AVERAGE REWARD OPTIMIZATION THEORY FOR DENUMERABLE STATE SPACES

L. I. Sennott
Department of Mathematics
Illinois State University
Normal, IL 61790-4520
U.S.A.
email: sennott@math.ilstu.edu

1 Introduction

In this chapter we deal with certain aspects of average reward optimality. It is assumed that the state space X is denumerably infinite, and that for each $x \in X$, the set $A(x)$ 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

$$W(x) = w(x, \phi) \equiv W, \qquad x \in X. \tag{1}$$

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

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

When X is finite, it is known, e.g. Sennott [34, Proposition 6.2.3], that there exist $\beta_0 \in (0, 1)$ and a stationary policy ϕ that is both discount optimal for $\beta \in (\beta_0, 1)$, and average reward optimal. Note that in the general case, $W(x)$ may not be constant. The following result was shown in [34, Proposition 6.4.1] for the cost minimization framework. The proof may be recast into the reward maximization framework.

Proposition 1 *Let X be finite. The following are equivalent:*

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

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

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

2 Counterexamples

When X is denumerably infinite, the situation is quite different from that when X is finite. An average reward optimal stationary policy may not exist. The following example shows that, in fact, an optimal policy of any sort may not exist.

Example 2 The state space has two "layers." The top layer consists of states $\{1, 2, 3, \ldots\}$ and the bottom layer of states $\{1^*, 2^*, 3^*, \ldots\}$. There is a single action in each bottom state and these are all absorbing, i.e. $p_{x^*x^*} = 1$. There are actions a and b in each top state, with $1 = p_{xx+1}(a) = p_{xx^*}(b)$. There are no rewards in any top state, and $r(x^*) = 1 - \frac{1}{x}$.

Beginning in state x, to achieve a positive reward requires that b be eventually chosen. Assume that b is first chosen in state y, where $y \geq x$. Then from that point on, a reward of $1 - \frac{1}{y}$ per unit time is earned. It is clear (and can be proved) that the rewards earned up to this point do not affect the limiting average reward, which is thus $1 - \frac{1}{y} < 1$. Clearly, $W(x) = 1$. However, no policy achieves an average reward of 1.

The next example shows that, even if an average cost optimal policy exists, it may be nonstationary.

Example 3 Let $X = \{1, 2, 3, \ldots\}$. There are two actions in each state with $1 = p_{xx+1}(a) = p_{xx}(b)$. There is no reward under a and $r(x, b) = 1 - \frac{1}{x}$. In words, we may advance to the next higher state and earn nothing, or remain in x and earn $1 - \frac{1}{x}$. Let us assume that the process begins in state 1 and operates under a stationary policy. If this policy chooses b for the first time in state x, then it must continue to choose b, and the average reward under this stationary policy is $1 - \frac{1}{x}$.

However, consider the nonstationary policy π that operates as follows: Upon first entry into state x, it chooses b a total of x times and then chooses a. The sequence of rewards generated under π is

$$0, 0, \frac{1}{2}, \frac{1}{2}, 0, \frac{2}{3}, \frac{2}{3}, \frac{2}{3}, 0, \frac{3}{4}, \frac{3}{4}, \frac{3}{4}, \frac{3}{4}, 0, \ldots \tag{2}$$

It may be shown that $w(1, \pi) = 1$, and hence π is average reward optimal.

These examples appear in Ross [31]. In both examples, it is the case that there exists a stationary policy achieving an average reward that is within ϵ of optimality. If this were always the case, we would probably be satisfied to know that we could produce a stationary policy with any desired degree of closeness to the optimal value. However [31, p. 91], gives an example for which no stationary policy is within ϵ of optimality.

3 Assumptions for Validity of Eq. (1)

The examples above show that some assumptions are necessary to achieve the goal of (1). One possible form for those assumptions is suggested by the situation when X is finite. In this case, as we have observed, there exists a stationary policy that is discount optimal, for every discount factor sufficiently close to 1, and is also average optimal. We may seek to obtain a similar result when X is denumerable. Since we also want $W(x)$ to be constant, this suggests that the condition in Proposition 1(ii) might be a suitable assumption. It is shown in

Ross [31, p. 95] that this will indeed work, but it turns out that this condition is too strong to enable us to treat many important applications.

The following set of assumptions is based on a generalization of the condition in Proposition 1. These assumptions are the reward version of those in Sennott [34, p. 132] and are a slight modification of those in Puterman [29, Section 8.10.2]. A version of these assumptions is also discussed in Arapostathis, et al [2] and Kitaev and Rykov [26]. Keep in mind the fact that quantities that are automatically finite for X finite may become infinite when X is infinite. Expectations exist because we are assuming that rewards are nonpositive, but they may be $-\infty$.

Let z be a distinguished state.

Assumption B *The quantity* $(1-\beta)V(z,\beta)$ *is bounded, for* $\beta \in (0,1)$.

This implies that $V(z,\beta) > -\infty$ and hence we may define the function $h(x,\beta) =: V(x,\beta) - V(z,\beta)$ without fear of introducing an indeterminate form.

Assumption C *There exists a (finite) nonnegative function* M *such that* $h(x,\beta) \geq -M(x)$, *for* $x \in X$ *and* $\beta \in (0,1)$.

The final assumption is:

Assumption D *There exists a (finite) nonnegative constant* L *such that* $h(x,\beta) \leq L$, *for* $x \in X$ *and* $\beta \in (0,1)$.

Note that $h(z,\beta) \equiv 0$ and hence we may always take $M(z) = 0$. As we will see, Assumption B is related to the requirement that the average reward be finite. Assumptions C and D are basically the condition in Proposition 1(ii), but modified to allow the lower bound to be a function, rather than a constant. Section 10 discusses a weaker set of assumptions that allows the upper bound to also be a function. It turns out that for many applications, the upper bound is constant, and this assumption simplifies the theory.

One may also wonder whether the distinguished state z plays a special role in the assumptions. The answer is no. For the cost minimization framework, it is shown in [34, Proposition 7.2.4] that if Assumptions A - D hold for z, then they hold when z is replaced by any other state.

Here is an important lemma.

