this model, the average optimality equations are

$$\begin{aligned} h(s) &= \max\{\lambda R - w - f(s) + \lambda h(s+1) + \mu h((s-1)^+), \\ &\quad -w - f(s) + \lambda h(s) + \mu h((s-1)^+)\} \end{aligned} \tag{1.48}$$

eq:gain_opt

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 $s \geq L$. Denote the stationary policy that uses control limit L by L. It is known that there exists a Blackwell optimal policy within this set. The following lemma asserts the intuitive idea that it is better to start with fewer customers in the system. We will use this result in the sample path arguments to follow.

prop:delh1

Lemma 13 *Suppose (w^*, h) satisfy the optimality equations. For $s \in S$, $\Delta h(s) < 0$.*

sample

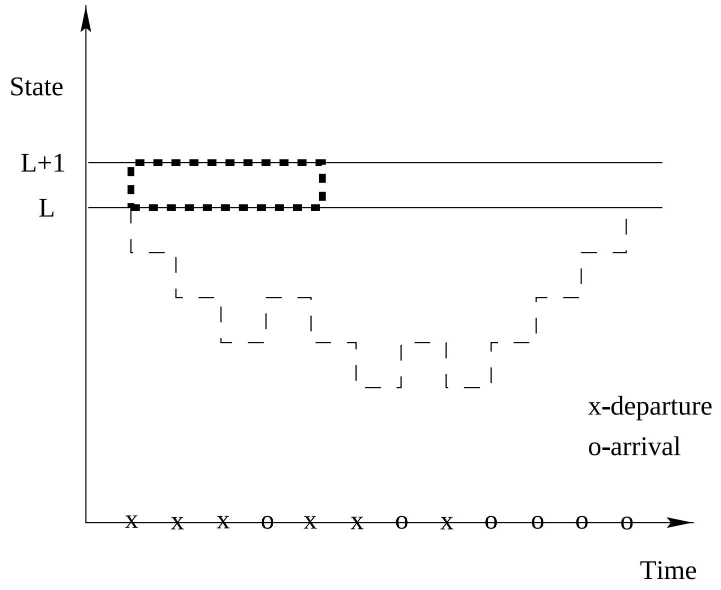

Figure 1.3: We would like to compare control limits L and $L+1$.

Haviv and Puterman [4] show that if there are two gain optimal control limits, only the higher one is bias optimal. Let L and $L+1$ be gain optimal control limits. Notice that since $L+1$ accepts customers in state L and control limit L rejects them, the two policies have different recurrent classes. The gain optimality equations are

$$h(s) \;=\; \begin{cases} \lambda R - g + \lambda h(s+1) + \mu h(s) & s = 0 \\ \lambda R - g - f(s) + \lambda h(s+1) + \mu h(s-1) & 0 < s \leq L \\ -g - f(s) + \lambda h(s) + \mu h(s-1) & s \geq L \end{cases} \qquad (1.49)$$

eq:h^{rv}

At $s = L$ we have equality in the optimality equations since both L and $L+1$ are gain optimal. Furthermore, note that state L is recurrent for both policies. Let c_L be the bias constant for control limit L. Similarly for c_{L+1}. Choose $\alpha = L$ as the reference state. Suppose we follow the sample paths of two processes on the same probability space both starting 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 are the same. This is denoted by the lighter dashed line of Figure 1.5. 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 1.5. Hence, while the total cost before returning to state L 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. Let $p = \frac{\mu}{\lambda+\mu}$ be the probability that the first event is a departure, t be the total expected number of decision epochs on the cycle given that the first event is a departure, and let M be the total expected cost accrued on that cycle. We have

