3 BIAS OPTIMALITY

Mark E. Lewis
Martin L. Puterman

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" reward. This reward, called the *bias* aids in distinguishing among multiple gain optimal policies. This chapter focuses on establishing the usefulness of the bias in distinguishing among multiple gain optimal policies, computing it and demonstrating the implicit discounting captured by bias on recurrent states.

3.1 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. We consider a special case of the sensitive optimality criteria which are considered in Chapter 8 of this volume.

Suppose the 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 *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 a policy which maximizes the bias.

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. Consider the following simple example:

Example 3.1 Let $\mathbb{X} = \{1, 2\}$, $A_1 = \{a, b\}$, and $A_2 = \{c\}$. Furthermore, let $p(2|1, a) = p(2|1, b) = p(2|2, c) = 1$ and $r(1, a) = 100$, $r(1, b) = 1$, and $r(2, c) = 1$. It is easy to see that an average reward maximizing decision-maker would be indifferent which action is chosen in state 1, but any rational decision-maker would clearly prefer action a in state 1. The analysis in this chapter will show, among other things, that using bias will resolve this limitation of the average reward criterion.

Unfortunately, this example gives an oversimplified perspective of the meaning of bias. 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?