Lemma 4 *Let* ϕ *be a stationary policy. Assume that there exist a (finite) constant* W *and a (finite) function* h*, that is bounded above, such that*

$$W + h(x) \leq T^{\phi}h(x), \qquad x \in X, \tag{3}$$

where

$$T^{\phi}h(x) = r(x,\phi(x)) + \sum_{y \in X} p_{xy}(\phi(x))h(y). \tag{4}$$

Then $w(x,\phi) \geq W$, *for* $x \in X$.

Proof. Assume that the process starts in state x, operates under ϕ, and let $X_0 = x, X_1, X_2, \ldots$ be the sequence of values. Then from (3) it follows that

$$W + h(X_t) \leq T^\phi h(X_t), \qquad t \geq 0. \tag{5}$$

We now show that $E_x^\phi[h(X_t)] > -\infty$. In fact, we prove by induction on t that the expectation is bounded below by $tW + h(x)$. This is clearly true for $t = 0$. Now assume that it is true for t . Then from (5) it follows that $E_x^\phi[h(X_{t+1})|X_t] \geq W + h(X_t)$. Taking the expectation of both sides, using a property of expectation (i.e. $E(E[X|Y]) = E[X]$), together with the induction hypothesis, we find that $E_x^\phi[h(X_{t+1})] \geq W + E_x^\phi[h(X_t)] \geq W + tW + h(x) = (t+1)W + h(x)$. This completes the induction.

Now take the expectation of both sides of (5) and rearrange, to obtain

$$E_x^\phi[r(X_t)] \geq W + E_x^\phi[h(X_t)] - E_x^\phi[h(X_{t+1})], \quad t \geq 0. \tag{6}$$

What has just been proved assures us that we have not created the indeterminate form $-\infty + \infty$. Add the terms in (6), for $t = 0$ to $n - 1$, and divide by n to obtain

$$\begin{aligned} \frac{v(x,\phi,n)}{n} &\geq W + \frac{h(x) - E_x^\phi[h(X_n)]}{n} \\ &\geq W + \frac{h(x) - L}{n}. \end{aligned} \tag{7}$$

Here L is the upper bound on h. Taking the limit infimum of both sides of (7) yields the result. ∎

The proper generalization of the finite state space result deals with sequences of discount factors rather than sufficiently large discount factors. The following definition sets up the basic concepts. (It is independent of the assumptions.) We will be taking subsequences of sequences, and for notational convenience, each time this is done, the subsequence will be indexed by n.

Definition 5 (i) *Let z be a distinguished state and assume that the function $h(x,\beta) =: V(x,\beta) - V(z,\beta)$ involves no indeterminate form. Let β_n be a sequence of discount factors converging to 1. (All sequences are assumed to converge to 1 from the left.) If there exist a subsequence δ_n and a function h such that*

$$\lim_{n\to\infty} h(x,\delta_n) = h(x), \qquad x \in X, \tag{8}$$

then h is a limit function (of the sequence $h(-,\beta_n)$).

(ii) *Let $\phi(\beta)$ be a stationary policy realizing the β discount optimality equation, and let $\beta_n \to 1$. Assume that there exist a subsequence δ_n and a stationary policy ϕ such that $\lim_{n\to\infty} \phi(\delta_n) = \phi$. This means that for a given x and sufficiently large n (dependent on x) we have $\phi(\delta_n)(x) = \phi(x)$. Then ϕ is a limit point (of $\phi(\beta_n)$).*

(iii) *Let ϕ be a limit point. The limit function h is associated with ϕ if there exists a sequence β_n such that $\lim_{n\to\infty} h(-, \beta_n) = h$ and $\lim_{n\to\infty} \phi(\beta_n) = \phi$.*

4 The Existence Theorem

The following existence theorem is our major result. It is the average reward counterpart of senb [34, Theorem 7.2.3].

Theorem 6 *Assume that Assumptions A-D hold. Then:*

(i) *There exists a finite constant $W =: \lim_{\beta\to 1}(1-\beta)V(x,\beta)$, for $x \in X$.*

(ii) *There exists a limit function. Any such function h satisfies $-M \le h \le L$ and*

$$W + h(x) \le Th(x), \qquad x \in X. \tag{9}$$

Let ψ be a stationary policy realizing the maximum in (9). Then ψ is average reward optimal with (constant) average reward W and

$$\lim_{n\to\infty} \frac{E_x^{\psi}[h(X_n)]}{n} = 0, \qquad x \in X. \tag{10}$$

(iii) *Any limit point ϕ is average reward optimal. There exists a limit function associated with ϕ. Any such function h satisfies*

$$W + h(x) \le T^{\phi}h(x), \qquad x \in X. \tag{11}$$

and

$$\lim_{n\to\infty} \frac{E_x^{\phi}[h(X_n)]}{n} = 0, \qquad x \in X. \tag{12}$$

(iv) *The average reward under any optimal policy is obtained as a limit.*

Proof. Observe that the theorem encompasses two viewpoints. It says that an optimal stationary policy may be obtained from (9), which is constructed by first obtaining a limit function. Or an optimal stationary policy may be obtained as a limit of discount optimal stationary policies. In any case, the average reward under any optimal policy (stationary or not) is obtained as a limit, rather than a limit infimum.

We first prove (ii). Fix a sequence $\beta_n \to 1$. It follows from Assumptions C - D and [34, Proposition B.6] that there exists a limit function of the sequence $h(-, \beta_n)$, and that any such function h satisfies $-M \leq h \leq L$.