$$\begin{aligned}
\frac{c_{L+1}}{c_L} &= \frac{\mathbb{E}_\alpha^L(\tau_\alpha)\left(\mathbb{E}_\alpha^{L+1}\sum_{n=0}^{\tau_\alpha-1} h^{rv(\alpha)}(X_n, d(X_n))\right)}{\mathbb{E}_\alpha^{L+1}(\tau_\alpha)\left(\mathbb{E}_\alpha^{L}\sum_{n=0}^{\tau_\alpha-1} h^{rv(\alpha)}(X_n, d(X_n))\right)} \\
&= \frac{\mathbb{E}_\alpha^L(\tau_\alpha)\left(pM + (1-p)(\frac{1}{\mu}+1)h^{rv(\alpha)}(L+1)\right)}{\mathbb{E}_\alpha^{L+1}(\tau_\alpha)\left(pM + (1-p)\cdot 0\right)} \\
&= \frac{(pt + (1-p))\left(pM + (1-p)(\frac{1}{\mu}+1)h^{rv(\alpha)}(L+1)\right)}{(pt + (1-p)(\frac{1}{\mu}+1))pM} \\
&= \frac{(pt + (1-p))\left(pM + (1-p)(\frac{1}{\mu}+1)h^{rv(\alpha)}(L+1)\right)}{(pt + (1-p))pM + (1-p)\frac{1}{\mu}pM}
\end{aligned}$$

Recall, that we have assumed that $\lambda + \mu = 1$ so $p/\mu = 1$. Thus,

$$\frac{c_{L+1}}{c_L} = \frac{\mathbb{E}_\alpha^L(\tau_\alpha)pM + (1-p)\,\mathbb{E}_\alpha^L(\tau_\alpha)(\frac{1}{\mu}+1)h^{rv(\alpha)}(L+1)}{\mathbb{E}_\alpha^L(\tau_\alpha)pM + (1-p)M} \tag{1.50}$$

fraction

Using (1.50) we need only compare $\mathbb{E}_\alpha^L(\tau_\alpha)(\frac{1}{\mu}+1)h^{rv(\alpha)}(L+1)$ and M. From Lemma 13, $h^{rv(\alpha)}(L+1) < h^{rv(\alpha)}(s)$ for all $s \leq L$. Since M is the total expected cost given that the first event is a departure, we know M consists only of costs as measured by $h^{rv(\alpha)}(s)$ for $s \leq L$. Hence, we have $\mathbb{E}_\alpha^L(\tau_\alpha)h^{rv(\alpha)}(L+1) < M$; 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 between gain optimal policies. This analysis begs the question, why is the higher control limit preferred? In essence, the choice the decision-maker must make is whether to add more waiting space. If optimal gain is the primary objective, it is clear that if adding this server reduces the gain, it should not be added. On the other hand, if adding the waiting space, does not change the gain, but decreases the average cost as measured by $h^{rv(\alpha)}$ the decision-maker would prefer to add the space. The question of why the relative value functions measure cost remains open. However, we can make the observation that with L as the reference state (so $h^{rv(\alpha)}(L) = 0$), Lemma 13 implies that $h^{rv(\alpha)}(s) < 0$ for $s > L$ while $h^{rv(\alpha)}(s) > 0$ for $s < L$. Thus, the average cost is decreased by time spent with more than L customers in the system. The bias based decision-maker prefers negative relative value functions.

Suppose now we consider Example 6 except that rewards are received upon service completion instead of upon acceptance to the system. Using the bias optimality equation (1.37), the authors [11] showed that if there are two gain optimal control limits, it is in fact the **lower** control limit that is bias optimal. By formulating this problem as in the previous example it is not difficult to show that with L as the reference state, $h^{rv(\alpha)}(s) > 0$ for $s > L$ while $h^{rv(\alpha)}(s) < 0$ for $s < L$. Using precisely the same sample path argument as Example 6, we get that the lower control limit is bias optimal.

This leads us to two conclusions. First, the intuitive idea of the bias as the total reward is quite restrictive since in total reward models the decision-maker is indifferent to when rewards are received. And, secondly the interpretation of the bias as an average cost problem is not sufficient either, since discounting is usually lost in the long-run analysis. In fact, some of the properties of discounting are retained after the limit is taken in the definition of $0-$discount optimality (see (1.10)). This line of discussion is pursued next.

