5 AVERAGE REWARD OPTIMIZATION THEORY FOR DENUMERABLE STATE SPACES

Linn I. Sennott

5.1 INTRODUCTION

In this chapter we deal with certain aspects of average reward optimality. It is assumed that the state space $\mathbb{X}$ is denumerably infinite, and that for each $x \in \mathbb{X}$, the set $\mathbb{A}(x)$ of available actions is finite. It is possible to extend the theory to compact action sets, but at the expense of increased mathematical complexity. Finite action sets are sufficient for digitally implemented controls, and so we restrict our attention to this case.

For initial state x, the quantity $W(x)$ is the best possible limiting expected average reward per unit time (*average reward*, for short). This is an appropriate measure of the largest expected reward per unit time that can possibly be achieved far into the future, neglecting short-term behavior. Many interesting applications have the property that the average reward is independent of the initial state, i.e. $W(x)$ is a constant.

This chapter develops a theory to guarantee the existence of a stationary policy ϕ and finite constant W such that

$$W(x) = w(x, \phi) \equiv W, \qquad x \in \mathbb{X}. \tag{5.1}$$

Such a policy is an *average reward optimal* stationary policy. In this chapter a stationary policy means a nonrandomized (pure) stationary policy. Implementing such a policy requires the controller to know only the current state x of the system. Table look-up may then determine the fixed action $\phi(x)$ appropriate in that state.

The development takes place under the assumption that there exists a nonnegative (finite) constant R such that $r(x,a) \le R$, for all $x \in \mathbb{X}$ and $a \in \mathbb{A}(x)$. Note that rewards may be unbounded below. In some applications, the actual reward is a random quantity. In these cases, $r(x,a)$ is to be interpreted as an expected reward.

In a typical reward maximization setting, it may be possible to incur costs as well as earn rewards. Costs can be built into the system as negative rewards. For example, to minimize over the set $\{5,2,8\}$ of costs, we may calculate $\max\{-5,-2,-8\} = -2$, and then the answer is $-(-2) = 2$. Our framework allows rewards to be unbounded below, thereby handling the common case of costs unbounded above. For example, queueing control problems may involve holding costs that are linear in the number of customers. If the buffers are unlimited (able to hold all arriving customers), then this would entail costs unbounded above. The theory does not allow the controller to earn arbitrarily large positive rewards. This is not a severe limitation in queueing control problems and other applications. For example, assume that the controller earns a unit reward each time a customer is admitted to the system. If the number of customers that can arrive in any slot is bounded, then the assumption will hold. If the distribution on customer batch sizes is unbounded, then we may allow the controller to earn a reward that is a function of the mean batch size.

We may define a new reward structure by subtracting R from the rewards in the original system. By so doing, the optimal policy will not be affected, and it will be the case that all rewards are nonpositive. Let us assume that this has already been done, so that for the rest of the chapter we make the following assumption.

Assumption 5.1 *We have $r(x,a) \le 0$, for all $x \in \mathbb{X}$ and $a \in \mathbb{A}(x)$.*

Note that to recover the average reward in the original setting, it is only necessary to add R to W.

To motivate our approach, let us consider the situation when $\mathbb{X}$ is finite. In this case, it is well-known that there exist $\beta_0 \in (0,1)$ and a stationary policy ϕ that is discount optimal for $\beta \in (\beta_0, 1)$. Such a policy is called *Blackwell optimal*, and it must also be average optimal. These claims are proved in the chapter by Hordijk and Yushkevich in this volume; also see Sennott [37, Proposition 6.2.3]. Note that in the general case, $W(x)$ may not be constant. To motivate the assumptions to be introduced in Section 3, we give the following result. It was stated in [37, Proposition 6.4.1] for the cost minimization framework, and the proof may be recast into the reward maximization framework.

Proposition 5.1 *Let $\mathbb{X}$ be finite. The following are equivalent:*

(i) $W(x) \equiv W$, *for* $x \in \mathbb{X}$.

(ii) *There exists $z \in \mathbb{X}$ and a finite constant L such that $|V(x,\beta) - V(z,\beta)| \le L$, for all $x \in \mathbb{X}$ and $\beta \in (0,1)$.*

(iii) *Given $y \in \mathbb{X}$, there exists a finite constant L such that $|V(x,\beta) - V(y,\beta)| \le L$, for all $x \in \mathbb{X}$ and $\beta \in (0,1)$.*