Now fix a limit function h as in (8). Using Assumption A we see that $(1-\delta_n)V(z, \delta_n)$ is a bounded sequence of real numbers. Any such sequence has a convergent subsequence. Hence there exist a subsequence ϵ_n and a (finite) number W such that

$$\lim_{n\to\infty} (1-\epsilon_n)V(z, \epsilon_n) = W. \tag{13}$$

Note that $(1-\beta)V(x,\beta) = (1-\beta)h(x,\beta) + (1-\beta)V(z,\beta)$. Let $\beta = \epsilon_n$ and let $n \to \infty$. The last term approaches W. It follows from (8) and the finiteness of h that the second term approaches 0. Hence

$$\lim_{n\to\infty} (1-\epsilon_n)V(x, \epsilon_n) = W, \qquad x \in X. \tag{14}$$

The discount optimality equation $V = T_\beta V$ may be rewritten as

$$(1-\beta)V(z,\beta) + h(x,\beta) = T_\beta h(x,\beta), \quad x \in X. \tag{15}$$

Now fix a state x and consider the sequence $\phi(\epsilon_n)(x)$ of discount optimal actions in x. Because the action set $A(x)$ is finite, it is the case that there exist an action $a(x)$ and a subsequence γ_n (dependent on x) such that $\phi(\gamma_n)(x) \equiv a(x)$. For the fixed state x and $\beta = \gamma_n$, (15) becomes

$$(1-\gamma_n)V(z,\gamma_n) + h(x,\gamma_n) = T_{\gamma_n}^{a(x)} h(x,\gamma_n). \tag{16}$$

Take the limit supremum of both sides of (16) as $n \to \infty$. Use (8), (13), and the limit supremum version of Fatou's lemma [34, Proposition A.2.1] to obtain

$$\begin{aligned} W + h(x) &\leq T^{a(x)} h(x) \\ &\leq Th(x).* \end{aligned} \tag{17}$$

Because this argument may be repeated for each x, it follows that (9) holds.

Now let ψ be a stationary policy realizing the maximum in (9). Then (3) holds for ψ. To prove that ψ is optimal, let π be an arbitrary policy and fix an initial state x. Then

$$w(x,\psi) \geq W \geq \liminf_{\beta\to 1} (1-\beta)V(x,\beta) \geq \liminf_{\beta\to 1} (1-\beta)v(x,\pi,\beta) \geq w(x,\pi). \tag{18}$$

The leftmost inequality follows from Lemma 4. The next inequality follows from (14) and the definition of the limit infimum. The next inequality follows since $V(-,\beta) \geq v(-,\pi,\beta)$, and the rightmost inequality follows from (33) in the chapter Appendix. This proves that ψ is average reward optimal. Moreover by setting$\pi = \psi$ we see that $w(x,\psi) \equiv W$ and hence W is the maximum average reward.

Recall that the whole argument was carried out with respect to the sequence β_n. Given this sequence we obtained a subsequence such that (14) holds for the maximum average reward W. This means that given any sequence, there exists a subsequence such that (14) holds for the fixed value W. This implies that the limit exists and hence (i) holds.

We prove (iv) and then return to the proof of (10). To prove (iv) let π be an arbitrary average reward optimal policy. Note that all that is assumed is that $W \equiv w(x,\pi)$. We have

$$\begin{aligned} W = \lim_{\beta\to 1}(1-\beta)V(x,\beta) &\geq \limsup_{\beta\to 1}(1-\beta)v(x,\pi,\beta) \\ &\geq \liminf_{\beta\to 1}(1-\beta)v(x,\pi,\beta) \geq W. \end{aligned} \tag{19}$$

The leftmost equality follows from (i). The rightmost inequality follows from (33) and the optimality of π. Hence all the terms in (19) are equal to W and it follows that $\lim_{\beta\to 1}(1-\beta)v(x,\pi,\beta)$ exists. Then (iv) follows from Proposition 22 in the Appendix.

Let us now prove (10). Using the optimality of ψ and (iv) it follows that we may take the limit of both sides of (7) to obtain (10).

The proof of (iii) is similar to the proof of (ii) and we omit it. (See [34, p. 137].) ∎

No implication concerning the structure of the Markov chain induced by an optimal stationary policy can be drawn from the assumptions. To see that this is the case, consider a process with any desired transition structure whatsoever and with identically 0 rewards. Then the assumptions hold and all policies are optimal.

The cost minimization analog of the assumptions are denoted (SEN) in [34]. A related set (SCH) of assumptions was introduced by Schal [32]. Problem 7.6 of [34] claims that (SEN) $\Leftrightarrow$ (SCH), and hence these assumptions sets may be shown to be equivalent.

Eq. (9) is the *average reward optimality inequality* (AROI). The next section explores conditions under which it will be an equality, yielding the *average reward optimality equation* (AROE).

5 The Average Reward Optimality Equation

It is possible for the inequality in (9) to be strict. Cavazos-Cadena [9] presents an example for which this is the case. However, in "normal situations", (9) is an equality. This section gives very weak conditions for the AROE to be valid.

We first develop some notation. Let G be a nonempty subset of X. Then $\mathcal{R}(x, G)$ is the set of policies π satisfying the following: Beginning in x and following π, the process will enter G at some time $t \geq 1$ with probability 1, and the expected time $m_{xG}(\pi)$ of a first passage from x to G is finite. We let $\mathcal{R}^*(x, G)$ be the subset of $\mathcal{R}(x, G)$ consisting of policies π such that the expected (total) reward $r_{xG}(\pi)$ of the first passage is also finite. If $G = \{y\}$, then $\mathcal{R}(x, G)$ (respectively, $\mathcal{R}^*(x, G)$) is denoted $\mathcal{R}(x, y)$ (respectively, $\mathcal{R}^*(x, y)$).

Proposition 7 *Assume that Assumptions A - D hold and let ϕ be a stationary policy realizing the maximum in (9). The AROI is an equality at a fixed state x under any of the following conditions:*

(i) *There exists a finite set G such that $\phi \in \mathcal{R}(x, G)$. This also implies that $\phi \in \mathcal{R}^*(x, G)$ and $h(x) = r_{xz}(\phi) - Wm_{xz}(\phi) + E_x^\phi[h(X_T)]$, where T is the time of a first passage to G.*

(ii) *We have $\phi \in \mathcal{R}(x, z)$. This also implies that $\phi \in \mathcal{R}^*(x, z)$ and $h(x) = r_{xz}(\phi) - Wm_{xz}(\phi)$.*

(iii) *The Markov chain induced by ϕ is positive recurrent at x.*

(iv) *We have $p_{xy}(\phi(x)) > 0$ for only finitely many values of y.*

Proof. We omit the proof of (i). A generalization of (i), with proof, is given in [34, Theorem 7.4.3] for the average cost framework. If we grant the truth of (i), then (ii) follows immediately by setting $G = \{z\}$. Similarly, (iii) follows by setting $G = \{x\}$, and (iv) follows by setting $G = \{y | p_{xy}(\phi(x)) > 0\}$. $\blacksquare$

Assume that the process operates under an optimal stationary policy determined by (9). From Proposition 7, we see that the AROI can be strict at x only if the process does not reach any finite set in a finite expected amount of time. A system with this property is unlikely to arise in applications. The impetus for Proposition 7 came from Cavazos-Cadena [8].