1.5.1 Bias and Implicit Discounting

Implicit discounting in bias allows us to explain why bias prefers control limit L or $L+1$. First note that the decision to accept or reject a customer in Example 6 for the long-run average reward based decision-maker is in essence based on whether or not the reward offered is higher than the cost of having the customer in the system. Thus, when the decision-maker is indifferent, the rewards and costs must balance. When rewards are received at arrivals the reward is received before the cost of having the customer in the system is accrued. On the other hand, when rewards are received upon service completion the decision-maker must accrue the cost of having a customer in the system before receiving the reward. The decision-maker only chooses to increase the amount of waiting space if the reward is received before the cost and the discounting is apparent. The following theorem explains how the bias based decision-maker discounts future rewards.

prop:discount **Theorem 14** *Suppose that α is a positive recurrent state for a fixed policy $d^\infty \in D^\infty$. 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*

$$c_d \;=\; -\frac{\mathbb{E}_\alpha^d \sum_{n=0}^{\tau_\alpha - 1}(n+1)[r(X_n) - w_d]}{\mathbb{E}_\alpha^d \, \tau_\alpha} \qquad (1.51)$$ eq:discount

So

$$h_d \;=\; h_d^{rv(\alpha)} - \frac{\mathbb{E}_\alpha^d \sum_{n=0}^{\tau_\alpha - 1}(n+1)[r(X_n) - w_d]}{\mathbb{E}_\alpha^d \, \tau_\alpha} \qquad (1.52)$$

The bias based decision-maker attempts to maximize (eq:discount 1.51). The factor "$n + 1$" discounts the excess rewards received later in the cycle. Thus, if r exceeds w it is better if it occurs earlier in the cycle when it is multiplied by a smaller factor.

Example 7

Again return to Example ex:main 3 and compute the bias constant using (eq:discount 1.51).

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

Similarly,

$$c_\psi \;=\; -(-1 + 2 \cdot (1) + 3 \cdot (0))/3 = -1/3$$

Notice that when the excess reward is received earlier it is worth more (-1 compared to -2), and when the cost is received earlier, it is more costly than later (1 compared to $2 \cdot 1$). Suppose we write "$d_1 \succ d_2$" if the bias based decision-maker prefers d_1 to d_2. In this example, $\delta \succ \gamma$ since the decision-maker chooses to receive the immediate reward and accrue the later cost. Similarly, $\psi \succ \gamma$; the decision-maker prefers not to accrue the immediate cost, despite the fact that there is a reward to be received later. Finally, comparing ψ to δ yields the following relation $\delta \succ \psi \succ \gamma$.

Precisely the same logic can be applied to the prior queueing example. When the reward is received upon acceptance, the decison-maker is willing to accept the arriving customer and $L+1 \succ L$. On the other hand, when the reward is received upon service completion and therefore discounted, the decision-maker chooses not to accept the customer and the relation is reversed.

1.6 Conclusions and future research

When we began our study of bias we presented a sequence of questions. The relationship of bias to the total reward problem was considered on two levels. First, if there is but one recurrent state, on which the reward is zero, the gain is clearly zero, and the bias is equivalent to the total reward until entering the recurrent state. Suppose now that there is a single recurrent class, but there is more than one state in this class. One might immediately conjecture that we need only subtract the gain from each of the rewards on the transient states, and again compute the total reward until entering the recurrent class. While this leads to a function which satisfies the AEE, the relative value function, it does not equal the bias. The reason this is so, is that the bias includes implicit discounting. Hence, instead of simply computing total reward, we must consider when these rewards are received.

Computationally, we have shown that in the unichain case the bias can be computed by any of the methods used to compute the gain. By noticing that the form of the BOE is exactly the same as that of the AOE, we need not introduce any new methods for computation. This also leads to sample path methods which illuminate the fact that the bias based decision-maker prefers policies that spend more time in states with negative relative values on the recurrent class. The interpretation of this fact is open. However, since this also leads to implicit discounting we may shed a little light on the subject.