6 A Sufficient Condition for Assumption C

This section presents a sufficient condition for Assumption C to hold and then uses this to show how the assumptions can be verified in an example.

Proposition 8 *Assume that Assumption A holds. Let z be the distinguished state, and assume that $V(z, \beta) > -\infty$, for $\beta \in (0, 1)$. Given $x \neq z$, assume that there exists a policy $\pi_x \in \mathcal{R}^*(x, z)$. Then $h(x, \beta) \geq r_{xz}(\pi_x)$, and hence Assumption C holds with $M(x) = -r_{xz}(\pi_x)$.*

Proof. If the process begins in state $x \neq z$ and follows π_x, it will reach state z at some time in the future. Let T be a random variable denoting this time.

Let the policy π follow π_x until z is reached, and then follow a discount optimal policy $\phi(\beta)$. Then

$$\begin{aligned} V(x,\beta) &\geq v(x,\pi,\beta) \\ &= E_x^\pi\left[\sum_{t=0}^{T-1}\beta^t r(X_t,A_t)\right] + E_x^\pi[\beta^T]V(z,\beta) \\ &\geq E_x^\pi\left[\sum_{t=0}^{T-1} r(X_t,A_t)\right] + V(z,\beta) \\ &= r_{xz}(\pi_x) + V(z,\beta). \end{aligned} \tag{20}$$

The validity of the second inequality follows from Assumption A. The result that follows by subtracting $V(z,\beta)$ from both sides. ∎

We now give an example. All the examples take place in discrete time.

Example 9 Figure 1 shows the structure. We have a single server queue, and the service time of a customer is geometrically distributed with success parameter $\mu \in (0,1]$. There are Bernoulli arrival processes $k = 1, 2, \ldots, K$, and process k has parameter $p_k \in (0,1]$.

The state of the system is the number $x \geq 0$ of customers currently in the system. In each state and each time slot, the action set is $\{0, 1, \ldots, K\}$. Action $k, 1 \leq k \leq K$, allows arrival process k to operate. In this case, with probability p_k, one customer will enter the system, and with probability $1 - p_k$, no customer arrives . Let us adopt the convention that the customer arrives sometime during the slot, so that, if it arrives to an empty buffer, it will enter service at the beginning of the next slot. Choosing action 0 bars the system from new customers for that slot.

There is an increasing holding cost $H(x)$, where $H(0) = 0$. If process k is chosen, then a positive reward $R(k)$ is earned. If action 0 is chosen, then we set $R(0) = 0$. If R is the maximum possible reward, then we may set this up as a reward maximization problem by defining $r(x,k) = -H(x) + R(k) - R, 0 \leq k \leq K$.

Let us argue informally that Assumptions B - D hold, with distinguished state 0. Let ϕ be the (stationary) policy that always chooses 0. In this case, no new customers can enter the system and eventually the process will reach state 0 and remain there. In state 0 the reward is $-R$, and hence $v(0,\phi,\beta) = -\frac{R}{1-\beta}$. Then $0 \geq (1-\beta)V(0,\beta) \geq (1-\beta)v(0,\phi,\beta) = -R$, and Assumption B holds.

For $x \geq 1$, it is easy to see that $v(x,\phi,\beta) > -\infty$, and hence $V(x,\beta)$ is finite. Moreover, $m_{x0}(\phi) = \frac{x}{\mu}$ and $r_{x0}(\phi) = -[H(x) + H(x-1) + \ldots + H(1) + xR]/\mu$, and hence $\phi \in \mathcal{R}^*(x,0)$. It follows from Proposition 8 that Assumption C holds.

We may prove by induction on the finite horizon that $V(0,\beta) \geq V(x,\beta)$. This is intuitively clear since every state has the same choices and the holding costs are increasing in the state. Granted this, we see that Assumption D holds with $L = 0$.

It then follows from Theorem 6 that an average reward optimal stationary policy may be determined from the AROI. Since Proposition 7 (iv) holds, the AROE is valid. We may conjecture that an optimal policy will choose 0 for sufficiently large x.

Example 10 This is a modification of Example 9. Figure 2 shows the structure. The arrival streams are as in Example 9. However, each stream has its own queue and server, with stream k being served at geometric rate μ_k.

The state of the system is the vector $\mathbf{x}$ of buffer occupancies. In each state and each time slot, an allowable action is a vector $\mathbf{a}$, such that a_k equals 1 if the kth process is activated and 0 if it is not.

There is an increasing holding cost $H_k(x_k)$ on the content of buffer k, with zero cost for an empty buffer. Let $H(\mathbf{x}) = \sum_k H_k(x_k)$. If $a_k = 1$, then a positive reward R_k is earned. Let $R = \sum_k R_k$ be the maximum possible reward, and let $R(\mathbf{a}) = \sum_k R_k a_k$ be the reward earned under action $\mathbf{a}$. We may set this up as a reward maximization problem by defining $r(\mathbf{x}, \mathbf{a}) = -H(\mathbf{x}) + R(\mathbf{a}) - R$.

Let the distinguished state be $\mathbf{0}$, and let ϕ be the (stationary) policy that always chooses $\mathbf{a} = \mathbf{0}$. In this case, no new customers can enter the system and eventually the process will enter state $\mathbf{0}$ and remain there. The argument showing that Assumptions B - C hold is similar to that for Example 9.

It is intuitively clear that the best situation for the system is to be in state $\mathbf{0}$. This is so because there are the same actions in each state, and the holding cost is minimized in $\mathbf{0}$. So we have $V(\mathbf{0}, \beta) \geq V(\mathbf{x}, \beta)$. A formal induction proof on the finite horizon may be given to justify this claim. Granted this, we see that Assumption 3 holds with $L = 0$.

7 A Sufficient Condition for Assumptions B - C

In Example 9, the presence of the non-admit action aided us in verifying the assumptions. What if the controller must always choose among the active arrival streams? In this section, we give a sufficient condition for Assumptions B and C to hold, and then show how this condition is useful in the modified version of Example 9.

Definition 11 *Let z be a distinguished state. A z standard policy is a (randomized) stationary policy ϕ such that $m_{xz}(\phi)$ and $r_{xz}(\phi)$ are finite, for all $x \in X$.*

The implications of Definition 11 are: (1) The Markov chain induced by ϕ has a single positive recurrent class S containing z; (2) The expected time and reward to reach the class from any $x \notin S$ are finite; (3) The average reward is a constant $w(\phi)$, for all initial states, and

$$w(\phi) = \sum_{y \in S} q(y, \phi) r(y, \phi(y)), \tag{21}$$

where $q(-,\phi)$ is the steady state distribution. If ϕ is randomized, then the last term in (21) is the expected reward associated with state y. For details on this concept and other results concerning Markov chains used throughout the chapter, see [34, Appendix C].

This concept provides a sufficient condition for Assumptions B and C to hold.

Proposition 12 *Assume that Assumption A holds, and let z be the distinguished state. If there exists a z standard policy, then Assumptions B - C hold.*

Proof. Let ϕ be a z standard policy. Then

$$w(\phi) = \sum_{y \in S} q(y,\phi) E_y^{\phi}[r(X_n)], \qquad n \geq 0. \tag{22}$$

This is easily proved by induction on n. It holds for $n = 0$ by (21). For full details, see [34, p. 299].

Multiplying both sides of (22) by β^n and summing over n yields

$$w(\phi) = (1-\beta) \sum_{y \in S} q(y,\phi) v(y,\phi,\beta). \tag{23}$$

It follows from (23) that $w(\phi) \leq (1-\beta) q(z,\phi) v(z,\phi,\beta)$. Hence

$$w(\phi) q^{-1}(z,\phi) \leq (1-\beta) v(z,\phi,\beta) \leq (1-\beta) V(z,\beta) \leq 0. \tag{24}$$

and Assumption B holds. The quantity on the left of (24) equals $r_{zz}(\phi)$.

The validity of Assumption C now follows from Proposition 8. ∎

Example 13 This is Example 9, modified to remove the 0 action, so that the controller must choose among the active streams. We assume that $p_1 < \mu < 1$, and note that $c =: [(1-\mu)p_1]/[\mu(1-p_1)] < 1$. Assume that

$$\sum_{x=1}^{\infty} H(x) c^x < \infty. \tag{25}$$

Assumptions A and D will hold as before. Let ϕ be the stationary policy that always chooses the first stream. If we can show that ϕ is 0 standard, then the validity of Assumptions B - C will follow from Proposition 12.

It is clear that ϕ induces an irreducible Markov chain on X. If this chain is positive recurrent with finite average reward, then it will follow that ϕ is 0 standard. It is shown in [34, Proposition 8.5.1] that

$$q(0,\phi) = 1 - \frac{p_1}{\mu}, \qquad q(x,\phi) = \left(\frac{q(0,\phi)}{1-\mu} \right) c^x, \quad x \geq 1. \tag{26}$$

Then $w(\phi) = R_1 - R - \left(\frac{q(0,\phi)}{1-\mu}\right) \sum H(x)c^x > -\infty$, where the finiteness follows from (25). Hence the assumptions hold.

senb If the arrival streams are not Bernoulli, various Lyapunov techniques (e.g. [34, Appendix C]) may be used to prove that ϕ is 0 standard. In particular, if the mean of the first arrival stream is less than $\mu(< 1)$, the $(n+1)$th moment of this stream is finite, and $H(x)$ is bounded by a polynomial of degree n, then the result will still hold.

8 Verifying Assumption D

In Examples 9, 10, and 13, we verified Assumption D by finding a state that maximized the value of V. Choosing that state as the distinguished state then verified Assumption D with $L = 0$. Here is a generalization of this technique.

Proposition 14 *Assume that Assumption A holds, and that Assumptions B - C hold for distinguished state z. Assume:*

(i) *There exists a finite set G of states, containing z, such that, for $x \notin G$ and $\beta \in (0,1)$, there exists $y \in G$ with $V(y,\beta) \geq V(x,\beta)$.*

(ii) *Given $y \in G - \{z\}$, there exists a policy $\pi_y \in \mathcal{R}^*(z,y)$.*

Then Assumption D holds.

Proof. We claim that the nonnegative quantity

$$L =: \max_{y \in G - \{z\}} [-r_{zy}(\pi_y)] < \infty \tag{27}$$

will work to verify Assumption D. To see this, fix $x \neq z$ and $\beta \in (0,1)$. If $x \notin G$, let y be as in (i). Whereas, if $x \in G$, let $y = x$. Let the process start in z and follow the policy π_y until y is reached, and then follow a β discount optimal stationary policy. Using similar reasoning to that in (20), we have

$$\begin{aligned} V(z,\beta) &\geq r_{zy}(\pi_y) + V(y,\beta) \\ &\geq -L + V(x,\beta). \end{aligned} \tag{28}$$

and hence Assumption D holds. ∎

It is possible to generalize Proposition 14 to allow G to be infinite, if we can show that L defined in (27) as a supremum, is finite. Here is an example using Proposition 14.

Example 15 This is a modification of Example 10. In this case, if $x_k = 0$, then we must set $a_k = 1$. That is, if a buffer is empty, then we must activate the arrival process for that buffer. Otherwise, the model remains the same. Notice that the allowable set of actions now depends on the state.

Let ϕ be the stationary policy that chooses $a_k = 0$ when $x_k \geq 1$. When a buffer is empty, it must admit, but as soon as a customer enters the system and begins service, arrivals are rejected until that service is finished. Hence each buffer contains either 0 or 1 customer. Indeed, the positive recurrent class is $S = \{\mathbf{x} | x_k = 0, 1\}$. It is readily seen that ϕ is $\mathbf{0}$ standard, and hence by Proposition 12, it follows that Assumptions B - C hold.