Since the relative value function is zero on the chosen reference state, if a state entered after leaving the reference state has a negative relative value, an equivalent, positive excess reward must have previously been received in order to balance the rewards and the gain on the cycle. On the other hand, if a state entered before returning to the reference state has positive relative value, a reward less than the gain must be earned prior to entering that state. This is the crucial point of our analysis of implicit discounting and implies that higher rewards received earlier in the cycle are preferred.

We have restricted ourselves to the unichain finite state and action space case. We feel that it is clear that there is a need to extend each of these ideas to multichain, countable, and general state space cases. It is also important to notice that the discounting bias captures, is only captured on the recurrent states. For discounting on the transient states, one would need to use the next term in the Laurent series expansion. Why this is so also remains unanswered.

Bibliography

article:ara_et_al [1] Aristotle Arapostathis, Vivek S. Borkar, Emmanuel Fernandez-Gaucherand, Mrinal K. Ghosha, and Steven I. Marcus. Discrete-time controlled Markov processes with average cost criterion: A survey. *Siam Journal on Control and Optimization*, 31(2):282–344, March 1993.

dis_opt:black [2] David Blackwell. Discrete dynamic programming. *Annals of Mathematical Statistics*, 33:719–726, 1962.

article:dena [3] Eric V. Denardo. Computing a bias optimal policy in a discrete-time Markov decision problem. *Operations Research*, 18:279–289, 1970.

bias_opt:hav_put [4] Moshe Haviv and Martin L. Puterman. Bias optimality in controlled queueing systems. *Journal of Applied Probability*, 35:136–150, 1998.

book:her_las [5] Onésimo Hernández-Lerma and Jean B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer-Verlag Inc., New York, 1996.

article:her_las [6] Onésimo Hernández-Lerma and Jean B. Lasserre. Policy iteration in average cost Markov control processes on Borel spaces. *Acta Applicandae Mathematicae*, 47:125–154, 1997.

book:her_las2 [7] Onésimo Hernández-Lerma and Jean B. Lasserre. *Further Topics on Discrete-Time Markov Control Processes*. Springer-Verlag Inc., New York, 1999.

book:howard [8] Ronald A. Howard. *Dynamic Programming and Markov Processes*. John Wiley & Sons, Inc., New York, 1960.

article:bias1 [9] Mark E. Lewis, Hayriye Ayhan, and Robert D. Foley. Bias optimality in a queue with admission control. *Probability in the Engineering and Informational Sciences*, 13:309–327, 1999. to appear.

article:bias3 [10] Mark E. Lewis and Martin L. Puterman. A probabilistic analysis of bias optimality in unichain Markov decision processes. 1999. submitted.

article:bias2 [11] Mark E. Lewis and Martin L. Puterman. A note on bias optimality in controlled queueing systems. *Journal of Applied Probability*, 37(1), 2000. to appear.

unif:lipp [12] Steven A. Lippman. Applying a new device in the optimization of exponential queueing systems. *Operations Research*, 23(4):687–712, 1975.

article:mann [13] Elke Mann. Optimality equations and sensitive optimality in bounded Markov decision processes. *Optimization*, 16(5):767–781, 1985.

article:meyn [14] Sean 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.

mdp_book:put [15] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley and Sons, Inc., New York, 1994.

article:sch_fed [16] P. Schweitzer and A. Federgruen. The functional equations of undiscounted Markov renewal programming. *Mathematics of Operations Research*, 3:308–321, 1977.

ticle:vein_nodisc [17] Arthur F. Veinott. On finding optimal policies in discrete dynamic programming with no discounting. *Annals of Mathematical Statistics*, 37(5):1284–1294, October 1966.

article:vein_disc [18] Arthur F. Veinott, Jr. Discrete dynamic programming. *Annals of Mathematical Statistics*, 40(5):1635–1660, October 1969.

chapter:vein [19] Arthur F. Veinott, Jr. Markov decision chains. In *Studies in Optimization*, volume 10 of *Studies in Mathematics*, pages 124–159. Mathematics Association of America, 1974.