A bit of thought convinces us that V takes on its maximum value in the finite set S. The reason is that, for any state $\mathbf{x}$ with $x_1 > 1$, the controller is in a better position as x_1 decreases to 1, while holding the other coordinates constant. Once this has been done, the same reasoning can be given for x_2, etc. until we reach a state in S. We may set $\pi_{\mathbf{y}} = \phi$, for $\mathbf{y} \in S - \{\mathbf{0}\}$. Hence it follows from Proposition 14 that Assumption D holds.

9 A Stronger Set of Assumptions

This section presents a sufficient "non-structural" set of conditions for the assumptions.

Proposition 16 *Assume that Assumption A holds, and let z be a distinguished state. Assume:*

(i) *There exists a z standard policy ϕ with positive recurrent class S.*

(ii) *There exists $\epsilon > 0$ such that $G = \{x | r(x, a) \geq w(\phi) - \epsilon$ for some $a\}$ is a finite set.*

(iii) *Given $y \in G - S$, there exists a policy $\pi_y \in \mathcal{R}^*(z, y)$.*

Then Assumptions B - D hold. Moreover:

1. *The AROE holds.*

2. *The Markov chain induced by an optimal stationary policy ψ has at least one positive recurrent state in the set $G(\psi) = \{x | r(x, \psi(x)) \geq W - \epsilon\}$. Let $S(\psi)$ be the set of positive recurrent states. The number of positive recurrent classes making up $S(\psi)$ cannot exceed $|G(\psi)|$, and there are no null recurrent classes.*

3. *If ψ realizes the maximum in the AROE, then $\psi \in \mathcal{R}^*(x, G(\psi) \cap S(\psi))$, for all x. Hence, if $S(\psi)$ consists of a single class, then ψ is y standard, for $y \in S(\psi)$.*

Proof. It follows from (i) and Proposition 12 that Assumptions B - C hold. Suppose we can show

(*): V takes on its maximum in the finite set G given in (ii).

We may then apply Proposition 14 to show that Assumption D holds. (Add z to the set G if necessary. For $y \in (G \cap S) - \{z\}$, set $\pi_y = \phi$.)

To show (*), let us fix $\beta \in (0,1)$ and suppress it in our notation. We first let σ be any (randomized) stationary policy and choose $x \notin G$. Let T be the time of a first passage from x to G, under σ. For notational purposes, let $\alpha = \frac{w(\phi)-\epsilon}{1-\beta}$. Then we have

$$v(x,\sigma) \le E_x^\sigma[\alpha I(T=\infty) + \{\alpha(1-\beta^T) + \beta^T v(X_T,\sigma)\} I(T<\infty)]. \tag{29}$$

This follows since $w(\phi)-\epsilon$ is an upper bound on the rewards outside of G.

Since the expression on the right of (23) is a convex combination, it follows from (23) that there exists $y \in S$ such that $w(\phi) \le (1-\beta)v(y,\phi)$. We claim that

$$w(\phi) \le (1-\beta)v(i,\phi), \text{ for some } i \in G. \tag{30}$$

This is proved by contradiction. Assume that (30) fails. Use (29) with $\sigma = \phi$ and $x = y$ to obtain a contradiction. (Note that $I(T=\infty) = 0$.)

Since G is finite, it follows that there exists $j \in G$ that maximizes the value of V for initial states in G. Then from (30) it follows that

$$\frac{w(\phi)}{1-\beta} \le V(j). \tag{31}$$

Let us now begin the process in $x \notin G$ and operate under the discount optimal stationary policy $\phi(\beta)$. Applying (29) with $\sigma = \phi(\beta)$ and using (31) yields $V(x) \le V(j)$. This shows that V takes on its maximum in G, and proves that the assumptions hold.

Proposition 16 is the average reward version of [34, Theorem 7.5.6]. The rather lengthy argument for the validity of (1)-(3) is given there for the average cost case, and we omit the proof. ∎

The following remarks discuss background and also some subtle ramifications of the conditions in Proposition 16.

Remark 17 *Proposition 16 (i-iii) is a version of an assumption set originally developed by Borkar [3, 4, 5, 6] and denoted (BOR) in [34]. The proof that (BOR) $\Rightarrow$ (SEN) originally appeared in Cavazos-Cadena and Sennott [11]. Assume that (BOR) holds, and let ψ be an optimal stationary policy. From (2) it follows that the Markov chain induced by ψ has at least one positive recurrent state in $G(\psi)$, and no null recurrent classes. Thus all non-positive recurrent states must be transient. It is shown in Sennott [33] that the probability of reaching a positive recurrent class from any transient state is 1. However, an example is given showing that the expected time of such a first passage may be infinite. Of course, by (3) this behavior cannot occur if ψ realizes the AROE.*

Remark 18 *There is a slightly weaker assumption set due to Stidham and Weber [36], and denoted (WS) in [34]. The only change is that (ii) appears without the ϵ. It is the case that (BOR) $\Rightarrow$ (WS) $\Rightarrow$ (SEN). Under (WS) it is still possible to prove the corresponding version of (2) which has the ϵ omitted. However, interestingly enough, (3) may fail. Problem 7.7 in [34] asks the reader for such a construction, and this is available from the author.*

cava, cavb, cavd

The following corollaries of Proposition 16 are due to Cavazos-Cadena [7, 8, 10]. In each case, it is easily shown that conditions (i-iii) of Proposition 16 hold. These results are very useful when the rewards are unbounded outside of finite sets.

Corollary 19 *Assume that Assumption A holds, and let z be a distinguished state. Assume:*

(i) *There exists a z standard policy ϕ with positive recurrent class S.*

(ii) *Given a positive number U, the set $G(U) = \{x | r(x, a) \geq -U$ for some $a\}$ is finite.*

(iii) *Given $y \in X - S$, there exists a policy $\pi_y \in \mathcal{R}^*(z, y)$.*

Then the conclusions of Proposition 16 hold.

Corollary 20 *Assume that Assumption A holds. Assume that there exists a standard policy ϕ such that $S = X$. If, for each positive number U, the set $G(U) = \{x | r(x, a) \geq -U$ for some $a\}$ is finite, then the conclusions of Proposition 16 hold.*

The last example shows how Corollary 20 may be applied.

Example 21 Consider a polling system as in Figure 3. Stations $1, 2, \ldots, K$ are arranged in a ring. We will be dealing with a number of distributions, and for each one, we assume that the parameter of that distribution lies in the interval $(0, 1]$. Each station has an infinite buffer and the arrival stream to station k is Bernoulli with parameter p_k. The service time of a customer at station k follows a geometric distribution with rate μ_k. The server travels around the ring counterclockwise from station 1 to station 2, etc., and finally back to station 1. The walking time for the server to get from station $k - 1$ to station k is geometrically distributed with rate ω_k. Note that station 0 is station K. The set-up time at station k (which occurs after a walk terminating at k) is geometrically distributed with rate δ_k. The arrival processes, service times, walking times, and set-up times are all independent.

The state of the system is a vector $(\mathbf{x}, k, z)$. Here $\mathbf{x}$ is the K vector of buffer occupancies. The quantity k indicates the number of the station currently involved, and $z = 0, 1, 2$ indicates the condition of the server. Here $z = 0$ means that the server is already set-up at k (and ready to serve customers if the buffer is nonempty); $z = 1$ means that the server is setting up at k; and $z = 2$ means that the server is walking from $k - 1$ to k.

A choice of action is available only when $z = 0$ or 1. The action set is $A = \{a, b\}$, where $a =$ remain at the present station, and $b =$ initiate a walk. We are assuming that if the server is walking to a station, then the walk must be completed. However, when the station is reached and set-up is begun, it may be aborted at any time.

A holding cost of $H_k x$ is incurred on a buffer content of x at station k, where $H_k > 0$. The objective is to minimize the expected average holding cost. This may be cast as a reward maximization problem by setting the reward equal to $-\sum H_k x_k$.

We assume that

$$\sum_{k=1}^{K} \frac{p_k}{\mu_k} < 1. \tag{32}$$

This is the condition for stability of the polling system under a stationary policy known as *exhaustive service*, denoted ϕ. This policy operates as follows. If the server arrives to an empty station, it immediately initiates a walk to the next station. If it arrives to a station with customers, it sets up and serves customers at that station until the buffer empties, and it then walks to the next station. Note that when the system is empty the server will continually cycle until a customer enters the system.

It is easy to see that ϕ induces an irreducible Markov chain on X. Moreover, it may be shown that the chain is positive recurrent with finite average reward. It then follows immediately from Corollary 20 that the conclusions of Proposition 16 hold.

Most of the work on polling systems has been done for continuous time systems. Under the assumptions of Poisson arrivals and exponential service times, the stability of the exhaustive service policy under (32) was derived heuristically by Takagi [37] and rigorously by Altman et.al [1] and Georgiadis and Szpankowski [18]. See also Fricher and Jaibi [17]. Takagi [38, 39] are useful survey articles containing many references.

10 Weakening the Assumptions

It is possible to weaken the assumptions by allowing the constant L in Assumption D to be a function. This necessitates several additional assumptions. This development, for the cost minimization framework, is given in [34, Sec. 7.7]. The resulting set of assumptions is denoted (H). They are related to a line of development due to Hordijk [23, 24] and also to Hu [25]. Also see Spieksma [35]. Example 7.7.4 of [34] is a priority queueing system satisfying (H) but for which (SEN) may not hold.

11 Appendix

For a given policy π and initial state x, let $w^*(x, \pi)$ be the limit supremum of the expected average rewards. That is, $w^*(x, \pi)$ is defined as in (–) but with lim inf replaced by lim sup. The following result provides a crucial link between the discounted value function under π and the lim inf and lim sup expected average rewards under π.

Proposition 22 *For any policy π and initial state x we have*

$$w(x,\pi) \le \liminf_{\beta\to 1}(1-\beta)v(x,\pi,\beta) \le \limsup_{\beta\to 1}(1-\beta)v(x,\pi,\beta) \le w^*(x,\pi). \quad (33)$$

For any $x \in X$ the following are equivalent:

(i) *All the terms in (33) are equal and finite.*

(ii) $w(x,\pi) = w^*(x,\pi) > -\infty$, *and hence the quantity in ( ) is obtained as a limit.*

(iii) $\lim_{\beta\to 1}(1-\beta)v(x,\pi,\beta)$ *exists and is finite.*

Proof. This result is stated for the cost minimization case as Proposition 6.1.1 in [34] and a complete proof appears in Sec. A.4 of [34]. The statement of (33) for the continuous case appears in Widder [41]. Granting (33), it is easy to see that the only non-trivial implication is (iii) $\Rightarrow$ (ii). The original proof of this is due to Karamata (see Titchmarsch [40]). This proof is adapted and fully explicated in [34]. ∎

12 Bibliographic Notes

Important early work was done by Derman [12, 13]. Stronger assumptions than those in this chapter were developed by Federgruen and others [14, 15, 16]. Also see Lippman [27] and Wijngaard [42].

A line of development has extended some of these results to the general state space case, see Hernandez-Lerma and others [19, 20, 21, 22, 28]. Also see Ritt and Sennott [30].

Detailed discussions of prior work occur in Arapostathis et al [2], Puterman [29] and Sennott [34].

Bibliography

alt [1] E. Altman, P. Konstantopoulos, and Z. Liu, "Stability, monotonicity and invariant quantities in general polling systems," *Queueing Sys.* **11**, 35-57, 1992.

ara [2] A. Arapostathis, V. Borkar, E. Fernandez-Gaucherand, M. Ghosh, and S. Marcus,"Discrete-time controlled Markov processes with average cost criterion: a survey," *SIAM J. Control Optim.* **31**, 282-344, 1993.

bora [3] V. Borkar, "On minimum cost per unit time control of Markov chains,"*SIAM J. Control Optim.* **22**, 965-978, 1984.

borb [4] V. Borkar, "Control of Markov chains with long-run average cost criterion," in *Stochastic Differential Systems, Stochastic Control Theory and Applications*, edited by W. Fleming and P. L. Lions, Springer-Verlag, New York, 1988.

borc [5] V. Borkar, "Control of Markov chains with long-run average cost criterion: the dynamic programming equations," *SIAM J. Control Optim.* **27**, 642-657, 1989.

bord [6] V. Borkar, *Topics in Controlled Markov Chains*, Pitman Research Notes in Mathematics No. 240, Longman Scientific-Wiley, New York, 1991.

cava [7] R. Cavazos-Cadena, "Weak conditions for the existence of optimal stationary policies in average Markov decision chains with unbounded costs," *Kybernetika* **25**, 145-156, 1989.

cavb [8] R. Cavazos-Cadena, "Solution to the optimality equation in a class of Markov decision chains with the average cost criterion," *Kybernetika* **27**, 23-37, 1991.

cavc [9] R. Cavazos-Cadena, "A counterexample on the optimality equation in Markov decision chains with the average cost criterion," *Sys. Control Letters* **16**, 387-392, 1991.

cavd [10] R. Cavazos-Cadena, "Recent results on conditions for the existence of average optimal stationary policies," *Ann. Op. Res.* **28**, 3-27, 1991.

cave [11] R. Cavazos-Cadena and L. Sennott, "Comparing recent assumptions for the existence of average optimal stationary policies," *Op. Res. Letters* **11**, 33-37, 1992.

dera [12] C. Derman, "Denumerable state Markovian decision processes–average cost criterion," *Ann. Math. Stat.* **37**, 1545-1553, 1966.

derb [13] C. Derman, *Finite State Markovian Decision Processes*, Academic, New York, 1970.

feda [14] A. Federgruen and H. Tijms, "The optimality equation in average cost denumerable state semi-Markov decision problems, recurrency conditions and algorithms," *J. Appl. Prob.* **15**, 356-373, 1978.

fedb [15] A. Federgruen, A. Hordijk, and H. Tijms, "Denumerable state semi-Markov decision processes with unbounded costs, average cost criterion," *Stoc. Proc. Appl.* **9**, 223-235, 1979.

fedc [16] A. Federgruen, P. Schweitzer, and H. Tijms, "Denumerable undiscounted semi-Markov decision processes with unbounded costs," *Math. Op. Res.* **8**, 298-313, 1983.

fri [17] C. Fricker and M. Jaibi, "Monotonicity and stability of periodic polling models," *Queueing Sys.* **15**, 211-238, 1994.

geo [18] L. Georgiadis and W. Szpankowski, “Stability of token passing rings,” *Queueing Sys.* **11**, 7-33, 1992.

hera [19] O. Hernandez-Lerma and J. Lasserre, “Average cost optimal policies for Markov control processes with Borel state space and unbounded costs,” *Sys. Control Letters* **15**, 349-356, 1990.

herb [20] O. Hernandez-Lerma, “Average optimality in dynamic programming on Borel spaces–unbounded costs and controls,” *Sys. Control Letters* **17**, 237-242, 1991.

herc [21] O. Hernandez-Lerma, “Existence of average optimal policies in Markov control processes with strictly unbounded costs,” *Kybernetika* **29**, 1-17, 1993.

herd [22] O. Hernandez-Lerma and J. Lasserre, *Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1996.

hora [23] A. Hordijk, “Regenerative Markov decision models,” *Math. Prog. Study* **6**, 49-72, 1976.

horb [24] A. Hordijk, *Dynamic Programming and Markov Potential Theory* Second Ed., Mathematisch Centrum Tract 51, Amsterdam, 1977.

hu [25] Q. Hu, “Discounted and average Markov decision processes with unbounded rewards: new conditions,” *J. Math. Anal. Appl.* **171**, 111-124, 1992.

kit [26] M Kitaev and V. Rykov,*Controlled Queueing Systems*, CRC Press, Boca Raton, 1995.

lip [27] S. Lippman, “On dynamic programming with unbounded rewards,” *Man. Sci.* **21**, 1225-1233, 1975.

mon [28] R. Montes-de-Oca and O. Hernandez-Lerma, “Conditions for average optimality in Markov control processes with unbounded costs and controls,” *J. Math. Sys. Estimation and Control* **4**, 1-19, 1994.

put [29] M. Puterman, *Markov Decision Processes*, Wiley, New York, 1994.

rit [30] R. Ritt and L. Sennott, “Optimal stationary policies in general state space Markov decision chains with finite action sets,” *Math. Op. Res.* **17**, 901-909, 1992.

ros [31] S. Ross, *Introduction to Stochastic Dynamic Programming*, Academic, New York, 1983.

sch [32] M. Schal, “Average optimality in dynamic programming with general state space,” *Math. Op. Res.* **18**, 163-172, 1993.

sena [33] L. Sennott, “The average cost optimality equation and critical number policies,” *Prob. Eng. Info. Sci.* **7**, 47-67, 1993.

senb [34] L. Sennott, *Stochastic Dynamic Programming and the Control of Queueing Systems*, Wiley, New York, 1999.

spi [35] F. Spieksma, *Geometrically Ergodic Markov Chains and the Optimal Control of Queues*, Ph.D. thesis, Leiden University, 1990.

sti [36] S. Stidham, Jr. and R. Weber, "Monotonic and insensitive optimal policies for control of queues with undiscounted costs," *Op. Res.* **87**, 611-625, 1989.

taka [37] H. Takagi, *Analysis of Polling Systems*, MIT, Cambridge, 1986.

takb [38] H. Takagi, "Queueing analysis of polling models: an update," in *Stochastic Analysis of Computer and Communication Shystems*, edited by H. Takagi, North Holland, New York, 1990.

takc [39] H. Takagi, "Queueing analysis of polling models: progress in 1990-1994," in *Frontiers in Queueing*, edited by J. Dshalalow. CRC Press, Boca Raton, 1997.

tit [40] E. Titchmarsh, *Theory of Functions*, Second Ed., Oxford University Press, Oxford, 1939.

wid [41] D. Widder, *The Laplace Transform*, Princeton University Press, Princeton, 1941.

wij [42] J. Wijngaard, "Existence of average optimal strategies in Markovian decision problems with strictly unbounded costs," in *Dynamic Programming and Its Applications*, edited by M. Puterman, Academic, New York, 1978.